Modèles et algorithmes dédiés pour la résolution de problèmes inverses parcimonieux en traitement du signal et de l'image

Sébastien Bourguignon

Habilitation à Diriger des Recherches de l'Université de Nantes

École doctorale Math-STIC
(Mathématiques et Sciences et Technologies de l'Information et de la Communication)
Spécialité Signal, Image, Vision

Modèles et algorithmes dédiés pour la résolution de problèmes inverses parcimonieux en traitement du signal et de l'image

Sébastien Bourguignon

École Centrale de Nantes
Laboratoire des Sciences du Numérique de Nantes (UMR 6004 CNRS)

Habilitation soutenue le 6 février 2020 devant le jury composé de :

Pierre Chainais	Professeur, École Centrale de Lille	Président
Laure Blanc-Féraud	Directrice de Recherche CNRS, Laboratoire I3S	Rapporteure
Stéphane Canu	Professeur, INSA Rouen	Rapporteur
Olivier Michel	Professeur, INP Grenoble	Rapporteur
Christine De Mol	Professeure, Université Libre de Bruxelles	Examinatrice
Nicolas Gillis	Associate Professor, Université de Mons	Examinateur
Jérôme Idier	Directeur de Recherche CNRS, LS2N	Tuteur

Avant-propos

Ce document, rédigé dans le but d'obtenir l'Habilitation à Diriger des Recherches de l'Université de Nantes, décrit les activités de recherche que j'ai menées postérieurement à ma thèse de doctorat.

Il est organisé en trois parties. La première est une présentation générale composée, au Chapitre 1, d'un curriculum vitae étendu retraçant mon parcours et mes activités d'enseignement et de recherche. Un bilan des différentes responsabilités scientifiques que j'ai exercées, de l'encadrement d'étudiants et de mes publications y est détaillé. Une synthèse de l'ensemble de mon activité de recherche.

J'ai choisi, dans la deuxième partie, de présenter de manière détaillée trois contributions dans des thématiques ayant particulièrement structuré mes travaux de recherche. L'ensemble de ces travaux relève de la méthodologie des problèmes inverses, où la notion de parcimonie occupe une place centrale. Une introduction à ce cadre de travail est d'abord proposée au Chapitre 3. Le Chapitre 4 est consacré à des méthodes de restauration de données hyperspectrales pour l'observation astrophysique. Si mes recherches dans ce domaine ont démarré lors de mon séjour post-doctoral à l'Observatoire de la Côte d'Azur, l'imagerie hyperspectrale et le traitement de données astrophysiques occupent toujours une place importante dans mes recherches actuelles. Le Chapitre 5 concerne la résolution de problèmes de déconvolution et de reconstruction d'images ultrasonores dans le contexte du contrôle non destructif (CND), thème sur lequel j'ai commencé à travailler dès mon arrivée à l'École Centrale de Nantes et sur lequel je compte poursuivre mes recherches, dans un contexte régional où le CND est très présent. Enfin, je présente au Chapitre 6 des travaux plus génériques sur l'optimisation exacte de critères parcimonieux impliquant la « norme » ℓ_0. Cette thématique, dont je suis l'initiateur au sein de mon équipe de recherche, constitue désormais un élément central de mes recherches, concrétisée par le démarrage du projet ANR « *Jeunes Chercheurs* » MIMOSA que je porte. Elle témoigne également d'une certaine prise d'autonomie dans mon parcours scientifique.

La dernière partie présente au Chapitre 7 quelques pistes de travaux de recherche que je souhaite poursuivre dans les années à venir, où chaque thème abordé pourrait donner lieu à la définition d'un futur sujet de thèse.

Une annexe vient clôturer ce manuscrit, contenant la sélection de cinq publications que j'estime représentatives de mon activité de recherche.

Table des matières

Première partie

Présentation générale

Chapitre 1

Curriculum Vitae

1.1 État civil

Sébastien Bourguignon
LS2N, École Centrale de Nantes
1 rue de la Noë, BP 92101
44321 Nantes Cedex 3

Né le 5 juillet 1977
Nationalité française
Vie maritale, 2 enfants

Tél : 02.40.37.69.15
Messagerie : Sebastien.Bourguignon@ec-nantes.fr
Web : http://pagesperso.ls2n.fr/~bourguignon-s/

1.2 Situation actuelle

- Maître de Conférences à l'École Centrale de Nantes, 61ème section (Informatique industrielle, Automatique et Traitement du Signal) depuis le 01/09/2011.
- 7ème échelon de la classe normale depuis le 16 juillet 2019.
- Enseignement au département Automatique et Robotique de l'École Centrale de Nantes.
- Recherche dans l'équipe Signal, Image et Son [1] au Laboratoire des Sciences du Numérique de Nantes [2], anciennement IRCCyN), UMR 6004 CNRS – Université de Nantes – École Centrale de Nantes – École des Mines de Nantes.

1. SIMS, https://sims.ls2n.fr
2. LS2N, http://ls2n.fr

1.3 Parcours académique

2011-. . . : **Maître de Conférences** à l'École Centrale de Nantes.

2009–11 : **Post-doctorant** au laboratoire Cassiopée (désormais Lagrange), UMR 6202 CNRS – Université de Nice – Observatoire de la Côte d'Azur.

2007–08 : **Chercheur permanent** à l'Institut Français de Recherche pour l'Exploitation de la Mer (Ifremer), Centre de Brest, Laboratoire de Technologie Halieutique.

2006–07 : **Post-doctorant** au Laboratoire d'Astrophysique de Toulouse et de Tarbes (LATT, désormais Institut de Recherche en Astrophysique et Planétologie), UMR 5572 – CNRS – Université Toulouse III Paul Sabatier.

2005–06 : **Attaché Temporaire d'Enseignement et de Recherche** à temps complet, 61$^{\text{ème}}$ section, Université Toulouse III Paul Sabatier. Recherche dans l'équipe *Signal, Image, Instrumentation* du LATT.

2002–05 : **Allocataire de recherche**, Université Toulouse III Paul Sabatier. Recherche dans l'équipe *Signal, Image, Instrumentation* du LATT.

2002–05 : **Moniteur de l'enseignement supérieur**, CIES de Midi-Pyrénées.

1.4 Diplômes universitaires

2005 : **Doctorat en Traitement du Signal**, Université Toulouse III Paul Sabatier : *Analyse spectrale à haute résolution de signaux irrégulièrement échantillonnés : application à l'Astrophysique.* Directeur de thèse Hervé Carfantan.

2002 : **DEA Signal, Image et Acoustique**, Université Toulouse III Paul Sabatier – INP Toulouse.

2001 : **Ingénieur Supélec** (École Supérieure d'Électricité, désormais CentraleSupélec).

Ingénieur en Télécommunications (ETSIT, Escuela Técnica Superior de Ingenieros de Telecomunicación), Université Polytechnique de Madrid, Espagne.

1.5 Activité d'enseignement

Cette section résume mon activité d'enseignement depuis mon arrivée à l'École Centrale de Nantes (ECN) en septembre 2011. J'ai également enseigné auparavant, de mon entrée en DEA en 2002 à 2011, pour un volume d'environ 500 h équivalent TD, dans des disciplines relevant essentiellement de la 61$^{\text{ème}}$ section. De 2011 à 2018, mon service d'enseignement à l'ECN a varié entre 194 h et 370 h équivalent TD annuelles, avec une moyenne de 280 h (décharges pédagogiques incluses). La figure 1.1 montre la ventilation de mon service dans les différentes formations de l'École Centrale de Nantes, année par année. Les intitulés des formations et des cours ont pu changer depuis 2011 ; j'utilise ici

les dénominations actuelles et les abréviations d'usage : CM = cours magistral, TD = travaux dirigés, TP = travaux pratiques.

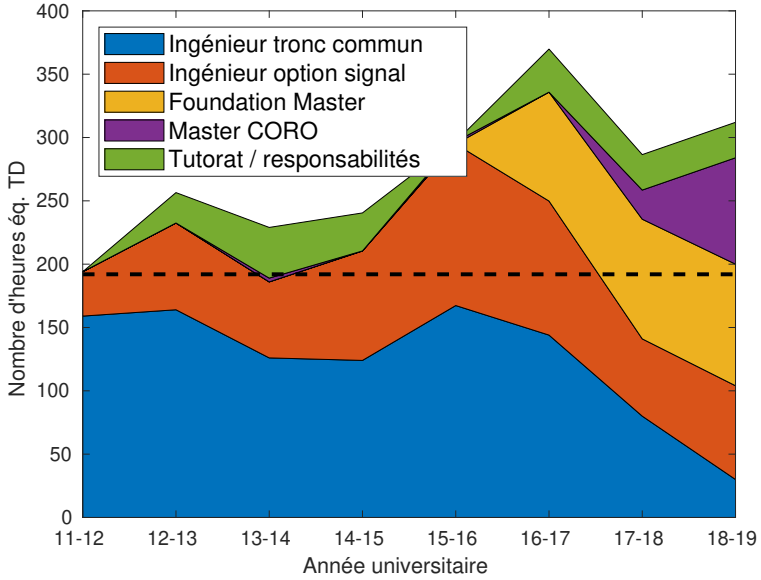


FIGURE 1.1 – Évolution de mon service d'enseignement depuis mon recrutement à l'École Centrale de Nantes. La ligne horizontale marque la limite statutaire de 192 h éq. TD.

1.5.1 Formation ingénieur de l'École Centrale de Nantes

- Tronc commun de première année :
  - *Instrumentation et Capteurs.* TD et TP.
  - *Signaux et Systèmes.* TD et TP.
  - *Commande.* TD et TP.
  - *Mathématiques.* TD et TP.
- Option disciplinaire de deuxième et troisième années DATASIM (Données, Algorithmes, Traitement et Applications en Signal et IMage) :
  - *Analyse et Représentation des Signaux et des Images.* Module complet (CM, TD, TP).
  - *Imagerie et Méthodes Inverses.* Module (CM, TD, TP) partagé avec Jérôme Idier.
  - *Outils Statistiques.* Module (CM, TD, TP) partagé avec Éric Le Carpentier.
  - Encadrement de *Projets en Traitement du Signal et de l'Image.*

1.5.2 Master CORO (Control and Robotics), parcours SIP (Signal and Image Processing), cours en anglais

- M1 CORO-SIP : *Spectral and Time-Frequency Analysis.* Module complet (CM, TD, TP).
- M2 CORO-SIP :
  - *Design of Signal and Image Representations.* Module complet (CM, TD, TP).

- *Signal and Image Restoration, Inversion Methods.* Module (CM, TD, TP) partagé avec Jérôme Idier.

1.5.3 Formation *Foundation Master*

Cette formation dispensée à l'ECN depuis 2016 s'adresse à des étudiants étrangers en dernière année de premier cycle (*bachelor*) et constitue une année préparatoire aux formations de Master de l'École Centrale de Nantes. J'y ai monté et j'y dispense les enseignements suivants :

- *Mathematics* (cours complet sous forme de CM).
- *Computer Programming and Data Analysis* (cours complet sous forme de TD sur ordinateur).

1.5.4 Responsabilités pédagogiques

- Responsable du parcours *Signal and Image Processing* du Master CORO (depuis 2019).
- Responsable de l'option professionnelle *Recherche et Développement* (filière métier) pour les élèves-ingénieurs de troisième année à l'ECN (depuis 2016).
- Suivi et tutorat pédagogique d'élèves-ingénieurs en alternance (depuis 2013).
- Participation au jury de concours d'étudiants intégrant les Écoles Centrales *via* la filière CAST'Ing (admission sur titres d'étudiants titulaires d'une Licence universitaire), 2018.

1.6 Activité de recherche

La description détaillée de mes travaux de recherche constitue la matière première de ce manuscrit et fait l'objet des Chapitres 2 à 7. Sans rentrer dans le contenu scientifique, cette section présente les différents marqueurs (responsabilités, expertise, collaborations) qui ont jalonné mon activité de recherche.

1.6.1 Participation à la vie collective, animation

- Membre élu au Conseil Scientifique de l'École Centrale de Nantes (2014–2018), réélu en 2018.
- Membre du Conseil Scientifique et correspondant au LS2N du Groupement d'Intérêt Scientifique (GIS) ECND-PdL, *Évaluation et Contrôle Non Destructifs en Pays de la Loire*, créé en 2017 [3].
- Organisation des séminaires de l'équipe SIMS du LS2N (fréquence mensuelle).

1.6.2 Organisation de congrès

- Membre du Comité d'Organisation de la conférence iTWIST'20 prévue à Nantes en juin 2020.

3. ecnd-pdl.fr

- Membre du Comité Scientifique du 2nd Franco-Chinese Acoustic Conference (FCAC), Le Mans, 2018.
- Membre du Comité de Programme de la 3rd IEEE International Conference on Image Processing, Applications and Systems (IPAS), 2018.
- Organisation de sessions spéciales *Problèmes Inverses* au Congrès Français d'Acoustique (2014, 2016).
- Jury pour le Best Student Paper Award au 7th European Workshop on Structural Health Monitoring (EWSHM), 2014.
- Organisation d'une session spéciale *Astrophysics and Planetary science* au 3rd Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (IEEE WHISPERS), 2011.

1.6.3 Expertise scientifique

- Relecteur auprès des journaux :

  Digital Signal Processing

  ICES Journal of Marine Science

  IEEE Journal of Selected Topics in Signal Processing

  IEEE Transactions on Computational Imaging

  IEEE Transactions on Image Processing

  IEEE Transactions on Information Theory

  IEEE Transactions on Signal Processing

  Machine Learning

  Signal Processing

  Transactions de la Société Canadienne de Génie Mécanique

- Relecteur pour le *Congrès Français d'Acoustique* et les conférences *EUSIPCO, EWSHM, GRETSI, IEEE ICASSP, IEEE SSP, IEEE WHISPERS.*
- Expertise de projet auprès de l'Agence Nationale de la Recherche (ANR).

1.6.4 Projets de recherche et activité contractuelle

Coordination de projets de recherche

- **Projet JCJC (jeunes chercheurs) de l'ANR** (appel à projets 2016), 2017–2020. *MIMOSA : Mixed-Integer programming Methods for Optimization of Sparse Approximation criteria.* **Porteur**, cinq chercheurs permanents impliqués (LS2N, Lab-STICC Brest, IRAP Toulouse, ENAC Toulouse) + un post-doctorant + un doctorant (non financé par le projet). Budget 191 k€.
- **Projet « Recherche, Formation, Innovation » de la région Pays de la Loire** *via* l'Institut Le Mans Acoustique, 2014–2018. Collaboration avec le Laboratoire d'Acoustique de l'Université du Maine (LAUM), *DECIMAP : Modèles de propagation et déconvolution de signaux ultrasonores dans les matériaux biphasiques.* **Co-**

porteur, quatre chercheurs permanents impliqués + un post-doctorant + un doctorant. Budget 150 k€.

- **Projet JCJC (jeunes chercheurs) du GdR ISIS et du GRETSI** (appel à projets 2013), 2014-2015. *Optimisation globale pour la résolution de problèmes parcimonieux en norme ℓ_0.* **Porteur**, deux chercheurs impliqués. Budget 6 k€.

Activité contractuelle

- **Contrat de collaboration associé à une thèse CIFRE, société DB-SAS**, 2017-2020. *Méthodes d'imagerie ultrasonore avancées et rapides pour le contrôle non destructif de matériaux atténuants et diffusants.* Budget : 36 k€.
- **Prestation pour l'Institut de Recherche Technologique Jules Verne**, 2016-2017. *Traitement d'un signal issu d'un profilomètre laser pour des applications de soudage.* Budget : 7 k€.

Implication dans d'autres projets de recherche collaboratifs

- **Projet Exploratoire Premier Soutien (PEPS) CNRS 2018 Astro-Informatique** (renouvelé en 2019) *OSIS : Outils statistiques pour l'imagerie hyperspectrale du milieu interstellaire.* Porteur Maryvonne Gerin (laboratoire LERMA, Observatoire de Paris). Budget : 4 k€ + 4 k€.
- **Défi CNRS 2015 Imag'In** (renouvelé en 2016) *MultiPlanNet : Réseau pour l'analyse et la fusion de données multimodales des surfaces planétaires.* Porteur Frédéric Schmidt (laboratoire GEOPS, Université Paris Sud). Budget : 5 k€ + 5 k€.

1.7 Encadrement de stagiaires, doctorants et post-doctorants

1.7.1 Stages de Master 2

Avant ma prise de poste à l'École Centrale de Nantes, j'ai co-encadré deux stagiaires du Master *Signal, Image, Acoustique et Optimisation* de l'Université Toulouse III Paul Sabatier (un dès la fin de ma thèse, puis un lors de mon séjour à l'Observatoire de la Côte d'Azur). J'ai ensuite encadré trois stagiaires du Master *Automatique, Robotique et Informatique Appliquée* (désormais *Control and Robotics*) de l'École Centrale de Nantes. J'ai également encadré deux stagiaires de Master provenant de recrutements extérieurs, à l'Université de Monastir en Tunisie (*Mastère de Recherche en Informatique*) puis à l'Université de Limoges (*Master ACSYON de Mathématiques Appliquées*).

- Jiayi Hou (2019) : **Sparse unmixing methods for hyperspectral imaging**. Master *Control and Robotics*, parcours *Signal and Image Processing*, École Centrale de Nantes.
- Ghandy Ajib (2018) : **Contributions to a branch-and-bound algorithm for ℓ_0-norm sparse optimization**. Master *ACSYON*, Université de Limoges. Co-encadrement avec Ramzi Ben Mhenni (doctorant, École Centrale de Nantes).
- Ramzi Ben Mhenni (2016) : **Algorithme de type Branch-and-Bound pour l'optimisation exacte de critères parcimonieux en norme ℓ_0**. Mastère de Recherche en *Informatique, spécialité Systèmes de Raisonnement Automatique*, Uni-

versité de Monastir, Tunisie. Co-encadrement avec Evgeny Gurevsky (maître de conférences, Université de Nantes).

- Yan Li (2013) : **Algorithmes de déconvolution parcimonieuse par pénalisation ℓ_0**. Master *Automatique, Robotique et Informatique appliquée*, parcours Automatique, Signal et Image, École Centrale de Nantes. Co-encadrement avec E. Carcreff (doctorant, Université du Maine).
- Julien Picaud (2012) : **Restauration de cubes hyperspectraux en Astrophysique**. Master *Automatique, Robotique et Informatique Appliquée*, parcours Automatique, Signal et Image, École Centrale de Nantes.
- Benjamin Trémoulhéac (2010) : **Représentations parcimonieuses de données hyperspectrales en astrophysique : aspects informationnels et algorithmiques**. Master *Signal, Image, Acoustique et Optimisation*, Université Toulouse III Paul Sabatier. Co-encadrement avec David Mary (maître de conférences, Université de Nice).
- Ali Khazaal (2006) : **Algorithmes d'optimisation pour une représentation parcimonieuse de signaux bruités**. Master *Signal, Image, Acoustique et Optimisation*, Université Toulouse III Paul Sabatier. Co-encadrement avec Hervé Carfantan (maître de conférences, Université Toulouse 3).

1.7.2 Encadrement de doctorants

Depuis mon arrivée à l'École Centrale de Nantes, j'ai participé à l'encadrement de **six thèses de doctorat** : cinq de l'École Centrale de Nantes (dont une en cotutelle avec l'École Polytechnique de Montréal) et une de l'Université du Maine (Le Mans). Trois de ces thèses ont été soutenues (taux d'encadrement cumulé 110 %) et trois sont en cours (taux d'encadrement cumulé 150 %). Si j'ai été ou suis co-encadrant de cinq de ces thèses, j'ai bénéficié d'une décharge d'habilitation à diriger des recherches pour diriger la thèse de Ramzi Ben Mhenni, dont le sujet émane de mon initiative et s'adosse au projet ANR MIMOSA que je porte.

Thèses en cours

- Nans Laroche, **Méthodes d'imagerie ultrasonore avancées et rapides pour le contrôle non destructif de matériaux atténuants et diffusants**. École Centrale de Nantes. Encadrement à 40 %. Directeur de thèse Jérôme Idier (30 %), co-encadrant Aroune Duclos (30 %). Financement industriel (convention CIFRE). Thèse démarrée en novembre 2017.
- Marie-Antoinette AlHajj, **Inversion conjointe de mesures non destructives pour découpler des gradients multiples dans le béton d'enrobage de structures en béton armé**. IFSTTAR, École Centrale de Nantes. Encadrement à 30 %. Directrice de thèse Géraldine Villain (40 %), co-encadrant Sérgio Palma Lopes (30 %). Financement IFSTTAR. Thèse démarrée en octobre 2017.
- Ramzi Ben Mhenni, **Programmation mixte en nombres entiers pour l'optimisation parcimonieuse en traitement du signal.** École Centrale de Nantes. Directeur de thèse (80 %), co-encadrement Jordan Ninin (20 %). Financement *via* une allocation ministérielle. Thèse démarrée en octobre 2016.

Thèses soutenues

- Penghuan Liu, **Statistical and numerical optimization for speckle blind structured illumination microscopy.** École Centrale de Nantes. Encadrement à 30 %. Directeur de thèse Jérôme Idier (70 %). Financement du gouvernement chinois *via* le *China Scholarship Council (CSC)*. Thèse soutenue en mai 2018.
  Penghuan Liu est rentré en Chine après avoir soutenu sa thèse.
- Corentin Friedrich, **Méthodes de reconstruction en tomographie de diffraction 3-D.** École Centrale de Nantes, thèse en co-tutelle avec Polytechnique Montréal. Encadrement à 30 %. Directeur de thèse Jérôme Idier (40 %), co-directeur Yves Goussard (Polytechnique Montréal, 30 %). Financement École Centrale de Nantes et Polytechnique Montréal. Thèse soutenue en septembre 2016.
  Corentin Friedrich est ingénieur en CDI au sein de la société Brainchip holdings, Toulouse.
- Ewen Carcreff, **Déconvolution adaptative de signaux ultrasonores pour le contrôle non destructif par ultrasons**. Université du Maine. Encadrement à 50%. Directeur de thèse Laurent Simon (20%), co-directeur Jérôme Idier (30%). Financement de la région Pays de la Loire. Thèse soutenue en novembre 2014.
  Ewen Carcreff est ingénieur R&D en CDI au sein de la société DB-SAS, Nantes.

1.7.3 Post-doctorants

J'ai collaboré avec un post-doctorant recruté sur le projet DECIMAP ()*Modèles de propagation et déconvolution de signaux ultrasonores dans les matériaux biphasiques*), recruté au LS2N pendant 17 mois. J'ai également travaillé avec un post-doctorant recruté sur le projet ANR MIMOSA, de novembre 2018 à mai 2019 [4].

- Ngoc Nguyên Tran : *Développement d'algorithmes d'optimisation exacte pour des problèmes de parcimonie en traitement du signal*, novembre 2018 – mai 2019. Financement *via* le projet ANR MIMOSA.
- Nizar Bouhlel : *Évaluation non-destructive de matériaux biphasiques par des méthodes ultrasonores*, septembre 2015 – janvier 2017. Financement de la région Pays de la Loire *via* l'Institut Le Mans Acoustique.

1.8 Quelques éléments de reconnaissance scientifique

- Bénéficiaire de la **Prime d'Encadrement Doctoral et de Recherche** (PEDR) depuis 2016.
- **Invitations pour des séminaires** (depuis 2010) :
  - Séminaire de l'équipe *Optimisation Globale et Résolution Ensembliste*, laboratoire LS2N, Nantes : *Optimisation globale pour des problèmes de moindres carrés à faible cardinalité*, mars 2019.
  - Séminaire de l'équipe *Modélisation, Optimisation, Dynamique*, laboratoire XLIM, Limoges : *Global optimization for least squares problems with low cardinality*, février 2019.

4. Le contrat était initialement prévu pour deux ans, mais la personne recrutée l'a interrompu après sept mois pour raisons personnelles.

  - Séminaires *Optimisation exacte de critères parcimonieux en norme ℓ_0 par programmation mixte en nombres entiers* :
    - Laboratoire LISIC, Calais, juillet 2017.
    - Séminaire *Signal Image*, Laboratoire IMS, Bordeaux, janvier 2016.
    - Séminaire *Mathématiques pour l'Analyse des Données*, Laboratoire Lagrange, Nice, juin 2015.
    - Séminaire *Traitement du Signal – Apprentissage Statistique*, Laboratoires I2M-LIF, Marseille, janvier 2015.
  - Journée scientifique *Parcimonie*, Fédération de recherche Charles Hermite, Nancy : *Restauration de données hyperspectrales astrophysiques par approximation parcimonieuse*, juin 2014.
  - Journée *Optimisation pour le traitement et l'analyse de données*, Plan Pluri Formation *Pôle Signal Image*, Observatoire Midi-Pyrénées, Toulouse : *Optimisation parcimonieuse en l'absence de transformées rapides : application à la restauration de spectres astrophysiques*, novembre 2010.
  - Séminaire *parcimonie*, CEA-LETI, Grenoble : *Restauration de spectres astrophysiques avec contraintes de parcimonie : aspects informationnels et algorithmiques*, octobre 2010.
- Participation à des **jurys de thèse** (examinateur hors encadrement)
  - Alina-Georgiana Merecescu, *Inverse problems of deconvolution applied in the fields of Geoscience and Planetology.* Université Paris-Saclay, septembre 2018.
  - Antoine Crinière, *Contribution au développement d'outils d'analyse de séquences d'images infrarouges : Application au contrôle non destructif de structures de génie civil.* École Centrale de Nantes, IFSTTAR, Bouguenais, octobre 2014.
- Participation à des **comités de sélection** (recrutement de maître de conférences)
  - Poste de Maître de Conférences en 65$^{\text{ème}}$ section à l'École Centrale de Nantes et au CHU de Nantes (laboratoire ITUN), 2016.
  - Poste de Maître de Conférences en 61$^{\text{ème}}$ section à l'ENSIM (École Nationale d'Ingénieurs du Mans) et au LAUM (Laboratoire d'Acoustique de l'Université du Maine), 2014.
- Participation à des **comités de suivi de thèse**
  - Mahdi Jammal, INSA Rouen, 2019 – ...
  - Bian Xiong, IFSTTAR, Nantes, 2019 – ...
  - Nathalie Freycenon, Université du Maine, LAUM, Le Mans, 2017 – ...
  - Thomas Paviet-Salomon, ENSTA Bretagne, Lab-STICC, Brest, 2017 – ...
  - Matti Niskanen, Université du Maine, LAUM, Le Mans, 2015 – ...
  - Borys Shchukin, Université du Maine, LAUM, Le Mans, 2014 – 2017
  - Nicolas Le Touz, École Centrale de Nantes, IFSTTAR, Nantes, 2016
  - Maxime Legendre, École Centrale de Nantes, Nantes, 2014
- **Papier finaliste** pour le *EuCAP 2015 Propagation Best Paper Award* : C. Friedrich, S. Bourguignon, J. Idier and Y. Goussard, *Faster resolution of the 3-D forward*

problems in microwave imaging by a partial-block BiCGStab algorithm, in EuCAP, The 9th European Conference on Antennas and Propagation, Lisbon, Portugal, April 2015

1.9 Principales collaborations en cours et récentes

- Marc Allain, Institut Fresnel, Marseille
- Mickaël Bazot, New York University at Abu Dhabi
- Nizar Bouhlel, Institut d'Electronique et de Télécommunications de Rennes (IETR)
- Ewen Carcreff, Société DB-SAS, Nantes
- Hervé Carfantan, Institut de Recherche en Astrophysique et Planétologie (IRAP), Toulouse
- Aroune Duclos, Laboratoire d'Acoustique de l'Université du Maine (LAUM), Lee Mans
- Yves Goussard, Polytechnique Montréal
- Jérôme Idier, Laboratoire des Sciences du Numérique de Nantes (LS2N)
- David Mary, Laboratoire Lagrange, Nice
- Marcel Mongeau, École Nationale de l'Aviation Civile (ENAC), Toulouse
- Jordan Ninin, Laboratoire des Sciences et Techniques de l'Information, de la Communication et de la Connaissance (Lab-STICC), Brest
- Sérgio Palma Lopes, Institut français des sciences et technologies des transports, de l'aménagement et des réseaux (IFSTTAR), Nantes
- Jérôme Pety, Institut de Radioastronomie Millimétrique (IRAM), Grenoble
- Frédéric Schmidt, Laboratoire Géosciences Paris Sud (GEOPS), Orsay
- Laurent Simon, LAUM, Le Mans
- Charles Soussen, Laboratoire des Signaux et des Systèmes, Gif / Yvette
- Géraldine Villain, IFSTTAR, Nantes

1.10 Liste des publications

Les articles de journaux [S1,J5,J8,J11] et l'acte de conférence [ACIC3] sont joints en annexe de ce document.

Soumissions en cours

[S1] R. Ben Mhenni, S. Bourguignon et J. Ninin, « Global optimization for sparse solution of least squares problems », *Soumis à Optimization Methods and Software*, 2019.

Articles de journaux

[J1] F. Schmidt et S. Bourguignon, « Efficiency of BRDF sampling and bias on the average photometric behavior », *Icarus*, vol. 317, pp. 10–26, Jan 2019.

[J2] J. Idier, S. Labouesse, M. Allain, P. Liu, S. Bourguignon et A. Sentenac, « On the superresolution capacity of imagers using unknown speckle illuminations », *IEEE Transactions on Computational Imaging*, vol. 4, 1, pp. 87–98, March 2018.

[J3] S. Labouesse, A. Negash, J. Idier, S. Bourguignon, T. Mangeat, P. Liu, A. Sentenac et M. Allain, « Joint reconstruction strategy for structured illumination microscopy with unknown illuminations », *IEEE Transactions on Image Processing*, vol. 26, 5, pp. 1–14, mai 2017.

[J4] D. Testa, H. Carfantan, M. Albergante, P. Blanchard, S. Bourguignon, A. Fasoli, A. Goodyear, A. Klein, J. Lister et T. Panis, « Sparse representation of signals : from astrophysics to real-time data analysis for fusion plasmas and system optimization analysis for ITER and TCV », *Plasma Physics and Controlled Fusion*, vol. 58, 12, 2016.

[J5] S. Bourguignon, J. Ninin, H. Carfantan et M. Mongeau, « Exact sparse approximation problems via mixed-integer programming : Formulations and computational performance », *IEEE Transactions on Signal Processing*, vol. 64, 6, pp. 1405–1419, 2016.

[J6] C. Friedrich, S. Bourguignon, J. Idier et Y. Goussard, « Reconstruction of 3-D microwave images based on a Block-BiCGStab algorithm », *Journal of Physics : Conference Series*, vol. 657, pp. 012014, 2015.

[J7] E. Carcreff, S. Bourguignon, A. Duclos, L. Simon et J. Idier, « Detection of flat bottom holes using sparse deconvolution », *Physics Procedia*, vol. 70, pp. 558–561, 2015.

[J8] E. Carcreff, S. Bourguignon, J. Idier et L. Simon., « A linear model approach for ultrasonic inverse problems with attenuation and dispersion », *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, vol. 61, 7, pp. 1191–1203, 2014.

[J9] M. Bazot, S. Bourguignon et J. Christensen-Dalsgaard, « A Bayesian approach to the modelling of α Cen A », *Monthly Notices of the Royal Astronomical Society*, vol. 427, 3, pp. 1847–1866, 2012.

[J10] S. Bourguignon, D. Mary et É. Slezak, « Processing MUSE hyperspectral data : Denoising, deconvolution and detection of astrophysical sources », *Statistical Methodology*, vol. 9, 1, pp. 32–43, 2012.

[J11] S. Bourguignon, D. Mary et É. Slezak, « Restoration of astrophysical spectra with sparsity constraints : Models and algorithms. », *IEEE J. Sel. Topics Signal Processing*, vol. 5, 5, pp. 1002–1013, 2011.

[J12] D. Serre, E. Villeneuve, H. Carfantan, L. Jolissaint, V. Mazet, S. Bourguignon et A. Jarno, « Modeling the spatial PSF at the VLT focal plane for MUSE WFM data analysis purpose », *Proc. SPIE 7736*, p. 773649, 2010.

[J13] M. Vannier, D. Mary, F. Milour, R. G. Petrov, S. Bourguignon et C. Theys, « Spectral regularization and sparse representation bases for interferometric imaging », *Proc. SPIE 7734*, p. 77342J, 2010.

[J14] P. Gaulme, M. Vannier, T. Guillot, B. Mosser, D. Mary, W. W. Weiss, F.-X. Schmider, S. Bourguignon, H. J. Deeg, C. Régulo, S. Aigrain, J. Schneider, H. Bruntt, S. Deheuvels, J.-F. Donati, T. Appourchaux, M. Auvergne, A. Baglin, F. Baudin, C. Catala, E. Michel et R. Samadi, « Possible detection of phase changes from the non-transiting planet HD 46375b by CoRoT », *Astronomy and Astrophysics*, vol. 518, pp. L1 53, 2010.

[J15] V. Trenkel, L. Berger, S. Bourguignon, M. Doray, R. Fablet, J. Masse, V. Mazauric, C. Poncelet, G. Quemeneur, C. Scalabrin et H. Villalobos, « Overview of recent progress in fisheries acoustics made by Ifremer with examples from the Bay of Biscay », *Aquatic Living Resources*, vol. 22, 4, pp. 433–446, 2009.

[J16] S. Bourguignon, L. Berger, C. Scalabrin, R. Fablet et V. Mazauric, « Methodological developments for improved bottom detection with the ME70 multibeam echosounder », *ICES Journal of Marine Science*, vol. 66, 6, pp. 1015–1022, 2009.

[J17] S. Bourguignon et H. Carfantan, « New methods for fitting multiple sinusoids from irregularly sampled data », *Statistical Methodology*, vol. 5, 4, pp. 318–327, 2008.

[J18] S. Bourguignon, H. Carfantan et J. Idier, « A sparsity-based method for the estimation of spectral lines from irregularly sampled data », *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, 4, pp. 575–585, 2007.

[J19] S. Bourguignon, H. Carfantan et T. Boehm, « Sparspec : a new method for fitting multiple sinusoids with irregularly sampled », *Astronomy and Astrophysics*, vol. 462, pp. 379–387, 2007.

Chapitres d'ouvrage

[O1] S. Bourguignon et H. Carfantan, « Estimation de spectres de raies pour des signaux irrégulièrement échantillonnés en astrophysique », in *Méthodes d'inversion appliquées au traitement du signal et de l'image*, J.-F. Giovannelli et J. Idier, Eds., pp. 163–190. Hermès - Lavoisier, 2013.

[O2] S. Bourguignon et H. Carfantan, « Line spectra estimation for irregularly sampled signals in Astrophysics », in *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing*, J.-F. Giovannelli et J. Idier, Eds., pp. 141–168. ISTE-Wiley, 2015.

Conférences internationales avec actes et comité

[ACIC1] N. Laroche, E. Carcreff, S. Bourguignon, J. Idier et A. Duclos, « Fast inverse approach for the deconvolution of ultrasonic TFM images using a spatially varying PSF in NDT », in *IEEE International Ultrasonics Symposium*, Glasgow, Royaume Uni, 2019.

[ACIC2] N. Laroche, E. Carcreff, S. Bourguignon, J. Idier et A. Duclos, « An inverse approach for ultrasonic imaging by total focusing point for close reflectors separation », in *IEEE International Ultrasonics Symposium (IUS)*, Kobe, Japan, 2018.

[ACIC3] R. Ben Mhenni, S. Bourguignon, J. Ninin et F. Schmidt, « Spectral unmixing with sparsity and structuring constraints », in *Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Amsterdam, The Netherlands, 2018.

[ACIC4] M. Boudineau, H. Carfantan et S. Bourguignon, « An l0 solution to sparse approximation problems with continuous dictionaries, », in *International Conference on Acoustic, Speech and Signal Processing (ICASSP'2018)*, Calgary, Alberta, Canada, 2018.

[ACIC5] E. Carcreff, N. Laroche, D. Braconnier, A. Duclos et S. Bourguignon, « Improvement of the total focusing method using an inverse problem approach », in *2017 IEEE International Ultrasonics Symposium (IUS)*, Sept 2017, pp. 1–4.

[ACIC6] S. Labouesse, M. Allain, J. Idier, S. Bourguignon, A. Negash, P. Liu et A. Sentenac, « Fluorescence blind structured illumination microscopy : a new reconstruction strategy », in *IEEE International Conference on Image Processing (ICIP)*, Phoenix, USA, 2016.

[ACIC7] M. Boudineau, H. Carfantan, S. Bourguignon et M. Bazot, « Sampling schemes and parameter estimation for nonlinear bernoulli-gaussian sparse models », in *IEEE Statistical Signal Processing Workshop*, Palma de Mallorca, Spain, june 2016.

[ACIC8] C. Friedrich, S. Bourguignon, J. Idier et Y. Goussard, « Faster resolution of the 3-D forward problems in microwave imaging by a partial-block BiCGStab algorithm », in *EuCAP, The 9th European Conference on Antennas and Propagation*, Lisbon, Portugal, 2015.

[ACIC9] E. Carcreff, S. Bourguignon, J. Idier, L. Simon et A. Duclos, « Including frequency-dependent attenuation for the deconvolution of ultrasonic signals », in *19th International Congress on Acoustics*, Montréal, Québec, Canada, 2013.

[ACIC10] E. Carcreff, S. Bourguignon, J. Idier et L. Simon, « Resolution enhancement of ultrasonic signals by up-sampled sparse deconvolution », in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, Vancouver, Canada, 2013.

[ACIC11] E. Carcreff, S. Bourguignon, J. Idier et L. Simon, « High-resolution deconvolution applied to nondestructive testing », in *Acoustics 2012*, Nantes, France, 2012.

[ACIC12] Y.-S. Shen, T.-H. Chan, S. Bourguignon et C.-Y. Chi, « Spatial-spectral unmixing of hyperspectral data for detection and analysis of astrophysical sources with the MUSE instrument », in *Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Shanghai, China, 2012.

[ACIC13] S. Paris, D. Mary, A. Ferrari et S. Bourguignon, « Sparsity-based composite detection tests. application to astrophysical hyperspectral data », in *EUSIPCO*, Barcelona, Spain, 2011.

[ACIC14] S. Bourguignon, C. Soussen, H. Carfantan et J. Idier, « Sparse deconvolution : comparison of statistical and deterministic approaches », in *IEEE Workshop on Statistical Signal Processing*, Nice, France, 2011, pp. 317–320.

[ACIC15] F. Schmidt, S. Bourguignon, S. Le Mouélic, N. Dobigeon, C. Theys et E. Tréguier, « Accuracy and performance of lineat unmixing techniques for detecting minerals on Omega / Mars Express », in *Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Lisbon, Portugal, 2011.

[ACIC16] S. Bourguignon, H. Carfantan, E. Slezak et D. Mary, « Sparsity-based spatial-spectral restoration of MUSE astrophysical hyperspectral data cubes », in *Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Lisbon, Portugal, 2011.

[ACIC17] S. Bourguignon, D. Mary et E. Slezak, « Sparsity-based denoising of hyperspectral astrophysical data with colored noise. application to the MUSE instrument », in *Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Reykjavik, Iceland, 2010.

[ACIC18] R. Lefort, R. Fablet, J.-M. Boucher, L. Berger et S. Bourguignon, « Automatic fish school classification for acoustic sensing of marine ecosystem », 2008, pp. 1 – 5.

[ACIC19] S. Bourguignon et V. Mazauric, « Methodological developments for improved bottom detection with the ME70 multibeam echosounder », in *6th Ecosystem Approach with Fisheries Acoustics and Complementary Technologies (SEAFACTS)*, Bergen, Norway, 2008.

[ACIC20] M. Bazot, S. Bourguignon et J. Christensen-Dalsgaard, « Estimation of stellar parameters using Monte-Carlo Markov Chains », in *XXI Century challenges for stellar evolution (Memorie della Societa Astronomica Italiana, vol. 79)*, S. Cassisi et M. Salaris, Eds., 2007.

[ACIC21] S. Bourguignon et H. Carfantan, « New methods for fitting multiple sinusoids from irregularly sampled data », in *Astronomical Data Analysis IV*, Marseille, France, 2006.

[ACIC22] S. Bourguignon et H. Carfantan, « Spectral analysis of irregularly sampled data using a Bernoulli-Gaussian model with free frequencies », in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, Toulouse, France, 2006, pp. 516–519.

[ACIC23] S. Bourguignon et H. Carfantan, « Bernoulli-Gaussian spectral analysis of unevenly spaced astrophysical data », in *IEEE Workshop on Statistical Signal Processing*, Bordeaux, France, 2005, pp. 811–816.

[ACIC24] S. Bourguignon, H. Carfantan et L. Jahan, « Regularized spectral analysis of unevenly spaced data », in *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, Philadelphia, PA, USA, 2005, pp. 421–424.

[ACIC25] S. Bourguignon, H. Carfantan et L. Jahan, « Regularized estimation of line spectra from irregularly sampled astrophysical data », in *PSIP : Physics in Signal and Image Processing*, Toulouse, France, 2005.

Conférences nationales avec actes et comité

[ACNC1] R. Ben Mhenni, S. Bourguignon, M. Mongeau, J. Ninin et H. Carfantan, « Algorithme *branch-and-bound* pour l'optimisation exacte en norme ℓ_0 », in *Actes du 27e colloque GRETSI*, 2019.

[ACNC2] N. Laroche, S. Bourguignon, E. Carcreff, J. Idier et A. Duclos, « Approche inverse rapide pour la déconvolution d'images ultrasonores par une PSF variable », in *Actes du 27e colloque GRETSI*, 2019.

[ACNC3] R. Ben Mhenni, S. Bourguignon, J. Ninin et F. Schmidt, « Démélange parcimonieux exact dans une approche supervisée en imagerie hyperspectrale », in *Actes du 26e colloque GRETSI*, Juan-les-Pins, sep. 2017.

[ACNC4] P. Liu, J. Idier, S. Bourguignon, S. Labouesse, M. Allain et A. Sentenac, « Minimum contrast estimation for super-resolution fluorescence microscopy using speckle patterns », in *Actes du 26e colloque GRETSI*, Juan-les-Pins, sep. 2017.

[ACNC5] S. Bourguignon, J. Ninin, H. Carfantan et M. Mongeau, « Optimisation exacte de critères parcimonieux en norme l_0 par programmation mixte en nombres entiers », in *Actes du 25e colloque GRETSI*, Lyon, France, 2015.

[ACNC6] S. Labouesse, M. Allain, J. Idier, S. Bourguignon, P. Liu et A. Sentenac, « Blind fluorescence structured illumination microscopy : A new reconstruction strategy », in *Actes du 25e colloque GRETSI*, Lyon, France, 2015.

[ACNC7] E. Carcreff, S. Bourguignon, J. Idier et L. Simon, « Algorithmes de déconvolution impulsionnelle à résolution augmentée », in *Actes du 24e colloque GRETSI*, Brest, France, 2013.

[ACNC8] J. Picaud et S. Bourguignon, « Restauration de données hyperspectrales astrophysiques par approximation parcimonieuse », in *Actes du 24e colloque GRETSI*, Brest, France, 2013.

[ACNC9] S. Bourguignon, H. Carfantan, D. Mary, E. Slezak et A. Ferrari, « Restauration des cubes hyperspectraux du spectro-imageur MUSE », in *Actes du 23e colloque GRETSI*, Bordeaux, France, 2011.

[ACNC10] E. Villeneuve, H. Carfantan, A. Jarno, D. Serre, V. Mazet et S. Bourguignon, « Modélisation et estimation de la psf d'un instrument hyperspectral au sol pour l'astrophysique », in *Actes du 23e colloque GRETSI*, Bordeaux, France, 2011.

[ACNC11] S. Bourguignon, H. Carfantan et J. Idier, « Minimisation de critères de moindres carrés pénalisés par la norme l1 dans le cas complexe », in *Actes du 21e colloque GRETSI*, Troyes, France, 2007.

Conférences internationales avec actes

[ACI1] S. Labouesse, A. Negash, J. Idier, S. Bourguignon, T. Mangeat, P. Liu, A. Sentenac et M. Allain, « Fast reconstruction in blind fluorescence structured illumination microscopy », in *OSA Conference : Imaging and Applied Optics*, San Francisco, CA, USA, juin 2017, p. MM3C.2.

[ACI2] S. Labouesse, J. Idier, P. Liu, M. Allain, S. Bourguignon et A. Sentenac, « Super-resolution capacity of imagers using random illuminations », in *OSA Conference : Imaging and Applied Optics*, Heidelberg, Allemagne, juil. 2016, p. MTh1H.5.

[ACI3] P. Liu, S. Bourguignon, J. Idier, M. Allain, S. Labouesse et A. Sentenac, « A marginal image reconstruction approach in fluorescence microscopy with pseudo-random illumination patterns », in *OSA Conference : Imaging and Applied Optics*, Heidelberg, Allemagne, juil. 2016, p. MTh1H.3.

Conférences nationales avec actes

[ACN1] E. Carcreff, N. Laroche, S. Bourguignon, J. Idier et A. Duclos, « Une approche inverse pour l'imagerie ultrasonore par focalisation en tout point », in *Actes du 14e Congrès Français d'Acoustique*, Le Havre, France, 2018.

[ACN2] E. Carcreff, R. Lallement, G. Dao et S. Bourguignon, « Imagerie rapide par focalisation en tout point pour le contrôle non destructif par ultrasons », in *Actes du 13e Congrès Français d'Acoustique*, Le Mans, France, 2016.

[ACN3] S. Bourguignon, E. Carcreff, N. Lurin et S. Moreau, « Déconvolution d'échos ultrasonores régulièrement espacés pour la mesure d'épaisseur de structures fines », in *Actes du 13e Congrès Français d'Acoustique*, Le Mans, France, 2016.

[ACN4] N. Bouhlel, S. Bourguignon, A. Duclos et J.-P. Groby, « Déconvolution des signaux ultrasonores dans les matériaux poreux », in *Actes du 13e Congrès Français d'Acoustique*, Le Mans, France, 2016.

[ACN5] E. Carcreff, S. Bourguignon, J. Idier et L. Simon, « Contrôle non destructif des matériaux atténuants et dispersifs par dconvolution impulsionnelle », in *Actes du 12e Congrès Français d'Acoustique*, Poitiers, France, 2014.

Conférences sans actes / Journées scientifiques

[C1] R. Ben Mhenni, S. Bourguignon, J. Ninin et F. Schmidt, « Méthodes exactes de démélange spectral en norme ℓ_0 et contraintes de parcimonie structurée à l'aide de mip », in *19e conférence ROADEF*, Lorient, France, 2018, Société Française de Recherche Opérationnelle et Aide à la Décision.

[C2] N. Bouhlel, S. Bourguignon, A. Duclos et J.-P. Groby, « Characterization of poroelastic materials through interface scattering and propagation models », in *5th symposium on the Acoustics of Poro-Elasic Materials (SAPEM)*, Le Mans, France, 2017.

[C3] N. Bouhlel et S. Bourguignon, « Deconvolution of ultrasonic signals in porous materials : Estimation of acoustic propagation parameters and wave separation », in *Review of Progress in Quantitaive Nondestructive Evaluation (QNDE)*, Atlanta, USA, 2016.

[C4] S. Bourguignon, E. Carcreff, C. Lin et J. Peng, « Blind sparse deconvolution of regularly spaced ultrasonic echoes for thickness measurement », in *Review of Progress in Quantitative Nondestructive Evaluation (QNDE)*, Atlanta, USA, 2016.

[C5] S. Bourguignon, J. Ninin, H. Carfantan et M. Mongeau, « Programmation mixte en nombres entiers pour l'approximation parcimonieuse de signaux », in *16e conférence ROADEF*, Marseille, France, 2015, Société Française de Recherche Opérationnelle et Aide à la Décision.

[C6] E. Carcreff, S. Bourguignon, A. Duclos et J.-P. Groby, « An ultrasonic propagation model suited to inversion methods (with some application to porous materials) », in *4th symposium on the Acoustics of Poro-Elasic Materials (SAPEM)*, Stockholm, Sweden, 2014.

[C7] E. Carcreff, S. Bourguignon, J. Idier et L. Simon, « Déconvolution pour le contrôle par ultrasons des matériaux atténuants et dispersifs) », in *Actes des journées de la Confédération Française des Essais Non Destructifs*, Bordeaux, France, 2014, COFREND.

[C8] S. Bourguignon, H. Carfantan et J. Idier, « Analyse spectrale de signaux à échantillonnage irrégulier en astrophysique », in *Journée "Échantillonnage irrégulier"*, Paris, France, 2008, GdR ISIS.

[C9] S. Bourguignon, H. Carfantan et J. Idier, « Optimisation parcimonieuse en l'absence de transformées rapides », in *Journée "Optimisation de critères convexes non différentiables pour la résolution de problèmes inverses"*, Paris, France, 2010, GdR ISIS.

[C10] S. Bourguignon, J. Ninin, H. Carfantan et M. Mongeau, « Optimisation globale déterministe pour la résolution de problèmes parcimonieux en norme ℓ_0 », in *Journée "Optimisation non-convexe"*, Paris, France, 2014, GdR ISIS.

Séminaires invités

[S1] S. Bourguignon, « Optimisation globale pour des problèmes de moindres carrés à faible cardinalité », in *Séminaire équipe Optimisation Globale et Résolution Ensembliste*, Nantes, France, 2019, Laboratoire LS2N.

[S2] S. Bourguignon, « Global optimization for least squares problems with low cardinality », in *Séminaire équipe Modélisation, Optimisation, Dynamique*, Limoges, France, 2019, Laboratoire XLIM.

[S3] S. Bourguignon, « Optimisation exacte de critères parcimonieux en norme ℓ_0 par programmation mixte en nombres entiers », in *Séminaire Mathématiques pour l'analyse des données*, Calais, France, 2015, Laboratoire LISIC.

[S4] S. Bourguignon, « Optimisation exacte de critères parcimonieux en norme ℓ_0 par programmation mixte en nombres entiers », in *Séminaire Mathématiques pour l'analyse des données*, Bordeaux, France, 2015, Laboratoire IMS.

[S5] S. Bourguignon, « Optimisation exacte de critères parcimonieux en norme ℓ_0 par programmation mixte en nombres entiers », in *Séminaire Mathématiques pour l'analyse des données*, Nice, France, 2015, Observatoire de la Côte d'Azur.

[S6] S. Bourguignon, « Exact minimisation of ℓ_0-based sparse approximation criteria through mixed integer programming », in *Séminaire Signal et Apprentissage*, Marseille, France, 2015, Laboratoires LIF et IMM.

[S7] S. Bourguignon, « Restauration de données hyperspectrales astrophysiques par approximation parcimonieuse », in *Journée Parcimonie*, Nancy, France, 2014, Fédération de Recherche Charles Hermite.

[S8] S. Bourguignon, « Restauration de spectres astrophysiques avec contraintes de parcimonie : aspects informationnels et algorithmiques », in *Séminaire Parcimonie*, Grenoble, France, 2010, CEA-LETI.

Thèse de doctorat

[T1] S. Bourguignon, *Analyse spectrale à haute résolution de signaux irrégulièrement échantillonnés : application à l'Astrophysique*, Thèse de Doctorat, Université Paul Sabatier - Toulouse 3, 2005.

Logiciel

[L1] S. Bourguignon, H. Carfantan, A. Khazaal et A. Lasfar, « Sparspec : Sparse modeling for the spectral analysis of unevenly spaced data, `http://www.ast.obs-mip.fr/sparspec` ».

Chapitre 2

Vue synthétique de mes travaux de recherche

Ce chapitre restitue une vision synthétique de mes activités de recherche postérieures à ma thèse de doctorat, dans l'ordre chronologique de leur démarrage. Les sujets que je considère les plus centraux feront l'objet de la seconde partie de ce manuscrit.

L'essentiel de mon activité de recherche relève de la méthodologie des problèmes inverses, développée dans différents contextes applicatifs. Une démarche commune a guidé l'ensemble de mes travaux, consistant, d'une part, en une réflexion autour du modèle reliant le processus d'acquisition des données aux quantités d'intérêt à estimer et du choix d'un estimateur approprié, et d'autre part en le développement d'algorithmes de calcul adaptés, voire spécifiquement construits pour le problème considéré. J'ai donc souvent travaillé au contact de chercheurs issus de différentes communautés scientifiques, avec un vocabulaire, une vision des problèmes et des habitudes de travail différents des miens. La recherche d'un langage commun et de pistes de recherche au croisement de compétences complémentaires, si elle s'est parfois révélée difficile et chronophage, s'est toujours montrée enrichissante scientifiquement et humainement, et je retire de cette ouverture un intérêt essentiel dans mon activité de chercheur. Les Sections 2.1 à 2.4 présentent ainsi les contributions que j'ai pu apporter dans différents contextes applicatifs : acoustique sous-marine (2.1), imagerie hyperspectrale en astrophysique (2.2), contrôle non destructif (2.3) et pour deux modalités d'imagerie (imagerie micro-ondes et imagerie de fluorescence, Section 2.4).

Je m'intéresse également au développement d'algorithmes pour le traitement du signal dans un contexte plus générique. Je développe donc également une activité de recherche « amont », plus éloignée (mais non déconnectée !) des différentes applications sur lesquelles j'ai pu travailler. C'est notamment le cas de mes travaux sur l'optimisation parcimonieuse exacte en norme ℓ_0, *via* le projet ANR MIMOSA que je porte, qui seront décrits en Section 2.5.

2.1 Traitement de données acoustiques pour l'imagerie sous-marine

J'ai occupé de mai 2007 à septembre 2008 un poste de chercheur permanent à l'Ifremer (centre de Brest, Laboratoire de Technologie Halieutique), dans le domaine de l'acoustique sous-marine. Mes recherches ont concerné essentiellement le traitement d'images acoustiques dites de la « colonne d'eau », dans l'objectif d'estimation de paramètres biologiques

et écosystémiques [Simmonds et MacLennan, 2005]. C'est par exemple à partir de campagnes de mesures acoustiques que sont réalisées les estimations de biomasse de différentes espèces, servant de base à la fixation des quotas de pêche par l'Union Européenne. L'acquisition de données est réalisée au moyen de sondeurs acoustiques disposés sous la coque d'un navire océanographique. L'émission d'ondes acoustiques et la réception du signal d'écho le long de la colonne d'eau, conjuguées à l'avancée du navire, permet alors de former des images. Ma prise de poste coïncidait avec l'arrivée d'une nouvelle génération de sondeurs multi-faisceaux, permettant d'acquérir simultanément des données acoustiques selon différentes directions angulaires, formant ainsi des images tri-dimensionnelles, et débouchant sur de nouveaux problèmes méthodologiques.

Mes travaux ont surtout concerné le problème de la détection du fond marin dans de telles images. Dans la perspective de l'évaluation de biomasse sous-marine, le poisson se trouvant souvent à proximité du fond, il est primordial de séparer, dans le signal rétro-diffusé, la contribution du fond de celle des cibles biologiques d'intérêt [MacLennan *et al.*, 2004]. Si, à la verticale, l'écho acoustique renvoyé par le fond est très net (front de montée rapide, rapport signal sur bruit élevé), ce n'est plus le cas pour les échos reçus dans les faisceaux obliques de l'instrument, où l'écho peut « rouler » sur le sol et produire une signature plus complexe dans les données. Une manière de régulariser le problème consiste alors à introduire d'autres sources d'information, basées notamment sur la continuité spatiale du sol, aussi bien latéralement (les données acquises simultanément dans des faisceaux angulaires adjacents) que longitudinalement (dans la direction de l'avancée du navire). J'ai ainsi développé une méthode, basée sur un algorithme de filtrage particulaire [Arulampalam *et al.*, 2001], prenant en compte cette régularisation, permettant une estimation en ligne et peu coûteuse en calcul. Des résultats satisfaisants ont été obtenus sur des jeux de données de test, la mise en œuvre opérationnelle de cette méthode n'a, à ma connaissance, pas été réalisée suite à ma démission de l'Ifremer à l'été 2008. Ces travaux ont néanmoins fait l'objet d'une présentation en conférence et de la publication d'un article dans un journal international.

Publications associées

- Articles de journaux : [Bourguignon *et al.*, 2009, Trenkel *et al.*, 2009].
- Conférences internationales : [Bourguignon et Mazauric, 2008, Lefort *et al.*, 2008].

2.2 Imagerie hyperspectrale en Astrophysique et Planétologie

L'imagerie hyperspectrale ou spectro-imagerie concerne l'acquisition simultanée d'images dans un grand nombre de bandes spectrales, produisant des cubes de données à deux dimensions spatiales et une dimension spectrale. J'ai commencé à m'intéresser au traitement de données hyperspectrales lors de mon arrivée en février 2009 à l'Observatoire de la Côte d'Azur, dans le cadre du projet ANR DAHLIA [1] (*Données Astronomiques HyperspectraLes : algorIthmes Avancés*), impliquant cinq laboratoires : le Centre de Recherche Astrophysique de Lyon (CRAL), le Laboratoire d'Astrophysique de Toulouse et de Tarbes (LATT, désormais IRAP), le Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection (LSIIT, désormais ICUBE) à Strasbourg et les laboratoires Fizeau et Cassiopée de l'Observatoire de la Côte d'Azur à Nice, désormais fusionnés dans le laboratoire Lagrange. Il concernait le développement de méthodes d'exploitation

1. https://dahlia.oca.eu

de données hyperspectrales en astrophysique, en perspective de l'arrivée d'une nouvelle génération d'instruments d'observation couplant imagerie et spectroscopie. L'instrument MUSE[2] (Multi Unit Spectroscopic Explorer), en construction à l'époque (projet piloté par le CRAL [Bacon *et al.*, 2006]) et en partie opérationnel depuis 2013, fait figure de pionnier dans le domaine. Installé sur le Très Grand Télescope (VLT) de l'Observatoire Austral Européen, au Chili, MUSE permet d'acquérir des images d'environ 90 000 pixels, sur près de 4 000 longueurs d'onde couvrant le domaine visible et proche infra-rouge, là où les instruments de génération antérieure délivraient des images dans quelques canaux spectraux à large bande ou, à l'inverse, des mesures spectroscopiques focalisées en un point de l'espace.

Le contexte de l'observation astronomique diffère sensiblement du cas classique de l'imagerie hyperspectrale de télédétection terrestre (dont les problématiques relèvent essentiellement de la classification et de la séparation de sources), requérant le développement de méthodes spécifiques. Tout d'abord, la nature des scènes observées est fondamentalement différente : les mesures correspondent ici à la lumière émise par des sources très lointaines, étendues d'au plus quelques pixels, aux propriétés méconnues, et dont les spectres d'émission diffèrent d'une source à l'autre. Il y a donc très peu de cohérence dans les données au niveau spatial. Ensuite, comme toute observation astronomique depuis un instrument situé au sol, les mesures subissent l'*étalement* spatial et spectral de la lumière émise, dû aux turbulences atmosphériques et à la bande passante limitée du dispositif d'acquisition [Villeneuve, 2012]. Enfin, en raison de leur éloignement, les sources astrophysiques émettent très peu de lumière en direction de l'observateur et sont polluées par de nombreuses émissions parasites ; les données présentent donc un très fort niveau de bruit.

J'ai abordé l'analyse de telles données hyperspectrales sous l'angle du débruitage et de la déconvolution. Afin de prendre en compte le niveau de bruit très élevé, nous avons proposé des modèles imposant de fortes contraintes au niveau spectral, où les spectres recherchés admettent une décomposition parcimonieuse dans un dictionnaire de formes élémentaires spécifiquement construit pour ces données, que nous avons défini en collaboration avec des astronomes. Contrairement aux modèles parcimonieux standard utilisant des transformées (en ondelettes par exemple [Mallat, 2008]), un tel dictionnaire s'avère plus judicieux, exploitant un maximum de connaissance physique sur les sources recherchées. La sélection d'une composante dans le dictionnaire est alors associée à la détection d'une composante ayant une interprétation physique (essentiellement, des raies d'émission ou d'absorption lumineuse, signalant la présence d'un élément chimique, ou encore une discontinuité dans le spectre correspondant à la cassure de Lyman [Tennyson, 2005]). Une telle finesse de modélisation s'effectue cependant au détriment du coût de calcul associé. Nous avons alors développé des algorithmes d'estimation parcimonieuse dédiés, reposant sur l'optimisation de critères pénalisés par la norme ℓ_1 et sur un algorithme glouton. Dans un premier temps, nous avons considéré la restauration des spectres pris séparément, pour des raisons évidentes de complexité calculatoire. Nous avons ensuite abordé la restauration conjointe spatiale et spectrale de cubes d'étendue spatiale limitée, afin de prendre en compte la réponse spatiale du système d'acquisition. L'ensemble de ces travaux sera détaillé au Chapitre 4.

Mes travaux sur l'imagerie hyperspectrale ont connu une pause, coïncidant avec la fin de mon post-doctorat à l'Observatoire de la Côte d'Azur et ma prise de poste à l'École Centrale de Nantes en septembre 2011. Ayant démarré d'autres thématiques de recherche

2. https ://www.eso.org/sci/facilities/develop/instruments/muse.html

et ne disposant plus de financement dédié, j'ai arrêté de travailler sur ce sujet en septembre 2012, avec l'encadrement d'un stage de Master encadré à l'École Centrale de Nantes. Plus récemment, je travaille cependant à nouveau sur le traitement de données hyperspectrales et en particulier sur des questions de démélange spectral parcimonieux, *via* le projet ANR MIMOSA. Le contexte est différent, puisqu'il vise des applications en planétologie, notamment par une collaboration que je mène avec Frédéric Schmidt du laboratoire Géosciences Paris Sud (GEOPS). Il y est également question de parcimonie et de dictionnaires, représentant cette fois des spectres de minéraux potentiellement présents dans la scène observée. Les problèmes étant de plus petite taille (la dimension des spectres n'excède pas quelques centaines de longueurs d'onde et le nombre de composants recherchés est de quelques unités), nous les abordons sous l'angle de l'optimisation parcimonieuse exacte en norme ℓ_0 ; ils seront donc abordés au Chapitre 6 dédié à cette thématique (voir plus particulièrement le § 6.3).

Enfin, depuis quelques mois, je m'intéresse de nouveau à des problèmes de spectro-imagerie astronomique, *via* une collaboration naissante avec l'Institut de Radio-Astronomie Millimétrique (IRAM, Grenoble) et le Laboratoire d'Études du Rayonnement et de la Matière en Astrophysique et Atmosphères (LERMA, Observatoire de Paris). Il s'agit cette fois d'ondes radio (longueurs d'onde millimétriques et centimétriques), permettant d'inspecter les nuages de gaz entourant les zones de formation d'étoiles. Le projet Orion-B [3], porté par ces deux laboratoires et dont je suis désormais partenaire, envisage ainsi d'acquérir des données contenant environ un million de pixels, avec plus de 200 000 canaux spectraux par pixel. Si les données sont d'une toute autre dimension par rapport à celles d'un instrument comme MUSE, des points communs existent : ici aussi, l'information à exploiter est avant tout dans la dimension spectrale, où l'on recherche des raies de positions et largeurs variables. L'exploitation de grandes quantités de données pose également des questions statistiques de détection en grande dimension, qui se posaient déjà à l'époque du projet ANR DAHLIA.

Supervision de stages de Master

- Jiayi Hou (2019) : **Sparse unmixing methods for hyperspectral imaging**. Master *Control and Robotics*, parcours *Signal and Image Processing*, École Centrale de Nantes.
- Julien Picaud (2012) : **Restauration de cubes hyperspectraux en Astrophysique**. Master *Automatique, Robotique et Informatique appliquée*, parcours Automatique, Signal et Image, École Centrale de Nantes.
- Benjamin Trémoulhéac (2010) : **Représentations parcimonieuses de données hyperspectrales en astrophysique : aspects informationnels et algorithmiques**. M2 Recherche *Signal, Image, Acoustique et Optimisation*, Université de Toulouse. Co-encadrement avec David Mary (Maître de Conférences, Université de Nice).

Encadrement doctoral

- Ramzi Ben Mhenni, **Programmation mixte en nombres entiers pour l'optimisation parcimonieuse en traitement du signal.** École Centrale de Nantes.

3. http://iram.fr/~pety/ORION-B/

Directeur de thèse (80 %), co-encadrement Jordan Ninin (20 %). Financement *via* une allocation ministérielle. Thèse démarrée en octobre 2016.

Publications associées

- Articles de journaux : [Schmidt et Bourguignon, 2019, Bourguignon *et al.*, 2012, Bourguignon *et al.*, 2011c, Serre *et al.*, 2010].
- Conférences avec actes et comité de lecture : [Ben Mhenni *et al.*, 2018, Ben Mhenni *et al.*, 2017, Picaud et Bourguignon, 2013, Shen *et al.*, 2012, Paris *et al.*, 2011, Schmidt *et al.*, 2011, Bourguignon *et al.*, 2011b, Bourguignon *et al.*, 2011a, Villeneuve *et al.*, 2011, Bourguignon *et al.*, 2010].

2.3 Évaluation et contrôle non destructifs : déconvolution, imagerie et tomographie

L'évaluation et le contrôle non destructifs (ECND) concernent l'ensemble des méthodes permettant de caractériser l'état d'une structure ou d'un matériau sans les dégrader. On parle de contrôle lorsqu'il s'agit de s'assurer du bon état de l'objet inspecté, ou au contraire d'y détecter et caractériser des défauts. L'évaluation vise plutôt à caractériser les propriétés de matériaux de structures complexes, de plus en plus utilisés dans des contextes industriels (matériaux composites, matériaux poreux, bétons, bio-matériaux, ...)

De nombreuses modalités d'inspection mettent en œuvre un phénomène physique (ultrasons, courants de Foucault, rayons X, RADAR, thermographie, ...) et requièrent une analyse relevant de la méthodologie de l'inversion : ayant observé la réponse du matériau à une sollicitation, il s'agit de remonter aux causes qui l'ont produite. Le CND suit ainsi une évolution que l'on retrouve dans la plupart des disciplines scientifiques impliquant un processus de mesure : les données disponibles étant de plus en plus nombreuses et complexes, voire résultant de mesures conjointes *via* plusieurs modalités, l'information disponible s'en retrouve profondément enrichie. L'exploitation de cette richesse requiert alors le développement de méthodes avancées et adaptées à la complexité des données.

J'ai commencé à travailler sur des problèmes de CND lors de ma prise de poste à l'École Centrale de Nantes en 2011, poste pour lequel le renforcement de l'activité de recherche en CND était affiché comme une priorité par mon laboratoire. Cet engagement s'inscrit dans un environnement scientifique régional où le CND est très présent, tant au niveau académique qu'industriel, la région Pays de la Loire comptant des acteurs en CND bien identifiés à l'échelle nationale et internationale. J'ai ainsi pu bénéficier, dès mon arrivée à l'École Centrale de Nantes, des moyens apportés par le programme de recherche ECND-PdL (Évaluation et Contrôle Non Destructifs en Pays de la Loire), financé par la région Pays de la Loire de 2010 à 2014, visant à fédérer les compétences d'une quinzaine de laboratoires régionaux. Ce programme a en particulier financé la thèse de doctorat d'Ewen Carcreff, qui a démarré concomitamment à mon arrivée à l'École Centrale de Nantes et que j'ai co-encadrée. L'animation scientifique liée à cet environnement m'a permis de rencontrer un certain nombre d'acteurs en CND à l'échelle régionale, ayant donné naissance à plusieurs collaborations pluri-disciplinaires, académiques et industrielles. Ce programme, achevé en 2014, a débouché sur la création d'un Groupement d'Intérêt Scientifique (GIS ECND-PdL [4]), dont je suis membre du comité de direction et du conseil scientifique et le correspondant pour le Laboratoire des Sciences du Numérique de Nantes.

4. http://ecnd-pdl.fr/

2.3.1 ECND ultrasonore

Le CND ultrasonore et l'imagerie acoustique constituent la plus grande part de mon activité de recherche autour du CND. Je collabore sur ce thème avec le Laboratoire d'Acoustique de l'Université du Maine (LAUM) depuis 2011 et le lancement de la thèse d'Ewen Carcreff. Le CND par ultrasons [Krautkramer et Krautkramer, 1990] repose sur l'émission d'une onde acoustique à l'intérieur d'un matériau et l'analyse du signal reçu : lorsque l'onde rencontre un changement d'impédance acoustique, une partie se transmet à travers le milieu et l'autre partie se réfléchit [Saniie et Nagle, 1989]. La détection d'échos permet alors de localiser des changements d'impédance dus à la présence de défauts ou d'hétérogénéités dans le milieu inspecté. Le signal ultrasonore reçu peut être modélisé comme la convolution de la forme de l'onde ultrasonore émise avec la séquence de réflectivité du milieu, traduisant les changements d'impédance acoustique. Si les défauts rencontrés sont isolés et peu nombreux, cette séquence est de nature parcimonieuse. La détection et la localisation des défauts peut alors être formalisée comme un problème de déconvolution parcimonieuse [O'Brien *et al.*, 1990, Zala, 1992].

Dans la thèse d'Ewen Carcreff [Carcreff, 2014a], nous avons cherché à raffiner le modèle convolutif classique. Nous avons d'abord proposé un modèle dit *à haute résolution*, où la séquence parcimonieuse recherchée est discrétisée à une période plus fine que la période d'échantillonnage des données. La plupart des méthodes de déconvolution repose en effet sur un modèle de convolution discrète, qui n'est en général qu'une approximation résultant de la discrétisation d'un modèle convolutif à temps continu à la période correspondant à l'échantillonnage. Une discrétisation plus fine permet alors de réduire l'erreur de modélisation, au prix de la perte de la structure convolutive du modèle permettant des calculs rapides dans les algorithmes d'estimation. Nous avons montré que ce modèle suréchantillonné pouvait s'interpréter comme un système MISO (*Multiple Inputs, Single Output*), pour lequel les calculs peuvent s'effectuer à partir de convolutions multiples. Nous avons alors étendu les algorithmes classiques d'optimisation parcimonieuse (approche en norme ℓ_1 et algorithmes gloutons) au cadre de la déconvolution MISO [Carcreff *et al.*, 2013a, Carcreff *et al.*, 2013b].

Nous avons par ailleurs étudié la prise en compte de modèles physiques de propagation acoustique dans des algorithmes de déconvolution. Le modèle convolutif suppose en effet une invariance temporelle de la forme d'onde transmise dans le matériau inspecté. Dans de nombreux types de matériaux cependant (plastiques, résines, matériaux poreux ou composites, ...), la propagation acoustique est soumise à des effets d'atténuation et de dispersion fréquentielle qui modifient cette forme d'onde, en général par la perte des hautes fréquences. Si l'on perd alors le caractère convolutif du modèle, la prise en compte d'un modèle plus précis (et toujours linéaire) dans des algorithmes de reconstruction parcimonieuse a permis d'améliorer la détection de défauts dans des matériaux atténuants [Carcreff *et al.*, 2014b]. Dans un autre volet de la thèse, nous nous sommes intéressés à l'optimisation parcimonieuse en norme ℓ_0 pour la déconvolution, en proposant un algorithme de type glouton permettant une exploration plus complète que les approches standard, tout en limitant l'augmentation du coût calculatoire. La méthode exploite la forme d'onde intervenant dans le modèle pour identifier, préalablement à tout calcul, les combinaisons d'atomes susceptibles d'interférer et de perturber la sélection à chaque itération de l'algorithme [Carcreff, 2014b].

Les travaux précédents ont essentiellement concerné le traitement de signaux monodimensionnels (A-scans). Cependant, l'avènement de nouvelles générations de dispositifs ultrasonores constitués de réseaux de capteurs permet désormais d'acquérir des données mul-

tidimensionnelles, correspondant aux signaux de chaque paire émetteur-récepteur (FMC, Full Matrix Capture), rendant possible la reconstruction de cartes spatiales de la réflectivité par calcul numérique [Holmes *et al.*, 2005]. Je poursuis donc des travaux abordant la reconstruction d'images à partir de données FMC sous l'angle des problèmes inverses, portant à la fois sur les modèles de données (prise en compte de la forme d'onde et de l'atténuation) et sur les algorithmes de reconstruction [Laroche *et al.*, 2018, Laroche *et al.*, 2019]. Ces recherches sont toujours menées en collaboration avec le LAUM, notamment par l'encadrement de la thèse en entreprise (CIFRE) de Nans Laroche démarrée à l'automne 2017, en partenariat avec la société nantaise DB-SAS.

De manière plus marginale, je me suis intéressé à la mesure par ultrasons de l'épaisseur de pièces ou de matériaux. Ce problème est rencontré par exemple pour contrôler l'épaisseur de conduits soumis à de fortes contraintes, comme les circuits d'alimentation en eau des centrales électriques ou les conduites de pipelines, ou encore pour la mesure d'épaisseur de revêtements dans l'industrie automobile ou aéronautique. Nous avons montré que de telles mesures ultrasonores pouvaient bénéficier d'une modélisation beaucoup plus contrainte : la séquence de réflectivité est alors composée d'impulsions régulièrement espacées et d'amplitudes à décroissance géométrique. Une approche paramétrique a alors été proposée, permettant de résoudre un problème de déconvolution myope, où les paramètres de la séquence de réflectivité et ceux de la forme d'onde ultrasonore sont estimés conjointement par l'optimisation d'un critère des moindres carrés non linéaires. Ces travaux ont été développés *via* des projets d'étudiants de l'École Centrale de Nantes depuis 2014 et présentés dans des congrès d'acoustique [Bourguignon *et al.*, 2016b] et de contrôle non destructif [Bourguignon *et al.*, 2016a].

Un autre axe de recherche dans la poursuite de la thèse d'Ewen Carcreff a concerné la caractérisation acoustique de matériaux poro-élastiques, fréquemment utilisés dans l'industrie (mousses pour l'isolation phonique ou thermique, mousses métalliques dans l'industrie automobile ou aéronautique, par exemple). D'un point de vue acoustique, de tels matériaux se comportent comme des milieux biphasiques, composés d'une phase solide et d'une phase fluide, dans lesquels deux types d'onde correspondant aux deux phases se propagent simultanément, chacune d'elles obéissant à des lois d'atténuation et de dispersion différentes. Nous avons donc travaillé sur des méthodes d'identification des paramètres de propagation de ces ondes, pouvant fournir de nouveaux outils d'évaluation non destructive de l'état du matériau. Ces recherches, également en collaboration avec le LAUM, ont essentiellement été financées par la région Pays de la Loire *via* le projet DECIMAP (DEConvolution Impulsionnelle pour les MAtériaux Poreux). Côté LS2N, ce projet a financé le post-doctorat de Nizar Bouhlel pour une durée de 17 mois. Des résultats préliminaires ont été présentés dans des congrès d'acoustique et de CND [Bouhlel *et al.*, 2016, Bouhlel et Bourguignon, 2016, Bouhlel *et al.*, 2017] et un article de journal est en cours de finalisation.

2.3.2 Tomographie électromagnétique pour l'auscultation d'ouvrages en béton armé

Je collabore depuis l'automne 2017 avec le Laboratoire « Auscultation, Modélisation, Expérimentation des Infrastructures de Transport » (LAMES) de l'IFSTTAR, *via* le co-encadrement de la thèse de Marie Antoinette Al Hajj portant sur le développement de méthodes non destructives pour l'auscultation d'ouvrages de génie civil en béton armé. Si le béton est un matériau réputé solide et stable dans le temps, les ouvrages en béton armé sont souvent soumis à des agents agressifs pouvant endommager les armatures, comme la

pénétration des ions chlorures en milieu marin ou des sels de déverglaçage [Raharinaivo *et al.*, 1998]. Le développement de méthodes non destructives d'auscultation de ces ouvrages est alors primordial pour le suivi de leur état de dégradation [Breysse et Abraham, 2005]. Cependant, peu de modalités de mesure sont sensibles à la pénétration de tels agents aux profondeurs des armatures et le développement de méthodes non destructives est encore essentiellement dans une phase de recherche « amont ».

Nous nous intéressons, dans la thèse de Marie Antoinette Al Hajj, à des méthodes exploitant des mesures de résistivité électrique et de permittivité di-électrique, toutes deux sensibles à la teneur en eau et en ions chlorures du matériau inspecté. Il s'agit, à partir de mesures électriques de surface, de remonter à un profil en profondeur de ces quantités à l'intérieur du béton. Ce problème inverse de type tomographique est particulièrement difficile car en pratique, le nombre de mesures est faible : quelques dizaines de mesures en résistivité et quelques unités en permittivité. Des travaux préalables ont cependant montré la possibilité d'estimer des gradients de teneur en eau à partir de mesures de résistivité d'une part, et de permittivité d'autre part [Fares, 2015, Fares *et al.*, 2016]. La thèse a donc pour objectif de coupler les deux modalités dans une procédure d'inversion conjointe. Une difficulté majeure provient également du coût de calcul associé à la procédure d'inversion : chaque évaluation du modèle direct, prédisant les valeurs des observables électriques pour un profil de teneur en eau donné, requiert la résolution d'équations aux dérivées partielles à trois dimensions – réalisées à l'aide d'un logiciel dédié – pouvant prendre plus d'une minute sur une machine de calcul standard. Compte tenu du faible nombre de données, nous avons opté pour une modélisation paramétrique du profil de teneur en eau recherché, permettant de contraindre la solution. Nous avons alors développé une procédure d'estimation par moindres carrés non linéaires, basée sur un schéma numérique de type Levenberg-Marquardt. Cette méthode est en cours de validation sur des données simulées et fera l'objet, à l'automne 2019, d'une campagne de validation sur données réelles et de la rédaction d'un article de journal.

2.3.3 Profilométrie laser

J'ai réalisé, au cours de l'année 2016, une prestation pour l'Institut de Recherche Technologique (IRT) Jules Verne, portant sur le traitement de signaux issus d'un profilomètre laser pour des applications de soudage. L'application visée concerne le déploiement d'un robot réalisant des opérations d'assemblage des plaques métalliques pour la construction de navires. Ce robot est équipé d'un profilomètre laser mesurant la forme du joint de soudure en construction, à partir de laquelle il s'agit de positionner une torche de soudage, en visant les angles du profil.

Je me suis pour cela orienté vers une modélisation parcimonieuse du signal de profilométrie, où le dictionnaire est constitué de fonctions rampe, dont le point initial peut être localisé en toute position du profil. Ce modèle revient à approcher le profil par un modèle linéaire par morceaux, comportant un faible nombre de points de changement de pente. Cette prestation a essentiellement consisté à transférer le savoir-faire sur les méthodes parcimonieuses aux ingénieurs en charge du projet à l'IRT, et d'adapter avec eux l'algorithme d'estimation parcimonieuse SBR [Soussen *et al.*, 2011] à cette problématique.

Supervision de stages de Master

- Yan Li (2013) : *Algorithmes de déconvolution parcimonieuse par pénalisation* ℓ_0, Master *Automatique, Robotique et Informatique appliquée*, parcours Automatique,

Signal et Image, École Centrale de Nantes. Co-encadrement avec E. Carcreff (doctorant, Université du Maine).

Encadrement doctoral

- Nans Laroche, *Méthodes d'imagerie ultrasonore avancées et rapides pour le contrôle non destructif de matériaux atténuants et diffusants*. École Centrale de Nantes. Encadrement à 40 %. Directeur de thèse Jérôme Idier (30 %), co-encadrant Aroune Duclos (30 %). Financement industriel (convention CIFRE). Thèse démarrée en novembre 2017.
- Marie-Antoinette AlHajj, *Inversion conjointe de mesures non destructives pour découpler des gradients multiples dans le béton d'enrobage de structures en béton armé*. IFSTTAR, École Centrale de Nantes. Encadrement à 30 %. Directrice de thèse Géraldine Villain (40 %), co-encadrant Sérgio Palma Lopes (30 %). Financement IFSTTAR. Thèse démarrée en octobre 2017.
- Ewen Carcreff, *Déconvolution adaptative de signaux ultrasonores pour le contrôle non destructif par ultrasons*, Université du Maine. Encadrement à 50%. Directeur de thèse Laurent Simon (20%), co-directeur Jérôme Idier (30%). Financement de la région Pays de la Loire. Thèse soutenue en novembre 2014.

Publications associées

- Articles de journaux : [Carcreff *et al.*, 2015, Carcreff *et al.*, 2014b].
- Conférences avec actes et comité de lecture : [Laroche *et al.*, 2019, Laroche *et al.*, 2018, Carcreff *et al.*, 2017, Carcreff *et al.*, 2013c, Carcreff *et al.*, 2013b, Bourguignon *et al.*, 2011d, Carcreff *et al.*, 2013a].
- Conférences : [Bouhlel *et al.*, 2016, Bouhlel *et al.*, 2017, Bouhlel et Bourguignon, 2016, Carcreff *et al.*, 2014a, Bourguignon *et al.*, 2016b, Bourguignon *et al.*, 2016a].

2.4 Reconstruction d'images *via* différentes modalités d'acquisition

Depuis mon arrivée à l'École Centrale de Nantes, j'ai également travaillé sur deux thématiques liées à la reconstruction d'images, initiées par Jérôme Idier, par le co-encadrement de deux thèses soutenues en 2016 et 2018.

2.4.1 Imagerie micro-ondes / tomographie de diffraction

La première thèse concerne l'imagerie micro-ondes et a été réalisée par Corentin Friedrich, de 2012 à 2016, en cotutelle entre l'École Centrale de Nantes et Polytechnique Montréal [Friedrich, 2016]. L'imagerie micro-ondes a fait l'objet d'un fort intérêt au cours des dix dernières années, en particulier en raison du faible coût des dispositifs d'acquisition et de sa faible nocivité, avec en particulier des applications en imagerie biomédicale, en géosciences et en contrôle non destructif. À partir d'illuminations, sous différents angles de vue, d'un objet par un champ électromagnétique incident et des mesures du champ diffusé par cet objet, il s'agit d'estimer la permittivité diélectrique (complexe) en tout point de l'objet. Lorsque la taille de l'objet est de l'ordre de la longueur d'onde utilisée,

des phénomènes de diffraction se produisent, rendant le modèle d'observation non linéaire en les quantités d'intérêt [Zhang *et al.*, 2003, Abubakar et van den Berg, 2004, Pastorino, 2010]. La reconstruction des propriétés diélectriques à l'intérieur de l'objet est un problème inverse difficile requérant des algorithmes itératifs très coûteux en temps. En particulier, l'évaluation du modèle direct (le calcul numérique des champs électriques diffractés pour une distribution de permittivité connue dans l'objet), requise à chaque itération, nécessite la résolution d'un grand nombre de systèmes linéaires de la taille du nombre d'inconnues, qui devient vite très grande pour des problèmes d'imagerie 3D.

Dans la thèse de Corentin Friedrich, nous avons travaillé sur des algorithmes de résolution par blocs de ces systèmes linéaires, exploitant le fait qu'ils possèdent tous la même matrice. Nous avons adapté un algorithme dit *block-BiCGStab* [El Guennouni *et al.*, 2003] au contexte de l'imagerie micro-ondes et inséré ces calculs dans un algorithme itératif d'optimisation d'un critère des moindres carrés régularisé. Le gain en temps de calcul s'est révélé d'autant plus significatif que le problème était difficile, permettant en particulier la reconstruction d'objets pour lesquels aucun résultat de qualité comparable n'a été trouvé dans la littérature [Friedrich *et al.*, 2015a, Friedrich *et al.*, 2015b].

2.4.2 Super-résolution pour la microscopie à éclairements structurés

La seconde thèse est celle de Penghuan Liu, réalisée de 2014 à 2018, sur la reconstruction d'images de microscopie de fluorescence à partir d'éclairements structurés [Liu, 2018]. Comme toute modalité d'imagerie requérant un microscope optique, la microscopie de fluorescence est limitée par la diffraction de la lumière, avec une limite de résolution de l'ordre de $\lambda/2\mathrm{NA} \sim 200$ nm, où λ est la longueur d'onde de la lumière et NA est l'ouverture numérique du système optique. La microscopie à éclairements structurés (*Structured Illumination Microscopy*, SIM) est une technique permettant, idéalement, de réduire de moitié cette limite. Elle consiste à éclairer l'objet avec des motifs structurés et non pas uniformes qui, par un effet de modulation sur la scène éclairée avant son passage dans le système optique, rendent la mesure sensible à des hautes fréquences non observables par un éclairement uniforme. L'illumination successive sous différents éclairements permet alors, par un calcul postérieur, d'améliorer la résolution des images reconstruites jusqu'à un facteur deux [Heintzmann et Gustafsson, 2009]. L'imagerie SIM standard repose cependant sur la génération parfaitement maîtrisée de motifs d'illumination harmoniques, ce qui est difficile et très coûteux en pratique. Une approche dite blind-SIM a alors été proposée [Mudry *et al.*, 2012], où l'acquisition est effectuée *via* un grand nombre de d'illuminations successives de nature aléatoire, connues uniquement à travers leurs propriétés statistiques.

Nous avons étudié les propriétés d'un estimateur dit *marginal*[5], reposant sur l'ajustement des statistiques du second ordre du modèle aux statistiques empiriques des données. Nous avons exploré, d'une part, la capacité théorique d'un tel dispositif à produire une image à haute résolution *via* les propriétés asymptotiques de l'estimateur à nombre infini de données. Nous avons montré que les statistiques d'ordre deux des données contiennent l'information suffisante pour identifier les composantes fréquentielles de l'objet jusqu'au double de la fréquence de coupure du dispositif optique, permettant d'obtenir les mêmes performances que l'approche SIM à éclairements harmoniques parfaitement contrôlés [Idier *et al.*, 2018]. D'un point de vue pratique, cependant, le calcul numérique de cet

5. Le terme *marginal* est utilisé par opposition à l'approche *conjointe* préalablement existante dans la littérature [Mudry *et al.*, 2012], laquelle cherche à estimer conjointement l'objet éclairé et les motifs d'illumination, qui sont en fait des paramètres de nuisance.

estimateur s'avère prohibitif pour des images de taille réaliste : la mise en œuvre d'une procédure d'inversion itérative requiert la résolution, à chaque itération, d'un système de taille $N \times N$, où N est le nombre de pixels dans l'image, dont le coût de calcul est de l'ordre de N^3. Dans un second volet de la thèse, nous avons alors construit une méthode d'estimation approchée, reposant sur une troncature de la matrice de covariance du modèle. Plus précisément, nous avons considéré la restriction de la covariance spatiale à des patches de faible étendue, négligeant ainsi les corrélations entre les pixels éloignés. La complexité calculatoire a ainsi pu être réduite à l'ordre de $N^2 \log N$, sans que la qualité de la solution en soit fortement affectée [Liu *et al.*, 2017].

Encadrement doctoral

- Penghuan Liu, **Statistical and numerical optimization for speckle blind structured illumination microscopy.** École Centrale de Nantes. Encadrement à 30 %. Directeur de thèse Jérôme Idier (70 %). Financement du gouvernement chinois *via* le *China Scholarship Council (CSC)*. Thèse soutenue en mai 2018.
- Corentin Friedrich, **Méthodes de reconstruction en tomographie de diffraction 3-D.** École Centrale de Nantes, thèse en co-tutelle avec Polytechnique Montréal. Encadrement à 30 %. Directeur de thèse Jérôme Idier (40 %), co-directeur Yves Goussard (Polytechnique Montréal, 30 %). Financement École Centrale de Nantes et Polytechnique Montréal. Thèse soutenue en septembre 2016.

Publications associées

- Articles de journaux : [Idier *et al.*, 2018, Labouesse *et al.*, 2017, Friedrich *et al.*, 2015b].
- Conférences avec actes et comité de lecture : [Labouesse *et al.*, 2016, Friedrich *et al.*, 2015a, Liu *et al.*, 2017, Labouesse *et al.*, 2015].

2.5 Optimisation exacte en norme ℓ_0

Si les activités précédemment décrites ont essentiellement concerné des applications en traitement du signal et de l'image, ce dernier thème est résolument plus « amont ». Il concerne le développement d'algorithmes de résolution *exacte* de problèmes d'optimisation faisant intervenir la « norme » ℓ_0 pour l'approximation parcimonieuse. Le terme exact signifie que le résultat est garanti de produire le minimum global du problème fonction de coût considérée. Ce sujet se situe en dernier dans l'exposé de mes recherches, traduisant l'évolution d'une partie de mon activité vers des questions avant tout algorithmiques. L'optimisation exacte de critères parcimonieux constitue une thématique qui m'est propre au sein de mon équipe de recherche et occupe désormais une place centrale de mon activité.

La notion de parcimonie a jalonné mon parcours de chercheur, et on la retrouve en bonne place dans les chapitres antérieurs de ce manuscrit. Si j'y ai développé des algorithmes spécifiques, la taille souvent grande des problèmes abordés (ou des contraintes de mise en œuvre rapide voire en temps réel dans le cas du contrôle non destructif), m'a plutôt orienté vers des algorithmes en norme ℓ_1 et des approches gloutonnes, efficaces en temps de calcul. La « norme ℓ_0 »[6] représente pourtant la mesure la plus exacte

6. Le terme de norme est clairement un abus de langage, puisque cette fonction ne vérifie pas la propriété d'homogénéité : on a $\forall c \neq 0, \|c\,\boldsymbol{x}\|_0 = \|\boldsymbol{x}\|_0$. Ce n'est pas non plus une quasi-norme ni une pseudo-norme.

de la parcimonie, qui compte le nombre de composantes non nulles dans un vecteur : $\|\boldsymbol{x}\|_0 := \text{Card}\{p|x_p \neq 0\}$. L'approximation d'un vecteur $\boldsymbol{y}$ dans un dictionnaire $\mathbf{A}$ *sous contrainte de parcimonie* peut alors se formuler par :

$$\min_{\boldsymbol{x}\in\mathbb{R}^P} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\|^2 \text{ sous la contrainte (s.c.) } \|\boldsymbol{x}\|_0 \leq K. \tag{2.1}$$

De nombreux articles en traitement du signal introduisent cette formulation pour s'en détourner au profit d'approches sous-optimales, pour des raisons évidentes de coût calculatoire, le problème étant NP-difficile [Natarajan, 1995, Bienstock, 1996]. En revanche, un argument souvent rencontré considère que la seule façon de résoudre le problème (2.1) de manière globale est de résoudre l'ensemble des $C_P^K = \frac{P!}{K!(P-K)!}$ problèmes possibles de moindres carrés à K composantes pour en prendre la meilleure [7]. Ce n'est pas le cas, et les problèmes d'optimisation *à cardinalité contrainte* ont donné lieu à de nombreux travaux en optimisation [Li *et al.*, 2006, Shaw *et al.*, 2008, Bertsimas et Shioda, 2009, Cui *et al.*, 2013, par exemple]. Un autre argument souvent avancé justifiant le recours, par exemple, à une formulation en norme ℓ_1, se réfère aux travaux autour de l'échantillonnage compressé [Eldar et Kutyniok, 2012] : sous certaines hypothèses imposant que les colonnes de la matrice $\mathbf{A}$ soient suffisamment décorrélées entre elles, la résolution d'un problème en norme ℓ_1 permet de résoudre (2.1). Cependant, dans la plupart des problèmes inverses difficiles – et c'est le cas de toutes les applications considérées dans ce manuscrit – ces conditions ne sont pas valides et les résultats expérimentaux montrent clairement la nature sous-optimale des approches en norme ℓ_1 et des algorithmes gloutons.

J'ai commencé à m'intéresser à l'optimisation globale du problème ℓ_0 suite à des discussions menées avec Hervé Carfantan (IRAP, Toulouse) et des collègues spécialistes de recherche opérationnelle, Jordan Ninin (Lab-STICC, Brest) et Marcel Mongeau (ENAC, Toulouse). Un premier travail a alors vu le jour, notamment grâce au soutien du GdR ISIS *via* un projet *jeunes chercheurs* (2013), où différents problèmes d'approximation parcimonieuse sont reformulés en programmes en nombres mixtes (MIPs, mixed integer programs). Les MIPs sont des problèmes d'optimisation intégrant des contraintes d'intégrité sur une partie des variables [Wolsey, 1998], qui s'avèrent particulièrement adaptés à l'optimisation ℓ_0 : en introduisant des variables de décision binaires $b_p \in \{0,1\}$ encodant la nullité des composantes de $\boldsymbol{x}$ ($b_p = 1 \Leftrightarrow x_p \neq 0$), la norme ℓ_0 s'écrit comme la simple somme $\sum_{p=1}^P b_p$, générant une contrainte linéaire dans le problème (2.1). Nous avons notamment montré dans [Bourguignon *et al.*, 2015, Bourguignon *et al.*, 2016c], en utilisant le logiciel commercial CPLEX pour la résolution de ces MIPs, que :

i) pour des problèmes inverses parcimonieux de taille modérée mais difficiles (avec au plus une dizaine de composantes non nulles dans un dictionnaire de 200 composantes), l'optimum global du problème (2.1) peut être calculé de manière exacte et garantie. Le temps de calcul certes bien plus élevé que les algorithmes gloutons ou en norme ℓ_1, mais sans commune mesure avec l'évaluation explicite des C_P^K solutions réalisables. Celui-ci s'est également avéré fortement dépendent de la qualité de l'approximation, les problèmes à fort niveau de bruit étant bien plus difficiles à résoudre.

7. Dans la littérature issue de la communauté du traitement du signal, on trouve cependant l'argument de réduire le nombre de combinaisons dans l'article de Tropp et Wright [Tropp et Wright, 2010] : « **Brute force.** Search through all possible support sets, *possibly using cutting-plane methods to reduce the number of possibilities.* »

ii) Lorsque son calcul est réalisable, la solution globale obtenue est meilleure que celle des algorithmes classiques en termes de localisation du *support* (l'identification des composantes non nulles).

Suite à ces premiers résultats encourageants, j'ai alors orienté mes recherches vers la construction d'algorithmes de résolution dédiés, reposant sur le principe de séparation et évaluation (*branch-and-bound*) à la base des solveurs MIP. L'argument principal ayant motivé cette orientation réside en ce que les solveurs commerciaux sont développés pour résoudre des classes de problèmes les plus générales possibles, alors que le problème ℓ_0 est au contraire un MIP très particulier. Cette ligne de recherche est au cœur du projet ANR *jeunes chercheurs* MIMOSA que je porte, démarré en janvier 2017, et de la thèse en cours de Ramzi Ben Mhenni.

La méthode de branch-and-bound [Wolsey, 1998] repose sur la construction implicite d'un arbre de décision binaire, où un problème est divisé en sous-problèmes disjoints (étape de *séparation*). Dans notre cas, cette séparation consiste à prendre une décision sur une variable : est-elle nulle ($b_p = 0$) ou non-nulle ($b_p = 1$) ? À chaque nœud de l'arbre correspond ainsi un sous-problème, pour lequel une partie des variables binaires est fixée et l'autre partie reste indéterminée. L'*évaluation* d'un sous-problème consiste alors à calculer, de façon peu coûteuse, une borne inférieure de l'ensemble des sous-problèmes non encore explicités qu'il contient. Une manière classique de procéder est de calculer la *relaxation continue* de l'ensemble des variables binaires non encore déterminées : $b_p \in \{0, 1\}$ devient $b_p \in [0, 1]$. Le problème d'optimisation associé relève alors de l'optimisation continue (et convexe dans notre cas). Si cette borne inférieure est supérieure à la meilleure valeur connue de la fonction objectif (qui est donc une borne supérieure sur la valeur optimale), alors il est garanti que l'optimum ne peut être obtenu dans un de ces sous-problèmes, qui peuvent alors être éliminés des combinaisons possibles. L'algorithme explore ainsi implicitement l'ensemble des solutions et fournit le minimum global en un nombre fini d'itérations. Dans le pire des cas (si aucune élimination n'est réalisée), sa complexité est celle d'une recherche combinatoire exhaustive.

Nous avons proposé des stratégies de branchement et d'exploration (étape de séparation) spécifiquement construites pour ce problème, en grande partie inspirées de la construction des algorithmes gloutons d'estimation parcimonieuse [Tropp et Wright, 2010]. Par ailleurs, nous avons montré que les problèmes de relaxation continue mis en jeu lors de l'étape d'évaluation sont équivalents à des problèmes d'optimisation convexe faisant intervenir un terme en norme ℓ_1 opérant sur une partie des variables et des contraintes supplémentaires de borne. Nous avons alors proposé plusieurs algorithmes d'optimisation dédiés à ce problème, reposant sur le principe des méthodes homotopiques [Donoho, 2006] ou sur celui des *ensembles actifs* [Osborne *et al.*, 2000, Lee *et al.*, 2007]. Nous avons montré dans [Ben Mhenni *et al.*, 2019a, Ben Mhenni *et al.*, 2019b] que ces approches dépassent largement les performances obtenues par l'utilisation d'un solveur MIP générique (CPLEX), pourtant réputé comme l'un des plus puissants.

Ma motivation initiale pour investir ce domaine de recherche était essentiellement méthodologique : dans quelle mesure peut-on résoudre des problèmes ℓ_0 de manière exacte, avec des algorithmes dédiés ? Au fil du déroulement de ces travaux, je me suis également rendu compte que, sur des problèmes particuliers, l'utilisation de reformulations MIP permettait de formuler *de manière exacte* d'autres contraintes qui sont habituellement difficiles à prendre en compte. C'est notamment le cas du démélange spectral, où l'on peut chercher à imposer, en plus de la parcimonie « au sens ℓ_0 », des contraintes de nature discrète ou logique afin de lever des ambiguïtés de modèle. L'application des

développements méthodologiques au cas du démélange spectral occupe donc une place prépondérante dans ces travaux [Ben Mhenni *et al.*, 2017, Ben Mhenni *et al.*, 2018].

Le Chapitre 6 sera consacré à la présentation détaillée de ces travaux, portant à la fois sur les développements algorithmiques et sur l ?exploitation de MIP pour le démélange spectral.

Supervision de stages de Master

- Jiayi Hou (2019) : **Sparse unmixing methods for hyperspectral imaging.** Master *Control and Robotics*, École Centrale de Nantes.
- Ghandy Ajib (2018) : **Contributions to a branch-and-bound algorithm for ℓ_0-norm sparse optimization.** Master *ACSYON*, Université de Limoges. Co-encadrement avec Ramzi Ben Mhenni (doctorant, École Centrale de Nantes).
- Ramzi Ben Mhenni (2016) : **Algorithme de type Branch-and-Bound pour l'optimisation exacte de critères parcimonieux en norme ℓ_0.** Mastère de Recherche en *Informatique, spécialité Systèmes de Raisonnement Automatique*, Université de Monastir. Co-encadrement avec Evgeny Gurevsky (Maître de Conférences, Université de Nantes).

Encadrement doctoral

- Ramzi Ben Mhenni, **Programmation mixte en nombres entiers pour l'optimisation parcimonieuse en traitement du signal.** École Centrale de Nantes. Directeur de thèse (80 %), co-encadrement Jordan Ninin (20 %). Financement *via* une allocation ministérielle. Thèse démarrée en octobre 2016.

Publications associées

- Article de journal soumis : [Ben Mhenni *et al.*, 2019].
- Article de journal : [Bourguignon *et al.*, 2016c].
- Conférences avec actes et comité de lecture : [Ben Mhenni *et al.*, 2019a, Ben Mhenni *et al.*, 2018, Boudineau *et al.*, 2018, Ben Mhenni *et al.*, 2017, Bourguignon *et al.*, 2015].

2.6 Références

[Abubakar et van den Berg, 2004] Abubakar, A. et van den Berg, P. M. (2004). Iterative forward and inverse algorithms based on domain integral equations for three-dimensional electric and magnetic objects. *J. Comput. Phys.*, 195(1):236–262.

[Arulampalam *et al.*, 2001] Arulampalam, S., Maskell, S., Gordon, N. et Clapp, T. (2001). A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. *IEEE Transactions on Signal Processing*, 50:174–188.

[Bacon *et al.*, 2006] Bacon *et al.*, R. (2006). Probing unexplored territories with MUSE : a second generation instrument for the VLT. *In Proc. SPIE*, volume 6269 de *Ground-based and Airborne Instrumentation for Astronomy.*

[Ben Mhenni *et al.*, 2019a] Ben Mhenni, R., Bourguignon, S., Mongeau, M., Ninin, J. et Carfantan, H. (2019a). Algorithme *branch-and-bound* pour l'optimisation exacte en norme ℓ_0. *In Actes du 27e colloque GRETSI*.

[Ben Mhenni *et al.*, 2019b] Ben Mhenni, R., Bourguignon, S. et Ninin, J. (2019b). Global optimization for sparse solution of least squares problems. *Soumis à Optimization Methods and Software*.

[Ben Mhenni *et al.*, 2019] Ben Mhenni, R., Bourguignon, S. et Ninin, J. (2019). Global optimization for sparse solution of least squares problems. Rapport technique, École Centrale de Nantes.

[Ben Mhenni *et al.*, 2017] Ben Mhenni, R., Bourguignon, S., Ninin, J. et Schmidt, F. (2017). Démélange parcimonieux exact dans une approche supervisée en imagerie hyperspectrale. *In Actes du 26e colloque GRETSI*, Juan-les-Pins.

[Ben Mhenni *et al.*, 2018] Ben Mhenni, R., Bourguignon, S., Ninin, J. et Schmidt, F. (2018). Spectral unmixing with sparsity and structuring constraints. *In Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Amsterdam, The Netherlands.

[Bertsimas et Shioda, 2009] Bertsimas, D. et Shioda, R. (2009). Algorithm for cardinality-constrained quadratic optimization. *Computational Optimization and Applications*, 43(1):1–22.

[Bienstock, 1996] Bienstock, D. (1996). Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, 74(2):121–140.

[Boudineau *et al.*, 2018] Boudineau, M., Carfantan, H. et Bourguignon, S. (2018). An l0 solution to sparse approximation problems with continuous dictionaries,. *In International Conference on Acoustic, Speech and Signal Processing (ICASSP'2018)*, Calgary, Alberta, Canada.

[Bouhlel et Bourguignon, 2016] Bouhlel, N. et Bourguignon, S. (2016). Deconvolution of ultrasonic signals in porous materials : Estimation of acoustic propagation parameters and wave separation. *In Review of Progress in Quantitaive Nondestructive Evaluation (QNDE)*, Atlanta, USA.

[Bouhlel *et al.*, 2016] Bouhlel, N., Bourguignon, S., Duclos, A. et Groby, J.-P. (2016). Déconvolution des signaux ultrasonores dans les matériaux poreux. *In Actes du 13e Congrès Français d'Acoustique*, Le Mans, France.

[Bouhlel *et al.*, 2017] Bouhlel, N., Bourguignon, S., Duclos, A. et Groby, J.-P. (2017). Characterization of poroelastic materials through interface scattering and propagation models. *In 5th symposium on the Acoustics of Poro-Elasic Materials (SAPEM)*, Le Mans, France.

[Bourguignon *et al.*, 2009] Bourguignon, S., Berger, L., Scalabrin, C., Fablet, R. et Mazauric, V. (2009). Methodological developments for improved bottom detection with the ME70 multibeam echosounder. *ICES Journal of Marine Science*, 66(6):1015–1022.

[Bourguignon *et al.*, 2016a] Bourguignon, S., Carcreff, E., Lin, C. et Peng, J. (2016a). Blind sparse deconvolution of regularly spaced ultrasonic echoes for thickness measurement. *In Review of Progress in Quantitative Nondestructive Evaluation (QNDE)*, Atlanta, USA.

[Bourguignon *et al.*, 2016b] Bourguignon, S., Carcreff, E., Lurin, N. et Moreau, S. (2016b). Déconvolution d'échos ultrasonores régulièrement espacés pour la mesure d'épaisseur de structures fines. *In Actes du 13e Congrès Français d'Acoustique*, Le Mans, France.

[Bourguignon *et al.*, 2011a] Bourguignon, S., Carfantan, H., Mary, D., Slezak, E. et Ferrari, A. (2011a). Restauration des cubes hyperspectraux du spectro-imageur MUSE. *In Actes du 23e colloque GRETSI*, Bordeaux, France.

[Bourguignon *et al.*, 2011b] Bourguignon, S., Carfantan, H., Slezak, E. et Mary, D. (2011b). Sparsity-based spatial-spectral restoration of MUSE astrophysical hyperspectral data cubes. *In Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Lisbon, Portugal.

[Bourguignon *et al.*, 2010] Bourguignon, S., Mary, D. et Slezak, E. (2010). Sparsity-based denoising of hyperspectral astrophysical data with colored noise. application to the MUSE instrument. *In Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Reykjavik, Iceland.

[Bourguignon *et al.*, 2011c] Bourguignon, S., Mary, D. et Slezak, É. (2011c). Restoration of astrophysical spectra with sparsity constraints : Models and algorithms. *IEEE J. Sel. Topics Signal Processing*, 5(5):1002–1013.

[Bourguignon *et al.*, 2012] Bourguignon, S., Mary, D. et Slezak, É. (2012). Processing MUSE hyperspectral data : Denoising, deconvolution and detection of astrophysical sources. *Statistical Methodology*, 9(1):32–43.

[Bourguignon et Mazauric, 2008] Bourguignon, S. et Mazauric, V. (2008). Methodological developments for improved bottom detection with the ME70 multibeam echosounder. *In 6th Ecosystem Approach with Fisheries Acoustics and Complementary Technologies (SEAFACTS)*, Bergen, Norway.

[Bourguignon *et al.*, 2015] Bourguignon, S., Ninin, J., Carfantan, H. et Mongeau, M. (2015). Optimisation exacte de critères parcimonieux en norme l_0 par programmation mixte en nombres entiers. *In Actes du 25e colloque GRETSI*, Lyon, France.

[Bourguignon *et al.*, 2016c] Bourguignon, S., Ninin, J., Carfantan, H. et Mongeau, M. (2016c). Exact sparse approximation problems via mixed-integer programming : Formulations and computational performance. *IEEE Transactions on Signal Processing*, 64(6):1405–1419.

[Bourguignon *et al.*, 2011d] Bourguignon, S., Soussen, C., Carfantan, H. et Idier, J. (2011d). Sparse deconvolution : comparison of statistical and deterministic approaches. *In IEEE Workshop on Statistical Signal Processing*, pages 317–320, Nice, France.

[Breysse et Abraham, 2005] Breysse, D. et Abraham, O. (2005). *Méthodologie d'évaluation non destructive de l'état d'altération des ouvrages en béton.* Presses de l'École Nationale des Ponts et Chaussées, Paris.

[Carcreff, 2014a] Carcreff, E. (2014a). *Déconvolution adaptative de signaux ultrasonores pour le contrôle non destructif par ultrasons.* Thèse de doctorat, Université du Maine.

[Carcreff, 2014b] Carcreff, E. (2014b). *Déconvolution adaptative de signaux ultrasonores pour le contrôle non destructif par ultrasons.* Thèse de doctorat, Université du Maine.

[Carcreff *et al.*, 2014a] Carcreff, E., Bourguignon, S., Duclos, A. et Groby, J.-P. (2014a). An ultrasonic propagation model suited to inversion methods (with some application to porous materials). *In 4th symposium on the Acoustics of Poro-Elasic Materials (SAPEM)*, Stockholm, Sweden.

[Carcreff *et al.*, 2015] Carcreff, E., Bourguignon, S., Duclos, A., Simon, L. et Idier, J. (2015). Detection of flat bottom holes using sparse deconvolution. *Physics Procedia*, 70:558–561.

[Carcreff *et al.*, 2013a] Carcreff, E., Bourguignon, S., Idier, J. et Simon, L. (2013a). Algorithmes de déconvolution impulsionnelle à résolution augmentée. *In Actes du 24e colloque GRETSI*, Brest, France.

[Carcreff *et al.*, 2013b] Carcreff, E., Bourguignon, S., Idier, J. et Simon, L. (2013b). Resolution enhancement of ultrasonic signals by up-sampled sparse deconvolution. *In Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, Vancouver, Canada.

[Carcreff *et al.*, 2014b] Carcreff, E., Bourguignon, S., Idier, J. et Simon., L. (2014b). A linear model approach for ultrasonic inverse problems with attenuation and dispersion. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 61(7):1191–1203.

[Carcreff *et al.*, 2013c] Carcreff, E., Bourguignon, S., Idier, J., Simon, L. et Duclos, A. (2013c). Including frequency-dependent attenuation for the deconvolution of ultrasonic signals. *In 19th International Congress on Acoustics*, Montréal, Québec, Canada.

[Carcreff *et al.*, 2017] Carcreff, E., Laroche, N., Braconnier, D., Duclos, A. et Bourguignon, S. (2017). Improvement of the total focusing method using an inverse problem approach. *In 2017 IEEE International Ultrasonics Symposium (IUS)*, pages 1–4.

[Cui *et al.*, 2013] Cui, X., Zheng, X., Zhu, S. et Sun, X. (2013). Convex relaxations and MIQCQP reformulations for a class of cardinality-constrained portfolio selection problems. *Journal of Global Optimization*, 56(4):1409–1423.

[Donoho, 2006] Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306.

[El Guennouni *et al.*, 2003] El Guennouni, A., Jbilou, K. et Sadok, H. (2003). A block version of BiCGSTAB for linear systems with multiple right-hand sides. *Electronic Transactions on Numerical Analysis*, 16(129-142):2.

[Eldar et Kutyniok, 2012] Eldar, Y. et Kutyniok, G. (2012). *Compressed Sensing : Theory and Applications*. Cambridge University Press.

[Fares, 2015] Fares, M. (2015). *Détermination des gradients de teneur en eau et en chlorures dans les bétons par méthodes électromagnétiques non destructives*. Thèse de doctorat, LUNAM Université, École Centrale de Nantes.

[Fares *et al.*, 2016] Fares, M., Fargier, Y., Villain, G., Derobert, X. et Palma Lopes, S. (2016). Determining the permittivity profile inside reinforced concrete using capacitive probes. *NDT & E International*, 79:150–161.

[Friedrich, 2016] Friedrich, C. (2016). *Méthodes de reconstruction en tomographie de diffraction 3D - Application à l'imagerie d'un réacteur chimique*. Thèse de doctorat, École Centrale de Nantes, France / École Polytechnique de Montréal, Québec, Canada.

[Friedrich *et al.*, 2015a] Friedrich, C., Bourguignon, S., Idier, J. et Goussard, Y. (2015a). Faster resolution of the 3-D forward problems in microwave imaging by a partial-block BiCGStab algorithm. *In EuCAP, The 9th European Conference on Antennas and Propagation*, Lisbon, Portugal.

[Friedrich *et al.*, 2015b] Friedrich, C., Bourguignon, S., Idier, J. et Goussard, Y. (2015b). Reconstruction of 3-D microwave images based on a Block-BiCGStab algorithm. *Journal of Physics : Conference Series*, 657:012014.

[Heintzmann et Gustafsson, 2009] Heintzmann, R. et Gustafsson, M. G. L. (2009). Subdiffraction resolution in continuous samples. *Nature Photonics*, 3:362–364.

[Holmes *et al.*, 2005] Holmes, C., Drinkwater, B. W. et Wilcox, P. D. (2005). Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation. *NDT & E International*, 38(8):701 – 711.

[Idier *et al.*, 2018] Idier, J., Labouesse, S., Allain, M., Liu, P., Bourguignon, S. et Sentenac, A. (2018). On the superresolution capacity of imagers using unknown speckle illuminations. *IEEE Transactions on Computational Imaging*, 4(1):87–98.

[Krautkramer et Krautkramer, 1990] Krautkramer, J. et Krautkramer, H. (1990). *Ultrasonic Testing of materials*. Springer-Verlag, Berlin.

[Labouesse *et al.*, 2015] Labouesse, S., Allain, M., Idier, J., Bourguignon, S., Liu, P. et Sentenac, A. (2015). Blind fluorescence structured illumination microscopy : A new reconstruction strategy. *In Actes du 25e colloque GRETSI*, Lyon, France.

[Labouesse *et al.*, 2016] Labouesse, S., Allain, M., Idier, J., Bourguignon, S., Negash, A., Liu, P. et Sentenac, A. (2016). Fluorescence blind structured illumination microscopy : a new reconstruction strategy. *In IEEE International Conference on Image Processing (ICIP)*, Phoenix, USA.

[Labouesse *et al.*, 2017] Labouesse, S., Negash, A., Idier, J., Bourguignon, S., Mangeat, T., Liu, P., Sentenac, A. et Allain, M. (2017). Joint reconstruction strategy for structured illumination microscopy with unknown illuminations. *IEEE Transactions on Image Processing*, 26(5):1–14.

[Laroche *et al.*, 2019] Laroche, N., Bourguignon, S., Carcreff, E., Idier, J. et Duclos, A. (2019). Approche inverse rapide pour la déconvolution d'images ultrasonores par une PSF variable. *In Actes du 27e colloque GRETSI*.

[Laroche *et al.*, 2018] Laroche, N., Carcreff, E., Bourguignon, S., Idier, J. et Duclos, A. (2018). An inverse approach for ultrasonic imaging by total focusing point for close reflectors separation. *In IEEE International Ultrasonics Symposium (IUS)*, Kobe, Japan.

[Lee *et al.*, 2007] Lee, H., Battle, A., Raina, R. et Ng, A. Y. (2007). Efficient sparse coding algorithms. *In Advances in Neural Information Processing Systems*, pages 801–808.

[Lefort *et al.*, 2008] Lefort, R., Fablet, R., Boucher, J.-M., Berger, L. et Bourguignon, S. (2008). Automatic fish school classification for acoustic sensing of marine ecosystem. pages 1 – 5.

[Li *et al.*, 2006] Li, D., Sun, X. et Wang, J. (2006). Optimal lot solution to cardinality constrained mean–variance formulation for portfolio selection. *Mathematical Finance*, 16(1):83–101.

[Liu, 2018] Liu, P. (2018). *Statistical and numerical optimization for speckle blind structured illumination microscopy*. Thèse de doctorat, École Centrale de Nantes.

[Liu *et al.*, 2017] Liu, P., Idier, J., Bourguignon, S., Labouesse, S., Allain, M. et Sentenac, A. (2017). Minimum contrast estimation for super-resolution fluorescence microscopy using speckle patterns. *In Actes du 26e colloque GRETSI*, Juan-les-Pins.

[MacLennan *et al.*, 2004] MacLennan, D. N., Copland, P. J., Armstrong, E. et Simmonds, E. J. (2004). Experiments on the discrimination of fish and seabed echoes. *ICES Journal of Marine Science*, 61:201–210.

[Mallat, 2008] Mallat, S. (2008). *A Wavelet Tour of Signal Processing : The Sparse Way*. Elsevier Science.

[Mudry *et al.*, 2012] Mudry, E., Belkebir, K., Girard, J., Savatier, J., Moal, E. L., Nicoletti, C., Allain, M. et Sentenac, A. (2012). Structured illumination microscopy using unknown speckle patterns. *Nature Photonics*, 6:312–315.

[Natarajan, 1995] Natarajan, B. (1995). Sparse approximate solutions to linear systems. *SIAM Journal of Computing*, 2(24):227–234.

[O'Brien *et al.*, 1990] O'Brien, M. S., Sinclair, A. N. et Kramer, S. M. (1990). High resolution deconvolution using least-absolute-values minimization. *In Ultrasonics Symposium*, volume 2, pages 1151–1156.

[Osborne *et al.*, 2000] Osborne, M. R., Presnell, B. et Turlach, B. A. (2000). A new approach to variable selection in least squares problems. *IMA J Numer Anal*, 20(3):389–403.

[Paris *et al.*, 2011] Paris, S., Mary, D., Ferrari, A. et Bourguignon, S. (2011). Sparsity-based composite detection tests. application to astrophysical hyperspectral data. *In EUSIPCO*, Barcelona, Spain.

[Pastorino, 2010] Pastorino, M. (2010). *Microwave imaging*, volume 208. John Wiley & Sons.

[Picaud et Bourguignon, 2013] Picaud, J. et Bourguignon, S. (2013). Restauration de données hyperspectrales astrophysiques par approximation parcimonieuse. *In Actes du 24e colloque GRETSI*, Brest, France.

[Raharinaivo *et al.*, 1998] Raharinaivo, A., Arliguie, G., Chaussadent, T., Grimaldi, G., Pollet, V. et Tache, G. (1998). *La corrosion et la protection des aciers dans le béton*. Presses de l'École Nationale des Ponts et Chaussées, Paris.

[Saniie et Nagle, 1989] Saniie, J. et Nagle, D. (1989). Pattern recognition in the ultrasonic imaging of reverberant multilayered structures. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 36(1):80–92.

[Schmidt et Bourguignon, 2019] Schmidt, F. et Bourguignon, S. (2019). Efficiency of BRDF sampling and bias on the average photometric behavior. *Icarus*, 317:10–26.

[Schmidt *et al.*, 2011] Schmidt, F., Bourguignon, S., Le Mouélic, S., Dobigeon, N., Theys, C. et Tréguier, E. (2011). Accuracy and performance of lineat unmixing techniques for detecting minerals on Omega / Mars Express. *In Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Lisbon, Portugal.

[Serre *et al.*, 2010] Serre, D., Villeneuve, E., Carfantan, H., Jolissaint, L., Mazet, V., Bourguignon, S. et Jarno, A. (2010). Modeling the spatial PSF at the VLT focal plane for MUSE WFM data analysis purpose. *Proc. SPIE 7736*, page 773649.

[Shaw *et al.*, 2008] Shaw, D. X., Liu, S. et Kopman, L. (2008). Lagrangian relaxation procedure for cardinality-constrained portfolio optimization. *Optimization Methods and Software*, 23(3):411–420.

[Shen *et al.*, 2012] Shen, Y.-S., Chan, T.-H., Bourguignon, S. et Chi, C.-Y. (2012). Spatial-spectral unmixing of hyperspectral data for detection and analysis of astrophysical sources with the MUSE instrument. *In Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Shanghai, China.

[Simmonds et MacLennan, 2005] Simmonds, E. J. et MacLennan, D. N. (2005). *Fisheries Acoustics. Theory and Practice, 2nd edn.* Blackwell Publishing, Oxford.

[Soussen *et al.*, 2011] Soussen, C., Idier, J., Brie, D. et Duan, J. (2011). From Bernoulli-Gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584.

[Tennyson, 2005] Tennyson, J. (2005). *Astronomical Spectroscopy.* Imperial College Press.

[Trenkel *et al.*, 2009] Trenkel, V., Berger, L., Bourguignon, S., Doray, M., Fablet, R., Masse, J., Mazauric, V., Poncelet, C., Quemeneur, G., Scalabrin, C. et Villalobos, H. (2009). Overview of recent progress in fisheries acoustics made by Ifremer with examples from the Bay of Biscay. *Aquatic Living Resources*, 22(4):433–446.

[Tropp et Wright, 2010] Tropp, J. A. et Wright, S. J. (2010). Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958.

[Villeneuve, 2012] Villeneuve, E. (2012). *Déconvolution de données hyperspectrales pour l'instrument MUSE du VLT.* Thèse de doctorat, Université Toulouse 3.

[Villeneuve *et al.*, 2011] Villeneuve, E., Carfantan, H., Jarno, A., Serre, D., Mazet, V. et Bourguignon, S. (2011). Modélisation et estimation de la psf d'un instrument hyperspectral au sol pour l'astrophysique. *In Actes du 23e colloque GRETSI*, Bordeaux, France.

[Wolsey, 1998] Wolsey, L. A. (1998). *Integer Programming.* Wiley, New York, NY, USA.

[Zala, 1992] Zala, C. (1992). High-resolution inversion of ultrasonic traces. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 39(4):458–463.

[Zhang *et al.*, 2003] Zhang, Z. Q., Liu, Q. H., Xiao, C., Ward, E., Ybarra, G. et Joines, W. T. (2003). Microwave breast imaging : 3-D forward scattering simulation. *IEEE Trans. Biomed. Eng.*, 50(10):1180–1189.

Deuxième partie

Modèles et algorithmes dédiés pour la résolution de problèmes inverses parcimonieux en traitement du signal et de l'image

Chapitre 3

Introduction

Cette deuxième partie de manuscrit est consacrée à la présentation de mes travaux sur trois axes majeurs de mon activité de recherche. Les deux premiers abordent la résolution de problèmes inverses, sous des contraintes de modèles parcimonieux, respectivement pour l'imagerie hyperspectrale en astronomie et pour le contrôle non destructif (CND) par ultrasons. Le troisième concerne le développement d'algorithmes d'optimisation exacte en norme ℓ_0. Contrairement aux deux premiers sujets, l'essentiel de mes contributions sur ce thème n'est pas guidé par une application spécifique ; nous y retrouverons cependant des exemples d'application sur des problèmes de déconvolution pour le CND et de démélange pour l'imagerie hyperspectrale.

Dans ces trois thématiques, j'ai, d'une part, mené une réflexion sur la définition de modèles appropriés à la problématique abordée et présentant un caractère original, que ce soit dans la description du processus reliant les données mesurées à des quantités d'intérêt à estimer, ou dans la formulation mathématique du problème d'estimation associé. D'autre part, j'ai travaillé sur la construction de solutions algorithmiques dédiées, relevant essentiellement de l'optimisation numérique.

La notion de *parcimonie* occupe une place centrale dans l'ensemble de ces travaux. Ce chapitre propose donc une introduction succincte aux principales questions méthodologiques jalonnant les recherches autour de la parcimonie et en présente les méthodes de référence que nous retrouverons aux chapitres ultérieurs. Le tableau en page 58 clôt ce chapitre en présentant de manière résumée les contributions que j'ai pu apporter, au niveau des modèles et des algorithmes, dans chacun des trois axes de recherche qui seront détaillés aux Chapitres 4 à 6.

3.1 Problèmes inverses et parcimonie

Dans le cadre des problèmes inverses, nous nous intéressons à l'estimation d'un objet d'intérêt (un signal, une image, un ensemble de données), représenté par le vecteur $\boldsymbol{x} \in \mathbb{R}^P$, à partir d'un ensemble de mesures constituant le vecteur $\boldsymbol{y} \in \mathbb{R}^N$, sous un modèle de la forme :

$$\boldsymbol{y} = \mathcal{H}(\boldsymbol{x}) + \boldsymbol{\epsilon},$$

où l'opérateur $\mathcal{H}(\cdot)$ modélise l'acquisition des données et le terme $\boldsymbol{\epsilon}$ représente une quantité aléatoire, décrivant les perturbations entachant le processus d'observation (le bruit de mesure) et de possibles erreurs dans la modélisation.

Nous considérons ici des problèmes inverses linéaires, où le modèle s'écrit :

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{\epsilon},$$

où $\mathbf{H}$ est une matrice à coefficients réels de taille $N \times P$. Cette hypothèse de linéarité, bien que restrictive, est cependant légitime dans de nombreux contextes (voir par exemple [Idier, 2008]) ; ce sera le cas dans les travaux des chapitres suivants. Dans la plupart des problèmes mettant en jeu un dispositif de mesure, l'estimation de $\boldsymbol{x}$ à partir des données $\boldsymbol{y}$ est un problème dit mal posé, dans le sens où le processus d'acquisition produit des données déficientes, manquant d'information permettant de reconstruire correctement l'objet observé. Un exemple classique est celui de la déconvolution, où l'objet $\boldsymbol{x}$ est observé à travers un filtre ($\mathbf{H}\boldsymbol{x}$ représente alors la convolution du signal $\boldsymbol{x}$ par la réponse impulsionnelle du filtre), résultant en la perte d'information sur l'objet dans certaines gammes de fréquences. La définition d'une solution satisfaisante passe alors par une étape de régularisation, où des contraintes sont ajoutées, traduisant une connaissance *a priori*, *i.e.,* non issue de la mesure, sur l'objet recherché.

Dans ce contexte, l'information de *parcimonie* consiste à supposer que l'objet peut être représenté, de manière exacte ou approchée, par un ensemble limité de coefficients. Le cas le plus direct est celui où le vecteur $\boldsymbol{x}$ lui même, représentant par exemple les échantillons de l'amplitude d'un signal temporel ou l'intensité des pixels dans une image numérique, est essentiellement composé de valeurs nulles. Nous retrouverons ce modèle dans nos travaux sur la déconvolution parcimonieuse de signaux et d'images ultrasonores au Chapitre 5. Un cas plus général est celui où l'hypothèse de parcimonie s'effectue au moyen d'un espace de représentation spécifique : $\boldsymbol{x} = \mathbf{D}\boldsymbol{u}$, où la matrice $\mathbf{D}$ constitue le *dictionnaire* de représentation et $\boldsymbol{u}$ est le vecteur de coefficients supposé parcimonieux. Un tel modèle sera construit pour la restauration de cubes hyperspectraux en astrophysique au Chapitre 4 et pour le démélange spectral parcimonieux au Chapitre 6. Nous pouvons regrouper les deux modèles précédents sous la forme générique (avec respectivement $\mathbf{A} = \mathbf{H}$ ou $\mathbf{A} = \mathbf{HD}$) :

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{\epsilon}, \text{ avec } \boldsymbol{x} \text{ parcimonieux.} \tag{3.1}$$

L'estimation des composantes non nulles de $\boldsymbol{x}$ revient alors à sélectionner des colonnes de $\mathbf{A}$ (les *atomes*) et l'ensemble des indices correspondant aux composantes sélectionnés est appelé le *support*. Remarquons ici que des problèmes similaires existent dans le domaine de la statistique : on parle alors de sélection de variables ou de sélection de sous-ensembles (voir par exemple [Miller, 2002]). Nous supposerons dans la suite de ce chapitre que les colonnes de $\mathbf{A}$ sont de norme unité.

Mathématiquement, l'estimation d'un vecteur parcimonieux $\boldsymbol{x}$ peut être formulée comme la recherche d'un ajustement au sens des moindres carrés sous la contrainte d'un faible nombre (supposé connu) de composantes non nulles :

$$\mathcal{P}_{2/0} : \min_{\boldsymbol{x} \in \mathbb{R}^P} \frac{1}{2} \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|_2^2 \text{ sous la contrainte (s. c.) } \|\boldsymbol{x}\|_0 \leq K,$$

où la « norme » ℓ_0 compte le nombre de composantes non nulles dans un vecteur : $\|\boldsymbol{x}\|_0 \triangleq \text{Card}\,\{p | x_p \neq 0\}$. Notons ici que le choix d'une norme ℓ_2 pour l'ajustement de modèle, très fréquemment adopté en traitement du signal, revient à une hypothèse statistique d'une distribution gaussienne, de moyenne nulle et de matrice de covariance proportionnelle à l'identité, du terme d'erreur $\boldsymbol{\epsilon} = \boldsymbol{y} - \mathbf{A}\boldsymbol{x}$ [Idier, 2008]. Inversement, l'estimation peut également être formulée *via* la minimisation du nombre de composantes de $\boldsymbol{x}$ parmi les solutions satisfaisant un niveau d'approximation donné :

$$\mathcal{P}_{0/2} : \min_{\boldsymbol{x} \in \mathbb{R}^P} \|\boldsymbol{x}\|_0 \text{ s. c. } \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \alpha.$$

Cette formulation est particulièrement intéressante en traitement du signal, où l'on possède souvent une information supplémentaire sur la qualité des données et du modèle, permettant de régler le paramètre α. C'est également la formulation appropriée pour aborder le problème de *représentation* parcimonieuse (la recherche de la solution au système $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$ possédant le moins de composantes non-nulles), correspondant alors au cas sans bruit $\alpha = 0$. Enfin, le problème pénalisé :

$$\mathcal{P}_{2+0} : \min_{\boldsymbol{x}\in\mathbb{R}^P} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu \|\boldsymbol{x}\|_0 ,$$

où $\mu > 0$ règle le compromis entre l'erreur d'approximation et le niveau de parcimonie, est souvent rencontré pour les problèmes inverses, par exemple en géophysique [Mendel, 1983] ou en contrôle non destructif ultrasonore [Zala, 1992, O'Brien *et al.*, 1994]. Si le réglage du paramètre μ est moins intuitif que celui des paramètres des deux autres formulations [1], cette dernière formulation est souvent utilisée pour des raisons à la fois historiques (la régularisation des problèmes inverses a très tôt été abordée par l'ajout d'un terme de pénalisation) et pratiques (permettant de formuler un problème sans contrainte).

Remarquons ici que, en raison de leur non convexité, ces trois problèmes d'optimisation ne sont pas équivalents, *i.e.*, les ensembles de solutions décrites en faisant varier leur paramètre respectif ne coïncident pas. En particulier, les solutions atteintes dans la partie non-convexe de la frontière de Pareto ne peuvent pas être atteintes par la version pénalisée [Das et Dennis, 1997].

3.2 Méthodes classiques d'estimation parcimonieuse

Les problèmes d'optimisation en norme ℓ_0 sont connus pour être NP-difficiles, en raison de la nature discrète de la fonction de comptage [Natarajan, 1995, Bienstock, 1996] rendant le problème de nature essentiellement combinatoire. Par conséquent, le développement de méthodes de résolution approchées, permettant d'aborder des problèmes de grande taille, a fait l'objet de très nombreuses contributions depuis l'émergence du principe de parcimonie dans les années 2 000. Notons cependant que des travaux bien antérieurs peuvent en être considérés comme les précurseurs. On retrouve ainsi le principe des méthodes gloutonnes dans des algorithmes de déconvolution en radioastronomie dès les années 1970 [Hogbom, 1974] et des approches régularisées par la norme ℓ_1 sont déjà proposées pour des problèmes de déconvolution en géophysique [Taylor *et al.*, 1979, Alliney et Ruzinsky, 1994] ou en contrôle non destructif [O'Brien *et al.*, 1990].

L'objectif de cette section n'est pas d'établir un panorama exhaustif des méthodes d'estimation parcimonieuse, mais de présenter deux classes d'algorithmes très fréquemment rencontrées en traitement du signal et pour la sélection de variables et sur lesquels s'appuieront nos travaux : les méthodes dites gloutonnes et les algorithmes d'optimisation de critères en norme ℓ_1.

3.2.1 Algorithmes gloutons

Les algorithmes de poursuite adaptative ou algorithmes *gloutons*, popularisés par les travaux de Mallat et Zhang [Mallat et Zhang, 1993], construisent une solution parcimonieuse de manière itérative. De nombreuses variantes ont été proposées, dont les performances s'évaluent sous l'angle du compromis entre qualité de la solution et complexité

1. Dans une interprétation probabiliste bayésienne, on peut par exemple montrer que μ dépend à la fois du niveau de bruit sur le modèle et de la parcimonie de la solution [Soussen *et al.*, 2011].

calculatoire. Cette section rappelle le principe de deux algorithmes gloutons couramment rencontrés : *Orthogonal Matching Pursuit* (OMP) et *Orthogonal Least-Squares* (OLS) et de l'algorithme *Single Best Replacement* (SBR).

Les méthodes gloutonnes construisent une solution pour laquelle, à l'itération t, le signal $\boldsymbol{y}$ est approché par la combinaison linéaire de t atomes du dictionnaire, indexés par le support noté $\Omega^{(t)}$ et de poids associés $\boldsymbol{x}_\Omega^{(t)}$:

$$\boldsymbol{y} \simeq \widetilde{\boldsymbol{y}}^{(t)}, \text{ avec } \widetilde{\boldsymbol{y}}^{(t)} = \sum_{k \in \Omega^{(t)}} x_k^{(t)} \boldsymbol{a}_k = \mathbf{A}_{\Omega^{(t)}} \boldsymbol{x}_\Omega^{(t)},$$

où $\mathbf{A}_{\Omega^{(t)}}$ est la matrice composée des vecteurs colonne $\boldsymbol{a}_k$, $k \in \Omega^{(t)}$ et $\boldsymbol{x}_\Omega^{(t)}$ est le vecteur colonne formé par les coefficients $x_k^{(t)}$, $k \in \Omega^{(t)}$. Partant d'une initialisation nulle $\widetilde{\boldsymbol{y}}^{(0)} = 0$, chaque itération consiste alors en :

i) la sélection d'un atome $\boldsymbol{a}_{k^*}$ à partir de l'approximation de l'itération précédente, définissant le nouveau support $\Omega^{(t)} = \Omega^{(t-1)} \cup \{k^*\}$;

ii) la mise à jour des amplitudes $\boldsymbol{x}_\Omega^{(t)}$ des atomes sélectionnés ;

iii) la mise à jour de l'approximation $\widetilde{\boldsymbol{y}}^{(t)}$ et de l'erreur résiduelle $\boldsymbol{r}^{(t)} = \boldsymbol{y} - \widetilde{\boldsymbol{y}}^{(t)}$.

Notons que les trois formulations $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$ peuvent être abordées par ce type d'algorithme, en fonction du critère d'arrêt utilisé : un nombre de composantes non nulles égale à K pour la première, la norme du résidu inférieure à α pour la deuxième et un critère pénalisé qui ne décroît plus pour la dernière.

Orthogonal Matching Pursuit (OMP)

L'algorithme OMP [Pati *et al.*, 1993] sélectionne, à chaque itération, l'atome le plus corrélé au résidu :

i) $k^* = \arg\max_k |\boldsymbol{a}_k^T \boldsymbol{r}^{(t-1)}|$,

puis estime l'ensemble des amplitudes non nulles en minimisant l'erreur d'approximation sur le support courant :

ii) $\boldsymbol{x}_\Omega^{(t)} = \arg\min_{\boldsymbol{x}_\Omega \in \mathbb{R}^t} \|\boldsymbol{y} - \mathbf{A}_{\Omega^{(t)}} \boldsymbol{x}_\Omega\|^2$.

Le qualificatif d'orthogonal provient du fait qu'à chaque itération, $\widetilde{\boldsymbol{y}}^{(t)}$ représente la meilleure approximation, au sens de la distance euclidienne, de $\boldsymbol{y}$ par une combinaison linéaire des atomes de $\mathbf{A}_{\Omega^{(t)}}$: $\mathbf{A}_{\Omega^{(t)}} \boldsymbol{x}_\Omega^{(t)}$ est la projection orthogonale des données $\boldsymbol{y}$ sur le sous-espace engendré par les colonnes de $\mathbf{A}_{\Omega^{(t)}}$.

La sélection à l'étape i) peut être interprétée comme l'estimation, au sens des moindres carrés, du meilleur couple (k, x_k) tel que

$$\boldsymbol{r}^{(t-1)} \simeq x_k \boldsymbol{a}_k.$$

En effet :

$$(k^*, x_k^*) = \arg\min_{k, x_k} \left\|\boldsymbol{r}^{(t-1)} - x_k \boldsymbol{a}_k\right\|^2 \Leftrightarrow (k^*, x_k^*) = \arg\min_k \left\{ \min_{x_k} \left\|\boldsymbol{r}^{(t-1)} - x_k \boldsymbol{a}_k\right\|^2 \right\}$$

$$\Leftrightarrow k^* = \arg\min_k \left\|\boldsymbol{r}^{(t-1)} - x_k^* \boldsymbol{a}_k\right\|^2 \text{ avec } x_k^* = \frac{\boldsymbol{a}_k^T \boldsymbol{r}^{(t-1)}}{\boldsymbol{a}_k^T \boldsymbol{a}_k}$$

$$\Leftrightarrow k^* = \arg\max_k \left(\boldsymbol{a}_k^T \boldsymbol{r}^{(t-1)}\right)^2 \text{ avec } x_k^* = \frac{\boldsymbol{a}_k^T \boldsymbol{r}^{(t-1)}}{\boldsymbol{a}_k^T \boldsymbol{a}_k},$$

la dernière ligne étant obtenue en injectant la valeur de x_k^* dans le terme en norme ℓ_2 et exploitant l'hypothèse que les colonnes du dictionnaire sont de norme unité. Statistiquement, ce choix revient donc à estimer le meilleur modèle $x_k \boldsymbol{a}_k$ approchant le résidu $\boldsymbol{r}^{(t-1)}$, sous l'hypothèse d'erreurs $\boldsymbol{\epsilon}$ gaussiennes, de moyenne nulle et de matrice de covariance proportionnelle à l'identité.

L'estimation du premier atome (avec $\boldsymbol{r}^{(0)} = \boldsymbol{y}$) suppose donc un tel modèle sur $\boldsymbol{y}$, alors que l'itération suivante recherche une nouvelle composante dans le résidu, lequel était supposé ne contenir que du bruit à l'itération précédente. Cette approche est donc sensible aux interférences entre les différents atomes du dictionnaire.

Orthogonal Least-Squares (OLS)

L'algorithme OLS [Chen *et al.*, 1989] est une méthode gloutonne pour laquelle l'étape de sélection est plus complexe : le nouvel atome est celui qui, ajouté aux précédents, minimise l'erreur d'approximation du nouveau modèle. La sélection revient alors à calculer les amplitudes optimales correspondant à tous les modèles intégrant chaque nouvel atome, puis à choisir l'atome fournissant la plus faible erreur :

$$\text{i) et ii) } \{k^*, \boldsymbol{x}_\Omega^{(t)}\} = \underset{k,\, \boldsymbol{x}_\Omega \in \mathbb{R}^t,\, \Omega = \Omega^{(t-1)} \cup \{k\}}{\arg\min} \|\boldsymbol{y} - \mathbf{A}_\Omega \boldsymbol{x}_\Omega\|^2 \text{ avec } \Omega^{(t)} = \Omega^{(t-1)} \cup \{k^*\}.$$

L'étape de sélection est plus efficace que celle d'OMP, assurant la plus forte décroissance du critère des moindres carrés à chaque itération, même si elle ne s'affranchit pas totalement du problème d'interférence entre les différents atomes évoqué plus haut. Elle est bien entendu plus coûteuse, requérant à chaque itération la résolution d'un nombre de systèmes linéaires de taille $t \times t$ égal au nombre d'inconnues non encore sélectionnées [2].

Single Best Replacement (SBR)

L'algorithme SBR [Soussen *et al.*, 2011] est une approche itérative visant à minimiser le critère pénalisé $\mathcal{P}_{2+0}$. Chaque itération propose l'ajout d'une nouvelle composante ou bien le retrait d'une composante précédemment sélectionnée, en choisissant l'opération réduisant le plus ce critère. Cet algorithme peut être vu comme une amélioration de l'algorithme OLS permettant de remettre en cause les détections réalisées aux itérations précédentes. Il présente donc en général de meilleurs performances en pratique – au prix bien entendu d'une augmentation du coût calculatoire [Soussen *et al.*, 2011].

3.2.2 Relaxation ℓ_1

La substitution de la norme ℓ_0 par la norme ℓ_1, soit $\|\boldsymbol{x}\|_1 \triangleq \sum_p |x_p|$, dans les problèmes $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$, a donné lieu à de très nombreux travaux de recherche, aussi bien sur le plan informationnel qu'au niveau de l'optimisation mathématique. Nous notons $\mathcal{P}_{2/1}$, $\mathcal{P}_{1/2}$ et $\mathcal{P}_{2+1}$ les problèmes correspondants, avec des notations évidentes. Cet engouement s'explique par au moins deux raisons :

i) les solutions de ces problèmes d'optimisation sont parcimonieuses. On montre par exemple que la minimisation du critère des moindres carrés pénalisé par une fonction de la forme $\mu \sum_p \varphi(|x_p|)$ avec φ croissante sur $\mathbb{R}^+$, fournit une solution parcimonieuse au sens strict (*i.e.*, avec des composantes nulles) si et seulement si $\varphi'(0) \neq 0$ [Moulin et Liu, 1999], ce qui est le cas de la fonction $\varphi(u) = u$;

2. Les mises à jour des quantités intermédiaires peuvent néanmoins être réalisées récursivement, voir par exemple [Soussen *et al.*, 2011].

ii) les problèmes d'optimisation résultants relèvent de l'optimisation convexe et peuvent donc être résolus de manière garantie par des algorithmes d'optimisation locale.

Remarquons ici que, le terme des moindres carrés et la norme ℓ_1 étant tous deux convexes, les trois problèmes deviennent *équivalents* : pour toute valeur de $K \in \mathbb{R}^+$, il existe une valeur de $\alpha \in \mathbb{R}^+$ (respectivement, une valeur de $\mu \in \mathbb{R}^+$) telle que $\mathcal{P}_{2/1}$ et $\mathcal{P}_{1/2}$ (respectivement, $\mathcal{P}_{2/1}$ et $\mathcal{P}_{2+1}$) aient la même solution.

S'il est illusoire de réaliser ici l'inventaire des différents algorithmes d'optimisation en norme ℓ_1 proposés depuis une vingtaine d'années, notons qu'une grande majorité des travaux récents est consacrée aux méthodes de descente itérative dites proximales [Combettes et Pesquet, 2011], abordant généralement le problème sous la forme pénalisée $\mathcal{P}_{2+1}$. Cependant, pour les problèmes qui nous intéressent dans ce manuscrit, nous nous sommes tournés vers d'autres structures d'algorithmes que nous mentionnons ici (ces choix seront argumentés dans les sections dédiées) :

- les algorithmes de descente coordonnée par coordonnée (ou ICD pour *Iterative Coordinate Descent*) fonctionnent sur le principe simple de l'optimisation successive et répétée du critère en chacune de ses variables. Cette stratégie est généralement peu valorisée en optimisation locale, réputée pour sa lenteur de convergence. Cependant, son efficacité a été démontrée dans le cas de critères de moindres carrés pénalisés par la norme ℓ_1 [Friedman *et al.*, 2007, Bourguignon *et al.*, 2007, Wu et Lange, 2008]. Celle-ci est due au très faible coût de mise à jour des itérés successifs, ainsi qu'à la possibilité de définir des règles de balayage de l'ensemble des variables et des étapes d'accélération particulièrement efficaces, sur lesquelles nous reviendrons au Chapitre 4. Dans le cas de critères en norme ℓ_1, la convergence asymptotique (*i.e.,* lorsque le nombre d'itérations tend vers l'infini) est garantie uniquement pour la version pénalisée $\mathcal{P}_{2+1}$ [Tseng, 2001].

- Les méthodes homotopiques ont été développées dans la communauté de la statistique [Tibshirani, 1996, Osborne *et al.*, 2000a, Efron *et al.*, 2004, Donoho et Tsaig, 2008]. L'algorithme homotopique construit itérativement l'ensemble des solutions du problème $\mathcal{P}_{2+1}$ en fonction du paramètre de pénalisation μ, en le faisant varier de manière décroissante jusqu'à atteindre sa valeur cible, en un nombre fini d'itérations. Il peut donc être utilisé de la même manière pour la résolution des problèmes $\mathcal{P}_{2/1}$ et $\mathcal{P}_{1/2}$, en adaptant le critère d'arrêt.

- Les méthodes dites d'ensembles actifs ou de contraintes actives [Osborne *et al.*, 2000a, Osborne *et al.*, 2000b, Lee *et al.*, 2007] exploitent les conditions d'optimalité de Karush-Kuhn-Tucker (KKT) étendues au cas de critères non différentiables [Rockafellar, 1970]. L'algorithme considère séparément les variables nulles ou inactives (pour lesquelles les conditions KKT s'expriment par des inégalités) des variables non-nulles (les conditions KKT étant alors des égalités), lesquelles forment l'ensemble actif. À chaque itération, une direction de descente est construite, permettant de ramener dans l'ensemble actif une variable nulle violant les conditions d'optimalité. Une recherche scalaire est alors effectuée dans cette direction, suite à laquelle la séparation entre variables actives et inactives est mise à jour. Par construction, cet algorithme converge également en un nombre fini d'itérations. À ma connaissance, seules des versions résolvant les problèmes $\mathcal{P}_{2/1}$ [Osborne *et al.*, 2000a, Osborne *et al.*, 2000b] et $\mathcal{P}_{2+1}$ [Lee *et al.*, 2007] ont été proposées avec des garanties d'optimalité. Cela semble naturel car les conditions KKT de $\mathcal{P}_{1/2}$, à contraintes quadratiques, ne sont pas aussi facilement exploitables que dans les deux autres cas.

Rappelons encore que, ces problèmes d'optimisation en norme ℓ_1 étant convexes, tout algorithme d'optimisation à convergence garantie fournira la même valeur optimale de la fonction de coût. La question de savoir si le minimiseur obtenu est également le même est relative à celle de son unicité. Cette question n'est pas triviale dans le cas (très fréquent) où la matrice $\mathbf{A} \in \mathbb{R}^{N \times P}$ possède plus de colonnes que de lignes, pour lequel le problème d'optimisation n'est pas *strictement* convexe. Une caractérisation assez générale de l'unicité du minimiseur repose sur la propriété de représentation unique (URP, Unique Representation Property) de la matrice $\mathbf{A}$, qui suppose que tout ensemble de N de ses colonnes est linéairement indépendant [Gorodnitsky et Rao, 1997]. Si $\mathbf{A}$ satisfait cette propriété, alors on peut montrer que le problème en norme ℓ_1 admet au plus une solution avec moins de $N/2$ valeurs non nulles [Bourguignon *et al.*, 2007], ce qui est suffisant pour garantir l'unicité du minimiseur s'il est suffisamment parcimonieux.

3.2.3 Sous-optimalité par rapport au problème en norme ℓ_0

De nombreux travaux en théorie de l'information ont cherché à dériver des conditions (suffisantes), portant sur la matrice $\mathbf{A}$, garantissant que les algorithmes gloutons et/ou la relaxation en norme ℓ_1 résolvent le problème initialement formulé en norme ℓ_0 (par exemple, une cohérence mutuelle faible [Fuchs, 2004, Soussen *et al.*, 2013], une constante d'isométrie restreinte faible [Candès *et al.*, 2006], ou un *Exact Recovery Coefficient* positif [Tropp, 2006]). Ces conditions traduisent, sous des formes différentes, que la matrice $\mathbf{A}$ doit être *relativement* proche d'une matrice orthogonale. Ainsi, dans des cas particuliers comme celui de l'échantillonnage compressif [Eldar et Kutyniok, 2012], le dictionnaire est *choisi* pour satisfaire ces propriétés, de manière garantie ou avec une forte probabilité. En revanche, dans la plupart des problèmes inverses difficiles où l'opérateur $\mathbf{A}$ est contraint par un modèle physique, ces conditions ne sont pas satisfaites. Ce sera le cas de tous les problèmes que nous rencontrerons dans la suite de ce manuscrit.

3.3 Positionnement de mes travaux de recherche

Le tableau de la page 58 récapitule les contributions que j'ai pu apporter, aussi bien dans la construction de modèles originaux et adaptés à la problématique considérée que dans le développement d'algorithmes d'optimisation associés, et qui seront développées dans les trois chapitres suivants. Les caractères gras soulignent les contributions les plus importantes à mes yeux.

Chapitre Problème	Modèle	Algorithme	§
Restauration de données hyperspectrales astrophysiques			
Déconvolution de spectres			
	Synthèse de spectres parcimonieux dans un **dictionnaire de formes spécifiques**	**ICD avec accélérations**	4.2, 4.3
Déconvolution spatiale-spectrale			
	Modèle d'observation avec **PSF spatiale et spectrale, mais régularisation uniquement spectrale**	**ICD avec accélérations**, OMP	4.4
Contrôle Non Destructif par ultrasons			
Déconvolution de signaux monodimensionnels (A-scans)			
	Discrétisation fine de l'équation intégrale $\rightsquigarrow$ **déconvolution à haute résolution**	OMP, OLS, SBR, homotopie **pour la déconvolution MISO**	5.2
	Modèle non invariant prenant en compte des **phénomènes de propagation acoustique (atténuation et dispersion)**	Filtre adapté, SBR	5.3
Déconvolution d'échos régulièrement espacés			
	Déconvolution aveugle : modèle spécifique de séquence impulsionnelle et modèle gaussien de forme d'onde	Levenberg-Marquardt	5.4
Imagerie ultrasonore			
	Formulation d'un problème de **reconstruction d'images à partir de données FMC**	Fast Iterative Shrinkage Algotihm (FISTA)	5.5.2
	Formulation d'un problème de **déconvolution à PSF variable à partir de données refocalisées**	FISTA	5.5.3
Optimisation parcimonieuse exacte en norme ℓ_0			
Problèmes génériques de type déconvolution et sélection de variables			
	Reformulation MIP des 3 formulations $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$, $\mathcal{P}_{2+0}$	Solveur CPLEX	6.1
		Méthodes branch-and-bound	6.2
	Problèmes en norme ℓ_1 avec variables libres et contraintes de borne	**Homotopie, contraintes actives**	6.2.5
Démélange spectral parcimonieux			
	Parcimonie en norme ℓ_0	Solveur CPLEX	6.3.2
	Contraintes structurantes	Solveur CPLEX	6.3.3

3.4 Références

[Alliney et Ruzinsky, 1994] Alliney, S. et Ruzinsky, S. A. (1994). An algorithm for the minimization of mixed l_1 and l_2 norms with application to Bayesian estimation. *IEEE Trans. Signal Process.*, 42(3):618–627.

[Bienstock, 1996] Bienstock, D. (1996). Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, 74(2):121–140.

[Bourguignon *et al.*, 2007] Bourguignon, S., Carfantan, H. et Idier, J. (2007). A sparsity-based method for the estimation of spectral lines from irregularly sampled data. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):575–585.

[Candès *et al.*, 2006] Candès, E. J., Romberg, J. K. et Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223.

[Chen *et al.*, 1989] Chen, S., Billings, S. A. et Luo, W. (1989). Orthogonal least squares methods and their application to non-linear system identification. *International Journal of control*, 50(5):1873–1896.

[Combettes et Pesquet, 2011] Combettes, P. L. et Pesquet, J.-C. (2011). *Proximal Splitting Methods in Signal Processing*, pages 185–212. Springer New York, New York, NY.

[Das et Dennis, 1997] Das, I. et Dennis, J. (1997). A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems. *Structural optimization*, 14(1):63–69.

[Donoho et Tsaig, 2008] Donoho, D. L. et Tsaig, Y. (2008). Fast solution of ℓ_1-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54:4789–4812.

[Efron *et al.*, 2004] Efron, B., Hastie, T., Johnstone, I. et Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

[Eldar et Kutyniok, 2012] Eldar, Y. et Kutyniok, G. (2012). *Compressed Sensing : Theory and Applications*. Cambridge University Press.

[Friedman *et al.*, 2007] Friedman, J., Hastie, T., Höfling, H. et Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332.

[Fuchs, 2004] Fuchs, J.-J. (2004). On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344.

[Gorodnitsky et Rao, 1997] Gorodnitsky, I. F. et Rao, B. D. (1997). Sparse signal reconstruction from limited data using FOCUSS : a re-weighted minimum norm algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616.

[Hogbom, 1974] Hogbom (1974). Aperture synthesis with a non-regular distribution of interferometer baselines. *Astron. Astrophys. Suppl.*, 15:417–426.

[Idier, 2008] Idier, J. (2008). *Bayesian Approach to Inverse Problems*. ISTE Ltd and John Wiley & Sons Inc.

[Lee *et al.*, 2007] Lee, H., Battle, A., Raina, R. et Ng, A. Y. (2007). Efficient sparse coding algorithms. *In Advances in Neural Information Processing Systems*, pages 801–808.

[Mallat et Zhang, 1993] Mallat, S. et Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415.

[Mendel, 1983] Mendel, J. M. (1983). *Optimal Seismic Deconvolution*. Academic Press.

[Miller, 2002] Miller, A. (2002). *Subset selection in regression*. Chapman and Hall/CRC.

[Moulin et Liu, 1999] Moulin, P. et Liu, J. (1999). Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. *IEEE Trans. Inf. Theory*, 45:909–919.

[Natarajan, 1995] Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234.

[O'Brien *et al.*, 1990] O'Brien, M. S., Sinclair, A. N. et Kramer, S. M. (1990). High resolution deconvolution using least-absolute-values minimization. *In Ultrasonics Symposium*, volume 2, pages 1151–1156.

[O'Brien *et al.*, 1994] O'Brien, M. S., Sinclair, A. N. et Kramer, S. M. (1994). Recovery of a sparse spike time series by L1 norm deconvolution. *IEEE Transactions on Signal Processing*, 42:3353–3365.

[Osborne *et al.*, 2000a] Osborne, M. R., Presnell, B. et Turlach, B. A. D. (2000a). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*.

[Osborne *et al.*, 2000b] Osborne, M. R., Presnell, B. et Turlach, B. A. D. (2000b). On the LASSO and its dual. *Journal of Computational and Graphical Statistics*, 9(2): 319–337.

[Pati *et al.*, 1993] Pati, Y., Rezaiifar, R. et Krishnaprasad, P. S. (1993). Orthogonal matching pursuit : recursive function approximation with applications to wavelet decomposition. *In Asilomar Conference on Signals, Systems and Computers*, pages 40–44 vol.1.

[Rockafellar, 1970] Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

[Soussen *et al.*, 2013] Soussen, C., Gribonval, R., Idier, J. et Herzet, C. (2013). Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares. *IEEE Transactions on Information Theory*, 59:3158–3174.

[Soussen *et al.*, 2011] Soussen, C., Idier, J., Brie, D. et Duan, J. (2011). From Bernoulli Gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584.

[Taylor *et al.*, 1979] Taylor, H., Banks, S. et McCoy, F. (1979). Deconvolution with the l_1 norm. *Geophysics*, 44(1):39–52.

[Tibshirani, 1996] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288.

[Tropp, 2006] Tropp, J. A. (2006). Just relax : Convex programming methods for identifying sparse signals. *IEEE Transactions on Information Theory*, 52(3):1030–1051.

[Tseng, 2001] Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494.

[Wu et Lange, 2008] Wu, T. T. et Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. App. Statist.*, 2(1):224–244.

[Zala, 1992] Zala, C. (1992). High-resolution inversion of ultrasonic traces. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 39(4):458–463.

Chapitre 4

Restauration de cubes hyperspectraux en Astrophysique

L'imagerie hyperspectrale s'est considérablement développée depuis une quinzaine d'années, aussi bien sur le plan méthodologique qu'au niveau de la diversité des domaines d'application, où l'arrivée de nouvelles technologies de mesure permet d'appréhender le champ scientifique sous de nouvelles formes. L'observation astronomique n'échappe pas à cette évolution, et de nombreux instruments récents et futurs acquièrent désormais des images dans un grand nombre de canaux spectraux [Léna *et al.*, 2008]. L'instrument MUSE, installé depuis 2013 sur le Très Grand Télescope (VLT) au Chili, est ainsi un précurseur de nouvelles générations de spectro-imageurs, produisant des cubes hyperspectraux d'environ 90 000 pixels sur 4 000 longueurs d'onde.

Par rapport à l'imagerie hyperspectrale classique de télédétection terrestre ou même planétaire, où les traitements sont centrés sur des questions de classification et de démélange spectral, le contexte de l'observation de l'Univers revêt un certain nombre de spécificités, justifiant le développement de méthodes dédiées. Tout d'abord, contrairement au cas classique où l'on mesure des spectres de réflectance (la portion de la lumière solaire réfléchie par la scène), nous observons ici des spectres d'émission lumineuse, dont la nature diffère d'une source à l'autre. Les sources étant très lointaines, le niveau de bruit dans les données est alors très élevé, la détection de la lumière émise étant perturbée par de nombreuses sources parasites, comme le fond de ciel. De plus, la taille de chaque source sur l'image est généralement restreinte à quelques pixels et de nombreuses zones de l'image ne contiennent que du bruit, ou bien des sources trop faibles pour être détectées. Enfin, pour des instruments d'observation situés au sol, comme c'est le cas de MUSE, les données subissent un certain nombre de dégradations dues en particulier aux turbulences atmosphériques, à l'optique du télescope et à l'instrument de mesure. L'exposé de ce contexte et des caractéristiques des données, menant à la formulation de problèmes inverses, fait l'objet de la Section 4.1.

Afin de pallier le déficit informationnel des données, nous avons construit un cadre méthodologique reposant sur la définition d'un dictionnaire de motifs élémentaires recherchés dans chaque spectre. Les données étant très bruitées, seules quelques composantes peuvent être détectées dans chaque spectre, traduisant une hypothèse de parcimonie des coefficients de décomposition dans le dictionnaire. Contrairement aux modèles parcimonieux classiques basés sur l'utilisation des espaces de représentation génériques (transformées en ondelettes par exemple), l'utilisation d'un dictionnaire de formes exploitant une connaissance physique sur les spectres recherchés permet d'associer une information de détection à chaque composante non nulle dans le modèle. Nous justifions ce choix et

détaillons la construction du dictionnaire en Section 4.2.

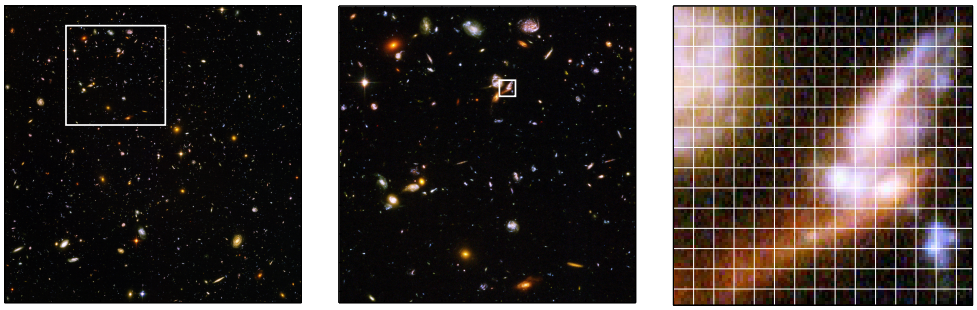
Nous avons abordé un premier problème de restauration de spectres, prenant en compte les spécificités du modèle observationnel (en particulier, la réponse instrumentale et le niveau de bruit varient avec la longueur d'onde), où chaque spectre d'un cube de données est approché par la combinaison linéaire d'un petit nombre d'éléments du dictionnaire. Cette approche formule un problème d'approximation parcimonieuse d'assez grande taille (environ 4 000 données pour 30 000 inconnues) et ne bénéficie pas d'algorithmes de calcul rapides, en raison de la construction *ad hoc* du dictionnaire. Nous avons alors proposé un algorithme d'optimisation en norme ℓ_1, basé sur un schéma de descente coordonnée par coordonnée (ICD) avec plusieurs étapes d'accélération, dont nous avons montré l'efficacité par rapport aux algorithmes concurrents [Bourguignon *et al.*, 2011c]. La Section 4.3 est consacrée à ces développements.

Nous avons ensuite cherché à exploiter ce modèle parcimonieux dans une approche visant à restaurer un cube d'images hyperspectrales, prenant en compte la réponse spatiale du dispositif d'acquisition. En effet, vu le très faible rapport signal sur bruit des données, la prise en compte d'un modèle d'étalement spatial vise à concentrer l'information « étalée » par l'instrument et ainsi à améliorer les performances de détection de sources faibles. Le verrou réside évidemment dans la complexité numérique due à la taille du problème. Nous avons proposé deux approches, décrites en Section § 4.4, basées sur un modèle parcimonieux en norme ℓ_1 et sur un algorithme glouton, incluant une phase préalable de réduction de dimension permettant d'alléger la complexité calculatoire.

Une conclusion et un certain nombre de poursuites possibles clôturent ce chapitre en Section 4.5.

4.1 Contexte astrophysique : mesures de champs lointains depuis le sol

Un objectif majeur de la nouvelle génération de spectro-imageurs concerne l'inspection de champs profonds de l'Univers. Un exemple de champ profond est représenté en figure 4.1, montrant une cartographie du champ ultra-profond de Hubble [1], résultant de



Champ total, 3 arcmin × 3 arcmin | Champ MUSE, 1 arcmin × 1 arcmin | Agrandissement, 3 arcsec × 3 arcsec

FIGURE 4.1 – Champ ultra-profond de Hubble, représenté à trois échelles différentes (source image NASA). L'image du centre (respectivement de droite) représente la zone matérialisée par le carré blanc sur l'image de gauche (respectivement du centre). Le quadrillage sur l'image de droite représente la taille des pixels des données MUSE.

l'acquisition d'un grand nombre d'images par le télescope spatial Hubble (*Hubble Space Telescope*, HST), où la couleur synthétique est obtenue à partir de trois images centrées sur des longueurs d'onde de 450, 606 et 814 nm. On peut y voir de nombreuses sources lumineuses, dont la couleur reflète la différence du contenu spectral. La taille de l'image du centre, de 1 arcmin × 1 arcmin, est représentative du champ de vue de MUSE.

Comparé à un instrument embarqué dans l'espace (comme le HST), un instrument tel que MUSE permet d'acquérir des données à très haute résolution spectrale, avec environ 4 000 longueurs d'onde couvrant le domaine visible et le proche infra-rouge du spectre électromagnétique, pour des longueurs d'onde de 365 à 930 nm. Les données produites souffrent en revanche d'un certain nombre de limitations, parmi lesquelles :

- la **résolution spatiale** : la taille d'un pixel sur le détecteur de l'instrument correspond à une surface d'environ 0.2 arcsec × 0.2 arcsec, bien plus grande que celle du télescope Hubble, comme le montre l'image de droite de la figure 4.1 ;
- le **niveau de bruit** : les sources observées, très lointaines, n'émettent que peu de lumière dans la direction de l'observateur. En résulte un rapport signal sur bruit très faible, en particulier en raison d'émissions parasites dues à l'atmosphère. La figure 4.2 montre ainsi des données simulées par les astronomes du consortium MUSE, dans une version non-bruitée (à gauche) et une version bruitée avec un niveau de bruit réaliste (à droite) ;

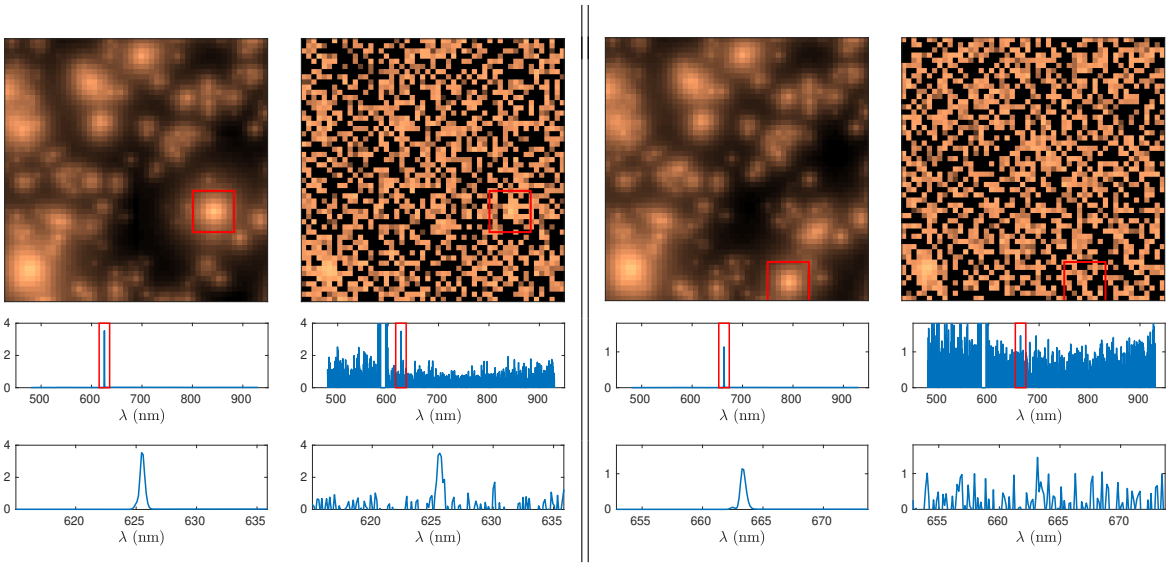


FIGURE 4.2 – Exemple de données simulées de type MUSE, correspondant à deux sources ponctuelles essentiellement constituées d'une raie d'émission. Pour chaque source, la colonne de gauche représente les données non bruitées, la colonne de droite les données bruitées. En haut, image à la longueur d'onde au centre de la raie d'émission (le carré rouge indique la position de la source), en échelle logarithmique d'amplitude. Au milieu et en bas, spectre correspondant à la position spatiale du centre de la source (le rectangle rouge indique la longueur d'onde centrale de la source). Les flux sont en erg.s^{-1}.cm^{-2} × 10^{-20}.

- l'**étalement des données** : en raison des effets conjugués de la turbulence atmosphérique, des limitations de l'optique du télescope et celles de l'instrument lui-même, l'observation d'une source ponctuelle et monochromatique va être étalée, à la fois spatialement et spectralement.

1. https://fr.wikipedia.org/wiki/Champ_ultra-profond_de_Hubble

L'ensemble de ces phénomènes a été étudié en profondeur (voir par exemple [Bacon *et al.*, 2006, Serre *et al.*, 2010, Villeneuve, 2012]). Nous en restituons ici les principales conclusions, dont la prise en compte s'avère primordiale pour les méthodes développées par la suite.

4.1.1 Un niveau de bruit variable en longueur d'onde

L'observation de sources astrophysiques depuis la Terre est dominée par l'émission parasite dite du *fond du ciel*, filtrée par l'atmosphère. Cette perturbation est particulièrement importante aux longueurs d'onde caractéristiques des composés chimiques présents dans l'atmosphère, induisant une variabilité spectrale du niveau de bruit. L'efficacité quantique de l'instrument (le rapport entre le nombre d'électrons collectés sur le capteur et le nombre de photons incidents) dépend également de la longueur d'onde. Par conséquent, le niveau de bruit entachant chaque donnée varie fortement selon l'axe spectral. La figure 4.3 montre un modèle de bruit fourni par les astrophysiciens du consortium MUSE. On y voit notamment des paquets de raies au-delà de 700 nm caractéristiques de la vapeur d'eau, l'effet de la variabilité en longueur d'onde de l'efficacité quantique (l'instrument est *meilleur* au centre), ainsi qu'un niveau de bruit « infini » autour de la longueur d'onde du sodium $\lambda_{\mathrm{Na}} = 589.2$ nm, en raison de l'émission parasite d'une étoile guide laser à cette longueur d'onde, nécessaire à l'utilisation de l'optique adaptative. Les astronomes sont

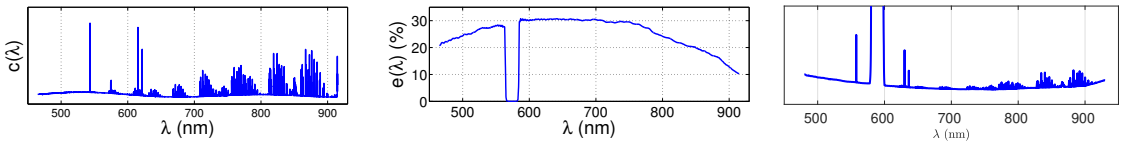


FIGURE 4.3 – Modèle de bruit utilisé pour des données hyperspectrales de type MUSE. À gauche, spectre d'émission du fond de ciel. Au centre, efficacité quantique du dispositif d'acquisition. À droite, spectre du niveau de bruit résultant (écart-type).

coutumiers de ce genre de modèle et, en pratique, une estimation de la variance du bruit est associée à chaque mesure. Nous supposerons par la suite que les différents échantillons de bruit sont indépendants, de variance différente, mais supposée connue, en tout point du cube de données.

4.1.2 Une réponse spatiale et spectrale, variable en longueur d'onde

La fonction d'étalement du point (*point spread function*, PSF) représente la signature, dans le cube de données, d'un modèle de source correspondant à une impulsion spatiale (source ponctuelle) et spectrale (source monochromatique) :

$$\delta(\boldsymbol{r} - \boldsymbol{r}_0, \lambda - \lambda_0) \xrightarrow{\text{observation}} \mathrm{PSF}_{\boldsymbol{r}_0,\lambda_0}(\boldsymbol{r}, \lambda).$$

Dans le cas de l'instrument MUSE, sa modélisation a été étudiée dans la thèse d'Emma Villeneuve [Villeneuve, 2012], dont nous reprenons certaines hypothèses. Nous la supposons séparable en le produit d'une réponse spatiale (*field spread function*, FSF) par une réponse spectrale (*line spread function*, LSF), lesquelles sont invariantes spatialement mais variables spectralement, l'étalement augmentant avec la longueur d'onde :

$$\mathrm{PSF}_{\boldsymbol{r}_0,\lambda_0}(\boldsymbol{r}, \lambda) = \mathrm{FSF}_{\lambda_0}(\boldsymbol{r})\ \mathrm{LSF}_{\lambda_0}(\lambda).$$

Selon [Villeneuve, 2012], la FSF peut être modélisée avec précision par une fonction de Moffat :

$$\mathrm{FSF}_{\lambda_0}(\boldsymbol{r}) \propto \left(1 + \frac{\|\boldsymbol{r}\|^2}{\alpha(\lambda_0)^2}\right)^{-\beta(\lambda_0)}$$

et la LSF peut être modélisée par une gaussienne :

$$\mathrm{LSF}_{\lambda_0}(\lambda) \propto \exp\left(-\frac{\lambda^2}{2\sigma(\lambda_0)^2}\right),$$

dont les paramètres $\alpha(\lambda_0)$, $\beta(\lambda_0)$ et $\sigma(\lambda_0)$ peuvent être calibrés à partir des données.

4.1.3 Formalisation d'un problème inverse

À partir de la modélisation ci-dessus et sous l'hypothèse de superposition des réponses instrumentales à des signatures élémentaires (hypothèse de linéarité), nous pouvons à présent relier les données hyperspectrales $y(\boldsymbol{r}, \lambda)$ à un cube hyperspectral « idéal » $x(\boldsymbol{r}, \lambda)$, sous la forme :

$$\begin{aligned} y(\boldsymbol{r}, \lambda) &\simeq \int_{\boldsymbol{r}_0, \lambda_0} x(\boldsymbol{r}_0, \lambda_0)\, \mathrm{PSF}_{\boldsymbol{r}_0, \lambda_0}(\boldsymbol{r}, \lambda)\, d\boldsymbol{r}_0\, d\lambda_0 \\ &= \int_{\lambda_0} \left(\int_{\boldsymbol{r}_0} x(\boldsymbol{r}_0, \lambda_0)\, \mathrm{FSF}_{\lambda_0}(\boldsymbol{r} - \boldsymbol{r}_0)\, d\boldsymbol{r}_0 \right) \mathrm{LSF}_{\lambda_0}(\lambda - \lambda_0)\, d\lambda_0. \end{aligned}$$

Considérant des données échantillonnées spatialement et spectralement, et en discrétisant de manière similaire le second membre, nous obtenons finalement un modèle :

$$\boldsymbol{y} = \mathbf{L}\mathbf{F}\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{4.1}$$

où les vecteurs colonne $\boldsymbol{y}$ et $\boldsymbol{x}$ représentent respectivement le cube de données et le cube « idéal », les produits par les matrices $\mathbf{L}$ et $\mathbf{F}$ représentent respectivement l'action de la LSF et de la FSF et $\boldsymbol{\epsilon}$ est un terme de perturbation modélisant le bruit et les erreurs de modèle.

4.2 Parcimonie spectrale dans un dictionnaire adapté

L'estimation de $\boldsymbol{x}$ à partir des données $\boldsymbol{y}$ dans le modèle (4.1) est un problème inverse particulièrement mal posé, en raison de la prédominance du bruit dans les données. Nous avons alors choisi d'introduire de fortes contraintes portant sur la dimension spectrale des données. Plusieurs raisons motivent ce choix. Tout d'abord, pour des données en champ profond, les sources ont une étendue spatiale limitée à quelques pixels (voire un seul pixel pour les plus lointaines) ; il n'est donc pas pertinent d'introduire de fortes contraintes structurant les données dans leur dimension spatiale. À l'inverse, le spectre observé en une position spatiale donnée correspond à l'émission lumineuse provenant d'un même objet. Les données présentent donc plus de cohérence selon l'axe spectral. C'est d'ailleurs dans la dimension spectrale que réside l'intérêt scientifique d'instruments comme MUSE : c'est en caractérisant une source par son spectre que l'on peut remonter à ses propriétés d'intérêt (composition chimique, distance, âge, ...).

4.2.1 Quelques éléments sur la physique des galaxies

Nous décrivons ici quelques propriétés de base des spectres des galaxies, à partir desquelles nous construirons un modèle permettant de régulariser le problème. Le spectre de la lumière émise par une galaxie est avant tout composé de raies, en émission et en absorption, caractéristiques de sa composition chimique [Tennyson, 2005]. Ces raies peuvent être de largeur variable, notamment en raison de la cinématique interne de la galaxie, produisant un étalement de la raie observée par effet Doppler Mais surtout, si les longueurs d'onde associées à la plupart des éléments chimiques sont bien caractérisées par des mesures en laboratoire, ce n'est plus le cas lorsque l'on observe une source astrophysique lointaine, dont la vitesse d'éloignement par rapport à l'observateur produit un *décalage vers le rouge*, c'est-à-dire une augmentation de la longueur d'onde observée. C'est à partir de mesures spectrales que les astrophysiciens peuvent remonter à la vitesse, et donc à la distance, de l'objet observé (plus l'objet est loin, plus il s'éloigne vite), par le paramètre de *redshift* z vérifiant $1+z = \lambda_{\text{obs}}/\lambda_{\text{lab}}$, quelle que soit la longueur d'onde. Des instruments comme MUSE visent ainsi la détection de galaxies atteignent des redshifts de plus de 6, correspondant à des distances de 10 milliards d'années-lumière.

Par ailleurs, un cas qui intéresse particulièrement les astronomes concerne la détection et la caractérisation de galaxies à cassure de Lyman [Tennyson, 2005]. La cassure de Lyman (*Lyman break*) correspond à une discontinuité dans le spectre d'émission lumineuse, due au fait que le rayonnement émis à des énergies supérieures à la limite de Lyman est presque complètement absorbé par les nuages d'hydrogène entourant les régions de formation d'étoiles des galaxies. Cette limite correspond aux longueurs d'onde inférieures à 91.2 nm. En raison du phénomène de décalage vers le rouge, cette discontinuité peut être observée dans la partie visible et proche infra-rouge pour des galaxies suffisamment lointaines.

Enfin, les spectres de galaxies présentent en général un continuum à variations lentes (la ligne de base, que l'on retrouve dans de nombreuses données de spectroscopie). Si sa compréhension et sa caractérisation présentent moins d'intérêt au niveau physique, sa prise en compte dans un modèle de spectre est essentielle afin de ne pas perturber la détection et l'estimation des autres composantes.

4.2.2 Construction d'un dictionnaire

Afin de construire un modèle de spectre prenant en compte ces différentes caractéristiques, je me suis orienté vers une approche parcimonieuse, où chaque spectre est modélisé à partir de motifs élémentaires pris dans un dictionnaire contenant des raies de centre et de largeur variables, des discontinuités de positions variables, et une modélisation paramétrée de la partie continue. Considérons l'axe $\{\lambda_n\}_{n=1,\dots,N_\lambda}$ de longueurs d'ondes échantillonnées défini pour des données de type MUSE : $\lambda_n \in [\lambda_{\min}, \lambda_{\max}] = [465\text{ nm}, 930\text{ nm}]$, avec un pas de discrétisation de $\delta_\lambda = 0.13$ nm. Nous construisons le dictionnaire de raies $\mathbf{D}_{\text{lines}}$ composé :

- d'impulsions en tout λ_n, modélisant les raies non résolues (*i.e.,* dont la largeur est inférieure à δ_λ) ;
- de splines [2] centrées en λ_n et de largeur variable entre $\Delta_{\min} = 3\delta_\lambda = 0.39$ nm et $\Delta_{\max} = 18$ nm, cette dernière valeur correspondant à une limite physique sur la largeur attendue des raies. Afin de limiter la taille du dictionnaire et d'éviter de trop

2. Nous avons choisi des splines car leur construction est simple et leur support est fini.

fortes corrélations entre ses différents éléments, nous avons défini 11 largeurs de raies et imposé un écartement minimal de $\sigma = \Delta/8$ entre deux raies de largeur Δ.

Le tableau 4.1 résume les paramètres associés au dictionnaire de raies.

Δ (nm)	0.13	0.39	0.65	1.17	1.43	2.21	3.25	4.55	6.37	8.97	12.61	17.94
($\times\delta_\lambda$)	1	3	5	9	11	17	25	35	49	69	97	138
σ (nm)	0.13	0.13	0.13	0.13	0.13	0.26	0.39	0.52	0.78	1.17	1.56	2.21
($\times\delta_\lambda$)	1	1	1	1	1	2	3	4	6	9	12	17

TABLE 4.1 – Constitution du dictionnaire de raies spectrales : largeurs de raie (Δ) et écart entre deux raies successives de même largeur (σ). Le pas d'échantillonnage en longueur d'onde est $\delta\lambda = 13$ nm.

Pour modéliser la possible présence de discontinuités, nous avons considéré un dictionnaire $\mathbf{D}_{\text{steps}}$ composé d'échelons localisés en tout point de l'axe de longueurs d'onde. Enfin, afin d'inclure la partie continue dans la modélisation parcimonieuse, nous avons construit un dictionnaire $\mathbf{D}_{\cos}$ composé de sinusoïdes de variations lentes : $c_{k,\ell}(\lambda) = \sin(2\pi f_k \lambda + \phi_\ell)$, avec $f_k = k/(\lambda_{\max} - \lambda_{\min})$, $k = 0, \ldots, 8$ et, pour $f_k \neq 0$, $\phi_\ell = \ell\pi/8, \ell = 0, \ldots, 7$. Cette partie du modèle peut sembler artificielle au niveau de l'interprétation physique, les spectres de galaxie n'ayant aucune raison de présenter des composantes oscillantes ! Rappelons cependant que notre objectif est ici de paramétrer la partie continue pour pouvoir l'estimer conjointement aux autres composantes, sans accorder trop d'importance à sa forme.

Le modèle pour un spectre $\boldsymbol{s} \in \mathbb{R}^N$ s'écrit alors finalement :

$$\boldsymbol{s} = \mathbf{D}\boldsymbol{u}, \text{ avec } \mathbf{D} = [\mathbf{D}_{\text{lines}}, \mathbf{D}_{\text{steps}}, \mathbf{D}_{\cos}] \text{ et } \boldsymbol{u} \text{ parcimonieux.} \tag{4.2}$$

Pour un spectre mesuré en 4 000 longueurs d'onde, la taille du dictionnaire est d'environ 30 000 éléments. La figure 4.4 représente un exemple de spectre synthétique en petite dimension.

4.3 Restauration de spectres

Dans une première exploitation du modèle spectral établi en Section 4.2, j'ai abordé la restauration des spectres considérés indépendamment les uns des autres. On considère alors le modèle d'observation :

$$\boldsymbol{y} = \mathbf{L}\boldsymbol{s} + \boldsymbol{\epsilon},$$

où le produit par $\mathbf{L}$ représente l'action de la LSF (la réponse spectrale du système d'acquisition, voir le § 4.1.3). Sous l'hypothèse du modèle (4.2), le problème revient à estimer un vecteur $\widehat{\boldsymbol{u}}$ parcimonieux satisfaisant le modèle :

$$\boldsymbol{y} = \mathbf{L}\mathbf{D}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{4.3}$$

puis à reconstruire le spectre sous la forme $\widehat{\boldsymbol{s}} = \mathbf{D}\widehat{\boldsymbol{u}}$. J'ai pour cela envisagé une approche basée sur l'optimisation d'un critère d'ajustement de modèle de type moindres carrés pénalisé en norme ℓ_1. Cependant, la prise en compte d'un modèle de bruit non *i.i.d.* requiert des précautions spécifiques, qui sont décrites au § 4.3.1. Une interprétation du paramètre de régularisation est proposée au § 4.3.2. Le § 4.3.3 décrit l'algorithme d'optimisation que j'ai proposé spécifiquement pour ce problème et quelques résultats sont présentés au § 4.3.4.

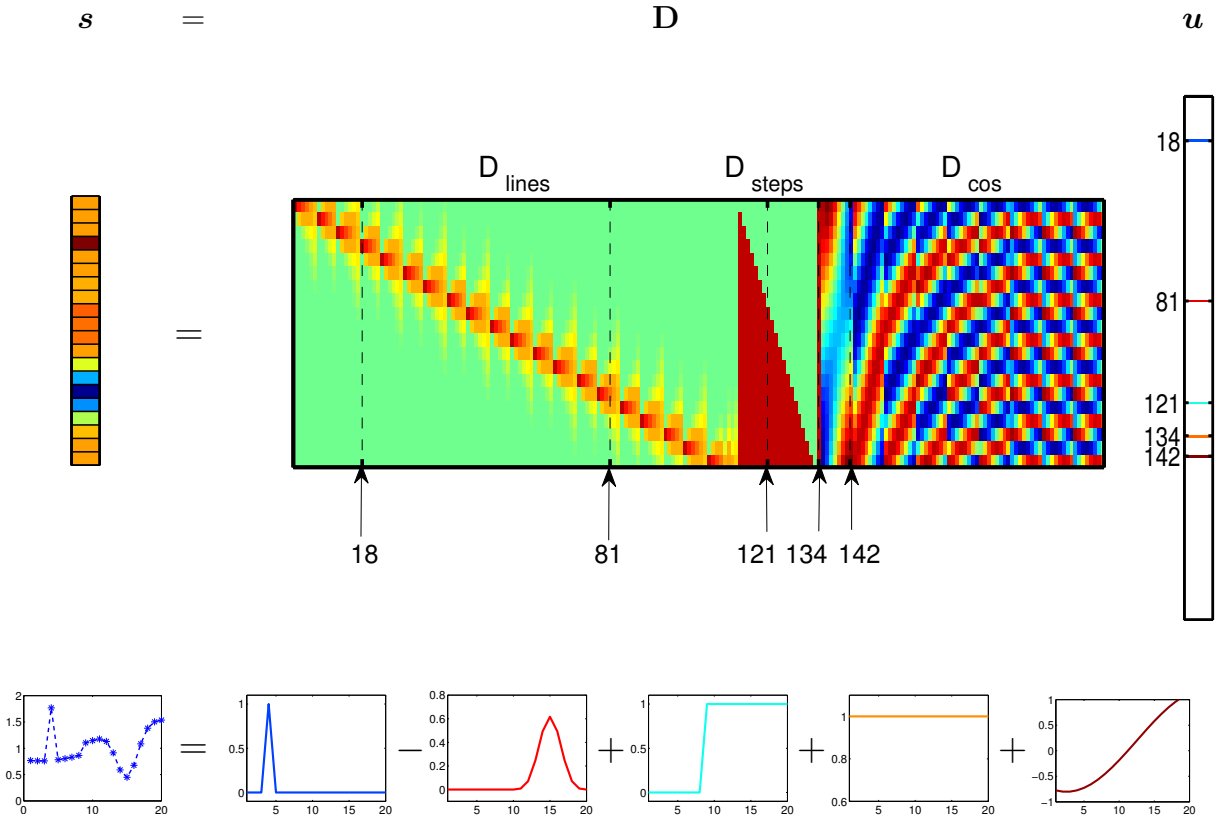


FIGURE 4.4 – Un exemple de spectre synthétisé à partir du dictionnaire $\mathbf{D} = [\mathbf{D}_{\text{lines}}, \mathbf{D}_{\text{steps}}, \mathbf{D}_{\text{cos}}]$. Sur la première ligne, l'axe vertical représente la longueur d'onde, et chaque colonne de la matrice $\mathbf{D}$ représente un motif élémentaire. La seconde ligne représente le spectre synthétique (à gauche) et les 5 atomes associés : d'une raie de largeur unité (atome 18 du dictionnaire), d'une raie plus large affectée d'un coefficient négatif (atome 81), d'une discontinuité (atome 121), d'une composante constante (atome 134) et d'un terme croissant (atome 142).

4.3.1 Prise en compte du bruit et dictionnaire équivalent

Nous avons vu au § 4.1.1 que le niveau de bruit variait fortement avec la longueur d'onde. Nous supposons par la suite que les différents échantillons de bruit sont indépendants et que la variance est connue en chaque point. Sous l'hypothèse que l'échantillon ϵ_n suit une loi normale, centrée et de variance σ_n^2, l'anti-log-vraisemblance du modèle (4.3) s'écrit, en notant $\mathbf{\Sigma} \triangleq \mathrm{diag}(\sigma_1^2, \ldots, \sigma_{N_\lambda}^2)$ la matrice de covariance du bruit :

$$\begin{aligned} \frac{1}{2}\|\boldsymbol{y} - \mathbf{LD}\boldsymbol{u}\|_{\mathbf{\Sigma}}^2 &= \frac{1}{2}(\boldsymbol{y} - \mathbf{LD}\boldsymbol{u})^T \mathbf{\Sigma}^{-1}(\boldsymbol{y} - \mathbf{LD}\boldsymbol{u}) \\ &= \frac{1}{2}\left\|\mathbf{\Sigma}^{-1/2}\boldsymbol{y} - \mathbf{\Sigma}^{-1/2}\mathbf{LD}\boldsymbol{u}\right\|^2, \end{aligned}$$

avec $\mathbf{\Sigma}^{-1/2} \triangleq \mathrm{diag}(\sigma_1^{-1}, \ldots, \sigma_{N_\lambda}^{-1})$. La seconde écriture ré-écrit le terme d'ajustement de modèle comme un terme en norme ℓ_2 standard, opérant sur des données *pré-blanchies* $\mathbf{\Sigma}^{-1/2}\boldsymbol{y}$, avec un *dictionnaire pré-blanchi* $\mathbf{\Sigma}^{-1/2}\mathbf{LD}$. Cette forme met en évidence un point important au moment d'envisager la régularisation parcimonieuse, et qui n'apparaît pas sous l'hypothèse classique d'un bruit *i.i.d.* : les différents atomes du dictionnaire pré-blanchi sont de norme différente, même si le dictionnaire initial $\mathbf{D}$ est normalisé. Par conséquent, il est impératif de normaliser les colonnes du dictionnaire équivalent, sans quoi les différents coefficients du vecteur $\boldsymbol{u}$ ne sont pas « à la même échelle » et les statistiques de détection des différents atomes en sont faussées. Cette étape est cruciale dans notre cas, où la variabilité du niveau de bruit est très forte (voir la figure 4.3). Un argumentaire plus complet et des simulations détaillées sur cet aspect peuvent être trouvés dans [Bourguignon *et al.*, 2010].

Nous considérons donc finalement le problème d'optimisation équivalent :

$$\widehat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} J(\boldsymbol{u}), \text{ où } J(\boldsymbol{u}) \triangleq \frac{1}{2}\|\boldsymbol{z} - \mathbf{A}\boldsymbol{u}\|^2 + \mu\,\|\boldsymbol{u}\|_1\,, \tag{4.4}$$
$$\text{avec } \boldsymbol{z} \triangleq \mathbf{\Sigma}^{-1/2}\,\boldsymbol{y} \text{ et le dictionnaire équivalent } \mathbf{A} \triangleq \mathbf{\Sigma}^{-1/2}\,\mathbf{LDN}^{-1},$$

où $\mathbf{N}$ est la matrice diagonale contenant les normes des colonnes de $\mathbf{\Sigma}^{-1/2}\,\mathbf{LD}$.

4.3.2 Interprétation et réglage de l'hyperparamètre μ

Le réglage des hyperparamètres)est toujours un point critique dans les approches pénalisées des problèmes inverses. Dans le cas d'une pénalisation en norme ℓ_1, nous avons proposé une règle simple et fournissant des résultats satisfaisants dans nos simulations. Elle consiste à exploiter les conditions d'optimalité de la solution $\widehat{\boldsymbol{u}}$ au problème (4.4). En indexant par NZ (respectivement Z) les variables non-nulles (respectivement nulles) de $\widehat{\boldsymbol{u}}$, celles-ci s'écrivent [Alliney et Ruzinsky, 1994] :

$$\widehat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} J(\boldsymbol{u}) \Leftrightarrow \begin{cases} \mathbf{A}_{\mathrm{NZ}}^T(\boldsymbol{z} - \mathbf{A}_{\mathrm{NZ}}\widehat{\boldsymbol{u}}_{\mathrm{NZ}}) = \mu\,\mathrm{sgn}(\widehat{\boldsymbol{u}}_{\mathrm{NZ}}) \\ \left|\mathbf{A}_{\mathrm{Z}}^T(\boldsymbol{z} - \mathbf{A}_{\mathrm{NZ}}\widehat{\boldsymbol{u}}_{\mathrm{NZ}})\right| \leq \mu \end{cases}. \tag{4.5}$$

Supposons que les données ne contiennent que du bruit : $\boldsymbol{z} = \mathbf{\Sigma}^{-1/2}\,\boldsymbol{\epsilon}$. Alors, on souhaite avoir $\widehat{\boldsymbol{u}} = \mathbf{0}$, *i.e.,* pas de fausses détections. Par construction, $\boldsymbol{z}$ suit une loi de moyenne nulle et de matrice de covariance égale à l'identité. Sous l'hypothèse supplémentaire que cette loi est gaussienne, les coefficients $\boldsymbol{a}_p^T\boldsymbol{z}$ sont également gaussiens, de moyenne nulle

et de variance unité (les colonnes $\boldsymbol{a}_p$ du dictionnaire $\mathbf{A}$ sont de norme unité). D'après la seconde condition ci-dessus, μ peut alors être vu comme la valeur maximale prise parmi p variables gaussiennes centrées réduites. Ainsi, sur ces données, où le nombre d'éléments dans le dictionnaire est de l'ordre de $P \simeq 30\,000$, un réglage de $\mu = 4$ s'est avéré efficace.

4.3.3 Optimisation dédiée : un algorithme de type ICD avec accélérations

De nombreux algorithmes ont été proposés pour l'optimisation de critères tels que (4.4), dans le cadre des représentations parcimonieuses [Tropp et Wright, 2010]. Si une vaste majorité des travaux se concentre aujourd'hui sur les méthodes proximales (voir par exemple [Combettes et Pesquet, 2011]), celles-ci requièrent l'évaluation répétée du gradient de la partie quadratique, reposant sur l'évaluation de produits matrice-vecteur de type $\mathbf{A}\cdot$ et $\mathbf{A}^T\cdot$. L'efficacité de ces calculs numériques dépend fortement de l'opérateur $\mathbf{A}$. Dans de nombreux travaux, l'opérateur $\mathbf{A}$ représente une décomposition dans des familles de fonctions pour laquelle des algorithmes de calcul rapides existent (Transformée en Cosinus Discrète, en ondelettes). Ici, nous avons fait le choix de la construction d'un dictionnaire *ad hoc*, plus adapté en termes de modèle de spectre, mais n'induisant pas de telles propriétés de calcul rapide.

Je me suis donc tourné vers des stratégies de descente coordonnée par coordonnée (ICD), sur lesquelles j'avais déjà travaillé lors de ma thèse. Bien que généralement déconseillée en optimisation, la stratégie ICD s'avère particulièrement adaptée pour des critères en norme ℓ_1. Il est intéressant de noter que ces algorithmes ont également été remis au goût du jour, à une époque proche de ces travaux, dans la communauté statistique [Friedman *et al.*, 2007, Wu et Lange, 2008].

L'algorithme ICD repose sur l'optimisation successive et répétée de la fonction de coût en chacune de ses variables scalaires. En notant $J(\boldsymbol{u}) = J(u_1, \ldots, u_P)$, l'itération t, correspondant à un balayage de l'ensemble des variables, s'écrit :

$$\forall p = 1, \ldots, P,\ u_p^{(t)} = \arg\min_{u_p} J(u_1^{(t)}, \ldots, u_{p-1}^{(t)}, u_p, u_{p+1}^{(t-1)}, \ldots, u_P^{(t-1)}).$$

On montre alors facilement que, dans le cas du critère (4.4), cette mise à jour s'écrit :

$$u_p^{(t)} = \mathcal{S}_\mu \left(\boldsymbol{a}_p^T (\boldsymbol{z} - \boldsymbol{e}_\ell^{(t)}) \right),\ \text{avec}\ \boldsymbol{e}_p^{(t)} = \sum_{q<p} u_q^{(t)} \boldsymbol{a}_q + \sum_{q>p} u_q^{(t-1)} \boldsymbol{a}_q, \tag{4.6}$$

où $\mathcal{S}_\mu$ est l'opérateur de *seuillage doux* [Donoho et Johnstone, 1994] :

$$\begin{cases} \mathcal{S}_\mu(x) = 0 \text{ si } |x| \leq \mu, \\ \mathcal{S}_\mu(x) = x - \ \mu \text{sgn}(x) \text{ sinon.} \end{cases}$$

Chaque mise à jour est donc très peu coûteuse, bénéficiant de la mise à jour récursive de $\boldsymbol{e}_p^{(t)}$. Par ailleurs, les itérés étant souvent parcimonieux, on peut se contenter de ne mettre à jour le plus souvent que les composantes non nulles, en ne balayant l'ensemble des composantes que périodiquement, sans altérer les garanties de convergence [Bourguignon *et al.*, 2011c].

J'ai également proposé différentes accélérations de cet algorithme, qui reposent sur les deux remarques suivantes.

1. Les itérés $\boldsymbol{u}^{(t)}$ deviennent rapidement parcimonieux au fil des itérations et le *support signé* de $\boldsymbol{u}^{(t)}$, défini comme la partition des variables en composantes positives, négatives et nulles, est rapidement proche de celui du minimiseur.

2. Il est possible de tester analytiquement l'optimalité d'un support signé. En effet, les conditions d'optimalité (4.5) s'écrivent encore [Fuchs, 2004] :

$$\widehat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} J(\boldsymbol{u}) \;\Leftrightarrow\; \begin{cases} \widehat{\boldsymbol{u}}_{\mathrm{NZ}} = \left(\mathbf{A}_{\mathrm{NZ}}^T \mathbf{A}_{\mathrm{NZ}}\right)^{-1} \mathbf{A}_{\mathrm{NZ}}^T (\boldsymbol{z} - \mu \operatorname{sgn}(\widehat{\boldsymbol{u}}_{\mathrm{NZ}})) & (i) \\ \left|\mathbf{A}_{\mathrm{Z}}^T (\boldsymbol{z} - \mathbf{A}_{\mathrm{NZ}} \widehat{\boldsymbol{u}}_{\mathrm{NZ}})\right| \leq \mu & (ii) \end{cases}$$

   Ainsi, si un itéré possède le bon support signé, la mise en œuvre de (i) permet d'optimiser les amplitudes associées. La condition (ii) permet alors de vérifier l'optimalité.

Les accélérations proposées dans [Bourguignon *et al.*, 2011c] consistent, d'une part, à tester régulièrement si le support signé de l'itéré courant est le bon, et d'autre part à tester l'optimalité de supports proches du support courant, en cherchant par exemple à annuler des composantes de faible amplitude.

4.3.4 Résultats

Nous présentons maintenant un exemple de résultat, repris de l'article [Bourguignon *et al.*, 2011c]. Il s'agit de données issues de simulations réalisées par les astronomes du Centre de Recherche en Astrophysique de Lyon – en particulier, elles n'ont pas été générées avec le modèle parcimonieux utilisé pour la reconstruction. La figure 4.5a) représente le spectre simulé. On peut y voir un certain nombre de raies en émission, ainsi qu'un continuum présentant une cassure vers $\lambda_{\mathrm{break}} = 670$ nm. Le spectre de variance du bruit est celui de la figure 4.3, et les données bruitées sont représentées en figure 4.5b). Hormis les quelques raies les plus intenses, il est très difficile de détecter d'autres composantes par inspection visuelle. Le spectre reconstruit par $\widehat{\boldsymbol{s}} = \mathbf{D}\widehat{\boldsymbol{u}}$, où $\widehat{\boldsymbol{u}}$ est le minimiseur du critère (4.4), est représenté en figure 4.5c). On peut y détecter plusieurs raies moins intenses, ainsi que la cassure.

Un éclairage sur ces détections est apporté par l'agrandissement du résultat autour de la raie détectée vers 625 nm, représenté en figure 4.6a). S'il y a bien un « pic » dans les données, celui-ci est d'amplitude faible, inférieure par exemple un autre pic produit vers 630 nm, dû au bruit. En revanche, les données contiennent clairement plusieurs valeurs positives consécutives autour de 625 nm et la probabilité que ces événements soient dus au bruit est très faible. C'est donc bien parce que l'on a cherché des raies étendues *via* le modèle (4.2) que l'on a pu détecter cette composante. Une analyse plus poussée en figure 4.6b) montre enfin que cette raie, asymétrique, est en fait estimée par la combinaison de deux atomes du dictionnaire. Cet exemple illustre à la fois l'intérêt d'une approche à base de dictionnaire (permettant de rechercher des formes spécifiques dans les données) et ses limites : il est illusoire de chercher à définir un dictionnaire représentant exactement les motifs recherchés.

La figure 4.7 présente enfin le comportement de plusieurs algorithmes d'optimisation du critère (4.4) sur ces mêmes données (deux premières lignes). Plusieurs variantes de l'approche ICD sont mises en œuvre et comparées à la méthode homotopique [Efron *et al.*, 2004, Donoho et Tsaig, 2008], à l'algorithme FISTA [Beck et Teboulle, 2009] et à deux méthodes proximales, GPSR [Figueiredo *et al.*, 2007] et SparSA [Wright *et al.*, 2009]. Ces trois dernières approches se révèlent très peu efficaces, requérant plusieurs dizaines de minutes pour atteindre l'optimum. Une analyse détaillée des itérations a pu montrer que, le dictionnaire étant très corrélé, les itérés « hésitent » entre les différentes composantes à activer dans la décomposition. La figure 4.7 en haut à droite montre ainsi que le support

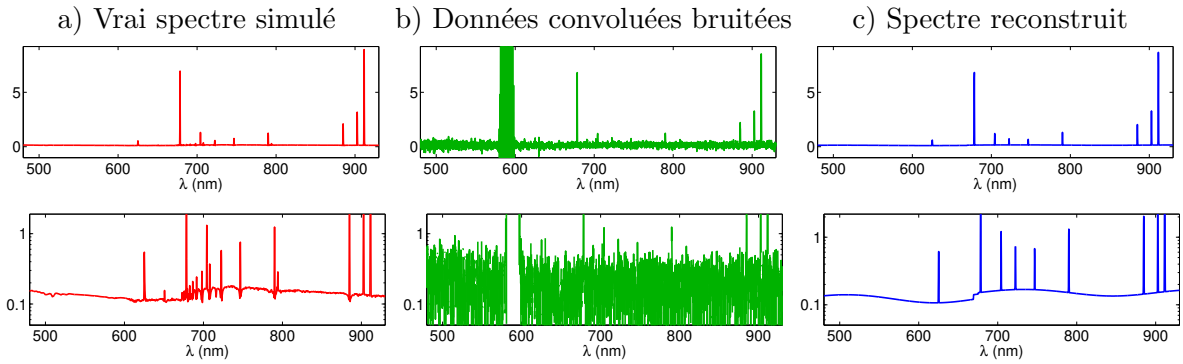


FIGURE 4.5 – Résultat d'estimation obtenu sur des simulations astrophysiques. À gauche, spectre non bruité. Au centre, données étalées par la LSF et bruitées. À droite, spectre estimé. Les amplitudes de flux lumineux sont en erg.s^{-1}.cm^{-2} $\times 10^{-20}$ et les spectres sont en échelles d'amplitude linéaire (en haut) et logarithmique (en bas). Figure reprise de [Bourguignon *et al.*, 2011c].

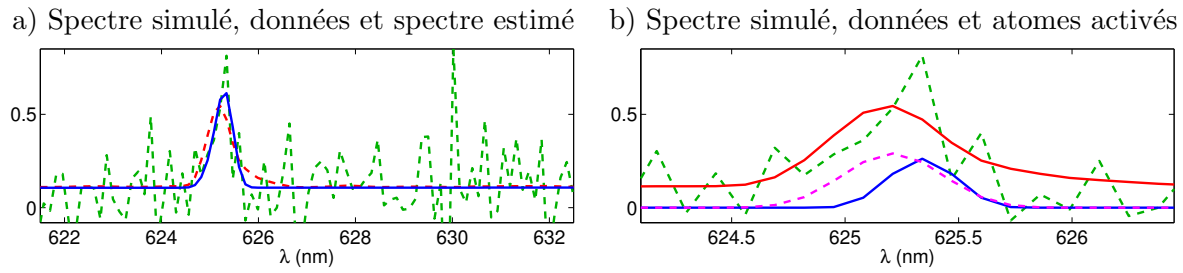


FIGURE 4.6 – Agrandissement des résultats de la figure 4.5 autour de la raie détectée vers 625 nm. À gauche, spectre simulé, données bruitées et spectre estimé (le code couleur reprend celui de la figure 4.5). À droite, détail des atomes ayant contribué à la modélisation (magenta et bleu). Les amplitudes de flux lumineux sont en erg.s^{-1}.cm^{-2} $\times 10^{-20}$. Figure reprise de [Bourguignon *et al.*, 2011c].

correct n'est atteint que peu d'itérations avant la convergence. Notons qu'il ne s'agit pas d'un problème de mise en œuvre : ces mêmes algorithmes, évalués sur des données simulées de même taille mais avec un dictionnaire aléatoire, se sont révélés bien plus efficaces, le temps de calcul se réduisant alors à quelques secondes [Bourguignon *et al.*, 2011c]. C'est bien la difficulté du problème, et en particulier la forte corrélation entre les éléments du dictionnaire, qui est à l'origine de leur mauvais fonctionnement.

La version brute de l'algorithme ICD, où chaque itération balaie l'ensemble des variables mises à jour par (4.6), n'est pas plus efficace. La restriction du balayage aux composantes non nulles (l'ensemble des variables étant balayé toutes les 250 itérations) permet de réduire très fortement le temps de calcul, à une dizaine de secondes. La figure en bas à droite montre enfin que le support correct avait été identifié après environ 4 secondes. L'ajout d'une étape testant l'optimalité du support, décrite au § 4.3.3, permet alors d'économiser plus de la moitié du temps de calcul (courbes rouge et verte). Enfin, la dernière ligne de la figure 4.7 représente le comportement des algorithmes les plus efficaces sur un autre jeu de données où, malgré la règle de balayage optimisée et l'étape de test du support, l'algorithme ICD s'avère assez lent, et en particulier moins efficace que la méthode homotopique. Dans ce cas, l'ajout d'étapes d'exploration locale du support (voir le § 4.3.3) permet de réduire significativement le temps de calcul. Au final, le temps de calcul requis pour la décomposition de spectres simulés dans des conditions observationnelles réalistes pour l'instrument MUSE était (en 2011) en général compris entre 5 et 10 secondes.

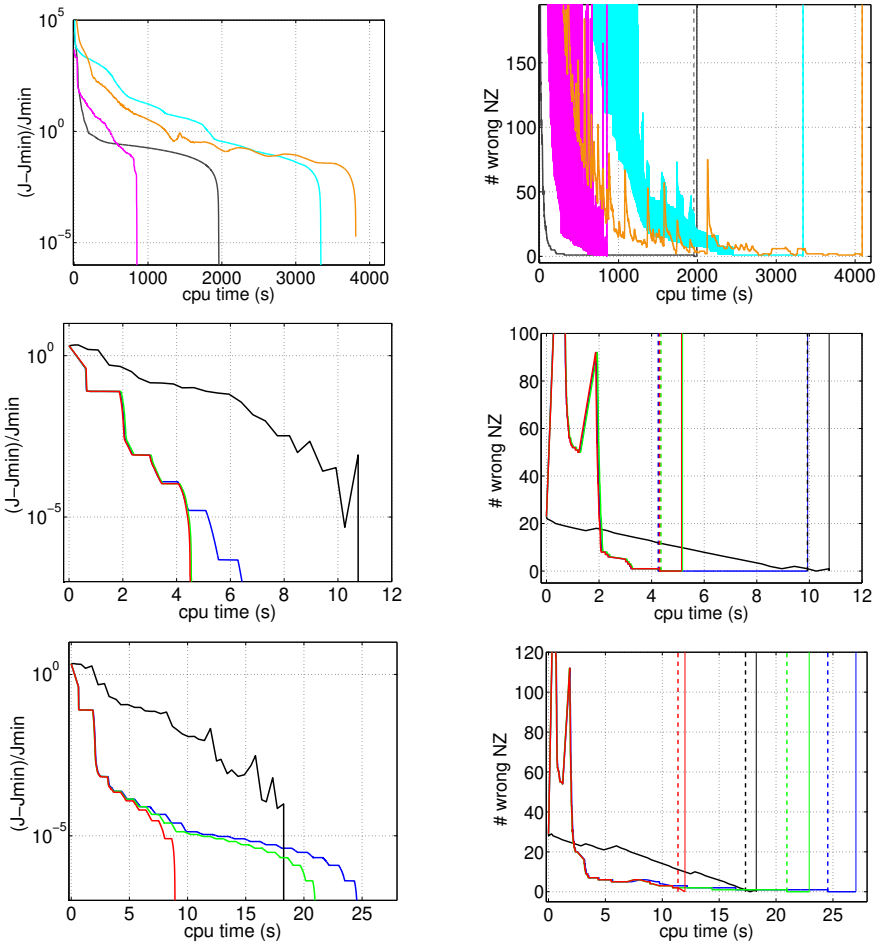


FIGURE 4.7 – Comportement de différents algorithmes d'optimisation du critère (4.4). Sur la ligne du haut, FISTA, GPSR, ICD brut et SpaRSA. Sur les lignes du milieu et du bas : homotopie (en noir), et plusieurs variantes d'ICD : avec balayage des composantes non nulles, auquel on inclut le test du support, et des explorations locales. À gauche, évolution du critère en fonction du temps. À droite, nombre d'erreurs sur le support. Les lignes verticales pleines marquent l'instant de convergence de chaque algorithme. Les lignes verticales en pointillé marquent l'instant où le support a été identifié. Les deux premières lignes correspondent au même jeu de données, qui est celui de la figure 4.5. La ligne du bas correspond à un autre spectre, sur lequel l'apport des accélérations proposées est mis en évidence. Figure reprise de [Bourguignon *et al.*, 2011c].

4.4 Restauration spatiale-spectrale

J'ai cherché à exploiter le modèle parcimonieux de spectres pour la restauration d'un cube d'images hyperspectrales, avec un modèle d'observation intégrant la réponse spatiale du dispositif d'acquisition (FSF, voir le § 4.1.3). Vu le fort niveau de bruit dans les données, il est clair que la prise en compte de la FSF doit permettre d'améliorer les performances de détection de sources faibles.

Reprenons le modèle d'observation (4.1) d'un cube hyperspectral : $\boldsymbol{y} = \mathbf{LF}\boldsymbol{x} + \boldsymbol{\epsilon}$, où $\boldsymbol{y}$ contient l'ensemble des Q spectres $\boldsymbol{y}_q$, $q = 1, \ldots, Q$ du cube, concaténés dans un vecteur colonne, et $\boldsymbol{x}$ représente similairement le cube à restaurer, sous la forme d'un vecteur colonne concaténant les spectres $\boldsymbol{x}_q$. Sous l'hypothèse que chaque spectre $\boldsymbol{x}_q$ admet une décomposition parcimonieuse sous la forme $\boldsymbol{x}_q = \mathbf{D}\boldsymbol{u}_q$, le problème revient alors à estimer conjointement les vecteurs parcimonieux $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_Q$ sous le modèle

$$\boldsymbol{y} = \mathbf{LFD}^{\text{bloc}}\boldsymbol{u} + \boldsymbol{\epsilon}, \tag{4.7}$$

où la notation $\mathbf{D}^{\text{bloc}}$ représente la matrice bloc-diagonale composée de Q blocs contenant $\mathbf{D}$ et $\boldsymbol{u}$ représente la concaténation en colonne des vecteurs $\boldsymbol{u}_q$. L'originalité de ce modèle est de formuler un problème de déconvolution à la fois spatiale et spectrale, mais au moyen d'information *a priori* portant uniquement sur la dimension spectrale.

Le verrou est évidemment au niveau de la taille des problèmes et du coût de calcul associé : pour chaque spectre $\boldsymbol{x}_q$ d'environ 4 000 points correspond un vecteur $\boldsymbol{u}_q$ d'environ 30 000 coefficients. Deux pistes ont été étudiées. La première considère une approche en norme ℓ_1 après une étape préalable de réduction de la dimension du problème *via* la pré-sélection d'atomes dans le dictionnaire. La seconde privilégie un algorithme glouton, limitant de fait le nombre de calculs par le fort degré de parcimonie attendu sur la solution.

4.4.1 Approche en norme ℓ_1

Nous avons proposé une approche en deux étapes, où :

i) une décomposition parcimonieuse de chaque spectre est d'abord réalisée, comme en Section 4.3. Le paramètre de régularisation μ dans le critère (4.4) est ici choisi suffisamment bas, afin de favoriser la présence de composantes de faible amplitude, au prix de plus nombreuses fausses alarmes.

ii) Nous considérons alors un modèle spatial-spectral de type (4.7), mais où le dictionnaire $\widetilde{\mathbf{D}}$ est restreint aux atomes sélectionnés dans au moins un spectre à l'étape i). Une représentation parcimonieuse $\widetilde{\boldsymbol{u}}_q$ de chaque spectre est alors recherchée dans $\widetilde{\mathbf{D}}$ en optimisant à nouveau un problème en norme ℓ_1. Par une démarche analogue à celle du § 4.3.1, celui-ci s'écrit :

$$\min_{\widetilde{\boldsymbol{u}}} \frac{1}{2} \left\| \boldsymbol{z} - \mathbf{B}\widetilde{\boldsymbol{u}} \right\|^2 + \alpha \left\| \widetilde{\boldsymbol{u}} \right\|_1, \text{ avec } \boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y} \text{ et } \mathbf{B} = \boldsymbol{\Sigma}^{-1/2}\mathbf{LF}\widetilde{\mathbf{D}}^{\text{bloc}}\mathbf{N}^{-1}, \tag{4.8}$$

où la notation $\widetilde{\mathbf{D}}^{\text{bloc}}$ est immédiate, $\boldsymbol{\Sigma}$ représente la matrice de covariance du bruit et $\mathbf{N}^{-1}$ correspond à la normalisation des colonnes de $\boldsymbol{\Sigma}^{-1/2}\mathbf{LF}\widetilde{\mathbf{D}}^{\text{bloc}}$.

On peut voir l'étape i) comme une phase de *screening* [Xiang *et al.*, 2011] permettant de supprimer des variables avant d'effectuer l'optimisation du problème en norme ℓ_1, même si nous ne disposons d'aucune garantie que les variables supprimées à l'étape i) soient absentes du problème complet (*i.e.,* avec $\mathbf{D}^{\text{bloc}}$ à la place de $\widetilde{\mathbf{D}}^{\text{bloc}}$ dans (4.8)).

Si le nombre d'inconnues dans (4.8) se trouve sensiblement réduit, le dictionnaire équivalent $\mathbf{B}$ reste de taille très importante en raison du nombre de données impliquées, et ne peut même pas être construit explicitement. Nous avons à nouveau choisi une stratégie d'optimisation basée sur un algorithme ICD. Chaque minimisation scalaire (voir l'équation (4.6)) ne requiert alors que le calcul numérique d'une colonne du dictionnaire, lequel est effectué « à la demande » au cours de l'algorithme.

Une « preuve de principe » a été réalisée, sur des problèmes simulés de très faible dimension spatiale, faisant l'objet d'une publication à la conférence du Gretsi [Bourguignon *et al.*, 2011a] et à la conférence IEEE WHISPERS [Bourguignon *et al.*, 2011b]. Les données, représentées en figure 4.8a), sont générées sous la forme de deux sources ponctuelles, localisées aux pixels de coordonnées (2,2) et (3,3) (en initialisant le système de coordonnées en bas à gauche). Les deux spectres associés sont issus de simulations astrophysiques. Le spectre du pixel (3,3) est celui des résultats de la section précédente (voir figure 4.5). Le spectre du pixel (2,2) présente essentiellement une raie d'émission et un continuum décroissant. Ce cube est convolué spatialement par une FSF de 3×3 pixels, et spectralement par une LSF de 11 points, produisant en chaque pixel un mélange des deux sources (figure 4.8b)). Du bruit d'un niveau réaliste, variable en longueur d'onde, est enfin ajouté, produisant les données de la figure 4.8c). La figure 4.8d) présente le résultat de la restauration de chaque spectre considéré séparément, fournissant des signaux certes débruités, mais qui restent mélangés. La restauration par une approche prenant en compte la FSF reconstruit un cube de 6×6 pixels afin de prendre en compte les effets de bord de la convolution, dont nous n'affichons que les 4×4 pixels centraux. Les résultats sont présentés en figure 4.5e), où les deux sources sont « séparées », les spectres associés étant assez proches de la vérité terrain, eu égard au fort niveau de bruit. On note également la présence de deux artefacts de faible amplitude sur les pixels (1,2) et (4,3).

4.4.2 Approche gloutonne

Lors de mon arrivée à l'École Centrale de Nantes, j'ai poursuivi la piste de la restauration spatiale-spectrale, notamment par l'encadrement d'un stage de Master [Picaud, 2012]. Si les travaux précédents envisageaient l'optimisation de critères en norme ℓ_1, nous avons cette fois opté pour l'algorithme glouton *Orthogonal Matching Pursuit* (OMP). Les solutions recherchées étant très parcimonieuses en raison du fort niveau de bruit (peu de composantes sont détectables), cet algorithme fournit une alternative intéressante à l'optimisation ℓ_1, sa complexité étant essentiellement fonction du nombre de composantes non nulles dans la solution. Chaque itération consistant essentiellement en l'évaluation d'un produit par la matrice $\mathbf{B}$ et un produit par $\mathbf{B}^T$, nous avons cherché à exploiter la structure des opérateurs impliqués dans ces opérations pour réduire le temps de calcul :

- la FSF étant supposée invariante spatialement (voir le § 4.1.2), les produits $\mathbf{F}\cdot$ et $\mathbf{F}^T\cdot$ peuvent être effectués par une série de convolutions spatiales de chaque image par la FSF associée ;
- les produits par $\mathbf{L}$ et $\mathbf{L}^T$, en revanche, ne sont pas des produits de convolution, la LSF variant spectralement. Cependant, celle-ci étant très peu étendue (au plus une dizaine de points dans le cas de MUSE), ces produits peuvent être réalisés efficacement grâce à un codage creux de l'opérateur de LSF ;
- les produits $\mathbf{D}\boldsymbol{u}$, opérant sur des vecteurs parcimonieux, peuvent être réalisés en exploitant la parcimonie de $\boldsymbol{u}$;
- pour les produits $\mathbf{D}^T\cdot$, nous exploitons la structure du dictionnaire (voir le para-

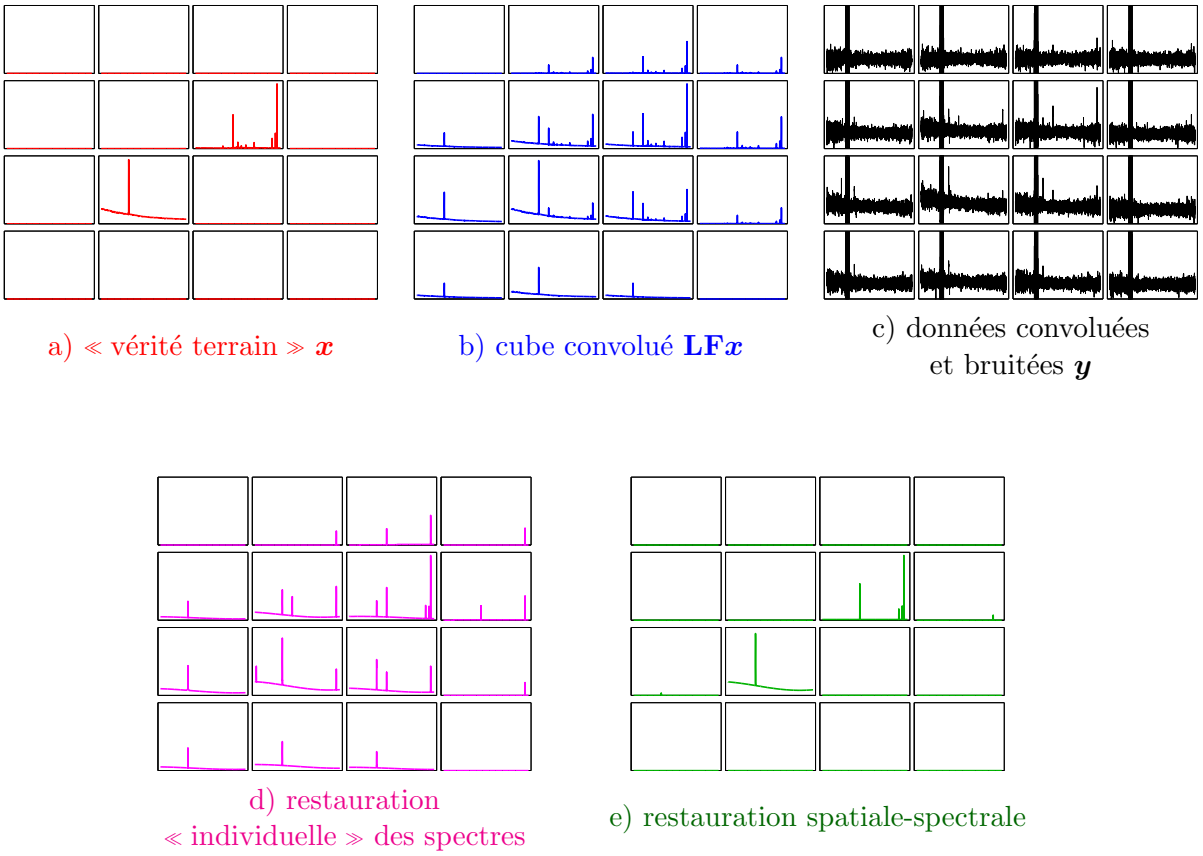


FIGURE 4.8 – Exemple de reconstruction par une approche spatial-spectrale, en très petite taille spatiale. Figure reprise de [Bourguignon *et al.*, 2011b].

graphe 4.2.2). Chaque sous-dictionnaire associé à une raie d'une largeur donnée est une matrice de Toeplitz (dont on a enlevé des colonnes pour les raies les plus larges, voir le tableau 4.1). Les calculs associés peuvent donc être réalisés par des algorithmes FFT [Golub et Van Loan, 1996]. Pour le sous-dictionnaire d'échelons $\mathbf{D}_{\text{steps}}$, le produit $\mathbf{D}_{\text{steps}}^T \boldsymbol{v}$ n'est rien d'autre que la somme cumulée des éléments de $\boldsymbol{v}$. Enfin, le sous-dictionnaire $\mathbf{D}_{\cos}$ composé de sinusoïdes étant de petite taille, les produits $\mathbf{D}_{\cos}^T \boldsymbol{v}$ sont calculés directement.

Ce travail a donné lieu à une publication à la conférence du Gretsi en 2013 [Picaud et Bourguignon, 2013]. Nous reproduisons en figure 4.4.2 un exemple de résultat obtenu sur un cube simulé par les astrophysiciens du projet MUSE, de 41×41 pixels et 3 600 longueurs d'onde. La FSF a un étalement spatial correspondant à 13×13 pixels, et la LSF occupe 11 canaux spectraux. L'algorithme OMP est arrêté lorsque le résidu d'estimation est statistiquement comparable au bruit. Au final, la solution possède seulement 309 atomes, ce qui représentait (en 2012) un temps de calcul de trois heures.

Les données contiennent en particulier deux sources quasi-monochromatiques (aux longueurs d'onde respectives de 721 nm et 866 nm), mais d'intensité faible : dans les deux cas, l'observation seule de l'image à la longueur d'onde associée et du spectre au pixel associé ne permet pas de détecter la source. Pour chacune des sources, en revanche, la reconstruction spatiale-spectrale fournit un atome modélisant très convenablement les spectres, à la position spatiale correcte. Les deux sources sont bien détectables et c'est bien la prise en compte de l'étalement spatial et spectral qui a permis de concentrer l'information utile.

4.5 Conclusions

Dans ces travaux, l'analyse d'images hyperspectrales astrophysiques a été abordée sous l'angle de la restauration d'un cube « idéal » ayant subi des dégradations dues au dispositif de mesure (prise en compte d'une fonction d'étalement spatiale et spectrale) et au bruit. Nous avons adopté l'hypothèse d'une représentation parcimonieuse, approchée, des spectres recherchés à l'aide d'un dictionnaire de formes synthétiques, construit à partir de connaissances en spectroscopie astronomique. Cette hypothèse représente une contrainte très forte, puisqu'« on ne trouve que ce que l'on cherche », limitant de fait les capacités de découverte de phénomènes inconnus, qui est aussi un objectif d'un instrument comme MUSE (Multi-Unit Spectroscopic Explorer). Cependant, elle permet de pallier le déficit d'information dans le cas d'observations très bruitées, où il s'avère déjà difficile de détecter des formes connues ! Par opposition à des représentations parcimonieuses génériques, souvent utilisées pour « débruiter » les données, ce modèle permet également d'associer à chaque atome présent dans la décomposition une information d'intérêt pour l'astrophysicien (raie en émission ou en absorption, cassure de Lyman).

Cette approche permet d'envisager la détection de composantes dans des spectres à très faible rapport signal sur bruit, par la prise en compte de formes appropriées dans le dictionnaire. L'exploitation d'un modèle d'observation spatial-spectral est évidemment plus ambitieuse, permettant d'envisager la détection de sources encore plus faibles en prenant en compte l'étalement spatial subi lors de l'acquisition des données. Pour gérer la complexité calculatoire, une approche gloutonne de type OMP me semble rétrospectivement préférable à l'optimisation en norme ℓ_1, car plus simple à mettre en œuvre. Afin de réduire davantage le coût calculatoire et / ou d'aborder des problèmes plus gros, il serait également possible de coupler les deux approches proposées au § 4.4, à savoir d'exécuter

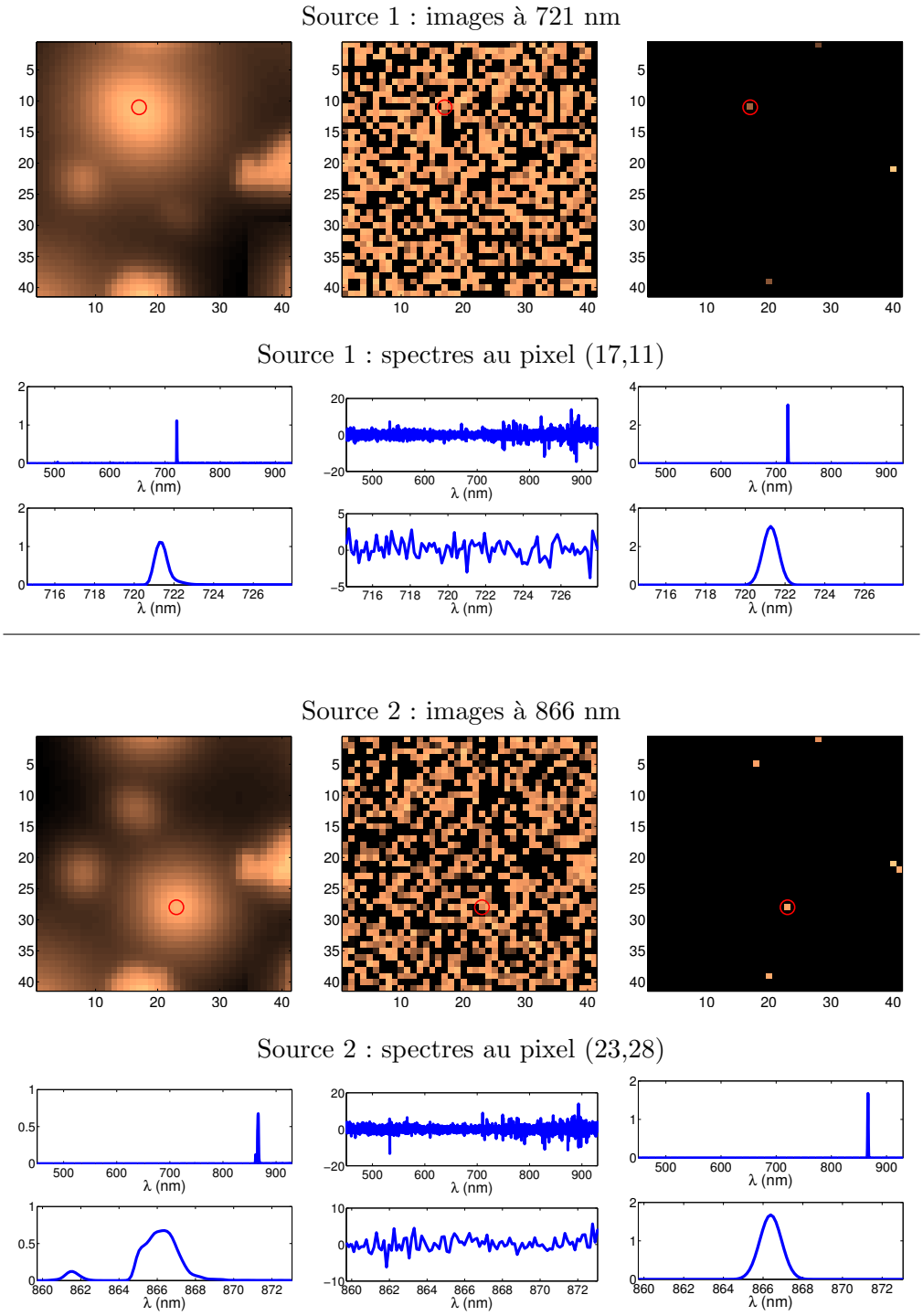


FIGURE 4.9 – Restauration spatiale-spectrale obtenue sur un cube simulé de 41 × 41 pixels × 3 600 longueurs d'onde : données et résultats correspondant à deux sources d'intérêt. À gauche, données convoluées non bruitées. Au centre, données bruitées. À droite, reconstruction après estimation parcimonieuse par algorithme OMP. Pour chaque source, sont représentés l'image à sa longueur d'onde centrale et le spectre à sa position spatiale (repérée par un cercle sur les images). Les amplitudes dans les images sont en échelle logarithmique, les pixels négatifs étant ramenés à 0. Pour chaque source, l'intégralité du spectre et un zoom autour de la longueur d'onde d'intérêt sont représentés. Figures reprises de [Picaud et Bourguignon, 2013].

un algorithme glouton sur un modèle dont on a préalablement réduit la dimension. Il me semble cependant important de garder à l'esprit que pour des projets instrumentaux de grande envergure, où l'acquisition des données requiert la mise en commun de plusieurs nuits d'observation [Bacon *et al.*, 2006], il est légitime d'envisager des méthodes d'analyse de complexité adaptée, quand bien même le temps de calcul associé prendrait, lui aussi, plusieurs jours.

Sur le plan algorithmique, nous avons construit un schéma d'optimisation de critères des moindres carrés pénalisés par la norme ℓ_1, qui s'est révélé très performant dans le cas de dictionnaires corrélés, rendant possible le calcul de chaque décomposition en quelques secondes. Partant d'un schéma de type *Iterative Coordinate Descent*, nous avons conçu des règles de balayage et des accélérations particulièrement efficaces. L'exploitation de ces accélérations dans d'autres structures d'algorithmes en norme ℓ_1 mériterait de ce fait d'être étudiée. Enfin, au-delà du problème applicatif ayant motivé ces travaux, cet algorithme pourrait avantageusement être utilisé dans d'autres problèmes d'optimisation en norme ℓ_1, dans le cas de dictionnaires corrélés et ne permettant pas l'exploitation de transformées rapides. Des perspective de recherche pus générales concernant l'imagerie hyperspectrale en astronomie ainsi que le démélange spectral seront abordées au Chapitre 7.

4.6 Références

[Alliney et Ruzinsky, 1994] Alliney, S. et Ruzinsky, S. A. (1994). An algorithm for the minimization of mixed l/sub 1/ and l/sub 2/ norms with application to bayesian estimation. *IEEE Transactions on Signal Processing*, 42(3):618–627.

[Bacon *et al.*, 2006] Bacon *et al.*, R. (2006). Probing unexplored territories with MUSE : a second generation instrument for the VLT. *In Proc. SPIE*, volume 6269 de *Ground-based and Airborne Instrumentation for Astronomy*.

[Beck et Teboulle, 2009] Beck, A. et Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202.

[Bourguignon *et al.*, 2011a] Bourguignon, S., Carfantan, H., Mary, D., Slezak, E. et Ferrari, A. (2011a). Restauration des cubes hyperspectraux du spectro-imageur MUSE. *In Actes du 23e colloque GRETSI*, Bordeaux, France.

[Bourguignon *et al.*, 2011b] Bourguignon, S., Carfantan, H., Slezak, E. et Mary, D. (2011b). Sparsity-based spatial-spectral restoration of MUSE astrophysical hyperspectral data cubes. *In Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Lisbon, Portugal.

[Bourguignon *et al.*, 2010] Bourguignon, S., Mary, D. et Slezak, E. (2010). Sparsity-based denoising of hyperspectral astrophysical data with colored noise. application to the MUSE instrument. *In Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Reykjavik, Iceland.

[Bourguignon *et al.*, 2011c] Bourguignon, S., Mary, D. et Slezak, É. (2011c). Restoration of astrophysical spectra with sparsity constraints : Models and algorithms. *IEEE J. Sel. Topics Signal Processing*, 5(5):1002–1013.

[Combettes et Pesquet, 2011] Combettes, P. L. et Pesquet, J.-C. (2011). *Proximal Splitting Methods in Signal Processing*, pages 185–212. Springer New York, New York, NY.

[Donoho et Johnstone, 1994] Donoho, D. L. et Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

[Donoho et Tsaig, 2008] Donoho, D. L. et Tsaig, Y. (2008). Fast solution of l1-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54:4789–4812.

[Efron *et al.*, 2004] Efron, B., Hastie, T., Johnstone, I. et Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499.

[Figueiredo *et al.*, 2007] Figueiredo, M. A. T., Bioucas-Dias, J. M. et Nowak, R. D. (2007). Majorization-minimization algorithms for wavelet-based image restoration. *IEEE Trans. Image Process.*, 16(12):2980–2991.

[Friedman *et al.*, 2007] Friedman, J., Hastie, T., Höfling, H. et Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332.

[Fuchs, 2004] Fuchs, J.-J. (2004). On sparse representations in arbitrary redundant bases. *IEEE Trans. Inf. Theory*, 50(6):1341–1344.

[Golub et Van Loan, 1996] Golub, G. et Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

[Léna *et al.*, 2008] Léna, P., Rouan, D. et Lebrun, F. (2008). *L'observation en astrophysique*. Savoirs actuels. Astrophysique. EDP Sciences.

[Picaud, 2012] Picaud, J. (2012). Restauration de cubes hyperspectraux en astrophysique. Mémoire de D.E.A., École Centrale de Nantes, France.

[Picaud et Bourguignon, 2013] Picaud, J. et Bourguignon, S. (2013). Restauration de données hyperspectrales astrophysiques par approximation parcimonieuse. *In Actes du 24e colloque GRETSI*, Brest, France.

[Serre *et al.*, 2010] Serre, D., Villeneuve, E., Carfantan, H., Jolissaint, L., Mazet, V., Bourguignon, S. et Jarno, A. (2010). Modeling the spatial PSF at the VLT focal plane for MUSE WFM data analysis purpose. *Proc. SPIE 7736*, page 773649.

[Tennyson, 2005] Tennyson, J. (2005). *Astronomical Spectroscopy*. Imperial College Press.

[Tropp et Wright, 2010] Tropp, J. A. et Wright, S. J. (2010). Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958.

[Villeneuve, 2012] Villeneuve, E. (2012). *Déconvolution de données hyperspectrales pour l'instrument MUSE du VLT*. Thèse de doctorat, Université Toulouse 3.

[Wright *et al.*, 2009] Wright, S. J., Nowak, R. D. et Figueiredo, M. A. T. (2009). Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7): 2479–2493.

[Wu et Lange, 2008] Wu, T. T. et Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. App. Statist.*, 2(1):224–244.

[Xiang *et al.*, 2011] Xiang, Z., Xu, H. et Ramadge, P. (2011). Learning sparse representations of high dimensional data on large scale dictionaries. *In Advances in Neural Information Processing Systems (NIPS)*.

Chapitre 5

Déconvolution et reconstruction d'images pour le contrôle non destructif ultrasonore

Le contrôle non destructif (CND) ultrasonore repose sur l'émission d'une onde acoustique par un transducteur positionné à distance ou à la surface du matériau inspecté [Krautkramer et Krautkramer, 1990]. La réception de l'onde après propagation dans le matériau permet alors d'obtenir des informations sur le milieu traversé. L'objectif le plus fréquent est alors de détecter et caractériser des défauts ou des hétérogénéités dans des pièces manufacturées au moment de leur fabrication, ou encore de contrôler l'usure de pièces soumises à des conditions de fonctionnement agressives. Par rapport à d'autres modalités de CND (rayons X, courants de Foucault, RADAR, ...), les mesures ultrasonores présentent l'avantage d'un faible coût et d'une grande simplicité et portabilité du dispositif de mesure, permettant son déploiement dans des environnements difficiles d'accès. Souvent, en revanche, les données recueillies ne délivrent pas directement l'information utile pour l'inspection et l'examen visuel des traces ultrasonores s'avère d'efficacité limitée : les données sont à bande passante relativement étroite (typiquement de quelques MHz dans les longueurs d'onde ultrasonores de 20 kHz à 1 GHz), limitant la résolution des signaux et des images bruts. En particulier, les signatures ultrasonores générées par d'éventuels défauts peuvent se mélanger dans les signaux reçus. Par ailleurs, le bruit entachant les mesures peut être particulièrement fort lors de l'inspection de matériaux complexes ou requérant une longues distances de propagation, le signal acoustique utile étant alors fortement atténué.

Les travaux que j'ai menés en CND ultrasonore se centrent avant tout sur la construction de modèles précis des signaux ultrasonores et le développement d'algorithmes d'estimation associés, dans le cadre des problèmes inverses régularisés. Nous détaillons tout d'abord en Section 5.1 le principe général d'une mesure de CND acoustique, dans le cas d'une acquisition réalisée en réflexion avec un seul capteur. Un problème inverse est alors formulé, s'apparentant à un problème de déconvolution parcimonieuse. Les sections suivantes présentent plusieurs travaux visant à raffiner le modèle convolutif classique, en proposant des solutions algorithmiques dédiées.

Une première contribution, présentée en Section 5.2, a porté sur la déconvolution dite « à haute résolution ». La déconvolution est souvent abordée sous la forme d'un problème purement numérique, où les données résultent de la convolution *discrète* d'une séquence numérique par un filtre de réponse impulsionnelle finie. Or, dans la plupart des problèmes

inverses mettant en jeu une opération de filtrage (c'est le cas pour le CND ultrasonore), le modèle physique sous-jacent est défini par une équation intégrale, à temps continu. Sa discrétisation à la période d'échantillonnage des données formule alors un modèle de convolution discrète, mais introduit des erreurs de modèle. Nous avons étudié l'apport d'une discrétisation plus précise de l'équation intégrale, en montrant notamment que le modèle résultant peut s'interpréter comme celui d'un système MISO (*Multiple Inputs, Single Output*) ; nous avons alors généralisé des algorithmes classiques de déconvolution parcimonieuse à ce cadre.

Nous présentons ensuite des travaux portant sur la modélisation de la propagation acoustique. En raison de phénomènes d'absorption et de diffusion de l'onde acoustique se propageant dans le matériau, celle-ci va subir des déformations, se traduisant par un filtrage passe-bas d'effet cumulatif avec la distance de propagation, correspondant aux phénomènes d'atténuation et de dispersion fréquentielles. Leur prise en compte dans un modèle de données linéaire (qui n'est alors plus un modèle convolutif au sens strict), la validation de ce dernier pour différents types de matériaux et son exploitation dans des algorithmes d'estimation parcimonieuse font l'objet de la Section 5.3.

Dans la Section 5.4, nous nous intéressons au cas particulier de la mesure ultrasonore de l'épaisseur de pièces accessibles d'un seul côté. C'est une problématique rencontrée dans différents domaines industriels, par exemple pour le contrôle de l'épaisseur de revêtements ou de couches de peinture, ou encore pour suivre l'état de dégradation de tuyaux soumis à de fortes corrosions. Dans ce contexte, la séquence de réflectivité peut être modélisée sous une forme plus contrainte, correspondant à des échos régulièrement espacés traduisant les multiples trajets aller-retour de l'onde réfléchie à chaque interface entre le matériau et le milieu extérieur. Lorsque l'épaisseur est fine devant la longueur d'onde, les échos se chevauchent et rendent l'interprétation visuelle impossible. Nous avons proposé un modèle adapté pour la séquence de réflectivité recherchée, ainsi qu'une paramétrisation de la forme des échos, dont les paramètres sont estimés conjointement dans une démarche de déconvolution myope.

Si l'ensemble des contributions précédentes concerne l'exploitation de signaux monodimensionnels, les modalités d'inspection ultrasonore évoluent désormais vers l'utilisation de sondes multi-éléments, mettant en œuvre quelques dizaines à quelques centaines de transducteurs fonctionnant conjointement et permettant d'envisager l'imagerie des matériaux, sur le principe de l'échographie médicale. Les modalités d'inspection sont cependant différentes, reposant en général sur l'acquisition de l'ensemble des réponses de chaque transducteur à l'onde émise successivement par chacun d'entre eux (données dites *Full Matrix Capture* ou FMC). Je me suis donc naturellement orienté vers ces problématiques où, dans la continuité des travaux précédents, la reconstruction d'une carte spatiale de la réflectivité ultrasonore en tout point du matériau est abordée sous l'angle des problèmes inverses. Ces travaux seront exposés en Section 5.5.

La Section 5.6 conclut ce chapitre en proposant plusieurs axes de travail pour des recherches à venir.

5.1 Principe de mesure et formulation d'un problème inverse

Expliquons tout d'abord le principe d'une mesure ultrasonore. Dans le cas de signaux mono-dimensionnels, on distingue les mesures en réflexion (le même transducteur jouant le rôle d'émetteur et de récepteur) des mesures en transmission (où deux capteurs sont utilisés, généralement disposés de part et d'autre de la pièce inspectée). Le schéma de la

figure 5.1 illustre ces deux modalités. Sans perte de généralité, nous considérons ici des

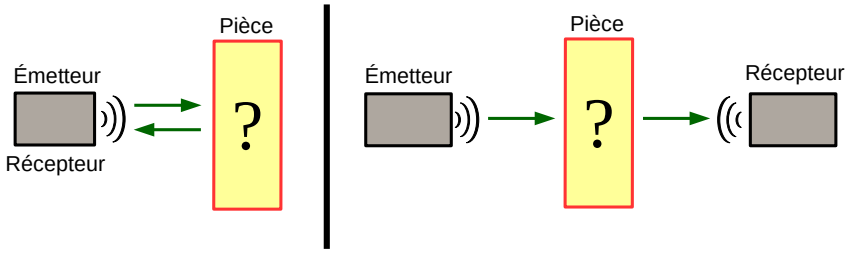


FIGURE 5.1 – Principe de mesures ultrasonores pour le CND. À gauche, mesure en réflexion. À droite, mesure en transmission. Figure reproduite de [Carcreff, 2014a].

mesures en réflexion, souvent préférées pour leur simplicité : un seul capteur est nécessaire, et la mesure ne requiert pas d'accès à la face arrière de la pièce inspectée. Lorsque l'onde rencontre une discontinuité d'impédance acoustique, due par exemple à la présence d'un défaut dans le matériau ou à un changement de milieu de propagation, une partie de l'onde est réfléchie et l'autre partie est transmise à travers l'interface [Saniie et Nagle, 1989]. Dans le cas d'un *réflecteur ponctuel* situé à une distance z du capteur, le signal reçu peut se modéliser sous la forme :

$$y^{\text{echo}}(t) = b(z)\, h_i(t - \tau(z)), \tag{5.1}$$

où l'amplitude $b(z)$ est fonction de la distance de propagation et du coefficient de réflexion à l'interface, $h_i(t)$ est la forme d'onde correspondant à l'impulsion acoustique générée et reçue par le transducteur et le retard $\tau(z) = 2z/c$ correspond au temps de trajet aller-retour de l'onde, à la vitesse c, entre le capteur et l'interface. En considérant maintenant un ensemble de K discontinuités situées à des distances z_k, $k = 1, \ldots, K$ et sous l'hypothèse de linéarité de la réponse du matériau et d'invariance temporelle, le principe de superposition permet alors de décrire le signal reçu par :

$$y(t) = \sum_{k=1}^{K} b(z_k) h_i(t - \tau(z_k)),$$

i.e., le produit de convolution de la réponse instrumentale $h_i(t)$ par une fonction parcimonieuse :

$$y(t) = (h_i * h_r)(t), \text{ où } h_r(t) = \sum_{k=1}^{K} b(z_k)\delta(t - \tau(z_k)) \tag{5.2}$$

est la *séquence de réflectivité*, représentant la signature acoustique du matériau traversé. Les données échantillonnées à la période T_e s'écrivent alors :

$$y_n = y(nT_e) = \int_{u\in\mathbb{R}} h_i(u) h_r(nT_e - u)\mathrm{d}u = \int_{u\in\mathbb{R}} h_i(nT_e - u) h_r(u)\mathrm{d}u, \tag{5.3}$$

où $h_r(t)$ est modélisée par l'équation (5.2), dans laquelle le nombre K de discontinuités, les amplitudes $b(z_k)$ et les temps d'arrivée $\tau(z_k)$ (*i.e.,* les positions z_k avec $\tau(z_k) = 2z_k/c$) sont inconnus. La figure 5.2 illustre le principe de formation des données.

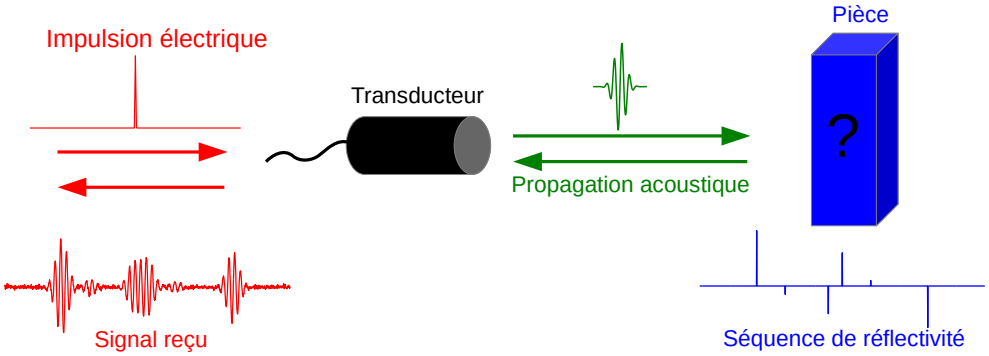


FIGURE 5.2 – Formulation d'un modèle de convolution parcimonieuse : l'impulsion ultrasonore transmise dans le matériau se réfléchit à chaque discontinuité d'impédance, représentée par la séquence de réflectivité, produisant une superposition d'échos retardés.

Nous avons abordé ce problème sous l'angle de la déconvolution parcimonieuse, problème ayant donné lieu à de nombreuses contributions, notamment pour le CND par ultrasons [Zala, 1992, O'Brien *et al.*, 1994] et pour la sismique-réflexion en géophysique [Mendel, 1983], mais aussi de manière plus académique en traitement du signal (voir par exemple [Kormylo et Mendel, 1982, Chi, 1987, Goussard *et al.*, 1990, Lavielle, 1993, Kaaresen, 1997, Cheng *et al.*, 1996, Soussen *et al.*, 2011, Selesnick et Bayram, 2014]). La plupart de ces travaux adopte le schéma classique de discrétisation de l'équation intégrale (5.3), où le pas de discrétisation correspond à la période d'échantillonnage. L'intégrale est alors approchée par la méthode des rectangles, permettant d'obtenir un modèle de convolution discrète :

$$y_n \simeq T_e \sum_{m\in\,\mathbb{Z}} h_i\big((n-m)T_e\big)\, h_r(mT_e) \quad \Leftrightarrow \quad \boldsymbol{y} = \mathbf{H}_i\boldsymbol{x} + \boldsymbol{\epsilon}, \tag{5.4}$$

en notant $\boldsymbol{y}$ le vecteur colonne collectant les données, $\mathbf{H}_i = [h_i(n-m)]_{n,m}$ la matrice de convolution discrète par la séquence $[h_i(nT_e)]_n$ et $\boldsymbol{x}$ le vecteur colonne contenant, à un facteur multiplicatif près, la séquence de réflectivité discrétisée à la période T_e. Enfin, $\boldsymbol{\epsilon}$ modélise un terme d'erreur, prenant en compte le bruit sur les mesures et l'inexactitude du modèle.

La détection et la localisation des discontinuités est donc reportée sur l'estimation d'un vecteur $\boldsymbol{x}$ parcimonieux, *i.e.*, comportant un grand nombre de valeurs nulles, à partir d'un modèle de convolution discrète : pour une discontinuité localisée en $z_k = 2p_k\Delta z$ avec p_k entier et $\Delta z = cT_e$, on aura $x_{p_k} = T_e b(z_k)$ et $x_p = 0$ sinon.

5.2 Déconvolution à haute résolution

La discrétisation, à la période d'échantillonnage T_e, du modèle convolutif à temps continu (5.3) induit une approximation des positions des possibles défauts sur une grille de période spatiale $\Delta z = cT_e$. Cette erreur de modélisation impacte non seulement la précision dans l'estimation des positions (erreur de l'ordre de Δz), mais peut s'avérer plus grave : des artefacts peuvent en particulier apparaître dans la séquence estimée afin de compenser l'inadéquation entre le modèle discrétisé et les données.

Nous avons proposé, dans la thèse d'Ewen Carcreff, de considérer un modèle discrétisé plus finement, dit à haute résolution, en introduisant un facteur de sur-échantillonnage L dans la discrétisation de l'équation intégrale (5.3), qui s'écrit alors :

$$y_n \simeq \frac{T_e}{L} \sum_m h_i\left(nT_e - m\frac{T_e}{L}\right) h_r\left(m\frac{T_e}{L}\right) \quad \Leftrightarrow \quad \boldsymbol{y} = \widetilde{\mathbf{H}}_i\, \widetilde{\boldsymbol{x}} + \boldsymbol{\epsilon}, \tag{5.5}$$

où, similairement aux notations de l'équation (5.4), $\widetilde{\mathbf{H}}_i$ est la matrice de terme général $\widetilde{h}_{m,n} = h_i\left(nT_e - m\frac{T_e}{L}\right)$ et $\widetilde{\boldsymbol{x}}$ le vecteur colonne contenant, à un facteur multiplicatif près, la séquence de réflectivité maintenant discrétisée à la période T_e/L.

Si ce modèle est plus précis que le précédent, il perd en revanche sa structure convolutive, qui est importante sur le plan calculatoire : la plupart des algorithmes d'optimisation parcimonieuse requiert l'évaluation répétée de produits matrice-vecteur impliquant l'opérateur du problème $\widetilde{\mathbf{H}}_i$ et son adjoint $\widetilde{\mathbf{H}}_i^T$. Dans le cas d'une matrice de convolution discrète, ces produits peuvent être réalisés de façon efficace à partir d'algorithmes FFT [Golub et Van Loan, 1996, p. 193]. Cette structure peut donc être exploitée pour l'optimisation parcimonieuse basée sur le modèle (5.4), mais ce n'est plus le cas pour le modèle (5.5).

Nous avons montré [Carcreff *et al.*, 2013a, Carcreff *et al.*, 2013b] que ce modèle pouvait être réécrit sous la forme d'un modèle MISO (*Multiple Inputs, Single Output*) :

$$\begin{aligned}
\sum_{m\in\mathbb{Z}} h_i\left(nT_e - m\frac{T_e}{L}\right) h_r\left(m\frac{T_e}{L}\right) &= \sum_{\ell=0}^{L-1} \sum_{\substack{m\in\mathbb{Z}\,|\\ (m \bmod L)=\ell}} h_i\left(nT_e - m\frac{T_e}{L}\right) h_r\left(m\frac{T_e}{L}\right) \\
\text{(avec } m = jL+\ell\text{)} \quad &= \sum_{\ell=0}^{L-1} \sum_{j\in\mathbb{Z}} h_i\left(nT_e - jT_e - \ell\frac{T_e}{L}\right) h_r\left(jT_e + \ell\frac{T_e}{L}\right) \\
&= \sum_{\ell=0}^{L-1} \sum_{j\in\mathbb{Z}} h_i^{(\ell)}\big((n-j)T_e\big)\, h_r^{(\ell)}\,(jT_e)
\end{aligned}$$

en considérant L réponses impulsionnelles $h_i^{(\ell)}(t) \triangleq h_i\left(t - \ell\frac{T_e}{L}\right)$, décalées d'un temps $\ell T_e/L$, $\ell = 0, \ldots, L-1$, discrétisées à la fréquence T_e et, similairement, L séquences de réflectivité $h_r^{(\ell)}(t) \triangleq h_r\left(t + \ell\frac{T_e}{L}\right)$. En notant $\mathbf{H}_i^{(\ell)}$ les L matrices de convolution de terme général $h_i^{(\ell)}(n-m)$ et $\boldsymbol{x}^{(\ell)}$ le vecteur colonne formé par les coefficients $\frac{T_e}{L} h_r^{(\ell)}(m)$, le modèle (5.5) s'écrit finalement :

$$\boldsymbol{y} = \sum_{\ell=0}^{L-1} \mathbf{H}_i^{(\ell)} \boldsymbol{x}^{(\ell)} + \boldsymbol{\epsilon},$$

où chaque produit $\mathbf{H}_i^{(\ell)}\boldsymbol{x}^{(\ell)}$ correspond à une convolution discrète. C'est donc un système MISO, représenté sur le schéma de la Figure 5.2.

Nous avons alors étendu plusieurs algorithmes classiques de déconvolution parcimonieuse au cas de systèmes MISO, en exploitant la structure des matrices de convolution

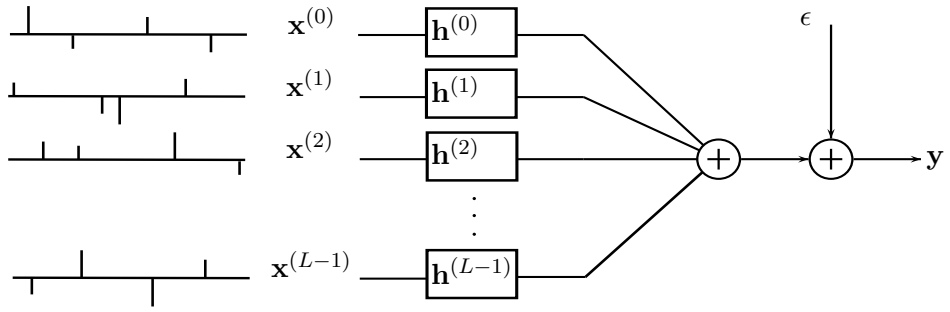


FIGURE 5.3 – Système MISO (Multiple Input, Single Output) : la sortie est la somme (bruitée) de L entrées (parcimonieuses dans notre cas) filtrées par L filtres différents.

$\mathbf{H}_i^{(\ell)}$ dans les calculs numériques, réalisés à partir d'algorithmes FFT. Les mises en œuvre proposées s'appliquent donc à n'importe quel système MISO à entrées parcimonieuses, même si le problème à l'origine de ces développements concernent un cas particulier, où les L filtres proviennent de l'échantillonnage d'une même réponse impulsionnelle.

Ces algorithmes ont déjà été présentés au Chapitre 3. Il s'agit des algorithmes gloutons Orthogonal Matching Pursuit (OMP), Orthogonal Least Squares (OLS) et Single Best Replacement (SBR) et de l'optimisation convexe du critère des moindres carrés pénalisé par la norme ℓ_1, réalisée ici par l'algorithme d'homotopie. Un exemple de résultat sur des données simulées est présenté en figure 5.4. La forme d'onde utilisée est de forme gaussienne modulée par une sinusoïde de fréquence 5 MHz, et les données sont échantillonnées à 25 MHz. Les amplitudes et les temps d'arrivée de chaque écho sont générés aléatoirement de manière continue, si bien qu'aucun instant τ_k ne correspond exactement aux valeurs discrètes utilisées dans les modèles (5.4) et (5.5). Les données sont entachées de bruit blanc additif centré gaussien avec un rapport signal sur bruit de 10 dB et présentent en particulier trois problèmes de chevauchement, approximativement aux instants 1, 3.5 et 6.5 μs. Les algorithmes sont réglés de manière à ce que l'erreur d'approximation soit de norme similaire à celle du bruit. Pour chaque algorithme, la solution obtenue avec le modèle classique (5.4) et la forme d'onde représentée en figure 5.4 a) est comparée avec la solution s'appuyant sur le modèle sur-échantillonné (5.5), avec les $L = 6$ formes d'onde représentées en figure 5.4 b).

Tous les algorithmes utilisant la résolution standard échouent à résoudre correctement les trois problèmes, le cas le plus fréquent étant celui où la position et le signe de l'amplitude associée sont mal estimés. Lorsque la résolution augmente ($L = 6$), toutes les méthodes produisent de meilleurs résultats. En particulier, OLS et SBR parviennent à résoudre correctement deux des trois problèmes sus-mentionnés. Notons qu'aucun algorithme ne parvient à résoudre le chevauchement autour de 3.5 μs, particulièrement difficile.

La Figure 5.5 présente enfin des résultats moyennés sur 1 000 réalisations aléatoires de données similaires à celles de la Figure 5.4, montrant respectivement, pour chaque algorithme, l'erreur sur la séquence estimée en fonction du facteur de sur-échantillonnage L et le coût de calcul associé. Afin de comparer les séquences estimées aux vraies séquences, dont les instants d'arrivée ne peuvent jamais coïncider exactement, nous utilisons une mesure de distance entre séquences impulsionnelles initialement proposée en neurosciences [Van Rossum, 2001] : les amplitudes sont d'abord binarisées en ± 1 afin d'accorder le même poids à toutes les détections. Les séquences sont ensuite convoluées par un noyau exponentiel $e^{-|t|/Te}$ où Te est la période d'échantillonnage, puis l'erreur quadratique entre les deux

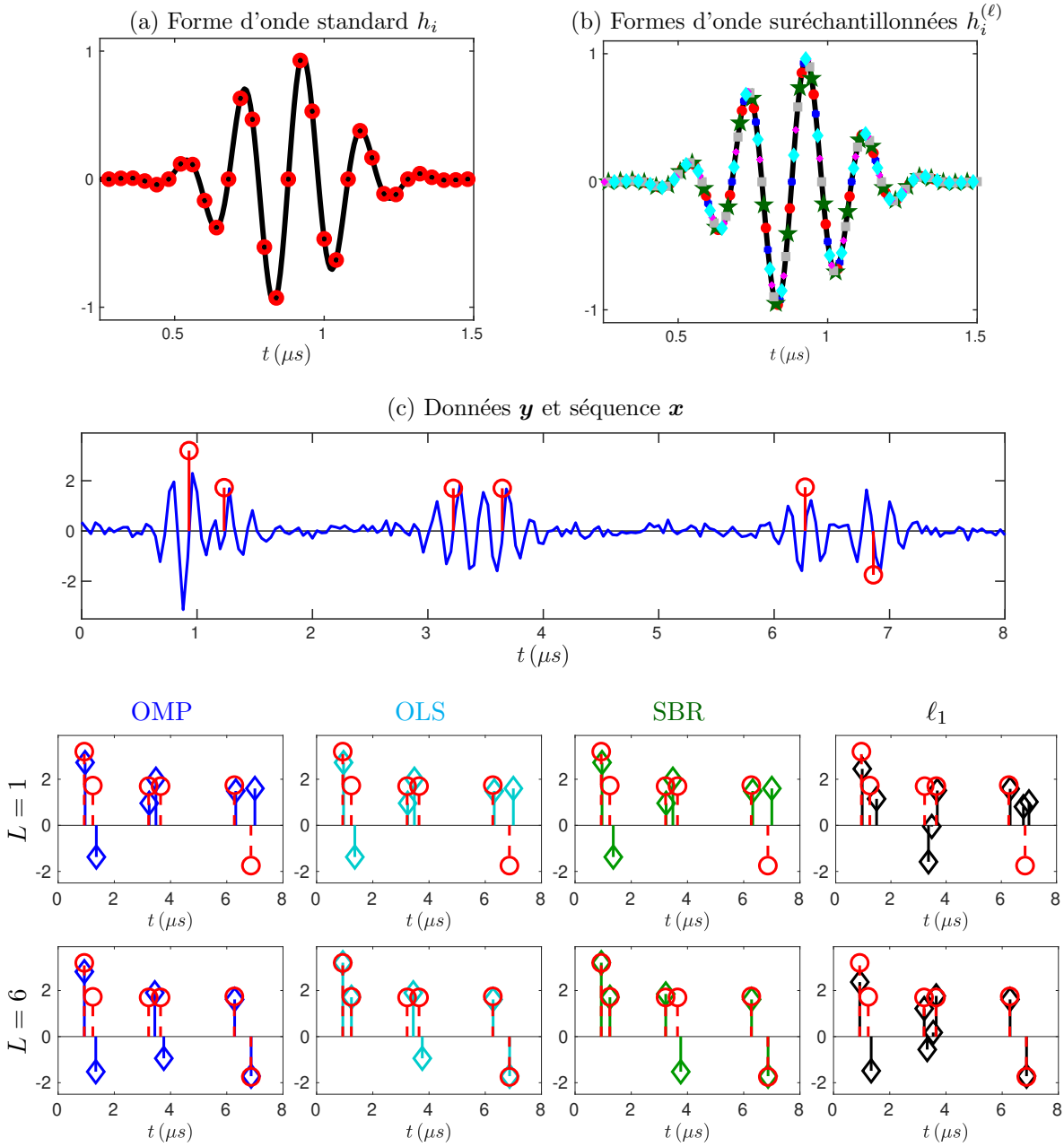


FIGURE 5.4 – Exemple de résultat de déconvolution à haute résolution sur un exemple simulé. (a) : Forme d'onde à temps continu (−) et forme discrète utilisée pour la déconvolution standard (○). (b) : Les $L = 6$ formes d'onde discrètes utilisées pour la déconvolution à haute résolution (marqueurs de différentes formes et couleurs). (c) Données et vraie séquence. Ligne « $L = 1$ » : résultats de plusieurs algorithmes à résolution standard. Ligne « $L = 6$ » : résultats des mêmes algorithmes à haute résolution.

séquences est calculée. Cette mesure permet d'associer une tolérance dans la précision de l'estimation de l'ordre de T_e. La comparaison entre les algorithmes confirme l'exemple de la Figure 5.4 et les résultats obtenus par exemple dans [Soussen *et al.*, 2011, Bourguignon *et al.*, 2011], à savoir un classement des méthodes par performances croissantes : OMP $<$ optimisation ℓ_1 $<$ OLS $<$ SBR. Les performances de tous les algorithmes s'améliorent lorsque L augmente. À partir de $L = 6$, cependant, l'amélioration devient négligeable. Une explication réside dans la variance intrinsèque sur l'estimation de temps d'arrivée, due à la présence de bruit [Quazi, 1981]. Les temps de calcul des différents algorithmes représentés sur la figure 5.5 à droite sont aussi en accord avec leur complexité et n'augmentent que linéairement avec L.

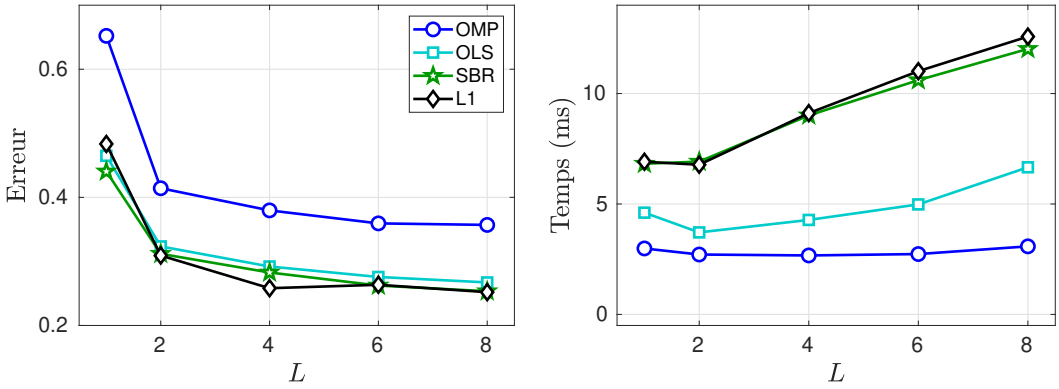


FIGURE 5.5 – Erreur moyenne sur les séquences estimées (gauche) et temps de calcul (droite) pour différents algorithmes de déconvolution parcimonieuse à haute résolution, en fonction du facteur de sur-échantillonnage L.

Nous présentons enfin un résultat expérimental, sur des données ultrasonores acquises au Laboratoire d'Acoustique de l'Université du Maine. Il s'agit de mesures en réflexion prises à partir d'une plaque d'aluminium d'épaisseur $d = 2$ mm, insonifiée en incidence normale. La fréquence centrale du transducteur est de 2.25 MHz et les données sont échantillonnées à 25 MHz. La séquence de réflectivité attendue dans ce cas correspond à une série d'impulsions, la première positive et les autres négatives, d'amplitudes décroissantes et régulièrement espacées de $\Delta\tau = 2d/c_{\text{Alu}}$, où $c_{\text{Alu}} \simeq 6380$ m.s^{-1} est la vitesse du son dans l'aluminium. Nous reviendrons en détail sur ce modèle en Section 5.4.

Les données et la séquence attendue sont représentées sur la figure 5.6 à gauche, montrant un fort recouvrement entre les différents échos dû à la finesse de la plaque. La réponse instrumentale h_i a été estimée sur un autre jeu de données plus simple, échantillonnée à $F_e = 25$ MHz puis interpolée avec un facteur de sur-échantillonnage de $L = 4$. La déconvolution est réalisée par l'algorithme SBR, dans le cas standard et dans le cas sur-échantillonné, correspondant alors à la reconstruction d'une séquence de réflectivité à 100 MHz. Dans le premier cas, les erreurs de détection sont importantes avec un dédoublement du pic principal, une erreur croissante pour les trois autres pics détectés en fonction du temps de vol, et même une erreur de signe sur le dernier pic. À l'inverse, l'algorithme exploitant le modèle sur-échantillonné parvient à détecter quatre pics en très bon accord avec la séquence théorique.

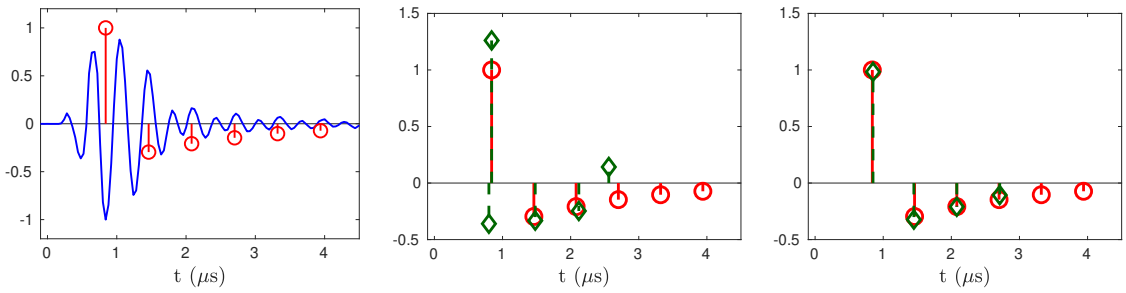


FIGURE 5.6 – Exemple de résultat expérimental exploitant un modèle de déconvolution à haute résolution : inspection ultrasonore d'une plaque fine, dont les échos se chevauchent. À gauche, données et séquence parcimonieuse recherchée (un train de pics régulièrement espacés, dont l'amplitude décroît géométriquement). Au centre, déconvolution par l'algorithme SBR avec un modèle standard discrétisé à la période d'échantillonnage $T_e = 0.04$ μs. À droite, déconvolution par l'algorithme SBR avec un modèle sur-échantillonné d'un facteur $L = 4$ (soit à la période 0.01 μs).

5.3 Modèle linéaire avec atténuation et dispersion

Dans un autre volet de la thèse d'Ewen Carcreff, nous avons cherché à raffiner la modélisation de signaux ultrasonores en prenant en compte des phénomènes de propagation acoustique. Le modèle convolutif (5.4) repose en effet sur l'hypothèse d'invariance de la forme d'onde $h_i(t)$ lors de sa propagation dans le milieu inspecté. Or, dans de nombreux matériaux, la propagation acoustique est soumise à des effets d'atténuation fréquentielle et de dispersion, qui modifient la forme d'onde au fil de la propagation. L'atténuation est due à l'absorption d'une partie du signal par les particules insonifiées, qui convertissent l'énergie acoustique en chaleur (phénomène de dissipation thermique, et à la diffusion de l'onde due à la présence d'hétérogénéités à petites échelles du matériau. Le phénomène d'atténuation acoustique augmente avec la fréquence, ce qui se traduit par un effet de filtrage passe-bas sur la forme d'onde transmise, dont l'effet augmente avec la distance de propagation. La dispersion traduit la dépendance en fréquence de la vitesse de phase de l'onde acoustique, laquelle va induire un déphasage, également cumulatif en fonction de la distance de propagation et propre à chaque fréquence, du signal transmis. L'atténuation et la dispersion sont deux phénomènes couplés, notamment par les relations de Kramers-Kronig [O'Donnell *et al.*, 1978], qui imposent la causalité du processus de propagation acoustique.

De nombreux travaux ont étudié les phénomènes d'atténuation et de dispersion (voir par exemple [Ophir et Jaeger, 1982, Narayana et Ophir, 1983]), où l'atténuation fréquentielle est souvent modélisée par une loi de puissance. Le modèle dispersif en est alors déduit afin de respecter la causalité du système physique [Gurumurthy et Arthur, 1982, Oppenheim et Schafer, 1989, Kuc, 1983]. La contribution principale de nos travaux a consisté à inclure les déformations prédites par des modèles d'atténuation et de dispersion dans un modèle linéaire de la forme (5.4) (qui sort alors du cadre convolutif strict), puis d'exploiter ce nouveau modèle dans des algorithmes d'estimation de la séquence de réflectivité associée.

Sous des hypothèses assez générales [Fink et Cardoso, 1984, Carcreff *et al.*, 2014], la propagation acoustique peut être décrite par l'application, sur la forme d'onde, d'une fonction de transfert de propagation qui s'écrit, pour une distance donnée z :

$$H_r(f, z) = b(z)e^{-\alpha(f)z}e^{-j\beta(f)z}, \tag{5.6}$$

où $b(z)$ représente la réflectivité, le terme de module $e^{-\alpha(f)z}$ représente l'atténuation et $e^{-j\beta(f)z}$ caractérise la vitesse de phase. Nous détaillons maintenant la construction de ces deux termes.

5.3.1 Modélisation de l'atténuation

La plupart des modèles d'atténuation rencontrés dans la littérature considère une loi de puissance [Ophir et Jaeger, 1982, Narayana et Ophir, 1983] :

$$\alpha(f) = \alpha_0 |f|^\gamma, \tag{5.7}$$

où les paramètres α_0 et γ caractérisent le matériau, avec $\alpha_0 \geq 0$ et en général $\gamma \in [1, 2]$. Le cas de l'atténuation linéaire (*i.e.*, $\gamma = 1$) a en particulier été validé pour de nombreux matériaux [Kak et Dines, 1978, Kuc, 1983]. En CND, des matériaux complexes peuvent présenter une atténuation non linéaire, comme certains polymères (polyéthylène, caoutchouc synthétique).

La propagation d'une onde ultrasonore sur une distance z est donc soumise à une atténuation fréquentielle représentée par la fonction de transfert de module $e^{-\alpha_0|f|^\gamma z}$, correspondant à un filtrage passe-bas. La figure 5.7 montre l'évolution de la réponse en fréquence en module de ce filtre, en fonction de la distance de propagation, dans le cas d'une atténuation linéaire. Le phénomène d'atténuation, qui s'amplifie lorsque la fréquence

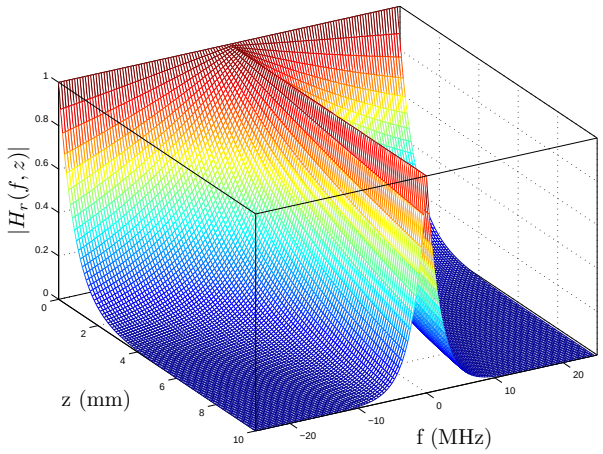


FIGURE 5.7 – Fonction de transfert de propagation (module), en fonction de la fréquence et de la distance de propagation Le modèle d'atténuation est linéaire ($\gamma = 1$) et $\alpha_0 = 50$ Np.MHz^{-1}.m^{-1}. [Carcreff, 2014b]

augmente, est donc à l'origine d'un compromis entre profondeur d'inspection et résolution : pour une même distance de propagation, des signaux ultrasonores émis à plus haute fréquence, permettant une meilleure résolution, seront plus fortement atténués. L'inspection de pièces de grande épaisseur impose donc d'utiliser des sondes à basse fréquence, moins performantes en termes de résolution, ce qui renforce l'intérêt pour la mise en œuvre de méthodes de déconvolution. Par ailleurs, l'atténuation déforme le signal émis dans le matériau en atténuant plus fortement ses hautes fréquences. C'est cette déformation que nous allons modéliser par la suite.

5.3.2 Modèle de phase dispersive associé

Plusieurs modèles ont été proposés dans la littérature, dans le cas d'un modèle d'atténuation linéaire, permettant de déduire le terme de phase $\beta(f)$ dans l'expression (5.6). Ce terme est tout d'abord décomposé en un terme de phase linéaire et un terme de phase dispersive :

$$\beta(f) = \frac{2\pi f}{c_\infty} + \xi(f), \text{ avec } \xi(f) \geq 0,$$

traduisant le fait que, pour une distance z, l'onde va subir un retard minimal correspondant au temps de vol des plus hautes fréquences, qui se propagent plus rapidement (à la vitesse c_∞), soit z/c_∞. La partie dispersive $\xi(f)$ va, elle, traduire la dépendance en fréquence de la vitesse de propagation. Nous avons retenu l'approche proposée dans [Kuc, 1984], qui impose l'expression de $\xi(f)$ par une contrainte de causalité sur la réponse impulsionnelle à temps discret, donnée par :

$$\xi(f) = -\frac{1}{F_e} \; \mathcal{P} \int_{g=-F_e/2}^{F_e/2} \alpha(g) \cot\left(\frac{\pi}{F_e}(f-g)\right) \mathrm{d}g,$$

où F_e est la fréquence d'échantillonnage et $\mathcal{P}$ correspond à la valeur principale de Cauchy de l'intégrale [1] (il faut en effet exclure la singularité en $g = f$). Dans le cas du modèle en loi de puissance (5.7), la phase dispersive s'écrit :

$$\xi(f;\alpha_0,\gamma) = -\frac{\alpha_0}{F_e} \; \mathcal{P} \int_{g=-F_e/2}^{F_e/2} |f|^\gamma \alpha(g) \cot\left(\frac{\pi}{F_e}(f-g)\right) \mathrm{d}g,$$

qui est en pratique calculée par intégration numérique. Nous avons cependant montré que, dans le cas fréquent de l'atténuation linéaire ($\gamma = 1$), cette intégrale possédait une expression analytique [Carcreff *et al.*, 2014].

L'estimation des paramètres des modèles de propagation pour un matériau donné (α_0, γ et c_∞) peut être réalisée au moyen d'un montage expérimental assez simple, visant à comparer les formes d'onde de la réponse impulsionnelle du dispositif et d'un écho isolé ayant traversé le matériau (par exemple, écho de surface et écho de fond acquis avec un échantillon suffisamment épais). Plus de détails peuvent être trouvés dans [Carcreff *et al.*, 2014, Carcreff, 2014b].

5.3.3 Prise en compte dans le modèle direct

La fonction de transfert de propagation, pour une distance z, s'écrit donc :

$$H_r(f,z) = b(z)e^{-j\alpha(f)z}e^{-j(2\pi fz/c_\infty+\xi(f))} = b(z)e^{-j2\pi fz/c_\infty}H_a(f,z), \tag{5.8}$$

où $H_a(f,z)$ regroupe les termes dus à l'atténuation et à la phase dispersive.

Revenons maintenant aux notations introduites en Section 5.1, où $h_i(t)$ représente la réponse impulsionnelle associée au couple émetteur-récepteur ultrasonores. Dans le cas

1. La valeur principale de Cauchy de l'intégrale d'une fonction $f(x)$ singulière en $x = a$ s'écrit, lorsqu'elle existe (c'est le cas ici) :

$$\mathcal{P}\int_{-\infty}^{+\infty} f(x)\mathrm{d}x = \lim_{\substack{u \to 0 \\ u>0}} \int_{-\infty}^{a-u} f(x)\mathrm{d}x + \int_{a+u}^{+\infty} f(x)\mathrm{d}x.$$

d'un réflecteur ponctuel situé à une distance z du capteur, en présence d'atténuation et de dispersion, le signal d'écho reçu s'écrit maintenant $y^{\text{echo}}(t) = h_i(t) * h_r(t, 2z)$, où $h_r(t, z)$ est la *réponse impulsionnelle de propagation*, associée à la fonction de transfert $H_r(f, z)$, et le facteur 2 traduit le trajet aller-retour de l'onde. On peut alors écrire d'après (5.8) :

$$y^{\text{echo}}(t) = b(z)\, h_i(t) * h_a(t - 2z/c_\infty, 2z), \tag{5.9}$$

où $h_a(t, z)$ est la réponse impulsionnelle associée à la fonction de transfert $H_a(f, z)$. Cette forme est à comparer à la forme (5.1) dans le cadre d'un modèle sans atténuation, *i.e.*, pour lequel $h_a(t - 2z/c_\infty, 2z) = \delta(t - 2z/c_\infty)$.

Par une démarche similaire à celle de la Section 5.1, nous construisons alors un modèle direct sous la forme

$$\boldsymbol{y} = \mathbf{H}_i \mathbf{H}_a \boldsymbol{x} + \boldsymbol{\epsilon}, \tag{5.10}$$

où, après discrétisation de l'axe temporel à la période T_e (la période d'échantillonnage des données) et de l'axe spatial à la période $\Delta_z = c_\infty T_e$:

- $\mathbf{H}_i$ est la matrice de convolution discrète, de terme général $\left[h_i\left((m-n)T_e\right)\right]_{m,n}$,
- $\mathbf{H}_a$ est la matrice de terme général $\left[h_a(nT_e, p\Delta_z)\right]_{n,p}$,
- $\boldsymbol{x}$ représente la séquence de réflectivité, à un facteur d'échelle près : $x_p = T_e \Delta_z\, b(p\Delta_z)$.

5.3.4 Quelques exemples de résultats

Nous présentons ici quelques résultats mettant en évidence l'apport d'une modélisation plus fine prenant en compte atténuation et dispersion fréquentielles.

Validation expérimentale du modèle

Nous avons réalisé plusieurs campagnes de mesures au Laboratoire d'Acoustique de l'Université du Maine, utilisant différents types de matériaux. À partir de la mesure d'une forme de référence, identifiée à $h_i(t)$, nous avons cherché à prédire le meilleur modèle d'écho en optimisant les paramètres de propagation associés. Le résultat a ensuite été comparé i) au modèle obtenu en injectant un simple retard dans la forme d'onde (correspondant au modèle convolutif standard (5.4)), ii) au modèle prenant uniquement en compte l'atténuation et la phase linéaire (*i.e.*, sans phase dispersive), fréquemment utilisé dans la littérature, et iii) au modèle proposé par Olofsson et Stepinski [Olofsson et Stepinski, 2001], que nous appellerons modèle d'Olofsson. Dans cette dernière référence, la déformation de l'onde entre deux échantillons successifs est modélisée par un filtrage de réponse impulsionnelle $[0, 1-a, a]$, où le paramètre $a > 0$ est à déterminer. Le modèle convolutif correspond alors au cas $a = 0$, où cette propagation comporte uniquement un retard pur d'un échantillon. À notre connaissance, ce modèle constitue l'unique « concurrent » à notre approche, permettant de former un modèle linéaire reliant un signal ultrasonore à la séquence de réflectivité sous une forme non exactement convolutive.

La figure 5.8 montre le modèle obtenu pour une forme d'onde s'étant propagée dans une plaque de polycarbonate de 10.2 mm d'épaisseur, matériau pour lequel l'atténuation est supposée linéaire en fréquence, *i.e.*, $\gamma = 1$ [Selfridge, 1985]. Le modèle invariant par translation ne colle que grossièrement aux données. La prise en compte de l'atténuation fréquentielle permet d'améliorer sensiblement le modèle. Mais surtout, la figure 5.8 c) met clairement en évidence l'apport du modèle avec atténuation et phase dispersive, par rapport au modèle à phase linéaire mais aussi au modèle d'Olofsson. Cette figure reproduit

des résultats de la thèse d'Ewen Carcreff [Carcreff, 2014a], dans laquelle figurent également des exemples de validation sur d'autres matériaux.

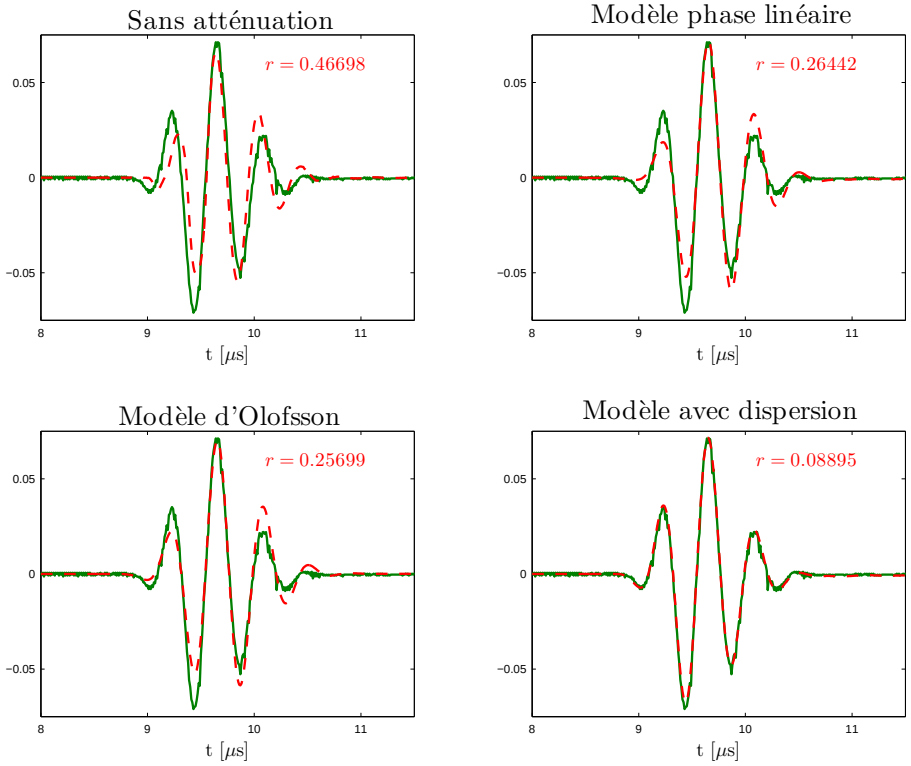


FIGURE 5.8 – Prédiction par différents modèles de l'écho de fond d'une onde ultrasonore ayant traversé une plaque de polycarbonate de 10.2 mm d'épaisseur. Dans chaque cas, les paramètres ont été optimisés pour fournir le résidu r le plus faible.

Exploitation pour le CND

Nous présentons pour conclure cette section un exemple d'application en CND, *via* une expérience menée au Laboratoire d'Acoustique de l'Université du Maine, consistant en l'inspection d'une plaque de polyméthacrylate de méthyle (PMMA ou Plexiglas®), dans laquelle a été usiné un trou à fond plat (Flat Bottom Hole, FBH) au fond de la pièce à une profondeur de 0.5 mm, simulant un manque de matériau. En raison de la faible profondeur du trou, les deux échos associés se chevauchent dans le signal ultrasonore reçu. La figure 5.9 montre les résultats obtenus par déconvolution parcimonieuse [2] s'appuyant sur le modèle (5.10). Avec le modèle convolutif standard, les deux composantes estimées sont de signe positif, ce qui est contraire à l'interprétation physique de la séquence de réflectivité. À l'inverse, le modèle prenant en compte atténuation et dispersion produit deux composantes négatives, conformes à la réalité physique, et améliorant la modélisation du signal, le résidu d'estimation étant plus faible.

2. Nous avons utilisé ici un algorithme de type SBR, généralisé à la recherche itérative de mouvements impliquant deux atomes. Cette variante, dont le coût calculatoire est plus élevé, est cependant préférable sur cet exemple où les deux échos interfèrent fortement.

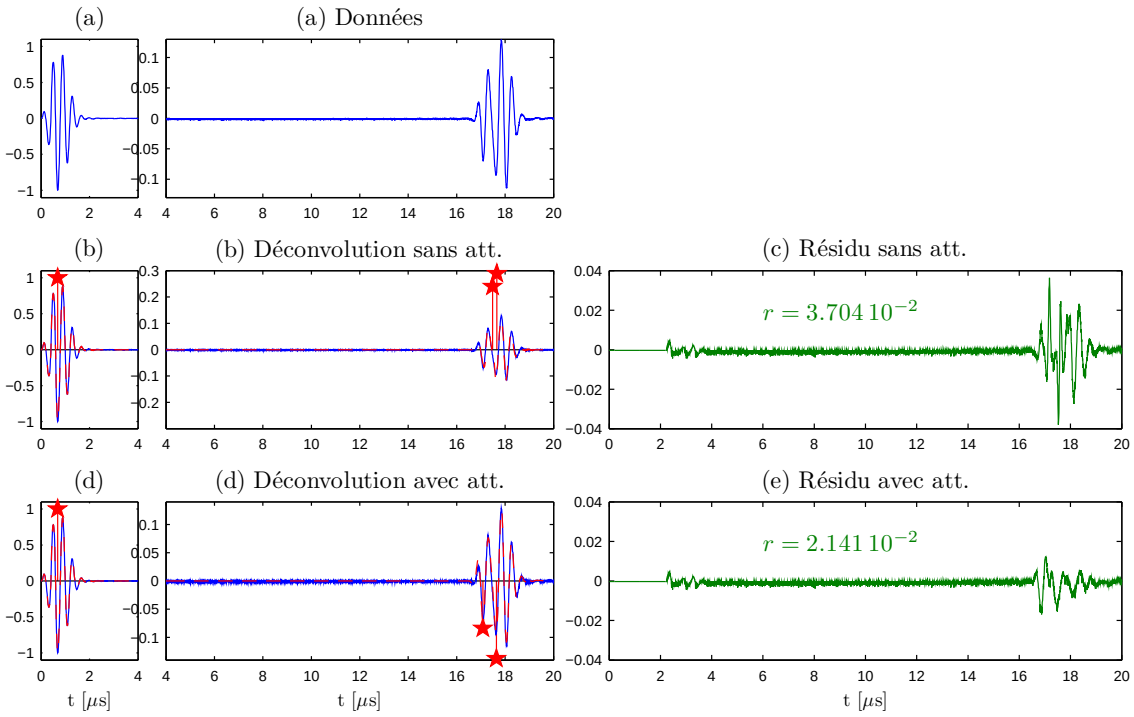


FIGURE 5.9 – Résultats de déconvolution obtenus pour la séparation de deux échos dus à la présence d'un trou proche du fond d'une pièce de PMMA. En haut, les données séparées entre l'écho de surface (à gauche) et le mélange des deux échos du trou et du fond de la pièce (à droite), agrandi en amplitude. Au milieu à gauche, résultat obtenu avec un modèle convolutif classique sans atténuation. En bas à gauche, résultat avec un modèle prenant en compte atténuation et dispersion. La trait plein bleu représente les données et les étoiles rouges les temps de vol et amplitudes des échos estimés (séquence $\boldsymbol{x}$ parcimonieuse). Le trait plein rouge représente le modèle associé : $\mathbf{H}_i\boldsymbol{x}$ en (b) et $\mathbf{H}_i\mathbf{H}_a\boldsymbol{x}$ en (c). Les figures de droite représentent le résidu entre les données et le modèle et indiquent sa norme.

5.4 Un cas particulier : mesures ultrasonores d'épaisseurs fines

Les travaux réalisés lors de la thèse d'Ewen Carcreff sur la déconvolution parcimonieuse m'ont amené à envisager le cas particulier de la mesure, par une méthode ultrasonore, de l'épaisseur de matériaux qui ne sont accessibles que d'un côté. Dans le cadre de projets étudiants ingénieurs de l'École Centrale de Nantes, nous avons abordé ce problème sous l'angle de la déconvolution, où la séquence de réflectivité recherchée, certes parcimonieuse, peut être modélisée de manière plus contrainte.

5.4.1 Contexte applicatif : mesures de pièces accessibles d'un seul côté

La mesure d'épaisseur de pièces ou de milieux accessibles d'un seul côté est une problématique rencontrée dans de nombreux domaines industriels. On peut citer le contrôle de l'épaisseur de revêtements dans le domaine de la fabrication industrielle (automobile, aéronautique) ou de la peinture, ou encore celui de tuyauteries soumises à des fortes contraintes de température et de pression, comme les circuits d'alimentation en eau des centrales électriques [O'Brien *et al.*, 1994] ou des conduites de pipelines [de Raad et Dijkstra, 1998]. L'utilisation de dispositifs ultrasonores est une modalité couramment utilisée dans ce contexte [Krautkramer et Krautkramer, 1990], de par son faible coût, sa facilité de déploiement dans des milieux difficiles d'accès, et son applicabilité à de nombreux types de matériaux.

Le principe de la mesure, schématisé en figure 5.10, repose sur l'évaluation du temps de vol entre les échos successifs d'une impulsion ultrasonore émise du côté accessible du matériau, subissant plusieurs allers-retours dans le matériau. Une limitation claire des méthodes ultrasonores réside alors dans le manque de résolution : si l'épaisseur du matériau est trop fine par rapport à la longueur l'impulsion, les échos se chevauchent et une simple inspection visuelle, ou une méthode basique d'estimation du temps de vol, s'avère inefficace, comme sur l'exemple de la Figure 5.10 à droite.

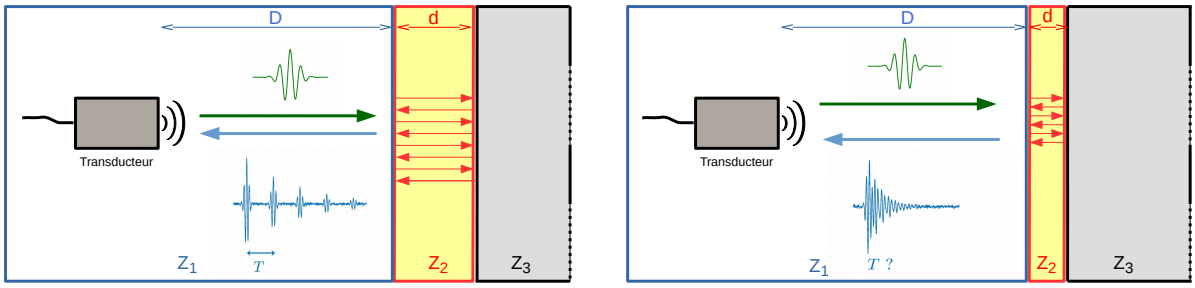


FIGURE 5.10 – Principe de mesure d'épaisseur par ultrasons : les aller-retours de l'onde dans le milieu d'impédance acoustique Z_2 génèrent des échos régulièrement espacés et d'amplitude décroissante. À gauche, l'épaisseur d est suffisamment grande devant la longueur d'onde pour que les échos soient séparés. À droite, l'épaisseur est plus fine et les échos se chevauchent dans les données.

5.4.2 Modèles de réflectivité et de réponse impulsionnelle

Considérons un dispositif ultrasonore selon le schéma de la Figure 5.10 : un matériau homogène d'épaisseur d inconnue (milieu 2) est insonifié par un transducteur émettant

et recevant les signaux ultrasonores, en incidence normale, éventuellement à une distance D du matériau et immergé dans un milieu 1. L'autre face du matériau est adossée à un milieu 3, d'épaisseur suffisamment grande devant d pour ne pas générer d'autres réflexions. Comme nous l'avons vu en § 5.1, lorsqu'une impulsion ultrasonore est émise par le transducteur, une partie de l'onde est réfléchie vers le transducteur et l'autre partie est transmise dans le matériau. Celle-ci va également se réfléchir partiellement à l'interface entre le matériau et le milieu 3, puis une partie sera transmise vers le transducteur et l'autre à nouveau réfléchie dans le matériau, etc. Par un modèle similaire à celui de l'équation (5.1), nous pouvons donc décrire le signal reçu par le transducteur :

$y(t) = \sum_{k=0}^{+\infty} y_k^{\text{echo}}(t)$, où $y_k^{\text{echo}}(t) = b_k\, h_i(t - t_0 - 2kd/c)$ est la contribution du $k^{\text{ème}}$ écho et :

- h_i représente la réponse impulsionnelle du transducteur ;
- t_0 est le temps de vol aller-retour de l'onde ultrasonore entre le transducteur et le matériau, supposé connu ($t_0 = 2D/c_1$, où c_1 est la vitesse du son dans le milieu 1) ;
- b_k est le coefficient de réflectivité associé au $k^{\text{ème}}$ écho.

Nous avons donc finalement un modèle convolutif :

$$y(t) = h_i * x(t), \text{ avec } x(t) = \sum_{k=0}^{+\infty} b_k \delta(t - t_0 - kT), \tag{5.11}$$

où $T = 2d/c$ est le temps de vol aller-retour dans le matériau, fonction de l'épaisseur recherchée d et de la vitesse du son dans le matériau.

Modélisation des coefficients de réflectivité

Le modèle (5.11) est un produit de convolution par une séquence parcimonieuse, mais où les temps de vol sont régulièrement espacés, d'une durée T inconnue. À T fixé, ce modèle est linéaire en les amplitudes b_k ; cependant, il est possible de contraindre ces amplitudes par un modèle plus précis. L'amplitude b_k traduit en effet deux phénomènes distincts : les pertes dues aux multiples réflexions et l'atténuation due à la propagation, que nous écrivons avec des notations évidentes :

$$b_k = R_k\, P_k.$$

Notons respectivement r_{ij} et t_{ij} les coefficients de réflexion et de transmission entre le milieu i et le milieu j, définis par [Krautkramer et Krautkramer, 1990] :

$$r_{ij} = \frac{Z_j - Z_i}{Z_i + Z_j} \text{ et } t_{ij} = \frac{2Z_j}{Z_i + Z_j} \tag{5.12}$$

où Z_i est l'impédance acoustique du milieu i. L'ensemble des transmissions et réflexions subies par le $k^{\text{ème}}$ écho se décompose alors en :

$$\begin{cases} R_0 &= r_{12}, \\ R_k &= t_{12}\, r_{23}^k\, r_{21}^{k-1}\, t_{21} = (1 - r_{12}^2)\, r_{23}^k\, r_{21}^{k-1} \text{ pour } k \geq 1, \end{cases} \tag{5.13}$$

la dernière égalité exploitant, par (5.12), les relations $r_{ji} = -r_{ij}$ et $t_{ij} = 1 + r_{ij}$.

L'atténuation due à la propagation se modélise classiquement sous une forme exponentielle (voir le § 5.3.1). Nous négligeons ici sa dépendance en fréquence [3]. Par des notations analogues à celles du § 5.3.1, on a alors :

$$P_k = e^{-2\alpha_1 D} e^{-2\alpha_2 (kd)}, \tag{5.14}$$

où les coefficients α_1 et α_2 dépendent du milieu traversé. En combinant les modèles (5.11) et (5.14), la séquence $x(t)$ s'écrit :

$$x(t) = r_{12} e^{-2\alpha_1 D} \delta(t - t_0) + \sum_{k=1}^{+\infty} (1 - r_{12}^2) r_{23}^k r_{21}^{k-1} e^{-2\alpha_1 D} e^{-2\alpha_2 (kd)},$$

soit, après reparamétrage en $a \triangleq r_{12} e^{-2\alpha_1 D}$, $\triangleq (1 - r_{12}^2) r_{23} e^{-2\alpha_1 D - 2\alpha_2 d}$ et $c \triangleq r_{23} r_{21} e^{-2\alpha_2 d}$:

$$x^{\boldsymbol{\theta}_x}(t) = a\delta(t - t_0) + b \sum_{k=1}^{+\infty} c^{k-1} \delta(t - t_0 - kT), \text{ où } \boldsymbol{\theta}_x \triangleq (a, b, c, T). \tag{5.15}$$

Modélisation de la réponse instrumentale

L'estimation des paramètres de la séquence de réflectivité, reposant sur le modèle convolutif (5.11), requiert la connaissance de la réponse impulsionnelle $h_i(t)$. C'est un point particulièrement critique ici, puisque l'estimation directe de $h_i(t)$ à partir du signal reçu est intrinsèquement liée à une nette séparation des échos dans le signal – cas pour lequel aucune méthode avancée d'estimation de temps de vol n'est nécessaire ! Afin de s'affranchir d'une étape préalable de calibrage, nous avons abordé l'estimation de l'épaisseur sous l'angle d'un problème de déconvolution myope, en imposant également un modèle paramétrique de la réponse $h_i(t)$ sous la forme d'une gaussienne modulée en fréquence [Demirli et Saniie, 2001] :

$$h_i^{\boldsymbol{\theta}_h}(t) = \cos(2\pi\nu t + \phi) e^{-\beta t^2}, \text{ où } \boldsymbol{\theta}_h = (\nu, \phi, \beta), \tag{5.16}$$

où le choix d'une amplitude unité permet de lever l'ambiguïté d'échelle, dans le modèle (5.11), entre la réponse impulsionnelle et la séquence de réflectivité. La figure 5.11 résume le modèle de données, conjuguant un modèle gaussien pour la réponse impulsionnelle et une séquence d'impulsions régulièrement espacées et à décroissance géométrique pour la réflectivité.

5.4.3 Déconvolution myope

L'ensemble des paramètres $\boldsymbol{\theta}_x$ de la séquence de réflectivité et $\boldsymbol{\theta}_h$ de la réponse impulsionnelle est alors estimé conjointement, au sens des moindres carrés :

$$(\widehat{\boldsymbol{\theta}}_x, \widehat{\boldsymbol{\theta}}_h) = \arg\min_{\boldsymbol{\theta}_x, \boldsymbol{\theta}_h} J(\boldsymbol{\theta}_x, \boldsymbol{\theta}_h), \text{ où } J(\boldsymbol{\theta}_x, \boldsymbol{\theta}_h) \triangleq \sum_{n=1}^{N} \Big(y(t_n) - (h_i^{\boldsymbol{\theta}_h} * x^{\boldsymbol{\theta}_x})(t_n) \Big)^2,$$

où $y(t_n)$ représentent les données échantillonnées aux instants $t_n = nT_e$ et les modèles de la séquence $x^{\boldsymbol{\theta}_x}$ et de la réponse impulsionnelle $h_i^{\boldsymbol{\theta}_h}$ sont respectivement donnés par (5.15)

3. Cette approximation est légitimée par le fait que nous considérons ici des faibles épaisseurs, donc des faibles distances de propagation.

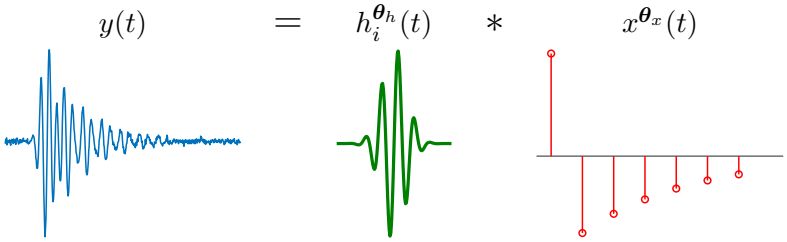


FIGURE 5.11 – Modèle de données pour l'estimation d'épaisseurs, résultant de la convolution d'une séquence d'impulsions régulièrement espacées par une forme d'onde gaussienne modulée en amplitude.

et (5.16). C'est un problème de moindres carrés non-linéaires, dont l'optimisation a été réalisée avec un algorithme itératif de Levenberg-Marquardt [Nocedal et Wright, 2006]. Le critère J n'étant pas convexe, la solution obtenue sera donc sensible au point initial. Une étude détaillée nous a amenés à envisager une initialisation des paramètres basée sur les éléments suivants.

Paramètres instrumentaux

- La fréquence ν correspond à la fréquence centrale du transducteur et peut être initialisée à la fréquence maximisant la transformée de Fourier (en module) des données.
- Le paramètre β, réglant la largeur de l'impulsion émise, peut être relié au facteur de bande passante (BWR, band width ratio) du transducteur, sous la forme approchée [Carcreff, 2014a] :

$$\beta \simeq \pi^2 \frac{\text{BWR}}{4\ln\sqrt{2}}\nu.$$

  Par conséquent, β est initialisé en fixant BWR $= 0.5$, valeur raisonnable pour des transducteurs courants.
- La phase φ est initialisée à 0. Notons que le critère J est $2\pi-$périodique en φ, mais il n'est pas souhaitable de contraindre la phase dans $[0, 2\pi[$: une estimation de φ modulo 2π donne le même modèle, et borner la phase dans un intervalle augmente le risque que l'optimisation vienne "buter" sur un bord de cet intervalle.

Paramètres de réflectivité

- Les paramètres a, b, c sont initialisés à 0.5.
- Le critère s'avère fortement multimodal en T, le temps de vol aller-retour de l'onde acoustique dans le matériau, notamment en raison de la nature oscillante de la forme d'onde. Afin de limiter le risque de tomber dans un mode local, l'optimisation est lancée successivement pour dix valeurs initiales de T régulièrement espacées dans un intervalle de valeurs admissibles $[T_{\min}, T_{\max}]$. Nous supposons ainsi que l'épaisseur recherchée est connue de manière grossière, entre $cT_{\min}/2$ et $cT_{\max}/2$, où c est la vitesse du son dans le milieu inspecté.

5.4.4 Résultats

Nous présentons ici une application sur des données expérimentales acquises au Laboratoire d'Acoustique de l'Université du Maine. Des plaques d'aluminium sont immergées dans l'eau et insonifiées, en incidence normale, par des mesures en réflexion avec un transducteur de fréquence centrale $f_0 = 2.25$ MHz (voir Figure 5.12). Notons que dans ce cas,

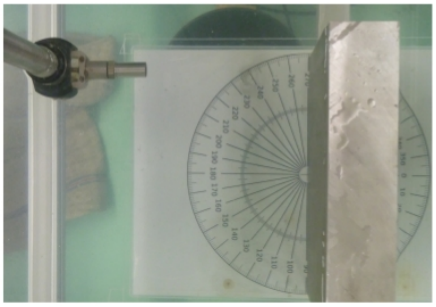


FIGURE 5.12 – Dispositif expérimental pour la mesure ultrasonore de plaques d'aluminium : la plaque est immergée dans de l'eau et insonifiée en réflexion.

les milieux de part et d'autre de la plaque sont les mêmes et, en négligeant l'atténuation de l'onde dans l'eau ($\alpha_1 = \alpha_2 = 0$), le modèle (5.15) est simplifié en

$$x^{\boldsymbol{\theta}_x}(t) = a\delta(t - t_0) - (1 - a^2)\sum_{k=1}^{+\infty} a^{2k-1}\delta(t - t_0 - kT), \text{ où } \boldsymbol{\theta}_x = (a, T).$$

Les données et les résultats associés pour des plaques d'épaisseur 4 mm, 2 mm et 1 mm sont présentés en Figure 5.13. La vitesse du son dans l'aluminium étant de $c_{\text{Alu}} \simeq 6\,380$ m.s^{-1}, la longueur d'onde équivalente vaut alors $\lambda = c_{\text{Alu}}/f_0 = 2.8$ mm et ces épaisseurs correspondent respectivement à $1.4\,\lambda$, $0.7\,\lambda$ et $0.35\,\lambda$, générant des chevauchements forts à très forts dans les données. Malgré cela, dans les trois cas, la déconvolution est rendue possible grâce aux très fortes contraintes imposées sur le modèle, permettant d'estimer très précisément l'épaisseur des plaques. Notons cependant que le coût de calcul reste élevé (de l'ordre de la dizaine de secondes), notamment en raison de la nécessité de lancer l'optimisation pour plusieurs initialisations du temps de vol recherché T.

5.5 Vers de nouvelles méthodes d'imagerie des matériaux

Les travaux présentés dans les sections précédentes concernent le CND à partir de signaux ultrasonores monodimensionnels (ou A-scans) qui, pour des raisons évidentes de simplicité et de portabilité des appareils de mesure, sont très fréquemment utilisés dans un contexte industriel. Le développement de dispositifs d'acquisition multi-éléments est cependant en plein essor et, s'il a d'abord concerné l'imagerie biomédicale, il touche désormais le CND [Holmes *et al.*, 2005], notamment grâce à l'usage de cartes graphiques permettant des calculs de reconstruction en temps réel [Sutcliffe *et al.*, 2012]. L'exploitation de données multi-éléments permet en effet la formation d'images, fournissant un degré d'information bien plus raffiné qu'avec une simple collection d'A-scans. Dans un contexte de CND, cependant, les méthodes développées restent en général assez simples afin de limiter le coût de calcul [Holmes *et al.*, 2005, Quaegebeur et Masson, 2012].

Nous nous intéressons, dans la continuité des travaux initiés par la thèse d'Ewen Carcreff, à des méthodes plus avancées exploitant des données multi-éléments, relevant de

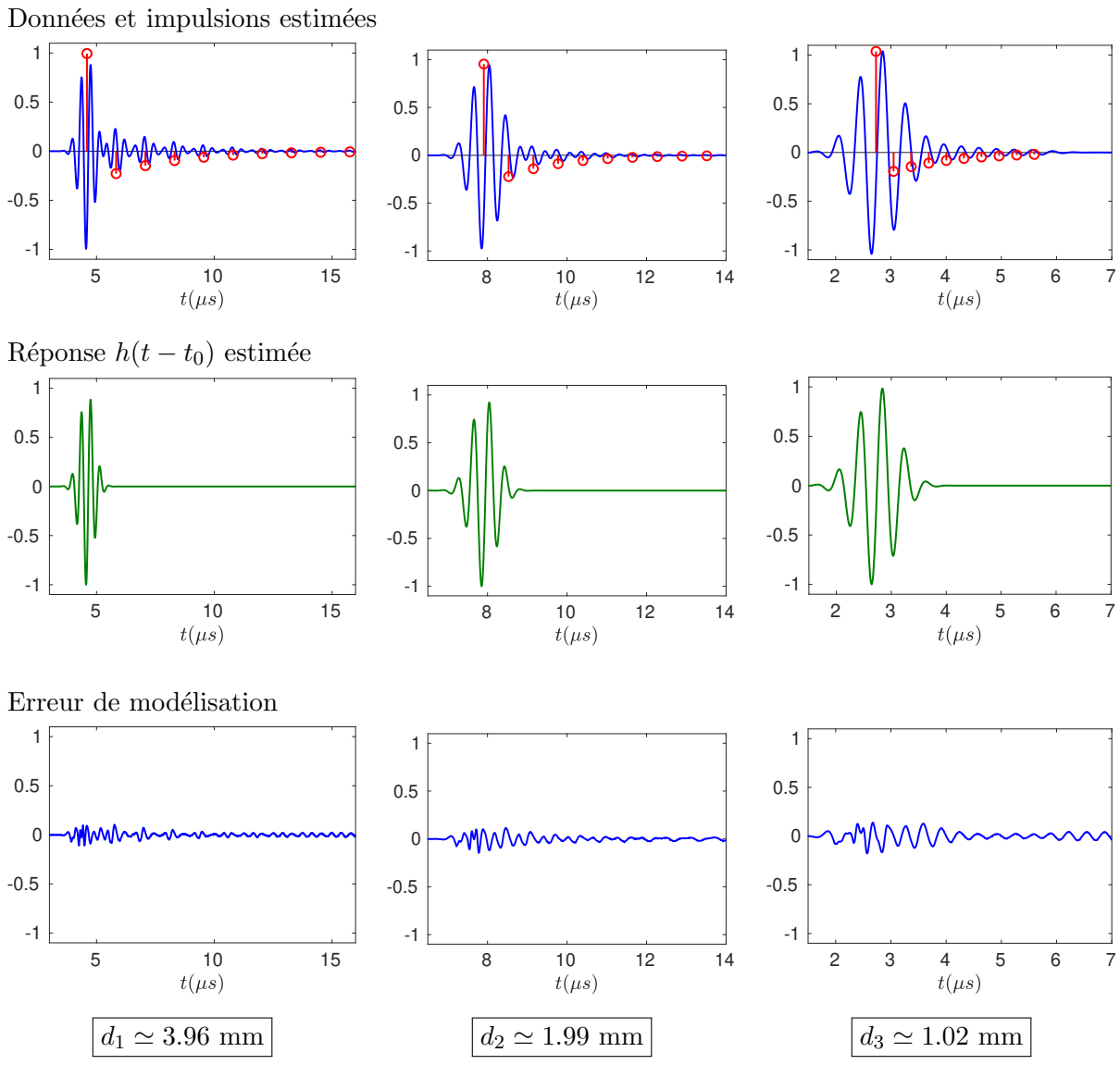


FIGURE 5.13 – Résultats de déconvolution myope de données acquises en réflexion pour l'estimation d'épaisseur. En haut, données (trait plein) et séquence de réflectivité estimée (∘). Au centre, réponse impulsionnelle estimée. En bas, écart entre les données et le modèle estimé. À gauche, au centre et à droite : plaques de 4 mm, 2 mm et 1 mm respectivement. Les valeurs d'épaisseur estimées sont renseignées en bas.

la méthodologie des problèmes inverses. Si elles peuvent être plus coûteuses en temps de calcul, nous visons des applications où la reconstruction en temps réel n'est pas forcément l'objectif prioritaire – tablant également sur l'évolution simultanée des machines de calcul, notamment à base de programmation sur GPU. Ces réflexions sont à la base de la thèse de Nans Laroche, démarrée à l'automne 2017, dans le cadre d'une convention CIFRE avec la société DB-SAS. Cette section présente quelques travaux engagés depuis le démarrage de cette thèse.

5.5.1 Données *Full Matrix Capture* et méthodes de refocalisation

Les données *Full Matrix Capture* (FMC), dites encore multi-statiques, consistent en l'ensemble des ${N_{\text{el}}}^2$ signaux produits par un réseau de N_{el} transducteurs, chacun émettant à tour de rôle et l'ensemble des transducteurs recevant la réponse. La figure 5.14 illustre ce principe. Considérons une barrette de transducteurs alignés le long d'un axe de co-

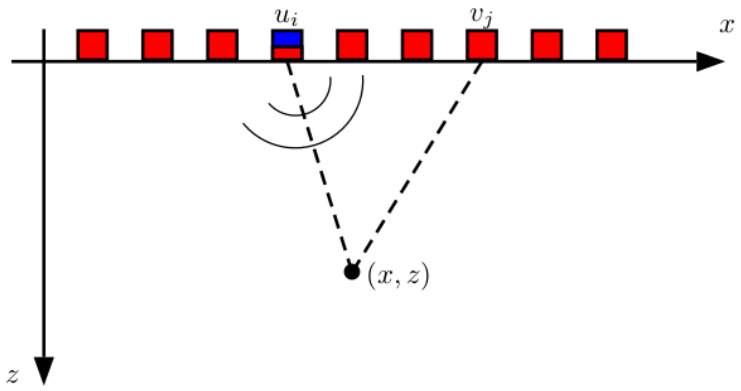


FIGURE 5.14 – Acquisition de données FMC : l'onde acoustique émise par le transducteur i se propage et se réfléchit dans le milieu, puis est reçue par le transducteur j.

ordonnées x (correspondant à la majorité des configurations existantes), inspectant un milieu dans une direction orthogonale représentée par un axe de coordonnées z. Nous notons $y_{ij}(t)$ le signal provenant de l'émission par le transducteur i, d'abscisse u_i et reçu sur le transducteur j, d'abscisse v_j.

La méthode d'imagerie standard à partir de ces données, dite TFM pour *Total Focusing Method* [Holmes *et al.*, 2005], consiste à *focaliser* en tout point (x, z) du milieu inspecté en sommant l ?ensemble des signaux y_{ij} « remis au bon temps de vol », soit :

$$O_{\text{TFM}}(x,z) = \sum_{i,j=1}^{N_{\text{el}}} y_{ij}(\tau(x,z,i,j)), \quad \text{avec } \tau(x,z,i,j) = \frac{\sqrt{(x-u_i)^2+z^2} + \sqrt{(x-v_j)^2+z^2}}{c}, \tag{5.17}$$

où c est la vitesse de l'onde acoustique dans le milieu, supposée constante. Si un réflecteur est localisé en (x, z), alors l'image TFM va superposer l'ensemble des contributions apportées par chaque signal en cet endroit et donc présenter une valeur élevée. L'imagerie TFM est un processus linéaire donc peu coûteux en ressources, les calculs pouvant de plus s'effectuer simultanément à l'acquisition. En revanche, l'image formée souffre en général d'un manque de résolution : de la même manière que les signaux monodimensionnels contiennent la forme de l'onde à chaque réflexion (voir par exemple la figure 5.4), l'image va contenir des formes oscillantes, limitant son interprétabilité. Ce manque de résolution

limite alors la capacité à imager correctement des défauts, et en particulier à séparer des défauts proches, qui est une problématique courante de CND.

Nos recherches visent donc à améliorer la résolution des images reconstruites où, de manière analogue à nos travaux sur la déconvolution de signaux de type A-scan, nous allons exploiter un modèle de formation de données prenant en compte la réponse des transducteurs.

5.5.2 Inversion de données FMC

Chacun des signaux $y_{ij}(t)$ peut être modélisé comme nous l'avons fait pour les signaux monodimensionnels dans les sections précédentes de ce chapitre [4]. En notant $\boldsymbol{y} \in \mathbb{R}^{{N_{\text{el}}}^2 N_t}$ le vecteur colonne obtenu en concaténant l'ensemble des signaux y_{ij} échantillonnés sur N_t points, nous pouvons alors écrire un modèle :

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{o} + \boldsymbol{\epsilon} \tag{5.18}$$

où $\boldsymbol{o} \in \mathbb{R}^{N_x N_z}$ correspond à la réflectivité (vectorisée en colonne) discrétisée sur une grille de $N_x \times N_z$ points, $\boldsymbol{\epsilon}$ représente classiquement le bruit et les erreurs de modélisation et la matrice $\mathbf{H}$ dépend de la forme d'onde due aux transducteurs et des temps de vol relatifs aux différentes positions. Même si la discrétisation en (x, z) peut être arbitrairement fine, un choix classique consiste à prendre un pas de discrétisation identique dans les deux dimensions, correspondant à la période d'échantillonnage temporel Δ_t des signaux y_{ij}, soit $\Delta_x = \Delta_z = c\Delta_t$. Le modèle (5.18) est en général largement surdéterminé, puisque l'on cherche à reconstruire une image à partir d'un ensemble de ${N_{\text{el}}}^2$ signaux. Le problème reste cependant mal posé car, comme dans le cas monodimensionnel, l'action de $\mathbf{H}$ revient à filtrer les composantes de l'image de réflectivité par une action de type passe-bande. En particulier, les données contiennent peu d'information sur les hautes fréquences recherchées, lesquelles impactent fortement la résolution de l'image reconstruite.

Nous avons donc proposé une approche régularisée, définissant l'image par la minimisation d'un critère pénalisé :

$$\widehat{\boldsymbol{o}} = \arg\min_{\boldsymbol{o}} J(\boldsymbol{o}), \text{ avec } J(\boldsymbol{o}) \stackrel{\triangle}{=} \frac{1}{2}\left\|\boldsymbol{y} - \mathbf{H}\boldsymbol{o}\right\|^2 + \mu_1 \left\|\boldsymbol{o}\right\|_1 + \mu_2 \left\|\mathbf{D}\boldsymbol{o}\right\|^2,$$

où le premier terme de pénalisation favorise la parcimonie de l'image de réflectivité et le second terme, opérant sur les différences $\mathbf{D}\boldsymbol{o}$ entre pixels voisins, permet de lisser spatialement la solution : pour des applications de CND, si le milieu traversé est essentiellement homogène (correspondant à l'absence de réflecteurs), les réflecteurs ne sont pas forcément toujours ponctuels. On montre alors que pour

$$\widehat{\boldsymbol{o}} = \mathbf{0} \Leftrightarrow \mu_1 \geq \mu_{\max}, \text{ avec } \mu_{\max} := \left\|\boldsymbol{H}^T\boldsymbol{y}\right\|_\infty.$$

En pratique, μ_1 est donc réglé comme une fraction de $\mu_{\max}$. Dans l'ensemble de nos expériences, nous avons réglé μ_2 à une petite valeur positive, *e.g.,* 10^{-2}.

Nous avons opté pour un algorithme standard d'optimisation en norme ℓ_1 : l'algorithme FISTA (*Fast Iterative Shrinkage Algorithm*) [Beck et Teboulle, 2009], qui s'étend sans problème au cas d'un second terme de pénalisation quadratique. S'agissant d'une méthode itérative de type gradient projeté, elle requiert de nombreuses évaluations du

4. Nous avions alors considéré des mesures en réflexion, correspondant au cas $i = j$. Dans le cas $i \neq j$, il s'agit plutôt de mesures en transmission.

gradient de la partie quadratique, impliquant des calculs matriciels de type $\mathbf{H}\boldsymbol{o}$ et $\mathbf{H}^T\boldsymbol{u}$. En raison de la géométrie d'inspection, cependant, l'opérateur $\mathbf{H}$ ne possède pas de structure facilement exploitable pour des calculs rapides et ne peut pas être explicitement stocké en mémoire. Nous avons donc proposé une mise en œuvre parallélisée à l'aide de GPU (*graphics processing unit*), où les différents éléments de chaque produit matrice-vecteur sont envoyés sur les différentes unités de calcul.

Ces travaux ont fait l'objet d'une publication de conférence [Laroche *et al.*, 2018] et de la préparation d'un article de journal en cours de finalisation. Nous en reproduisons ci-dessous quelques résultats, respectivement sur données simulées et données réelles, montrant le fort potentiel d'amélioration de cette approche par rapport à l'imagerie TFM.

La figure 5.15 montre un exemple sur des données simulées selon le modèle (5.18) – avec une résolution spatiale quatre fois plus fine que celle utilisée pour la reconstruction – où la scène est composée de deux réflecteurs ponctuels successivement séparés de λ, $\lambda/2$ et $\lambda/4$, λ désignant la longueur d'onde du signal ultrasonore, fixée ici à 1 mm. La forme d'onde utilisée pour simuler les données et pour la reconstruction est générée selon un modèle gaussien modulé en amplitude. Pour la méthode TFM, l'image brute (ligne du haut) présente une nature oscillante due aux oscillations de la forme d'onde. Nous représentons donc également (ligne du milieu) l'image obtenue en prenant l'enveloppe de chaque colonne de l'image brute. La méthode TFM parvient à séparer les deux défauts seulement dans la première configuration. À l'inverse, la méthode de reconstruction proposée (avec ici $\mu_1 = 0.6\mu_{\max}$ et $\mu_2 = 0$, les réflecteurs recherchés étant ponctuels) sépare très nettement les deux réflecteurs, avec une distance proche de la réalité. Notons cependant que le temps de calcul est ici de l'ordre de 500 s pour la reconstruction d'une image de 301×301 pixels, en raison de la grande taille des données (composées ici de $N_{\text{el}}^2 = 64^2$ signaux).

Enfin, la figure 5.16 présente un exemple sur des données réelles, où une pièce test en aluminium (représentée à gauche) a été usinée avec deux trous proches percés latéralement, de diamètre 1 mm, séparés entre eux d'1 mm environ. La fréquence des transducteurs est de 1.5 MHz, correspondant à une longueur d'onde de 4.2 mm et la pièce est inspectée par le côté supérieur. L'image TFM ne contient qu'une tache non résolue centrée au niveau des trous, alors que l'image obtenue par notre méthode produit deux traces séparées, permettant de détecter les deux trous. Le temps de calcul est ici de l'ordre de 2 minutes, pour la reconstruction d'une image 1001×1001 à partir de $N_{\text{el}}^2 = 32^2$ signaux.

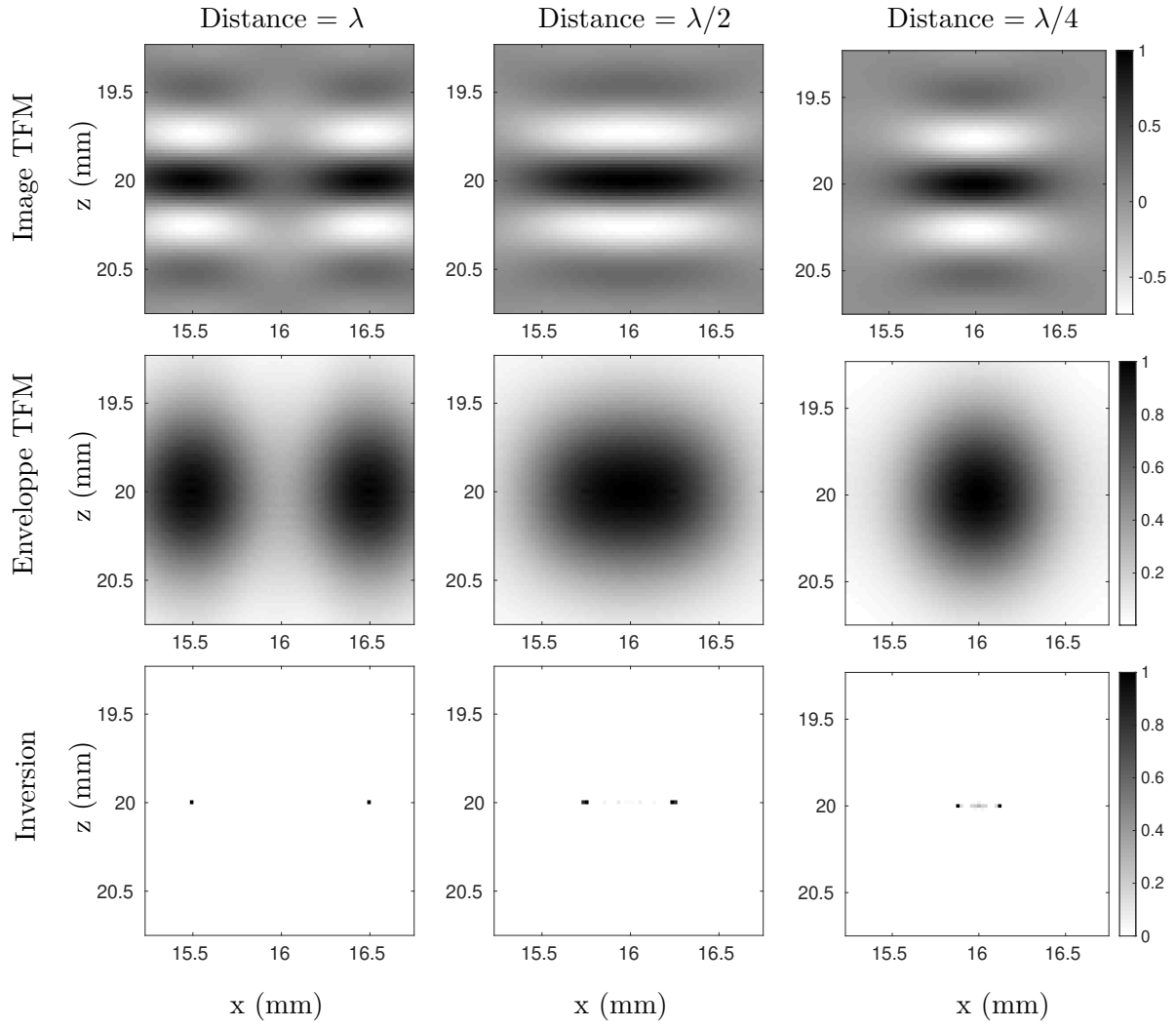


FIGURE 5.15 – Inversion de données FMC simulées pour une scène composée de deux réflecteurs ponctuels séparés de λ (gauche), $\lambda/2$ (centre) et $\lambda/4$ (droite), avec $\lambda = 1\,\text{mm}$. Image brute TFM (en haut), enveloppe (en colonne) de l'image TFM (milieu) et par méthode inverse (en bas, avec $\mu_1 = 0.6\mu_{\max}$ et $\mu_2 = 0$).

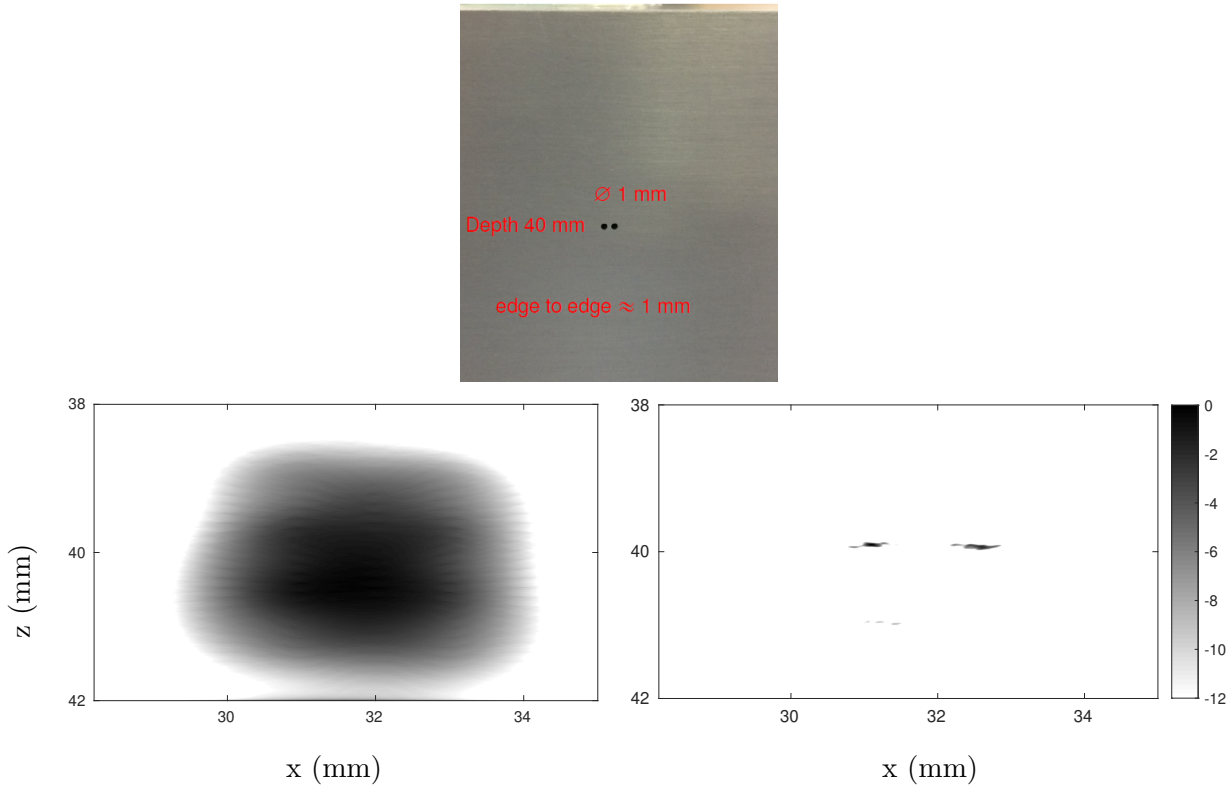


FIGURE 5.16 – Application de l'inversion de données FMC à des données réelles. En haut, photo de la pièce de test contenant deux trous proches et inspectée par la face supérieure. En bas, images reconstruites par TFM (enveloppe en colonne, à gauche) et par méthode inverse (à droite, avec $\mu_1 = 0.6 \left\| \mathbf{H}^T \boldsymbol{y} \right\|$ et $\mu_2 = 10^{-2}$), représentées au niveau des deux trous proches. L'amplitude des images est en échelle logarithmique. [Laroche *et al.*, 2018]

5.5.3 Inversion basée sur les données refocalisées

Si la méthode d'inversion prenant en entrée les données FMC est prometteuse, elle s'avère cependant coûteuse en temps de calcul, en raison de la très grande taille des données. Nous avons donc également envisagé des méthodes de reconstruction d'images prenant en entrée les données refocalisées dans l'image TFM. Après discrétisation de l'équation (5.17), l'image TFM peut s'écrire :

$$\boldsymbol{o}_{\mathrm{TFM}} = \mathbf{B}\boldsymbol{y},$$

où $\boldsymbol{o}_{\mathrm{TFM}}$ représente l'image TFM vectorisée et $\mathbf{B}$ est la matrice binaire de taille $N_x N_z \times {N_{\mathrm{el}}}^2 N_t$ sélectionnant le bon temps de vol dans chaque signal élémentaire. En y injectant le modèle des signaux (5.18), l'image TFM peut alors être reliée à l'image de réflectivité sous la forme :

$$\boldsymbol{o}_{\mathrm{TFM}} = \mathbf{H}_S\boldsymbol{o} + \boldsymbol{\epsilon}, \text{ avec } \mathbf{H}_S \triangleq \mathbf{BH}, \tag{5.19}$$

laquelle définit un nouveau modèle linéaire, avec cette fois un nombre de données réduit. L'idée sous-jacente est de considérer que l'image TFM, qui correspond à une rétroprojection des données brutes $\boldsymbol{y}$ sur la grille de reconstruction, en préserve l'essentiel de l'information, mais concentrée en un nombre de points bien plus faible. L'estimation de l'image de réflectivité s'apparente alors à un problème de type déconvolution, que nous abordons avec la même approche pénalisée qu'au § 5.5.2, en minimisant le critère :

$$J_S(\boldsymbol{o}) = \frac{1}{2}\left\|\boldsymbol{o}_{\mathrm{TFM}} - \mathbf{H}_S\boldsymbol{o}\right\|^2 + \mu_1\left\|\boldsymbol{o}\right\|_1 + \mu_2\left\|\mathbf{D}\boldsymbol{o}\right\|^2,$$

toujours au moyen de l'algorithme FISTA. La matrice $\mathbf{H}_S$ est maintenant carrée et chacune de ses colonnes définit une fonction d'étalement du point (PSF), correspondant à l'image TFM associée à un réflecteur ponctuel au point correspondant de la grille de reconstruction. Cette PSF est variable selon la position spatiale considérée ; la figure 5.17 en montre quelques exemples sur un problème simulé.

Afin de limiter le coût des calculs dans l'optimisation, nous avons alors proposé d'utiliser un modèle d'interpolation de PSF [Gentile *et al.*, 2013, Denis *et al.*, 2015], permettant d'approcher l'opérateur $\mathbf{H}_S$ par une combinaison linéaire de matrices de convolution :

$$\mathbf{H}_S \simeq \sum_{i=1}^{N_{\mathrm{PSF}}} \mathbf{H}_i\mathbf{W}_i, \tag{5.20}$$

où $\mathbf{H}_i$ représente la matrice de convolution spatiale par la PSF au pixel i et $\mathbf{W}_i$ est une matrice diagonale pondérant les différentes contributions. Nous avons utilisé le modèle *Inverse Distance Weighting* [Gentile *et al.*, 2013], pondérant la contribution de la $i^{\text{ème}}$ PSF en un pixel donné j par l'inverse de la distance entre les pixels i et j. Les produits de type $\mathbf{H}_S\cdot$ et $\mathbf{H}_S^T\cdot$ peuvent alors être réalisés à partir de N_{PSF} produits de convolution, lesquels peuvent être effectués à partir d'algorithmes FFT.

Les détails de cette approche peuvent être trouvés dans la référence [Laroche *et al.*, 2019], qui présente également des résultats préliminaires sur des données simulées que nous reproduisons ici. Une image de réflectivité synthétique est générée, comprenant 25 paires de réflecteurs ponctuels séparés de 0.5 mm selon l'axe x, réparties sur l'ensemble de l'image (tous les 5 mm selon les axes x et z). La sonde contient 64 éléments, la longueur d'onde utilisée vaut $\lambda = 1$ mm et les données FMC sont perturbées par un bruit blanc gaussien centré de rapport signal sur bruit 10 dB. Les pixels de l'image reconstruite sont

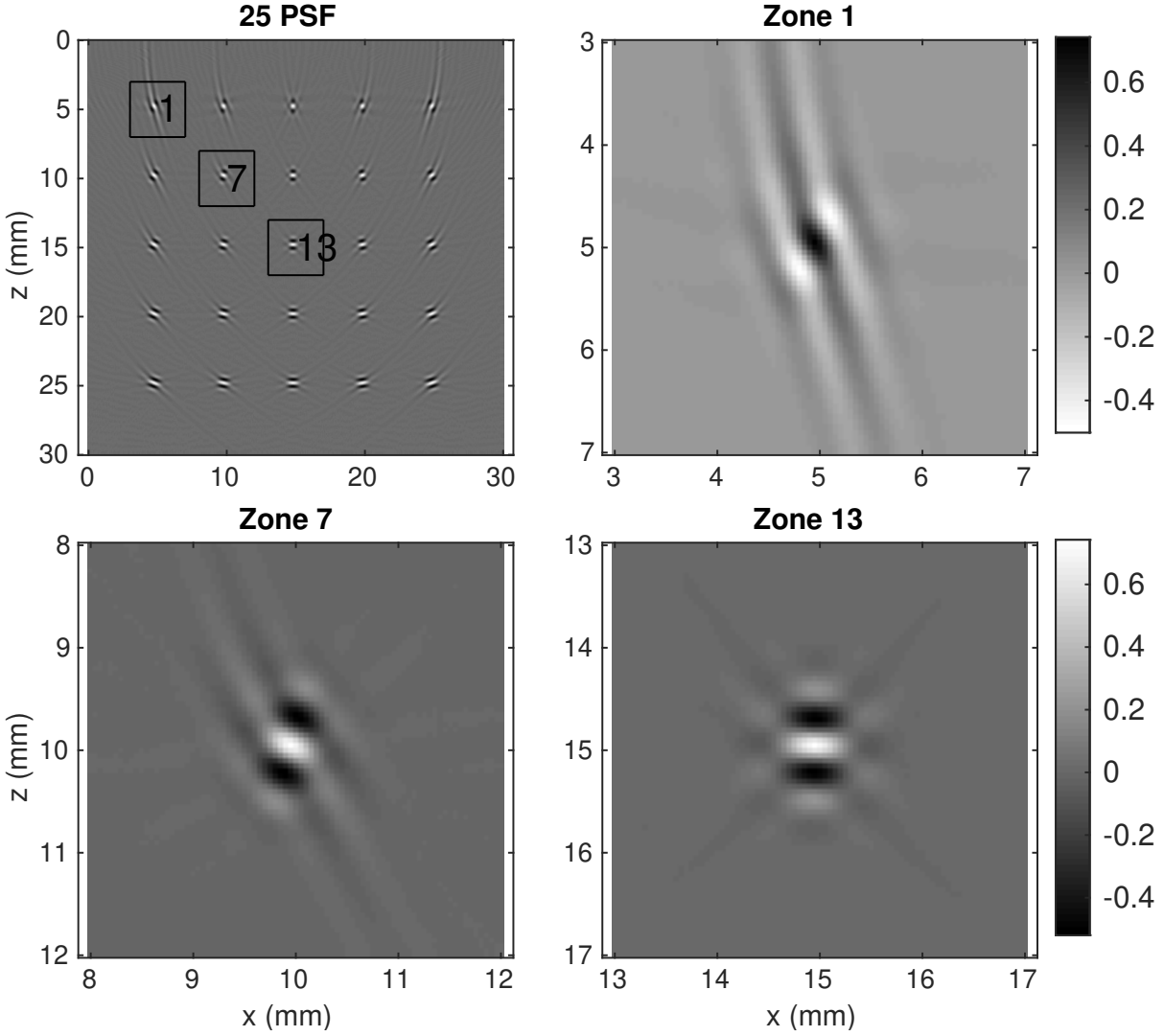


FIGURE 5.17 – Variabilité de la PSF du processus d'imagerie TFM en fonction de la position spatiale (données simulées). En haut à gauche, PSF obtenues en 25 positions de la grille de reconstruction. Les trois autres figures représentent trois cas de PSF dans les zones encadrées sur la figure (a). [Laroche *et al.*, 2019]

carrés, de taille 0.05 mm × 0.05 mm (l'image est alors de taille 251 ×251). L'image TFM obtenue est représentée sur la figure 5.18 à gauche et les profils correspondant à chaque ligne contenant les défauts sont tracés sur la figure 5.18 à droite. L'image TFM ne parvient pas à séparer les deux défauts lorsque la profondeur augmente.

Les profils de réflectivité estimés par notre méthode sont présentés sur la figure 5.18 à droite, en utilisant i) une seule PSF prise au centre de l'image et ii) un modèle interpolant (5.20) avec quatre PSF prises aux positions (x, z) (en mm) : $(10, 10)$, $(10, 20)$, $(20, 10)$ et $(20, 20)$. Les deux profils estimés permettent de résoudre plus de défauts proches que l'image TFM et le modèle interpolant fournit logiquement de meilleurs résultats que le modèle à PSF unique, notamment dans les coins de l'image.

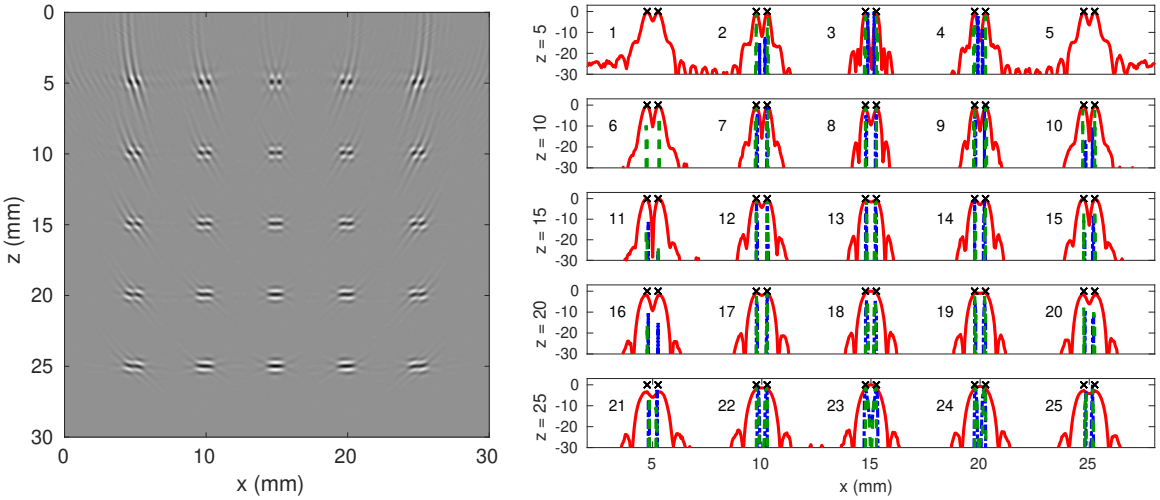


FIGURE 5.18 – Données simulées et résultats d'estimation. À gauche, image TFM d'une pièce synthétique composée de 25 paires de défauts proches. À droite, profils de réflectivité le long des cinq lignes contenant les défauts, correspondant à l'image TFM (rouge), à l'inversion par un modèle à une PSF (bleu) et par un modèle interpolant à quatre PSF (vert). Pour les approches inverses, $\mu_1 = 0.3 \left\| \mathbf{H}_S^T \boldsymbol{o}_{\mathrm{TFM}} \right\|_\infty$ et $\mu_2 = 0$. Les croix noires représentent les vraies positions. [Laroche *et al.*, 2019]

Enfin, les temps de calcul associés, représentés sur la figure 5.19, sont bien inférieurs à ceux de l'approche inversant directement les données FMC brutes du § 5.5.2, avec un surcoût limité pour l'approche utilisant quatre PSF au lieu d'une PSF unique. En particulier, le temps de calcul est invariant lorsque le nombre de transducteurs augmente, laissant entrevoir l'utilité de cette approche dans le cas de sondes à grand nombre d'éléments, dont l'usage devient de plus en plus fréquent en CND.

5.6 Conclusion

Dans mon activité de recherche en CND ultrasonore, j'ai cherché à développer des modèles précis de l'acquisition des données, puis à les exploiter dans des algorithmes d'inversion afin de fournir à l'utilisateur une information de meilleure qualité que celle produite par les données brutes ou les algorithmes classiquement utilisés. La prise en compte de la réponse impulsionnelle des transducteurs ultrasonores, voire de phénomènes de propagation acoustique, conjuguée à des modèles de séquence de réflectivité (en particulier *via* des contraintes de parcimonie), ont permis d'améliorer la résolution dans les échogrammes et

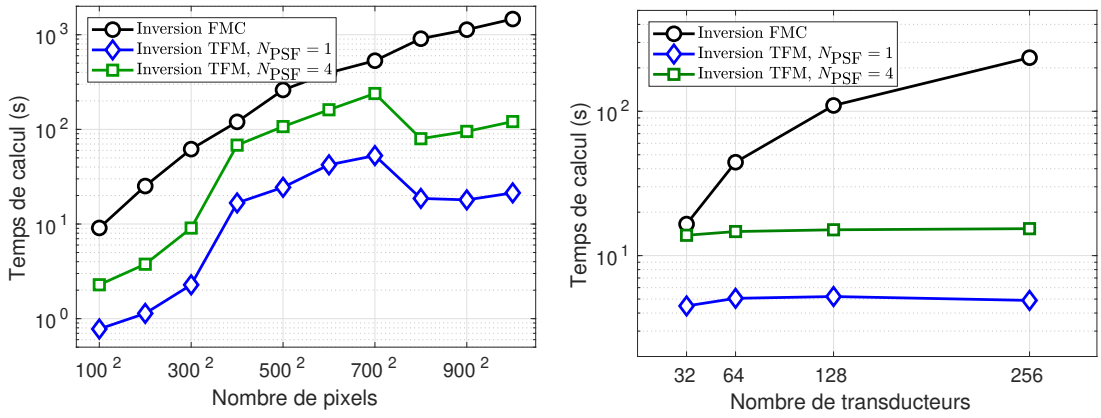


FIGURE 5.19 – Temps de calcul pour l'inversion basée sur les données FMC brutes (○), sur les données rétroprojetées TFM avec un modèle à une PSF (◇) et avec un modèle à quatre PSF (□). À gauche, en fonction du nombre de pixels de la grille de reconstruction (pour $N_{\mathrm{el}} = 64$ transducteurs). À droite, en fonction du nombre de transducteurs (pour une grille de reconstruction de 251×251 pixels). [Laroche *et al.*, 2019]

les images ultrasonores et ainsi de mieux détecter la présence de réflecteurs causés par des hétérogénéités d'impédance acoustique dans le matériau.

Dans le cas de signaux monodimensionnels, ces travaux ont porté sur la discrétisation de modèles convolutifs à haute résolution, la prise en compte de l'atténuation en fréquence et de la dispersion associée, puis sur un cas particulier d'échos régulièrement espacés. Nous avons commencé à étendre ces modèles au cas d'acquisitions multi-éléments, en prenant en compte la réponse impulsionnelle des transducteurs dans des approches de reconstruction d'images, opérant sur l'ensemble des signaux acoustiques bruts ou sur les données rétroprojetées sous la forme d'une image. Une intégration plus poussée des travaux initialement développés sur des signaux monodimensionnels pour la reconstruction d'images est un axe de recherche prioritaire ; nous y reviendrons dans les perspectives du Chapitre 7.

Enfin, si l'ensemble de ces travaux va généralement dans le sens de méthodes de traitement plus complexes que l'état de l'art (perte de structures convolutives dans les modèles, mise en place de méthodes itératives d'inversion parcimonieuse à la place de méthodes linéaires), ils s'accompagnent d'un accroissement significatif du temps de calcul. Afin de permettre leur utilisation dans un contexte de CND industriel, un effort de réflexion supplémentaire devrait être mené sur la mise en œuvre numérique. L'utilisation de GPU me semble particulièrement appropriée pour la reconstruction d'images, où de nombreuses opérations peuvent être réalisées en parallèle.

5.7 Références

[Beck et Teboulle, 2009] BECK, A. et TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202.

[Bourguignon *et al.*, 2011] BOURGUIGNON, S., SOUSSEN, C., CARFANTAN, H. et IDIER, J. (2011). Sparse deconvolution : comparison of statistical and deterministic approaches. *In IEEE Workshop on Statistical Signal Processing*, pages 317–320, Nice, France.

[Carcreff, 2014a] Carcreff, E. (2014a). *Déconvolution adaptative de signaux ultrasonores pour le contrôle non destructif par ultrasons*. Thèse de doctorat, Université du Maine.

[Carcreff, 2014b] Carcreff, E. (2014b). *Déconvolution adaptative de signaux ultrasonores pour le contrôle non destructif par ultrasons*. Thèse de doctorat, Université du Maine.

[Carcreff *et al.*, 2013a] Carcreff, E., Bourguignon, S., Idier, J. et Simon, L. (2013a). Algorithmes de déconvolution impulsionnelle à résolution augmentée. *In Actes du 24e colloque GRETSI*, Brest, France.

[Carcreff *et al.*, 2013b] Carcreff, E., Bourguignon, S., Idier, J. et Simon, L. (2013b). Resolution enhancement of ultrasonic signals by up-sampled sparse deconvolution. *In Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, Vancouver, Canada.

[Carcreff *et al.*, 2014] Carcreff, E., Bourguignon, S., Idier, J. et Simon., L. (2014). A linear model approach for ultrasonic inverse problems with attenuation and dispersion. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 61(7):1191–1203.

[Cheng *et al.*, 1996] Cheng, Q., Chen, R. et Li, T.-H. (1996). Simultaneous wavelet estimation and deconvolution of reflection seismic signals. *IEEE Transactions on Geoscience and Remote Sensing*, 34(2):377–384.

[Chi, 1987] Chi, C.-Y. (1987). A fast maximum likelihood estimation and detection algorithm for Bernoulli-Gaussian processes. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(11):1636–1639.

[de Raad et Dijkstra, 1998] de Raad, J. A. et Dijkstra, F. H. D. (1998). Mechanised UT on girth welds during pipeline construction : A mature alternative to radiography. *Insight*.

[Demirli et Saniie, 2001] Demirli, R. et Saniie, J. (2001). Model-based estimation of ultrasonic echoes. Part II : Nondestructive evaluation applications. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 48(3):803–811.

[Denis *et al.*, 2015] Denis, L., Thiébaut, E., Soulez, F., Becker, J. et Mourya., R. (2015). Fast approximations of shift-variant blur. *International Journal of Computer Vision*, 3:253–278.

[Fink et Cardoso, 1984] Fink, M. et Cardoso, J.-F. (1984). Diffraction effects in pulse-echo measurement. *IEEE Transactions on Sonics and Ultrasonics*, 31(4):313–329.

[Gentile *et al.*, 2013] Gentile, M., Courbin, F. et Meylan, G. (2013). Interpolating point spread function anisotropy. *Astronomy and Astrophysics*, 549:A1.

[Golub et Van Loan, 1996] Golub, G. et Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press.

[Goussard *et al.*, 1990] Goussard, Y., Demoment, G. et Idier, J. (1990). A new algorithm for iterative deconvolution of sparse spike trains. *In IEEE International Conference on Acoustic, Speech and Signal Processing*, pages 1547–1550, Albuquerque, NM, USA.

[Gurumurthy et Arthur, 1982] Gurumurthy, K. et Arthur, R. (1982). A dispersive model for the propagation of ultrasound in soft tissue. *Ultrasonic Imaging*, 4(4):355–377.

[Holmes *et al.*, 2005] Holmes, C., Drinkwater, B. W. et Wilcox, P. D. (2005). Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation. *NDT & E International*, 38(8):701 – 711.

[Kaaresen, 1997] Kaaresen, K. (1997). Deconvolution of sparse spike trains by iterated window maximization. *IEEE Transactions on Signal Processing*, 45(5):1173–1183.

[Kak et Dines, 1978] Kak, A. C. et Dines, K. A. (1978). Signal processing of broad-band pulsed ultrasound : Measurement of attenuation of soft biological tissues. *IEEE Transactions on Biomedical Engineering*, 25(4):321–344.

[Kormylo et Mendel, 1982] Kormylo, J. et Mendel, J. (1982). Maximum likelihood detection and estimation of bernoulli - gaussian processes. *IEEE Transactions on Information Theory*, 28(3):482–488.

[Krautkramer et Krautkramer, 1990] Krautkramer, J. et Krautkramer, H. (1990). *Ultrasonic Testing of materials*. Springer-Verlag, Berlin.

[Kuc, 1983] Kuc, R. (1983). Generating a minimum-phase digital filter model for the acoustic attenuation of soft tissue. *In Ultrasonics Symposium*, pages 794–796.

[Kuc, 1984] Kuc, R. (1984). Modeling acoustic attenuation of soft tissue with a minimum-phase filter. *Ultrasonic Imaging*, 6(1):24–36.

[Laroche *et al.*, 2019] Laroche, N., Bourguignon, S., Carcreff, E., Idier, J. et Duclos, A. (2019). Approche inverse rapide pour la déconvolution d'images ultrasonores par une PSF variable. *In Actes du 27^e colloque GRETSI*.

[Laroche *et al.*, 2018] Laroche, N., Carcreff, E., Bourguignon, S., Idier, J. et Duclos, A. (2018). An inverse approach for ultrasonic imaging by total focusing point for close reflectors separation. *In IEEE International Ultrasonics Symposium (IUS)*, Kobe, Japan.

[Lavielle, 1993] Lavielle, M. (1993). Bayesian deconvolution of bernoulli-gaussian processes. *Signal Processing*, 33(1):67 – 79.

[Mendel, 1983] Mendel, J. M. (1983). *Optimal Seismic Deconvolution*. Academic Press.

[Narayana et Ophir, 1983] Narayana, P. A. et Ophir, J. (1983). A closed form method for the measurement of attenuation in nonlinearly dispersive media. *Ultrasonic Imaging*, 5:117–21.

[Nocedal et Wright, 2006] Nocedal, J. et Wright, S. (2006). *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer New York.

[O'Brien *et al.*, 1994] O'Brien, M. S., Sinclair, A. N. et Kramer, S. M. (1994). Recovery of a sparse spike time series by L1 norm deconvolution. *IEEE Transactions on Signal Processing*, 42:3353–3365.

[O'Donnell *et al.*, 1978] O'Donnell, M., Jaynes, E. T. et Miller, J. G. (1978). General relationships between ultrasonic attenuation and dispersion. *The Journal of the Acoustical Society of America*, 63(6):1935–1937.

[Olofsson et Stepinski, 2001] Olofsson, T. et Stepinski, T. (2001). Minimum entropy deconvolution of pulse-echo signals acquired from attenuative layered media. *The Journal of the Acoustical Society of America*, 109(6):2831–2839.

[Ophir et Jaeger, 1982] Ophir, J. et Jaeger, P. (1982). Spectral shifts of ultrasonic propagation trough media with nonlinear dispersive attenuation. *Ultrasonic imaging*, 4:282–289.

[Oppenheim et Schafer, 1989] Oppenheim, A. et Schafer, R. (1989). *Discrete-time signal processing*. Prentice-Hall signal processing series. Prentice Hall.

[Quaegebeur et Masson, 2012] Quaegebeur, N. et Masson, P. (2012). Correlation-based imaging technique using ultrasonic transmit-receive array for non-destructive evaluation. *Ultrasonics*, 52(8):1056 – 1064.

[Quazi, 1981] Quazi, A. (1981). An overview on the time delay estimate in active and passive systems for target localization. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(3):527–533.

[Saniie et Nagle, 1989] Saniie, J. et Nagle, D. (1989). Pattern recognition in the ultrasonic imaging of reverberant multilayered structures. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 36(1):80–92.

[Selesnick et Bayram, 2014] Selesnick, I. W. et Bayram, I. (2014). Sparse signal estimation by maximally sparse convex optimization. *IEEE Transactions on Signal Processing*, 62(5):1078–1092.

[Selfridge, 1985] Selfridge, A. R. (1985). Approximate material properties in isotropic materials. *IEEE nTransactions on Sonics and Ultrasonics*, 32(3):381–394.

[Soussen *et al.*, 2011] Soussen, C., Idier, J., Brie, D. et Duan, J. (2011). From Bernoulli Gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584.

[Sutcliffe *et al.*, 2012] Sutcliffe, M., Weston, M., Dutton, B., Charlton, P. et Donne, K. (2012). Real-time full matrix capture for ultrasonic non-destructive testing with acceleration of post-processing through graphic hardware. *NDT & E International*, 51:16 – 23.

[Van Rossum, 2001] Van Rossum, M. C. W. (2001). A novel spike distance. *Neural Computation*, 13(4):751–763.

[Zala, 1992] Zala, C. (1992). High-resolution inversion of ultrasonic traces. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 39(4):458–463.

Chapitre 6

Optimisation exacte en norme ℓ_0 pour les problèmes inverses parcimonieux

Depuis une vingtaine d'années et l'article fondateur de Chen, Donoho et Saunders [Chen *et al.*, 1998], la thématique de l'approximation parcimonieuse a donné lieu à une très grande quantité de travaux en optimisation, où les algorithmes reposent soit sur la formulation d'un problème d'optimisation convexe utilisant une norme ℓ_1, soit sur une stratégie gloutonne. Ces approches ont été décrites au Chapitre 3, où nous avons également évoqué leurs limites à résoudre le problème en norme ℓ_0. Plus récemment, cependant, l'optimisation de fonctions de coût impliquant directement la norme ℓ_0 a connu un intérêt croissant. L'algorithme *Iterative Hard Thresholding* ou IHT [Herrity *et al.*, 2006, Blumensath et Davies, 2008], des algorithmes gloutons plus complexes objectivant une fonction ℓ_0 [Miller, 2002, Soussen *et al.*, 2011] ou couplés avec des stratégies de parcours d'arbre [Karahanoglu *et al.*, 2013], en sont quelques témoins. D'autres travaux ont étudié l'utilisation de fonctions continues et non-convexes, « intermédiaires entre la norme ℓ_1 et la norme ℓ_0 », produisant des algorithmes de type *Graduated Non-Convexity* [Mohimani *et al.*, 2009], *penalty decomposition* [Lu et Zhang, 2013], DC (*difference of convex functions*) programming [Le Thi *et al.*, 2015] ou encore PALM (Proximal Alternating Linearized Minimization, [Bolte *et al.*, 2014]). Si ces approches montrent souvent de meilleures solutions en pratique que les algorithmes en norme ℓ_1 ou les algorithmes gloutons basiques, les conditions de convergence vers le minimiseur ℓ_0 sont, au mieux, locales [1]. Dans [Soubies *et al.*, 2015], une relaxation *exacte* de la norme ℓ_0 (*i.e.,* produisant le même ensemble de minimiseurs) a été construite, mais aucun algorithme d'optimisation garanti n'a été proposé.

À l'inverse, des méthodes d'optimisation globale ont été proposées en recherche opérationnelle et en optimisation discrète pour le problème de moindres carrés sous contraintes de cardinalité :

$$\mathcal{P}_{2/0} : \min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \mathbf{A}\boldsymbol{x}\|^2 \text{ s.c. } \|\boldsymbol{x}\|_0 \leq K, \tag{6.1}$$

notamment pour des problèmes d'optimisation de portefeuille, où la contrainte de cardinalité impose un nombre maximum de placements. Les premiers travaux remontent, à ma connaissance, à un article de D. Bienstock en 1996 [Bienstock, 1996] suivi, une dizaine d'années plus tard, par un certain nombre de contributions [Li *et al.*, 2006, Jokar et

1. Des garanties d'optimalité globale ont été établies pour IHT [Herrity *et al.*, 2006, Blumensath et Davies, 2008], mais les conditions d'application sont extrêmement restrictives.

Pfetsch, 2008, Bertsimas et Shioda, 2009, Karahanoglu *et al.*, 2013, Burdakov *et al.*, 2016]. Ces travaux s'appuient sur une reformulation du problème $\mathcal{P}_{2/0}$ en un problème d'optimisation en nombres mixtes (MIP, *mixed-integer program*), mêlant variables continues et variables discrètes, pour lequel des algorithmes de résolution exacte ont été développés, reposant sur le principe des méthodes *branch-and-bound*.

Je me suis donc intéressé à ces formulations et à ces algorithmes, motivé par le fait que certains problèmes inverses rencontrés en traitement du signal restent de taille modérée mais suffisamment complexes pour légitimer la recherche de la *vraie* solution ℓ_0. C'est par exemple le cas de problèmes de déconvolution de signaux monodimensionnels ou du démélange spectral, que nous retrouverons au long de ce chapitre, qui restent limités par la taille du dictionnaire et le nombre d'inconnues non nulles recherchées.

Dans un premier temps, j'ai étudié la reformulation des trois problèmes d'approximation parcimonieuse introduits au Chapitre 3, à savoir

$$\begin{aligned}
\mathcal{P}_{2/0} &: \min_{\boldsymbol{x}\in\mathbb{R}^P} \frac{1}{2}\left\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\right\|_2^2 \quad \text{s. c.} \quad \left\|\boldsymbol{x}\right\|_0 \leq K, \\
\mathcal{P}_{0/2} &: \min_{\boldsymbol{x}\in\mathbb{R}^P} \left\|\boldsymbol{x}\right\|_0 \quad \text{s. c.} \quad \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2 \leq \alpha \\
\text{et } \mathcal{P}_{2+0} &: \min_{\boldsymbol{x}\in\mathbb{R}^P} \frac{1}{2}\left\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\right\|_2^2 + \mu\left\|\boldsymbol{x}\right\|_0,
\end{aligned}$$

sous la forme de MIP, et la capacité d'un solveur commercial à les résoudre de manière exacte. Nous montrons en particulier que la solution en norme ℓ_0 peut être calculée pour des problèmes de complexité limitée, et que dans ce cas elle fournit de meilleurs estimateurs de la séquence parcimonieuse recherchée. Ces travaux sont présentés en Section 6.1.

Ces résultats prometteurs m'ont alors incité à étudier les méthodes de résolution afin, d'une part, de s'affranchir de l'utilisation d'un logiciel et, surtout, de tirer parti des spécificités des problèmes en norme ℓ_0 par rapport à un MIP quelconque dans le but de réduire le temps de calcul et d'aborder des problèmes de plus grande dimension. C'est ici le cœur du projet ANR MIMOSA en cours et de la thèse de Ramzi Ben Mhenni, dont nous présentons les premiers résultats en Section 6.2 : la construction d'algorithmes branch-and-bound spécifiquement construits pour les trois problèmes en norme ℓ_0 permet de rivaliser avec, voire de largement surpasser, les performances d'un solveur MIP commercial.

En parallèle de ces développements algorithmiques, je suis également revenu vers l'imagerie hyperspectrale, en abordant le problème de démélange spectral parcimonieux [Iordache *et al.*, 2011, Greer, 2012, Drumetz *et al.*, 2019] avec une approche ℓ_0. Ce problème s'y prête en effet particulièrement bien, par sa complexité limitée (le nombre de composantes non nulles excède rarement quelques unités) mais également pour des raisons techniques, les contraintes de borne sur les inconnues facilitant la reformulation MIP. Nous verrons également que cette approche permet de prendre en compte de manière exacte des contraintes logiques visant à structurer l'espace des solutions, là où les approches classiques ne peuvent proposer que des relaxations inexactes. La Section 6.3 est consacrée à cette application.

Quelques éléments de conclusion seront enfin proposés en Section 6.4.

6.1 Reformulations MIP et résolution par solveur générique

L'optimisation de problèmes mêlant des variables continues et des variables entières est un problème récurrent en recherche opérationnelle, où ces problèmes sont connus sous le nom de programmes en nombres mixtes (MIPs, *mixed-integer programs*). La résolution de MIP a fait l'objet d'une quantité considérable de travaux et il est communément reconnu qu'en quinze ans, un gain d'un facteur 10^9 a été obtenu sur le temps de calcul, dû aussi bien à l'augmentation des capacités de calcul des processeurs, aux progrès réalisés dans l'optimisation de programmes linéaires et dans la mise en œuvre efficace de techniques mathématiques avancées et à la programmation informatique [Bixby, 2012]. Les méthodes de résolution s'appuient généralement sur le principe des algorithmes de séparation et évaluation *branch-and-bound*. Plusieurs solveurs existent maintenant, libres ou commerciaux [2], permettant l'optimisation globale de ce genre de problèmes, en particulier dans le cas de fonctions de coût linéaires ou quadratiques, sous des contraintes linéaires ou quadratiques [Belotti *et al.*, 2013]. Le principe de résolution sera abordé en Section 6.2. Dans cette section, nous abordons la reformulation des trois problèmes parcimonieux $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$ sous forme de MIP. Nous étudions ensuite les performances d'un solveur commercial, CPLEX, pour la résolution de problèmes de déconvolution et analysons la qualité des solutions obtenues.

6.1.1 Introduction de variables binaires et reformulations MIP

La reformulation de problèmes en norme ℓ_0 sous forme de MIP s'appuie sur une réécriture de la norme ℓ_0 à partir de variables de décision binaires $b_p \in \{0, 1\}$, $p = 1, \ldots, P$, encodant la nullité des composantes de $\boldsymbol{x}$:

$$b_p = 0 \Leftrightarrow x_p = 0. \tag{6.2}$$

La norme ℓ_0 s'écrit alors comme une simple somme : $\|\boldsymbol{x}\|_0 = \sum_p b_p$, formulant une partie linéaire dans la fonction de coût pour $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$ et une contrainte linéaire pour $\mathcal{P}_{2/0}$.

Afin de traduire cette condition en contraintes mathématiques, un choix classique [Jokar et Pfetsch, 2008] consiste à rajouter une hypothèse de borne sur la solution recherchée : $\forall p, |x_p| < M$ pour une valeur de M suffisamment grande. Dans ce cas, il est facile de voir que la condition (6.2) s'écrit :

$$-Mb_p \leq x_p \leq Mb_p, \tag{6.3}$$

qui représente un système d'inégalités linéaires. Cette reformulation, connue sous le nom de *bigM*, s'étend sans difficulté au cas de bornes inférieures (négatives) et supérieures (positives) différentes, et éventuellement spécifiques à chaque composante : $-M_p^{\text{inf}} \leq x_p \leq M_p^{\text{sup}}$, avec $M_p^{\text{inf}}, M_p^{\text{sup}} \geq 0$. Sauf cas particuliers, nous utiliserons une seule borne M dans la suite de ce chapitre par souci de simplicité. Si le problème initial n'est pas naturellement borné, la définition de ces bornes est un problème qui reste ouvert, et qui impacte l'efficacité de la résolution.

Nous avons ainsi reformulé les trois problèmes $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$ sous forme de MIP, récapitulés dans le tableau 6.1. Les détails menant à ces reformulations peuvent être trouvés dans [Bourguignon *et al.*, 2016]. Notons que $\mathcal{P}_{2/0}$ et $\mathcal{P}_{2+0}$ formulent des MIPs

2. Parmi les solveurs commerciaux, on peut notamment citer BARON, CPLEX, GUROBI, MOSEK, Xpress-MP, dont certains sont libres d'accès pour les usages académiques. Même Matlab propose désormais un solveur MIP, qui s'est révélé particulièrement inefficace... COUENNE et GLPK sont eux proposés en accès libre.

Problème	Reformulation MIP
$\mathcal{P}_{2/0} : \min_{\boldsymbol{x}\in\mathbb{R}^P} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\|^2 \text{ s.c. } \begin{cases} \|\boldsymbol{x}\|_0 \leq K \\ \|\boldsymbol{x}\|_\infty \leq M \end{cases}$	$\min_{\boldsymbol{x}\in\mathbb{R}^P, \boldsymbol{b}\in\{0,1\}^P} \frac{1}{2}\boldsymbol{x}^T\mathbf{A}^T\mathbf{A}\boldsymbol{x} - \boldsymbol{y}^T\mathbf{A}\boldsymbol{x}$ s.c. $\begin{cases} \mathbf{1}_P^T\boldsymbol{b} \leq K \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$
$\mathcal{P}_{0/2} : \min_{\boldsymbol{x}\in\mathbb{R}^P} \|\boldsymbol{x}\|_0 \text{ s.c. } \begin{cases} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\|^2 \leq \alpha \\ \|\boldsymbol{x}\|_\infty \leq M \end{cases}$	$\min_{\boldsymbol{x}\in\mathbb{R}^P, \boldsymbol{b}\in\{0,1\}^P} \mathbf{1}_P^T\boldsymbol{b}$ s.c. $\begin{cases} \frac{1}{2}\boldsymbol{x}^T\mathbf{A}^T\mathbf{A}\boldsymbol{x} - \boldsymbol{y}^T\mathbf{A}\boldsymbol{x} \leq alpha - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$
$\mathcal{P}_{2+0} : \min_{\boldsymbol{x}\in\mathbb{R}^P} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\|^2 + \mu\|\boldsymbol{x}\|_0 \text{ s.c. } \|\boldsymbol{x}\|_\infty \leq M$	$\min_{\boldsymbol{x}\in\mathbb{R}^P, \boldsymbol{b}\in\{0,1\}^P} \frac{1}{2}\boldsymbol{x}^T\mathbf{A}^T\mathbf{A}\boldsymbol{x} - \boldsymbol{y}^T\mathbf{A}\boldsymbol{x} + \mu\mathbf{1}_P^T\boldsymbol{b}$ s.c. $-M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b}$

TABLE 6.1 – Reformulations MIP de trois problèmes d'approximation parcimonieuse. La notation $\mathbf{1}_P$ désigne un vecteur colonne de taille P composé de 1.

quadratiques (MIQP, *Mixed Integer Quadratic Program*, à fonction de coût quadratique et contraintes linéaires), quand $\mathcal{P}_{0/2}$ formule un MIP à contraintes quadratiques et à fonction de coût linéaire (QCMIP, *Quadratically Constrained Mixed Integer Program*). Cette différentiation prend du sens au niveau de la résolution : les méthodes mises en œuvre dans les solveurs ne sont pas les mêmes, et nous verrons au paragraphe suivant que la classe de problème (MIQP ou QCMIP) impacte fortement l'efficacité de la résolution.

Nous avons également proposé des reformulations similaires pour des problèmes où l'ajustement du modèle est fait non plus au sens de la norme ℓ_2 mais avec une norme ℓ_1 ou une norme ℓ_∞ : ces normes étant linéaires par morceaux, les reformulations font alors apparaître des termes linéaires supplémentaires dans la fonction objectif ou les contraintes. Bien que moins fréquentes dans la littérature, ces normes peuvent trouver un intérêt dans le cas d'un modèle d'erreur sortant du cas gaussien représenté par l'ajustement ℓ_2 : la norme ℓ_1 traduit un ajustement plus robuste aux données aberrantes [Marks *et al.*, 1978, O'Brien *et al.*, 1994] et, à l'inverse, la norme ℓ_∞ traduit une hypothèse de bruit de distribution uniforme [Clason, 2012]. Les reformulations MIP correspondantes et des simulations légitimant l'usage de tels ajustements peuvent également être trouvées dans [Bourguignon *et al.*, 2016].

6.1.2 Résultats sur des problèmes de déconvolution

Nous avons évalué la capacité d'un solveur MIP à résoudre les formulations précédentes, puis analysé la qualité des solutions obtenues, en application à des problèmes inverses de déconvolution parcimonieuse. Un exemple est représenté en figure 6.1a) : des séquences parcimonieuses de 100 points sont générées aléatoirement et convoluées, puis bruitées.

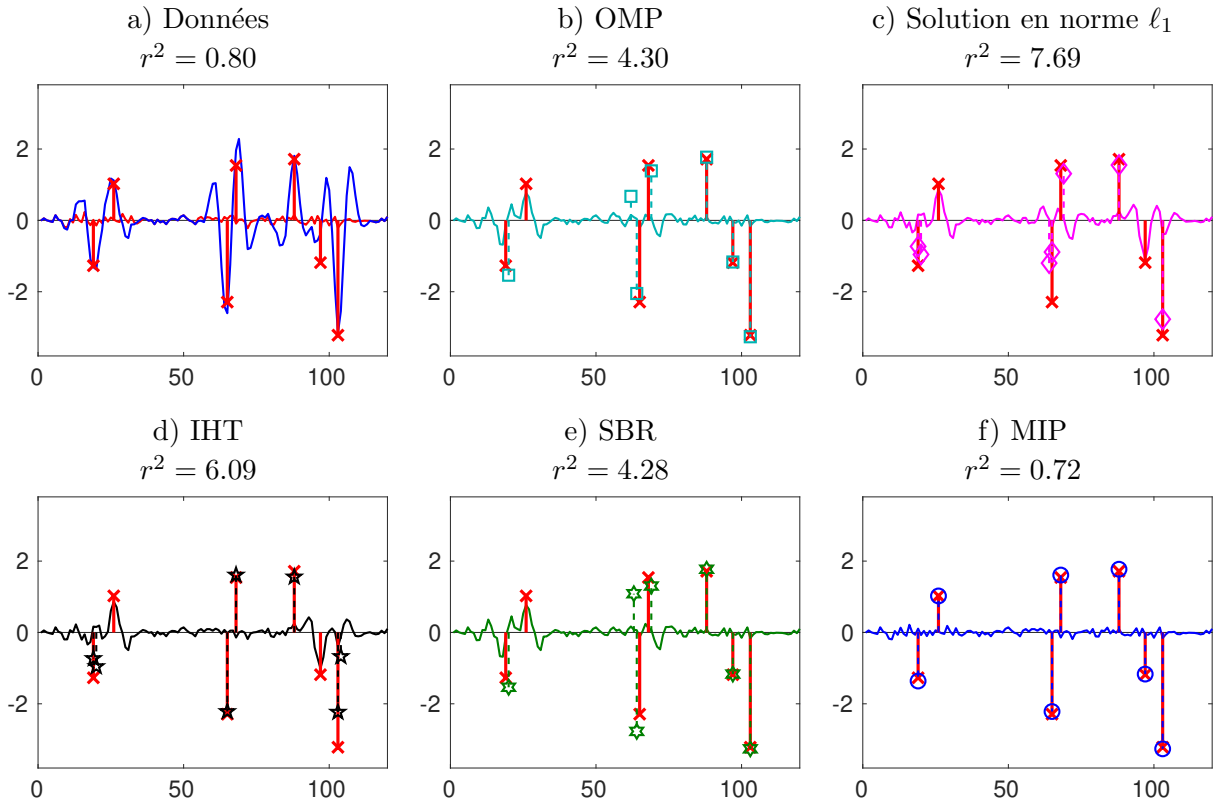


FIGURE 6.1 – Exemple de problème test de déconvolution parcimonieuse et résultats de plusieurs algorithmes. a) : vraie séquence parcimonieuse composée de $K = 7$ pics ($\times$) et données ($-$). Le trait fin rouge représente le bruit (RSB = 20 dB). b) à f) : estimation fournie par différents algorithmes, réglés de manière à fournir 7 composantes. Le trait fin représente le résidu d'estimation et la valeur r^2 correspond au carré de sa norme (de la norme du bruit pour les données en a)).

L'ensemble des simulations a été généré dans les conditions suivantes :

- les K positions des variables non nulles sont tirées aléatoirement, K variant entre 5 et 11 ;
- les amplitudes associées ont été tirées aléatoirement sous la forme $u + \text{sgn}(u)$, où u soit une loi normale centrée de variance unité. Rajouter ± 1 permet d'éviter de générer des amplitudes arbitrairement faibles, qui seraient trop difficiles à détecter en présence de bruit ;
- le bruit est additif, blanc, gaussien, centré et de variance σ^2 telle que le rapport signal sur bruit $\text{RSB}_{\text{dB}} = 10 \log \frac{\|\mathbf{A}\boldsymbol{x}\|^2}{N\sigma^2}$ égale 30, 20, ou 10 dB. Des instances sans bruit sont également générées.

Pour chaque problème, 50 instances aléatoires sont générées. L'ensemble des problèmes tests et les résultats d'estimation ont été publiés en matériel multimédia associé à l'article [Bourguignon *et al.*, 2016], et sont également disponibles sur ma page personnelle [3]. Malgré leur faible taille, ces problèmes sont difficiles en raison du fort chevauchement entre les échos. Le paramètre M est réglé à $M = 1.1\, x^1_{\text{max}}$, où $x^1_{\text{max}} = \left\|\mathbf{A}^T\boldsymbol{y}\right\|_\infty$ cor-

3. `http://pagesperso.ls2n.fr/~bourguignon-s/download_MIP.html`

respond à l'amplitude maximale des solutions à une composante non-nulle estimées par moindres carrés [4] (sans perte de généralité, nous supposons ici que les colonnes du dictionnaire $\mathbf{A}$ sont de norme unité). Si une composante p de l'optimum atteint la contrainte de borne $x_p = Mb_p$, alors l'optimisation est relancée avec une valeur de M plus grande ($M \leftarrow 1.1M$), et ce jusqu'à ce que ce ne soit plus le cas. Les temps de calcul donnés ici incluent le surcoût éventuel.

Nous avons mis en œuvre la résolution des problèmes MIP de la table 6.1 avec le solveur CPLEX (version 12.6), qui est unanimement reconnu comme un des meilleurs logiciels de résolution de MIP existants. C'est un logiciel commercial, mais utilisable librement à des fins académiques [5]. Les fonctions d'interfaçage Matlab/CPLEX sont également disponibles sur ma page web. L'optimisation est exécutée sur un ordinateur de bureau composé de huit processeurs cadencés à 3 GHz

La figure 6.1 montre un exemple de solutions obtenues par la résolution de MIP (sur la formulation $\mathcal{P}_{2/0}$) et par quelques algorithmes classiques d'estimation parcimonieuse : l'algorithme glouton *Orthogonal Matching Pursuit* (OMP), la solution du problème en norme ℓ_1 calculée ici par une méthode homotopique [Donoho et Tsaig, 2008], et deux algorithmes plus « orientés ℓ_0 » : *Iterative Hard Thresholding* ou IHT [Blumensath et Davies, 2008] et *Single Best Replacement* ou SBR [Soussen *et al.*, 2011]. Tous ces algorithmes sont réglés pour fournir le bon nombre de composantes non nulles et le rapport signal sur bruit est de 20 dB. Les algorithmes standard d'estimation parcimonieuse échouent tous à retrouver la bonne solution et seule la solution exacte en norme ℓ_0 fournit le bon support.

La figure 6.2 à gauche représente le taux d'identification correcte du support (la proportion de problèmes pour lesquels l'algorithme a localisé parfaitement les K composantes), sur les 50 instances de chaque problème, en fonction de K, pour ces mêmes algorithmes et où la résolution MIP est successivement limitée en temps de calcul à 1, 10, 100 et 1 000 secondes. Sans surprise, la recherche de l'optimum global du problème ℓ_0 fournit de loin les meilleures solutions, même si les performances se dégradent lorsque la complexité du problème augmente : en présence de bruit, même l'optimum global en norme ℓ_0 ne garantit pas de fournir la « vraie » séquence ayant généré les données. Il est également intéressant de noter que les solutions intermédiaires, obtenues en limitant le temps de calcul, sont aussi plus performantes que les solutions des algorithmes standard.

Au niveau du temps de calcul, nous avons montré dans [Bourguignon *et al.*, 2016] que la résolution exacte est possible, dans la mesure où la parcimonie est suffisamment forte, et est aussi d'autant plus efficace que le bruit est faible. Si la complexité combinatoire du problème dépend clairement du nombre de composantes non nulles recherchées, on comprend aussi l'impact du niveau de bruit en analysant la structure des algorithmes de résolution (la Section 6.2 leur sera consacrée) : plus les données sont bruitées, moins les différentes combinaisons parcimonieuses peuvent être discriminées entre elles et moins de branches pourront être élaguées rapidement dans la procédure de branch-and-bound. La figure 6.3 à gauche montre ainsi les temps de calcul de la résolution MIP sur l'ensemble des problèmes $\mathcal{P}_{2/0}$, pour un rapport signal sur bruit de 20 dB. Sans surprise non plus, ceux-ci sont bien plus élevés que pour les approches classiques, où le temps de calcul est de l'ordre de 10^{-2} s sur ces « petits » problèmes. Si l'ensemble des problèmes est résolu en moins d'1 s pour $K = 5$, les temps de calcul sont d'autant plus dispersés que la complexité augmente. Ainsi, pour $K = 5$ (respectivement $K = 7$), le temps varie de 0.2 s à 6 s (respectivement de 0.5 s à 491 s). Pour $K = 11$, une dizaine d'instances n'ont pas

4. On pourra se référer à la description de l'algorithme OMP pour la justification de cette propriété, voir page 54.

5. `https://www.ibm.com/products/ilog-cplex-optimization-studio`

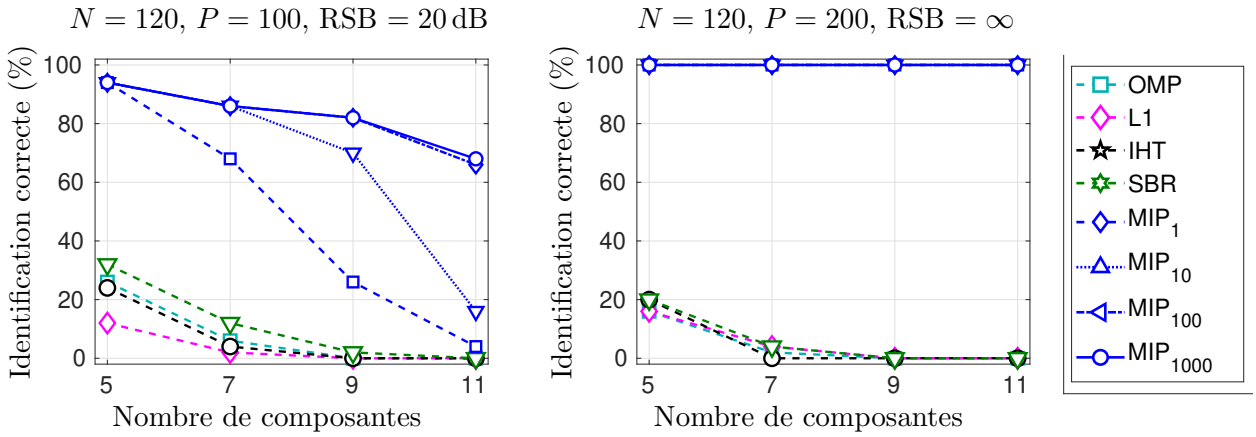


FIGURE 6.2 – Comparaison d'algorithmes sur des problèmes de déconvolution parcimonieuse (résultats moyennés sur 50 instances), avec $N = 120$ données : taux d'identification correcte du support en fonction de K. À gauche, $P = 100$ inconnues et rapport signal sur bruit de 20 dB. À droite, $P = 200$ inconnues et rapport signal sur bruit infini. Le temps maximum pour la résolution MIP (problème $\mathcal{P}_{2/0}$) est successivement limité à 1, 10, 100 et 1 000 s (voir la légende sur la figure).

abouti en moins de 1 000 s, reflétant la nature NP-difficile du problème. Enfin, si ces temps de calcul semblent élevés, ils doivent aussi être mis en rapport avec le coût de l'évaluation exhaustive de toutes les solutions (*i.e.,* la résolution de C_{100}^K problèmes de moindres carrés en dimension K), que nous avons estimé à environ 1 500 s pour $K = 5$, 4 jours pour $K = 7$ et plus d'un an pour $K = 9$.

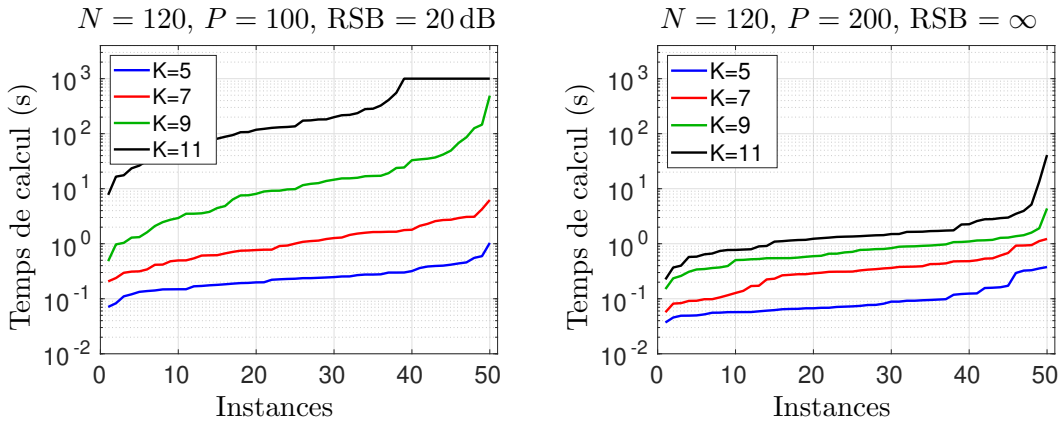


FIGURE 6.3 – Temps de calcul triés par ordre de croissant pour la résolution MIP des 50 instances du problème $\mathcal{P}_{2/0}$, pour K de 5 à 11. À gauche, $P = 100$ inconnues et rapport signal sur bruit de 20 dB. À droite, $P = 200$ inconnues et rapport signal sur bruit infini.

Nous avons également évalué les performances de l'approche MIP sur des problèmes non bruités, dits de *représentation* parcimonieuse. Afin de résoudre des problèmes non-triviaux avec $P > N$ (sinon, en l'absence de bruit, la solution des moindres carrés fournit l'optimum !), nous avons considéré un modèle convolutif sur-échantillonné d'un facteur deux par rapport aux données, comme proposé en Section 5.2. Le problème revient alors

à chercher la solution la plus parcimonieuse dans un système d'équations $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$ sous-déterminé (200 inconnues pour 120 données), qui est un cas particulier de la formulation $\mathcal{P}_{0/2}$ avec $\alpha = 0$[6]. La figure 6.2 à droite confirme un résultat attendu : sans bruit, la vraie séquence est la solution du problème ℓ_0 ; l'optimisation de $\mathcal{P}_{0/2}$ permet donc d'identifier parfaitement le support. C'est loin d'être le cas des autres méthodes, qui n'identifient quasiment jamais l'ensemble des composantes. La figure 6.3 à droite montre également que l'ensemble des problèmes est résolu de manière garantie en un temps bien plus faible que dans le cas bruité (au maximum 0.35 s, 1.2 s, 4.4 s et 40 s respectivement pour $K = 5$, 7, 9 et 11).

Enfin, nous avons mis en évidence dans [Bourguignon *et al.*, 2016] que le temps de calcul dépendait fortement de la formulation abordée : si les formulations $\mathcal{P}_{2/0}$ et $\mathcal{P}_{2+0}$, comportant le terme de moindres carrés dans la fonction objectif, sont résolues en des temps similaires, le solveur CPLEX est bien moins efficace sur la résolution du problème $\mathcal{P}_{0/2}$. Ce très fort écart de performance n'est cependant pas surprenant, le solveur gérant plus difficilement des contraintes quadratiques que des contraintes linéaires.

6.2 Construction d'algorithmes *branch-and-bound* dédiés

Les résultats précédents ont montré que la solution exacte de problèmes d'optimisation en norme ℓ_0 pouvait être calculée, sur des problèmes de complexité limitée (forte parcimonie, fort rapport signal sur bruit), à l'aide de reformulations MIP et de l'utilisation d'un solveur MIP générique. Je me suis donc intéressé au développement d'algorithmes de résolution dédiés aux problèmes $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$, pouvant remplacer l'utilisation d'un solveur. D'une manière générique, nous notons ici $\mathcal{P}^{\text{MIP}}$ la reformulation MIP d'une des trois formulations dans la colonne de droite de la table 6.1.

6.2.1 Principe de l'algorithme *branch-and-bound*

Le schéma de résolution d'un MIP exploite le principe dit de séparation et évaluation, ou *branch-and-bound* [Wolsey, 1998]. Nous pouvons en résumer le fonctionnement, sur nos problèmes, par les étapes suivantes :

o) Soit $f^\star$ la valeur optimale du problème MIP considéré. Supposons connue une configuration réalisable (*i.e.,* satisfaisant les contraintes) des variables d'optimisation $(\boldsymbol{b}, \boldsymbol{x})$, possiblement non optimale, pour laquelle la fonction objectif vaut $\overline{f}$, avec donc $\overline{f} \geq f^\star$.

i) Un arbre de décision binaire est créé, où à chaque nœud une décision est prise sur une variable : est-elle nulle ou non nulle ? Deux branches sont créées, correspondant aux cas $b_{p_i} = 1$ et $b_{p_i} = 0$, pour un indice p_i donné. Chaque branche génère ainsi un sous-problème, noté $\mathcal{P}^{(i)}$, où cette variable binaire est fixée.

ii) Au sous-problème $\mathcal{P}^{(i)}$ associé à un nœud donné i, correspond donc un ensemble de variables binaires fixées (notons-les $\boldsymbol{b}_{S_1} = \mathbf{1}$ et $\boldsymbol{b}_{S_0} = \mathbf{0}$), les autres étant indéterminées (notons-les $\boldsymbol{b}_{\overline{S}}$). Une *borne inférieure* $\underline{f}^{(i)}$ du minimum atteint sur l'ensemble des configurations des variables $\boldsymbol{b}_{\overline{S}}$ va alors être calculée, par exemple en relâchant les contraintes d'intégrité sur ces variables. On résout alors un problème d'optimisation

6. En pratique, cependant, la mise en œuvre est réalisée avec des contraintes égalité, évitant de recourir à des contraintes quadratiques inutiles.

continue, noté $\mathcal{R}^{(i)}$:

$$\mathcal{R}^{(i)} : \underline{f}^{(i)} = \min_{\boldsymbol{b}\in[0,1]^P, \boldsymbol{x}\in\mathbb{R}^P} \mathcal{P}^{\text{MIP}} \text{ s. c. } \boldsymbol{b}_{S_1} = \mathbf{1} \text{ et } \boldsymbol{b}_{S_0} = \mathbf{0}. \tag{6.4}$$

Autrement dit, au lieu de calculer explicitement la fonction de coût pour chaque configuration de ces variables, on va borner inférieurement l'ensemble des solutions atteintes sur les sous-problèmes issus de $\mathcal{P}^{(i)}$ par la relaxation continue de ces variables.

iii) Si $\underline{f}^{(i)} \geq \overline{f}$, alors le sous-problème ne peut contenir la solution optimale et on peut donc l'éliminer. Sinon, deux nouveaux sous-problèmes sont créés, en choisissant une nouvelle variable binaire dans $\overline{S}$.

iv) Si le calcul de la relaxation continue donne une configuration réalisable du problème (*i.e.,* si les variables relâchées sont binaires à l'optimum), l'ensemble des branches contenues dans ce sous-problème est résolu. On peut alors comparer cette solution à la meilleure solution connue, et le cas échéant mettre à jour $\overline{f}$. Des heuristiques de calcul de solutions réalisables peuvent également être mises en œuvre pendant l'algorithme afin d'améliorer la meilleure configuration connue, et donc de diminuer $\overline{f}$.

Par construction, le processus converge en un nombre fini d'étapes, et se termine lorsque l'ensemble de l'arbre de décision a été exploré. On a alors $\overline{f} = \min_i \underline{f}^{(i)} = f^\star$. La figure 6.4 illustre le parcours d'un tel arbre et la figure 6.5 montre l'évolution des deux bornes encadrant l'optimum.

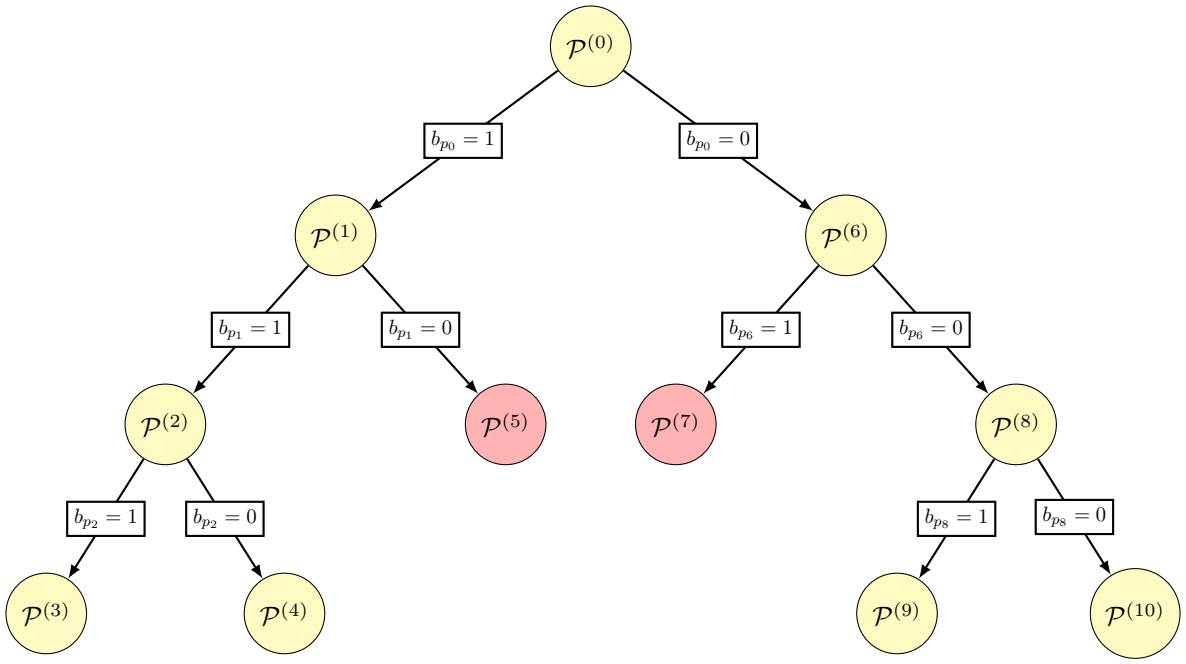


FIGURE 6.4 – Exemple de parcours d'un arbre binaire de décision : chaque nœud, correspondant à un problème d'optimisation $\mathcal{P}^{(i)}$, est divisé en deux nœuds fils obtenus en contraignant une variable binaire à valoir 0 ou 1. Le parcours se fait en profondeur d'abord, en branchant en priorité du côté $b_p = 1$ (les sous-problèmes sont numérotés dans leur ordre de visite). Les nœuds 5 et 7 ont été élagués.

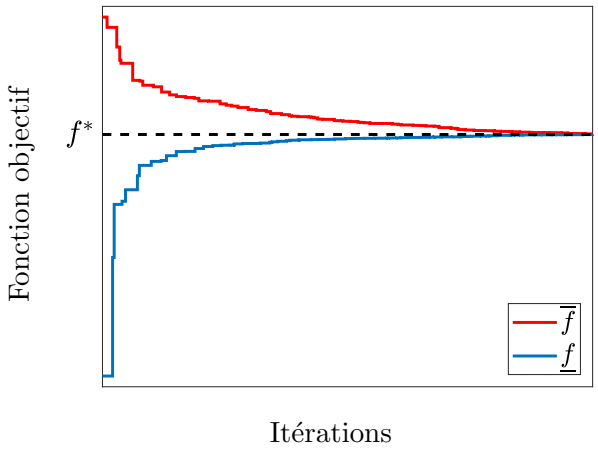


FIGURE 6.5 – Évolution (simulée) de la borne supérieure (meilleure solution connue à une itération donnée, $\overline{f}$) et de la borne inférieure globale (valeur minimale sur l'ensemble des problèmes relâchés, $\underline{f} = \min_i \underline{f}^{(i)}$).

6.2.2 Motivations pour des algorithmes dédiés

Ma motivation principale, qui a généré le projet ANR MIMOSA, résidait dans l'idée que ces problèmes ℓ_0 sont des MIP à la structure finalement assez simple, à laquelle un solveur générique, développé pour résoudre des problèmes plus complexes, est « indifférent ». En particulier, quelle que soit la formulation envisagée $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ ou $\mathcal{P}_{2+0}$:

- les variables binaires apparaissent sous la simple forme d'une somme ;
- chaque variable binaire est liée à une seule variable continue, sous la forme de contraintes linéaires (reformulation *bigM*) ;
- hormis cette reformulation, le problème sépare les variables continues dans la fonction de coût des variables binaires en contrainte, ou inversement ;
- les seuls termes quadratiques concernent les variables continues (il n'y a pas de produit de variables binaires, et pas de produit entre variables binaires et continues).

Éclairés par ces spécificités, nous détaillons dans les sections suivantes les différents choix effectués pour la construction d'une procédure branch-and-bound. Notons que des idées similaires apparaissent dans un article de 1996 [Bienstock, 1996], qui contient cependant peu de détails de mise en œuvre ni de résultats. Dans [Bertsimas et Shioda, 2009], les auteurs construisent également un algorithme de branch-and-bound dédié, mais dans un contexte inapplicable ici (problèmes de moindres carrés surdéterminés, à faible nombre de composantes nulles).

6.2.3 Stratégies d'exploration

Différentes questions se posent à l'heure de définir une stratégie d'exploration de l'arbre, sur le principe expliqué au § 6.2.1. Sur quelle variable b_{p_i} choisit-on de brancher (étapes i) et iii)) ? Considère-t-on d'abord le côté $b_{p_i} = 1$ ou $b_{p_i} = 0$? Quel sous-problème explorer en priorité ? Il n'est pas possible de répondre de manière générale à ces questions et de nombreuses possibilités coexistent dans la littérature et sont implantées dans les logiciels existants.

Nos choix s'inspirent de la construction des méthodes gloutonnes : nous choisissons une

stratégie de parcours de l'arbre en profondeur d'abord (*deep first search*) [Wolsey, 1998], qui consiste à subdiviser et descendre le plus profondément dans l'arbre de recherche avant de parcourir les nœuds en attente. Nous explorons alors en priorité la branche correspondant à activer des composantes, *i.e.,* à la décision $b_{p_i} = 1$. Puisque nous recherchons des solutions parcimonieuses, nous allons de fait limiter la profondeur de l'arbre de recherche. Une illustration est représentée en figure 6.4.

Afin de sélectionner l'indice p_i de la variable sur laquelle brancher, nous exploitons la solution fournie par la relaxation $\mathcal{R}^{(i)}$ en choisissant la variable binaire dans $\overline{S}$ dont la valeur est la plus proche de 1 :

$$p_i = \arg\max_{p \in \overline{S}} b_p^\star, \quad \text{où} \quad \boldsymbol{b}^\star \text{ est l'ensemble de variables continues minimisant (6.4).}$$

Notons que ces choix ne sont pas configurables dans l'utilisation standard d'un solveur comme CPLEX, où l'on peut par exemple choisir entre la variable relâchée la plus proche d'une valeur entière (indifféremment 0 ou 1) et, à l'inverse, la variable la plus proche de 0.5.

6.2.4 Les relaxations continues sont des problèmes ℓ_1

Ce paragraphe aborde une de nos contributions majeures sur la résolution, qui résulte de l'observation que les problèmes de relaxation continue de chaque sous-problème sont en fait des problèmes en norme ℓ_1. Plus exactement, nous montrons qu'à l'optimum $(\boldsymbol{b}^*, \boldsymbol{x}^*)$ du problème relâché $\mathcal{R}^{(i)}$, les contraintes (6.3) de reformulation *bigM* sont saturées : $\left|\boldsymbol{x}^*_{\overline{S}}\right| = M\boldsymbol{b}^*_{\overline{S}}$. Par conséquent, le terme correspondant à la norme ℓ_0 s'écrit

$$\sum_{p \in \overline{S}} b_p^* = \frac{1}{M} \sum_{p \in \overline{S}} |x_p^*| = \frac{1}{M} \left\|\boldsymbol{x}_{\overline{S}}\right\|_1 .$$

Les preuves peuvent être trouvées dans un article soumis à la revue *Optimisation Methods and Software* [Ben Mhenni *et al.*, 2019], reproduit en Annexe A.

Une interprétation géométrique connue de ce résultat [Chandrasekaran *et al.*, 2012] est donnée en figure 6.6, où la contrainte $\|\boldsymbol{x}\|_0 \leq K$ est relâchée en $\|\boldsymbol{x}\|_1 \leq MK$, *sous l'hypothèse supplémentaire que la solution est bornée.* On retrouve ici l'argument souvent avancé que « *la norme ℓ_1 est la relaxation convexe de la norme ℓ_0* ». Ceci n'est vrai, cependant, que si l'on rajoute des contraintes de borne sur $\boldsymbol{x}$. Cette limite est due au non-respect de la propriété d'homogénéité par la « norme » ℓ_0 : le domaine défini par $\|\boldsymbol{x}\|_0 \leq K$ ne définit pas un ensemble borné.

Le tableau 6.2 donne la reformulation des sous-problèmes rencontrés dans l'algorithme branch-and-bound, pour les trois formulations étudiées.

Ce résultat a deux conséquences importantes :

- Les relaxations continues de tous les sous-problèmes peuvent être reformulées *sans variables binaires.*
- Elles s'écrivent toutes comme des problèmes d'optimisation comprenant une terme d'ajustement quadratique, une norme ℓ_1 portant sur une partie des variables et des contraintes de borne sur l'ensemble des variables. Une optimisation dédiée peut donc être mise en œuvre, potentiellement plus efficace que la résolution du problème relâché initial – colonne de gauche dans la table 6.2 – par un algorithme générique de programmation quadratique. C'est l'objet du § 6.2.5.

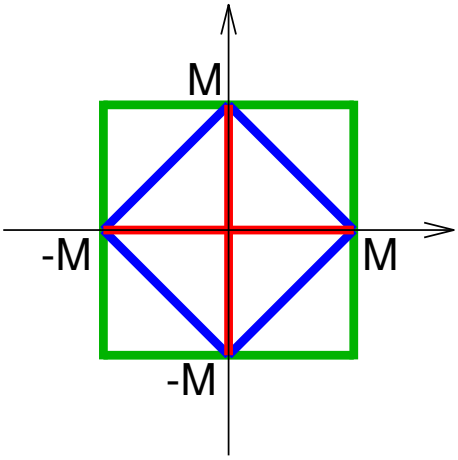


FIGURE 6.6 – La norme ℓ_1 est la relaxation convexe de la norme ℓ_0 **en présence de contraintes de borne**. À l'intérieur du carré vert, domaine admissible $\|\boldsymbol{x}\|_\infty \leq M$. L'intérieur du carré bleu correspond au domaine $\|\boldsymbol{x}\|_1 \leq M$, qui contient bien les deux axes rouges correspondant au domaine $\|\boldsymbol{x}\|_0 \leq 1$.

6.2.5 Optimisation en norme ℓ_1 avec contraintes de borne et variables libres

L'optimisation de critères impliquant un terme de moindres carrés et la norme ℓ_1 a donné lieu à de nombreux travaux, déjà évoqués dans ce manuscrit, notamment pour la forme pénalisée. Les problèmes qui nous intéressent ici sortent cependant de ce cadre standard. D'une part, ils incluent des contraintes de borne et le terme en norme ℓ_1 n'affecte qu'une partie des variables. D'autre part, nous cherchons à résoudre les trois classes de problèmes définis dans la colonne de droite de la table 6.2, où la norme ℓ_1 apparaît en contrainte, en fonction objectif, ou sous forme pénalisée. Rappelons par ailleurs que la résolution de ces problèmes vise à obtenir une borne inférieure sur un ensemble de solutions relatives à une branche de l'arbre de recherche (voir le § 6.2.1). Il est alors important d'envisager des algorithmes *exacts*, par opposition à des algorithmes de descente où, si l'on tronquait les itérations, la borne inférieure obtenue ne serait pas garantie.

Algorithme homotopique

Nous nous sommes orientés vers les méthodes homotopiques [Osborne *et al.*, 2000, Efron *et al.*, 2004, Donoho et Tsaig, 2008], dont la structure permet naturellement de prendre en compte ces contraintes, et qui permet de résoudre similairement les trois classes de problèmes. Notons qu'une version de l'algorithme homotopique pour des problèmes ℓ_1 avec contraintes de borne a été publiée concomitamment à nos travaux [Liang et Wang, 2017]. Nous l'avons généralisé au problème contenant à la fois des contraintes de borne et des variables « libres ». Les détails se trouvent dans la référence [Ben Mhenni *et al.*, 2019] ; nous en résumons ici le principe.

Considérons la forme pénalisée $\mathcal{P}_{2+1}^{(i)}$ de la table 6.2, en notant $\lambda = \mu/M$:

$$\min_{\boldsymbol{x}_{S_1}, \boldsymbol{x}_{\overline{S}}} \frac{1}{2} \|\boldsymbol{y} - \mathbf{A}_{S_1}\boldsymbol{x}_{S_1} - \mathbf{A}_{\overline{S}}\boldsymbol{x}_{\overline{S}}\|^2 + \lambda \|\boldsymbol{x}_{\overline{S}}\|_1 \text{ s.c. } \begin{cases} \|\boldsymbol{x}_{\overline{S}}\|_\infty \leq M \\ \|\boldsymbol{x}_{S_1}\|_\infty \leq M \end{cases}. \tag{6.5}$$

Le principe de la méthode homotopique est de calculer séquentiellement l'ensemble des solutions en fonction de λ. L'algorithme repose sur les propriétés suivantes.

Sous-problème relâché	Problème équivalent en norme ℓ_1
$\mathcal{R}^{(i)}_{2/0} : \min_{\boldsymbol{b}\in[0,1]^P, \boldsymbol{x}\in\mathbb{R}^P} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\|^2$ s.c. $\begin{cases} \sum_{p=1}^{P} b_p \leq K \\ \lvert\boldsymbol{x}\rvert \leq M\boldsymbol{b} \\ \boldsymbol{b}_{S_1}=1, \boldsymbol{b}_{S_0}=0 \end{cases}$	$\mathcal{P}^{(i)}_{2/1} : \min_{\boldsymbol{x}_{S_1}\in\mathbb{R}^{n_1}, \boldsymbol{x}_{\overline{S}}\in\mathbb{R}^{\overline{n}}} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}_{S_1}\boldsymbol{x}_{S_1}-\mathbf{A}_{\overline{S}}\boldsymbol{x}_{\overline{S}}\|^2$ s.c. $\begin{cases} \|\boldsymbol{x}_{\overline{S}}\|_1 \leq M(K-n_1) \\ \|\boldsymbol{x}_{\overline{S}}\|_\infty \leq M \\ \|\boldsymbol{x}_{S_1}\|_\infty \leq M \end{cases}$
$\mathcal{R}^{(i)}_{0/2} : \min_{\boldsymbol{b}\in[0,1]^P, \boldsymbol{x}\in\mathbb{R}^P} \sum_{p=1}^{P} b_p$ s.c. $\begin{cases} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\|^2 \leq \alpha \\ \lvert\boldsymbol{x}\rvert \leq M\boldsymbol{b} \\ \boldsymbol{b}_{S_1}=1 \\ \boldsymbol{b}_{S_0}=0 \end{cases}$	$\mathcal{P}^{(i)}_{1/2} : \min_{\boldsymbol{x}_{S_1}\in\mathbb{R}^{n_1}, \boldsymbol{x}_{\overline{S}}\in\mathbb{R}^{\overline{n}}} \frac{1}{M}\|\boldsymbol{x}_{\overline{S}}\|_1 \ (+n_1)$ s.c. $\begin{cases} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}_{S_1}\boldsymbol{x}_{S_1}-\mathbf{A}_{\overline{S}}\boldsymbol{x}_{\overline{S}}\|^2 \leq \alpha \\ \|\boldsymbol{x}_{\overline{S}}\|_\infty \leq M \\ \|\boldsymbol{x}_{S_1}\|_\infty \leq M \end{cases}$
$\mathcal{R}^{(i)}_{2+0} : \min_{\boldsymbol{b}\in[0,1]^P, \boldsymbol{x}\in\mathbb{R}^P} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\|^2 + \mu\sum_{p=1}^{P} b_p$ s.c. $\begin{cases} \lvert\boldsymbol{x}\rvert \leq M\boldsymbol{b} \\ \boldsymbol{b}_{S_1}=1 \\ \boldsymbol{b}_{S_0}=0 \end{cases}$	$\mathcal{P}^{(i)}_{2+1} : \min_{\boldsymbol{x}_{S_1}\in\mathbb{R}^{n_1}, \boldsymbol{x}_{\overline{S}}\in\mathbb{R}^{\overline{n}}} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}_{S_1}\boldsymbol{x}_{S_1}-\mathbf{A}_{\overline{S}}\boldsymbol{x}_{\overline{S}}\|^2 + \frac{\mu}{M}\|\boldsymbol{x}_{\overline{S}}\|_1 \ (+\mu\ n_1)$ s.c. $\begin{cases} \|\boldsymbol{x}_{\overline{S}}\|_\infty \leq M \\ \|\boldsymbol{x}_{S_1}\|_\infty \leq M \end{cases}$

TABLE 6.2 – Problèmes de relaxation continue mis en jeu pour chaque sous-problème dans l'algorithme branch-and-bound (colonne de gauche) et problèmes équivalents en norme ℓ_1 (colonne de droite), pour les trois formulations $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$. Les indices $1,\ldots,P$ sont partitionnés en S_0 (variables binaires fixées à 0, pour lesquelles $\boldsymbol{x}_{S_0}=\mathbf{0}$), S_1 (variables binaires fixées à 1, pour lesquelles $\boldsymbol{x}_{S_1}\neq\mathbf{0}$) et $\overline{S}$ (variables binaires non déterminées). Les dimensions de S_1 et $\overline{S}$ sont notées respectivement n_1 et $\overline{n}$.

- Soient

$$\boldsymbol{x}^{(0)}_{S_1} \triangleq \underset{\boldsymbol{x}_{S_1}\in\mathbb{R}^{n_1}}{\arg\min} \frac{1}{2}\|\boldsymbol{y}-\mathbf{A}_{S_1}\boldsymbol{x}_{S_1}\|^2 \ \text{ s. c. } \ \|\boldsymbol{x}_{S_1}\|_\infty \leq M$$

$$\text{et} \quad \lambda^{(0)} \triangleq \|\mathbf{A}^T_{\overline{S}}(\boldsymbol{y}-\mathbf{A}_{S_1}\boldsymbol{x}^{(0)}_{S_1})\|_\infty.$$

  On montre que pour $\lambda \geq \lambda^{(0)}$, les variables pénalisées $\boldsymbol{x}_{\overline{S}}$ sont identiquement nulles et $\boldsymbol{x}_{S_1} = \boldsymbol{x}^{(0)}_{S_1}$.

- On montre que la solution est continue et linéaire par morceaux en fonction de λ.

On peut alors former la suite de solutions $\boldsymbol{x}^{(t)}$ à chaque rupture de pente [7] :

$$\begin{cases} \boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)} + \gamma^{(t)} \boldsymbol{d}^{(t)} & (6.6) \\ \text{et } \lambda^{(t)} = \lambda^{(t-1)} - \gamma^{(t)}, & (6.7) \end{cases}$$

où $\boldsymbol{d}^{(t)}$ est le vecteur contenant les changements de pente et $\gamma^{(t)} > 0$ est la longueur de l'intervalle $[\lambda^{(t)}, \lambda^{(t-1)}]$.

- À une itération t donnée, on va rechercher le prochain point de rupture de pente en recherchant un changement dans la configuration des variables : une variable nulle devient non nulle (ou réciproquement), une variable non nulle atteint la borne $\pm M$ ou, inversement, une variable valant $\pm M$ quitte cette borne. Chaque cas produit une expression analytique de λ, obtenue en exploitant les conditions d'optimalité du problème (6.5). L'itération suivante est alors définie en sélectionnant l'événement correspondant au plus petit décrément de λ et en mettant à jour la configuration du support.

La figure 6.7 en haut représente un exemple de chemin de solutions obtenues en fonction de λ, où $\boldsymbol{x}^*$ désigne la solution du problème (6.5) pour une valeur donnée de λ.

Pour la version pénalisée, l'algorithme s'arrête à l'itération t telle que la valeur cible du paramètre de régularisation, notée $\lambda^\star$, soit dans l'intervalle $[\lambda^{(t}, \lambda^{(t-1)}]$. La solution est alors simplement calculée en remplaçant $\gamma^{(t)}$ par $\gamma^\star = \lambda^{(t-1)} - \lambda^\star$ dans (6.6). Lorsque λ décroît continûment, l'erreur $\frac{1}{2}\left\|\boldsymbol{y} - \mathbf{A}_{S_1}\boldsymbol{x}^*_{S_1} - \mathbf{A}_{\overline{S}}\boldsymbol{x}^*_{\overline{S}}\right\|^2$ décroît continûment et la norme ℓ_1 des variables pénalisées croît continûment car les deux fonctions sont convexes. La figure 6.7 en bas montre ainsi la courbe de Pareto $\left(\frac{1}{2}\left\|\boldsymbol{y} - \mathbf{A}_{S_1}\boldsymbol{x}^*_{S_1} - \mathbf{A}_{\overline{S}}\boldsymbol{x}^*_{\overline{S}}\right\|^2, \left\|\boldsymbol{x}^*_{\overline{S}}\right\|_1\right)$, représentant l'ensemble des solutions optimales en fonction de λ. L'algorithme peut donc, de manière similaire, résoudre les formulations contraintes de la table 6.2, en s'arrêtant à l'itération t telle que la contrainte *égalité* associée ($\left\|\boldsymbol{x}^*_{\overline{S}}\right\|_1 = M(K - n_1)$ pour $\mathcal{P}^{(i)}_{2/1}$ et $\frac{1}{2}\left\|\boldsymbol{y} - \mathbf{A}_{S_1}\boldsymbol{x}^*_{S_1} - \mathbf{A}_{\overline{S}}\boldsymbol{x}^*_{\overline{S}}\right\|^2 = \alpha$ pour $\mathcal{P}^{(i)}_{1/2}$) soit atteinte dans l'intervalle $[\lambda^{(t)}, \lambda^{(t-1)}]$. La solution d'intérêt est alors simplement calculée en remplaçant $\gamma^{(t)}$ par $\gamma^\star = \lambda^{(t-1)} - \lambda^\star$ dans (6.6). Dans les deux cas, la valeur $\gamma^\star$ correspondant aux contraintes égalité s'obtient analytiquement (voir [Ben Mhenni *et al.*, 2019]).

6.2.6 Évaluation des performances

L'efficacité des algorithmes branch-and-bound ainsi construits est évaluée sur les instances de problèmes de déconvolution parcimonieuse définies dans [Bourguignon *et al.*, 2016] et présentées au § 6.1.2, où $\boldsymbol{y} \in \mathbb{R}^{120}$, $\mathbf{A} \in \mathbb{R}^{100\times120}$. Le rapport signal sur bruit est ici de 10 dB (correspondant aux problèmes les plus difficiles et les plus coûteux en temps de calcul) et $M = 1.1\left\|\boldsymbol{A}^T\boldsymbol{y}\right\|_\infty$. La cardinalité des problèmes varie de $K = 5$ à $K = 9$ et la vraie valeur de K est utilisée pour les problèmes $\mathcal{P}_{2/0}$. Les paramètres respectifs des problèmes $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$ sont réglés de manière statistique en fonction du niveau de bruit et du degré de parcimonie : α est fixé de sorte que la probabilité $P(\|\boldsymbol{\epsilon}\|^2 \leq \alpha) = 95$ % et $\mu = 2\sigma^2\log(1/\rho - 1)$, où σ^2 est la variance du bruit et $\rho = K/P$ (voir [Soussen *et al.*, 2011] pour ce réglage). La qualité des solutions ayant déjà été étudiée en Section 6.1, nous nous intéressons ici uniquement aux temps de calcul associés.

7. À ma connaissance, les notations récursives utilisées ici sont dues à Donoho et Tsaig [Donoho et Tsaig, 2008] pour le problème pénalisé standard.

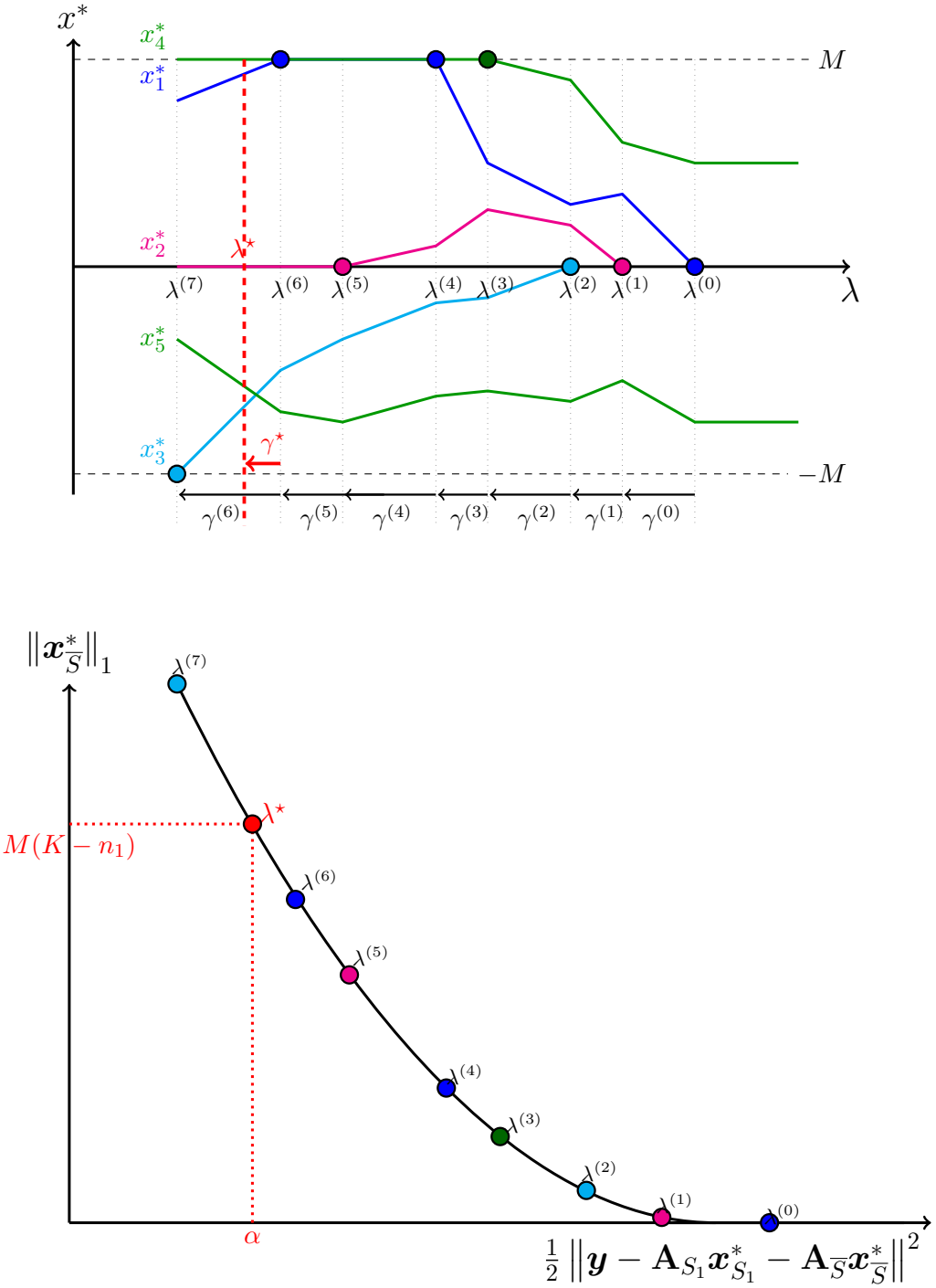


FIGURE 6.7 – Fonctionnement (simulé) de l'algorithme homotopique pour le problème (6.5). À gauche, exemple de chemin de solutions en fonction du paramètre de pénalisation λ, avec 4 variables non nulles à l'optimum : $\overline{S} = \{1, 3, 5\}$ et $S_1 = \{4\}$. Les cercles montrent les événements ayant changé la configuration des variables et les lignes verticales localisent les ruptures de pente. À droite, ensemble des solutions optimales dans un repère $\left(\frac{1}{2}\left\|\boldsymbol{y}-\mathbf{A}_{S_1}\boldsymbol{x}_{S_1}^*-\mathbf{A}_{\overline{S}}\boldsymbol{x}_{\overline{S}}^*\right\|^2, \left\|\boldsymbol{x}_{\overline{S}}^*\right\|_1\right)$ en fonction de λ (courbe de Pareto). Les cercles de couleur marquent les points de rupture correspondant à la figure de gauche.

Afin de dissocier les performances de la stratégie d'exploration de celles du calcul des relaxations à chaque nœud, les algorithmes développés (dénommés B&B$_{\text{R-HOM}}$ par la suite) sont comparés à :

- notre stratégie d'exploration par branch-and-bound, où les problèmes de relaxation continue (sous la forme de gauche dans la table 6.2) sont résolus par le solveur de programmation quadratique de CPLEX. Cette version est dénommée B&B$_{\text{R-CPLEX}}$ par la suite ;
- la résolution du problème MIP initial par le solveur MIP de CPLEX (version 12.8), nommée MIP$_{\text{CPLEX}}$.

Toutes les méthodes sont implémentées en C++ et exécutées sur une machine portable UNIX équipée de 31,1 Go de RAM et de quatre processeurs centraux (CPU) Intel Core i7-6600U cadencés à 2.6 GHz. Les calculs sont limités à un seul cœur afin de se concentrer sur la performance des algorithmes en désactivant leurs capacités de parallélisation. Le temps CPU maximum est fixé à 1 000 secondes et le paramétrage par défaut de CPLEX est utilisé.

Les temps de calcul obtenus, moyennés sur 50 instances de chaque problème, sont donnés en table 6.3. B&B$_{\text{R-HOM}}$ et B&B$_{\text{R-CPLEX}}$ mettant en œuvre la même stratégie de branchement, ils explorent le même nombre de nœuds — à de très légères différences près, probablement dues à la précision numérique. Cependant, le temps de calcul pour B&B$_{\text{R-HOM}}$ est plus faible d'un facteur de 8 à 25 pour $\mathcal{P}_{2/0}$ et $\mathcal{P}_{2+0}$, et d'un facteur de 60 à 120 pour $\mathcal{P}_{0/2}$, mettant en évidence le gain apporté par l'approche homotopique pour le calcul des relaxations. Il permet également de résoudre plus d'instances en 1 000 s.

Pour les problèmes $\mathcal{P}_{2/0}$ et $\mathcal{P}_{2+0}$, B&B$_{\text{R-HOM}}$ est plus efficace que MIP$_{\text{CPLEX}}$ sur les instances les plus faciles ($K = 5$ et $K = 7$), allant jusqu'à réduire le temps de calcul d'un facteur 4 sur les problèmes les plus simples ($\mathcal{P}_{2/0}$, $K = 5$). Pour $K = 9$, en revanche, MIP$_{\text{CPLEX}}$ résout un plus grand nombre d'instances en 1 000 s et est jusqu'à 30 % plus rapide (pour $\mathcal{P}_{2+0}$). Ce résultat n'est pas surprenant, la résolution de MIP quadratiques de CPLEX bénéficiant de nombreux raffinements qui ne sont pas mis en œuvre dans notre procédure de branch-and-bound (heuristiques, méthodes de coupe, etc.). Nous y reviendrons dans la conclusion de ce chapitre.

Les résultats sont très différents pour le problème $\mathcal{P}_{0/2}$, pour lequel B&B$_{\text{R-HOM}}$ est bien plus efficace que MIP$_{\text{CPLEX}}$ sur l'ensemble des instances, à la fois au niveau du nombre de nœuds explorés et du temps de calcul. En particulier, le temps de calcul par nœud est ici réduit d'un facteur d'au moins 6 et le nombre de nœuds explorés d'un facteur au moins 20. Au final, B&B$_{\text{R-HOM}}$ est 120 à 250 fois plus rapide que MIP$_{\text{CPLEX}}$.

6.3 Démélange spectral « ℓ_0 » et contraintes structurantes

Nous nous sommes également intéressés, dans la thèse de Ramzi Ben Mhenni, à l'exploitation de modèles en norme ℓ_0 pour le problème de démélange spectral parcimonieux. Le démélange est un problème classique en imagerie hyperspectrale de télédétection, où l'on cherche à décomposer un spectre de réflectance mesuré en une combinaison de spectres élémentaires « purs », et à estimer les coefficients (abondances) associés [Singer et McCord, 1979, Keshava et Mustard, 2002]. Le mélange est en particulier dû à l'éloignement de l'instrument de mesure – souvent embarqué à bord d'un avion ou d'un satellite – de la scène observée : la surface d'un pixel du détecteur a alors une empreinte au sol de grande taille, dans laquelle peuvent se superposer les contributions de différents éléments. Re-

Problème		Propre branch-and-bound						Solveur MIP CPLEX		
		B&B$_{\text{R-HOM}}$			B&B$_{\text{R-CPLEX}}$			CPLEX 12.8		
		Temps (s)	Nœuds ($\times 10^3$)	E	Temps (s)	Nœuds ($\times 10^3$)	E	Temps (s)	Nœuds ($\times 10^3$)	E
$\mathcal{P}_{2/0}$	$K=5$	**0.7**	1.28	0	7.7	1.28	0	3.0	1.71	0
	$K=7$	**11.6**	17.89	0	141.9	17.89	2	16.6	21.51	0
	$K=9$	**43.5**	57.37	9	448.1	57.46	30	53.8	72.04	6
$\mathcal{P}_{2+0}$	$K=5$	**1.8**	2.01	0	32.6	2.02	0	3.2	1.98	0
	$K=7$	**7.3**	10.20	0	187.3	10.22	7	7.4	9.61	0
	$K=9$	25.6	31.80	5	470.7	31.87	28	**17.3**	23.74	2
$\mathcal{P}_{0/2}$	$K=5$	**0.1**	0.21	0	6.0	0.21	0	25.7	6.71	0
	$K=7$	**0.9**	2.32	0	85.2	2.32	0	114.8	49.54	2
	$K=9$	**2.5**	5.22	0	296.9	5.22	18	328.2	101.07	17

TABLE 6.3 – Performances comparatives de B&B$_{\text{R-HOM}}$, B&B$_{\text{R-CPLEX}}$ et MIP$_{\text{CPLEX}}$ pour la résolution de problèmes simulés de déconvolution parcimonieuse : temps de calcul moyen (s), nombre moyen de nœuds explorés et nombre d'instances non résolues en 1 000 s (colonne 'E'). Les moyennes sont calculées sur les instances résolues en moins de 1 000 s par les trois algorithmes.

marquons que nous avons là une situation très différente de l'imagerie hyperspectrale en astronomie qui a fait l'objet du Chapitre 4, où dans un pixel on ne peut souvent détecter qu'une source au maximum.

Le démélange spectral me semble un problème particulièrement adapté pour des approches basées sur des formulations de MIP car la taille des problèmes est souvent limitée. Si l'on ne considère pas de contraintes spatiales reliant les différents spectres d'un cube d'images hyperspectrales, le problème reste unidimensionnel, où le nombre de données est limité à quelques centaines de longueurs d'onde. De plus, si la taille du dictionnaire peut être arbitrairement grande, en général le nombre de composantes recherchées n'excède pas quelques unités, limitant la combinatoire. Qui plus est, il est rarement nécessaire d'avoir des résultats de décomposition en temps réel. Par ailleurs, l'introduction de variables binaires encodant la présence ou l'absence de chaque composante fournit un cadre très flexible de modélisation dans ce contexte : outre la possibilité d'inclure des contraintes *exactes* de parcimonie (qu'on ne peut pas prendre en compte avec une norme ℓ_1 ou avec les algorithmes gloutons classiques, comme nous le verrons plus bas), ce cadre permet aussi la prise en compte de contraintes logiques permettant de structurer l'espace des solutions, afin de tenir compte des problèmes de variabilité spectrale [Zare et Ho, 2014].

6.3.1 Démélange spectral et parcimonie

Parmi les nombreuses approches abordant le démélange, nous nous focalisons sur des méthodes où l'on cherche à décomposer un spectre observé $\boldsymbol{y}$ en une combinaison linéaire

de spectres de référence :

$$\boldsymbol{y} \simeq \sum_{p=1}^{P} a_p \boldsymbol{s}_p = \mathbf{S}\boldsymbol{a}, \tag{6.8}$$

où $\boldsymbol{y} \in \mathbb{R}^N$ est un vecteur colonne représentant le spectre mesuré, $\mathbf{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_P]$ est un dictionnaire composé de P spectres élémentaires (avec en général $P \gg N$) et $\boldsymbol{a} = [a_1, \ldots, a_P]^T$ contient les coefficients du mélange recherché ou *abondances* qui, représentant des proportions, sont positives et de somme unité [8]. Le dictionnaire peut alors contenir une collection de spectres « purs » mesurés en laboratoire, ou résulter de l'identification préalable de spectres de référence (*endmembers*) à partir d'un jeu de données hyperspectrales. Dans les deux cas, l'hypothèse de parcimonie prend tout sons sens car on ne cherche en général qu'un faible nombre de composants présents à une position spatiale donnée [Iordache *et al.*, 2011, Bioucas-Dias *et al.*, 2012, Greer, 2012].

Une approche standard du problème de démélange, dite FCLS pour *Fully Constrained Least Squares* [Heinz et Chein-I-Chang, 2001], consiste à estimer le jeu d'abondances au sens des moindres carrés, sous des contraintes de positivité et de somme à 1 :

$$\text{FCLS} : \min_{\boldsymbol{a}\in\mathbb{R}^P} \|\boldsymbol{y} - \mathbf{S}\boldsymbol{a}\|^2 \quad \text{s. c.} \quad \forall p = 1, \ldots, P, a_p \geq 0 \quad \text{et} \quad \sum_{p=1}^{P} a_p = 1. \tag{6.9}$$

Remarquons que ce modèle introduit une certaine dose de parcimonie par les contraintes de positivité (à l'optimum, certains coefficients vont atteindre leur borne inférieure $a_p = 0$). Celles-ci peuvent cependant s'avérer insuffisantes, en particulier en présence de bruit et/ou d'erreurs dans le modèle. La figure 6.8 présente ainsi une simulation où les données sont générées selon le modèle (6.8). Le dictionnaire est constitué de $P = 481$ spectres de réflectance de minéraux issus de la base de données de l'*United States Geological Survey* (USGS), mesurés sur $N = 113$ longueurs d'onde entre 1 et 2.5 μm [Clark *et al.*, 2003]. Deux jeux de données sont générés contenant $K = 4$ composantes, sans bruit : $\boldsymbol{y} = \mathbf{S}\boldsymbol{a}$ (en haut à gauche) et avec bruit : $\boldsymbol{y} = \mathbf{S}\boldsymbol{a} + \boldsymbol{\epsilon}$ (en bas à gauche), pour un rapport signal sur bruit de $\text{RSB} = 10\log_{10}\frac{\|\mathbf{S}\boldsymbol{a}\|^2}{\|\boldsymbol{\epsilon}\|^2} = 40$ dB [9]. En l'absence de bruit, l'approche FCLS permet de retrouver les vraies composantes puisque, dans ce cas, le vrai modèle est solution du problème de l'équation (6.9). En revanche, sur des données bruitées, les abondances estimées se retrouvent « étalées » en de nombreux coefficients non nuls de faible amplitude. Si cet estimé produit des détections erronées (dont une partie pourrait être supprimée par seuillage), de manière plus critique, il ne permet pas de retrouver l'ensemble des composants ayant généré les données : le seuillage des composantes les plus significatives dans cet exemple « loupe » ainsi la détection des spectres 1 et 4.

Cette dispersion des abondances aurait pu être limitée si l'estimateur avait inclus une contrainte de parcimonie plus forte. L'introduction d'une contrainte en norme ℓ_1 n'est pas appropriée pour ce problème : en raison de la positivité et de la somme à 1, la norme ℓ_1 des abondances vaut 1 et ne peut donc constituer un levier pour imposer plus de parcimonie dans la solution. Un argumentaire plus développé peut être trouvé dans notre contribution [Ben Mhenni *et al.*, 2017], dont nous reproduisons l'illustration de la figure 6.9 : 300 réalisations aléatoires des données sont générées avec le dictionnaire

8. La légitimité de la contrainte de somme unité peut être discutée en pratique. Cependant, à l'inverse de nombreuses méthodes de démélange rejetant cette contrainte pour des raisons techniques, nous l'incluons dans les méthodes développées ici, car elle ne pose aucune difficulté.

9. Cette valeur de rapport signal sur bruit, en apparence très élevée, est due à la positivité des spectres (qui sont donc de moyenne non-nulle). Elle correspond cependant à un niveau de bruit non négligeable.

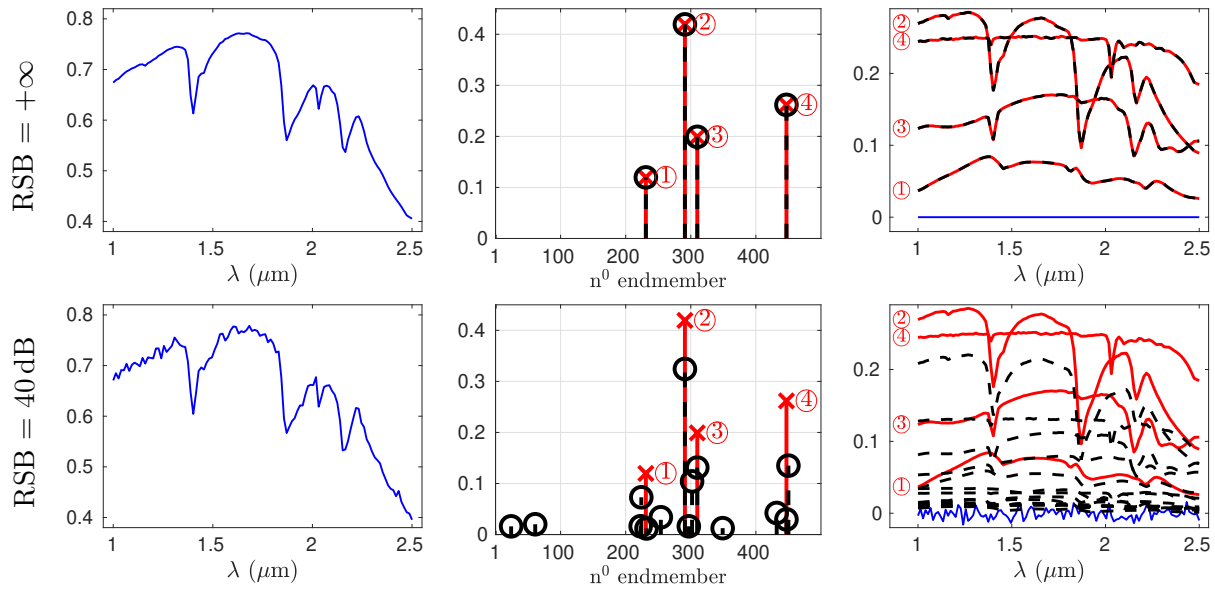


FIGURE 6.8 – Un exemple d'estimation par la méthode FCLS, sur un problème simulé à K = 4 spectres, sans bruit (ligne du haut) et avec bruit (ligne du bas). À gauche : spectre simulé ; au centre : abondances vraies (×) et estimées (○) ; à droite : spectres estimés pondérés par leurs abondances (vrai en trait plein rouge, estimation en tireté noir). En bleu, le bruit.

précédemment introduit, pour différents niveaux de bruit. Le nombre de composantes K varie entre 1 et 10 et les abondances sont positives et de somme à 1. La contrainte de somme à 1 dans (6.9) est remplacée par la contrainte en norme ℓ_1 : $\sum_{p=1}^{P} a_p = \tau$. La figure 6.9 représente, en fonction de τ, l'erreur moyenne sur le support, définie comme le nombre moyen d'erreurs sur les détections correspondant aux K plus grandes abondances estimées. L'introduction d'une contrainte en norme ℓ_1 n'améliore pas significativement le résultat de FCLS ($\tau = 1$) : la valeur de τ minimisant cette erreur passe ainsi de 0.998 pour un rapport signal sur bruit de 60 dB à 0.989 pour un RSB de 40 dB. Par conséquent, les solutions obtenues sont toujours très proches de celles de FCLS et présentent les mêmes limites évoquées ci-dessus.

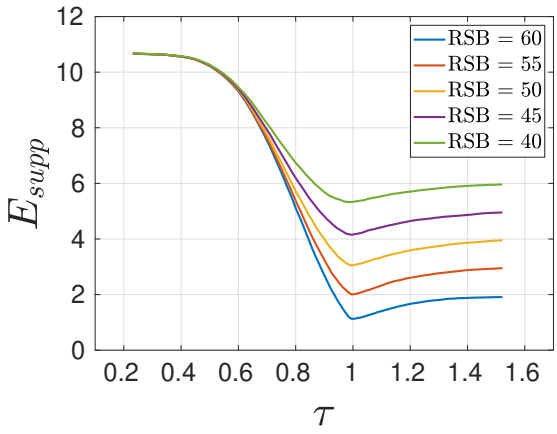


FIGURE 6.9 – Erreur moyenne sur le support (moyenne sur 300 réalisations aléatoires) pour une approche en norme ℓ_1 imposant $\|\boldsymbol{a}\|_1 = \sum_{p=1}^{P} a_p = \tau$, en fonction de τ, pour différents rapports signal sur bruit [Ben Mhenni *et al.*, 2017].

Notons que les travaux [Iordache *et al.*, 2011, Bioucas-Dias *et al.*, 2012, Iordache *et al.*, 2014], qui exploitent la norme ℓ_1 pour le démélange parcimonieux, relâchent au final la contrainte de somme à 1 dans leurs algorithmes, mais pour des raisons qui semblent moins relever d'un choix de modélisation que de l'incompatibilité technique ! D'autres travaux ont proposé de modéliser la parcimonie avec une norme ℓ_p, $p < 1$ [Drumetz *et al.*, 2019], plus pertinente, mais générant des problèmes d'optimisation non-convexe abordés par des méthodes locales. Les approches gloutonnes classiques s'avèrent également inefficaces pour gérer la contrainte de somme à un, notamment en raison de l'impossibilité de normaliser le dictionnaire [Greer, 2012]. L'approche dite *backward* proposée dans [Greer, 2012] produit une solution parcimonieuse en éliminant itérativement les composantes de faible amplitude dans la solution FCLS. C'est en quelque sorte une approche gloutonne « inversée », mais qui ne possède aucune garantie d'optimalité.

6.3.2 Reformulation MIP du problème de démélange parcimonieux

La formulation forte du problème de démélange parcimonieux, faisant intervenir la norme ℓ_0 :

$$\text{UNMIX}^{\ell_0} : \; ; \min_{\boldsymbol{a}\in\mathbb{R}^P} \|\boldsymbol{y} - \mathbf{S}\boldsymbol{a}\|^2 \quad \text{s.c.} \quad \boldsymbol{a} \geq \mathbf{0}, \; \sum_{p=1}^{P} a_p = 1 \quad \text{et} \quad \|\boldsymbol{a}\|_0 \leq K, \tag{6.10}$$

ne présente pas les inconvénients des approches parcimonieuses classiques évoqués au paragraphe précédent. D'une part, elle n'est pas incompatible avec les contraintes de positivité et de somme à 1. D'autre part, ce problème se prête plutôt bien à une reformulation MIP introduisant des variables binaires. En particulier, celle-ci ne requiert pas l'introduction de bornes *bigM* artificielles (voir le § 6.1.1) : puisque les abondances sont naturellement comprises entre 0 et 1, la condition $b_p = 0 \Leftrightarrow a_p = 0$ s'écrit ici $0 \leq a_p \leq b_p$. De manière similaire aux reformulations introduites au § 6.1.1, on montre alors que :

$$\text{UNMIX}^{\ell_0} \Leftrightarrow \min_{\boldsymbol{a}\in\mathbb{R}^P, \boldsymbol{b}\in\{0;1\}^P} \|\boldsymbol{y} - \mathbf{S}\boldsymbol{a}\|^2 \quad \text{s.c.} \quad \begin{cases} \mathbf{0} \leq \boldsymbol{a} \leq \boldsymbol{b} \\ \sum_{p=1}^{P} a_p = 1 \\ \sum_{p=1}^{P} b_p \leq K \end{cases}, \tag{6.11}$$

qui est un MIP quadratique à contraintes linéaires.

6.3.3 Introduction de contraintes structurantes

La prise en compte de la *variabilité spectrale* est une problématique récente en démélange spectral [Zare et Ho, 2014, Thouvenin *et al.*, 2018, Drumetz *et al.*, 2019]. Il s'agit de considérer que les spectres de référence utilisés pour le démélange peuvent subir des modifications par rapport à des spectres calibrés en laboratoire ou entre les différents pixels d'une même scène observée. Les raisons d'une telle variabilité sont multiples, comme la modification des conditions d'acquisition (notamment d'illumination), la présence d'aérosols dans l'atmosphère ou des changements fins de la composition des matériaux composant la scène [Zare et Ho, 2014].

Dans les approches de démélange à base de dictionnaire, une manière de prendre en compte cette variabilité consiste à inclure plusieurs spectres représentant le même minéral

recherché dans le dictionnaire, alors constitué de *groupes* de spectres de référence, et à incorporer des contraintes supplémentaires lors de la phase d'estimation. Dans [Meyer *et al.*, 2016, Drumetz *et al.*, 2019], par exemple, des contraintes de parcimonie structurée sont introduites, de type *group LASSO* [Yuan et Yin, 2006] (favorisant un faible nombre de groupes de spectres sélectionnés dans le mélange) ou, à l'inverse, de type *elitist LASSO* [Kowalski et Torrésani, 2009] (imposant la parcimonie à l'intérieur de chaque groupe de spectres du dictionnaire). Si, dans ces travaux, ces contraintes sont introduites sous la forme de pénalisations continues avec des normes mixtes [Kowalski et Torrésani, 2009], elles peuvent être prises en compte de manière plus forte, à l'aide de reformulations MIP, traduisant *de manière exacte* des contraintes de structuration difficiles. Nous présentons ci-dessous deux exemples.

Notons $G_1, \dots, G_J$ la partition de l'ensemble des indices des P éléments du dictionnaire en J groupes, et $\boldsymbol{a}_{G_j}$ le vecteur d'abondances associé.

Parcimonie dans dans chaque groupe

Lorsque le dictionnaire contient plusieurs représentants d'un même minéral, il peut être souhaitable d'imposer qu'un seul (ou un faible nombre) d'entre eux soit présent dans le mélange. Cette structuration est abordée dans [Meyer *et al.*, 2016, Drumetz *et al.*, 2019] sous la forme d'une norme mixte $\ell_{1,2}$, visant à imposer une contrainte de parcimonie à l'intérieur de chaque groupe :

$$\begin{aligned}\|\boldsymbol{a}\|_{G,1,2} = \left(\sum_{j=1}^{J} \left\|\boldsymbol{a}_{G_j}\right\|_1^2\right)^{1/2} &= \left(\sum_{j=1}^{J} \left(\sum_{k=1}^{\#G_j} a_{G_{jk}}\right)^2\right)^{1/2} \\ &= \sqrt{\left(a_{G_{11}} + \cdots + a_{G_{1\#G_1}}\right)^2 + \cdots + \left(a_{G_{J1}} + \cdots + a_{G_{J\#G_J}}\right)^2},\end{aligned}$$

où $\#G_j$ représente le nombre d'éléments dans le groupe j. Une formulation plus forte consiste à imposer explicitement un nombre maximum K_j d'abondances non nulles dans le groupe j, qui s'écrit :

$$\text{UNMIX}^{\text{G}\ell_0} : \min_{\boldsymbol{a}\in\mathbb{R}^P} \|\boldsymbol{y} - \mathbf{S}\boldsymbol{a}\|^2 \text{ s.c. } \begin{cases} \boldsymbol{a} \geq 0 \\ \sum_{p=1}^{P} a_p = 1 \\ \left\|\boldsymbol{a}_{G_j}\right\|_0 \leq K_j \end{cases}. \tag{6.12}$$

Comme nous l'avons fait au paragraphe précédent, ce problème peut être reformulé comme le MIP :

$$\text{UNMIX}^{\text{G}\ell_0} \Leftrightarrow \min_{\boldsymbol{a}\in\mathbb{R}^P, \boldsymbol{b}\in\{0;1\}^P} \|\boldsymbol{y} - \mathbf{S}\boldsymbol{a}\|^2 \text{ s.c. } \begin{cases} \mathbf{0} \leq \boldsymbol{a} \leq \boldsymbol{b} \\ \sum_{p=1}^{P} a_p = 1 \\ \forall j = 1, \dots, J, \sum_{p_j \in G_j} b_{p_j} \leq K_j \end{cases}.$$

Seuil minimal

Afin d'éviter la « dispersion » des abondances en de nombreuses composantes estimées de faible amplitude, nous pouvons encore imposer une valeur minimale τ aux amplitudes non nulles, formulant un problème à contraintes logiques :

$$\text{UNMIX}^{\text{seuil}} : \min_{\boldsymbol{a}\in\mathbb{R}^P} \|\boldsymbol{y}-\mathbf{S}\boldsymbol{a}\|^2 \text{ s.c. } \begin{cases} \boldsymbol{a} \geq 0 \\ \sum_{p=1}^{P} a_p = 1 \\ \forall p = 1,\dots,P, a_p \neq 0 \Rightarrow a_p \geq \tau \end{cases} .$$

Cette formulation peut être intéressante sur le plan méthodologique, évitant une opération de seuillage *a posteriori* des abondances faibles. À ma connaissance, elle n'est pas apparue dans la littérature de démélange, probablement en raison de la difficulté à la prendre en compte [10]. L'utilisation de variables binaires permet cependant une reformulation exacte :

$$\text{UNMIX}^{\text{seuil}} \Leftrightarrow \min_{\boldsymbol{a}\in\mathbb{R}^P, \boldsymbol{b}\in\{0;1\}^J} \|\boldsymbol{y}-\mathbf{S}\boldsymbol{a}\|^2 \text{ s.c. } \begin{cases} \mathbf{0} \leq \boldsymbol{a} \leq \tau\boldsymbol{b} \\ \sum_{p=1}^{P} a_p = 1 \end{cases} .$$

Notons que cette contrainte induit également une structure parcimonieuse « au sens ℓ_0 » du vecteur d'abondances puisque, en raison de la positivité et de la contrainte de somme à 1, le nombre d'abondances non-nulles ne peut excéder $1/\tau$.

Notons pour terminer que les différents types de contraintes introduits ci-dessus à base de variables binaires (parcimonie simple, parcimonie de groupe, seuil minimum) peuvent tout à fait être considérés conjointement, en accumulant les contraintes associées dans un nouveau problème, qui reste sous la forme d'un MIP quadratique à contraintes linéaires.

6.3.4 Quelques résultats de simulations

Nous présentons quelques résultats obtenus sur des problèmes simulés. Le dictionnaire est composé de $P = 481$ spectres en réflectance de minéraux issus de la base de données de l'USGS [Clark *et al.*, 2003], dans lequel nous avons formé des groupes de spectres correspondant au même minéral. Le dictionnaire ainsi structuré comprend 85 groupes contenant entre 2 et 18 spectres et 146 spectres avec un seul représentant. Les spectres actifs sont tirés aléatoirement dans le dictionnaire, avec au plus un représentant par groupe, et les abondances associées sont tirées uniformément au-dessus d'une valeur minimale de 0.1. Le nombre de composantes dans le mélange varie de 1 à 7 et un bruit gaussien est ajouté, pour un rapport signal sur bruit variant de 60 dB (données peu bruitées) à 40 dB (niveau de bruit moyen).

Les solutions de la méthode FCLS et de l'approche *backward* proposée dans [Greer, 2012] sont comparées aux solutions obtenues par les estimateurs suivants reposant sur la résolution de MIP :

- $\text{UNMIX}^{\text{G}\ell_0}$, où seule la parcimonie à l'intérieur de chaque groupe est imposée avec $K_j = 1$ (*i.e.,* au plus un spectre est actif dans chaque groupe),
- UNMIX^{ℓ_0}, où la norme ℓ_0 est réglée à sa vraie valeur,

10. On trouve cependant une contrainte similaire pour des problèmes d'optimisation de portefeuille, imposant un montant minimum de transactions [Bertsimas et Shioda, 2009, Cui *et al.*, 2013].

- UNMIX$^{\ell_0+\mathrm{G}\ell_0}$ cumule la contrainte de parcimonie dans chaque groupe avec une contrainte en norme ℓ_0,
- UNMIX$^{\mathrm{seuil}+\mathrm{G}\ell_0}$ cumule la contrainte de parcimonie dans chaque groupe avec un seuil minimal sur les abondances non nulles (réglé à $\tau = 0.1$).

Les estimations par UNMIX$^{\ell_0}$ et UNMIX$^{\ell_0+\mathrm{G}\ell_0}$ et la méthode *backward* sont réglées de manière à fournir le bon nombre de composantes non nulles.

La figure 6.10 montre les résultats obtenus pour un niveau de bruit (RSB) de 40 dB et un mélange à $K = 4$ spectres. Les données ont déjà été présentées sur la figure 6.8. La solution de FCLS, également présentée en figure 6.8, ne détecte correctement que les

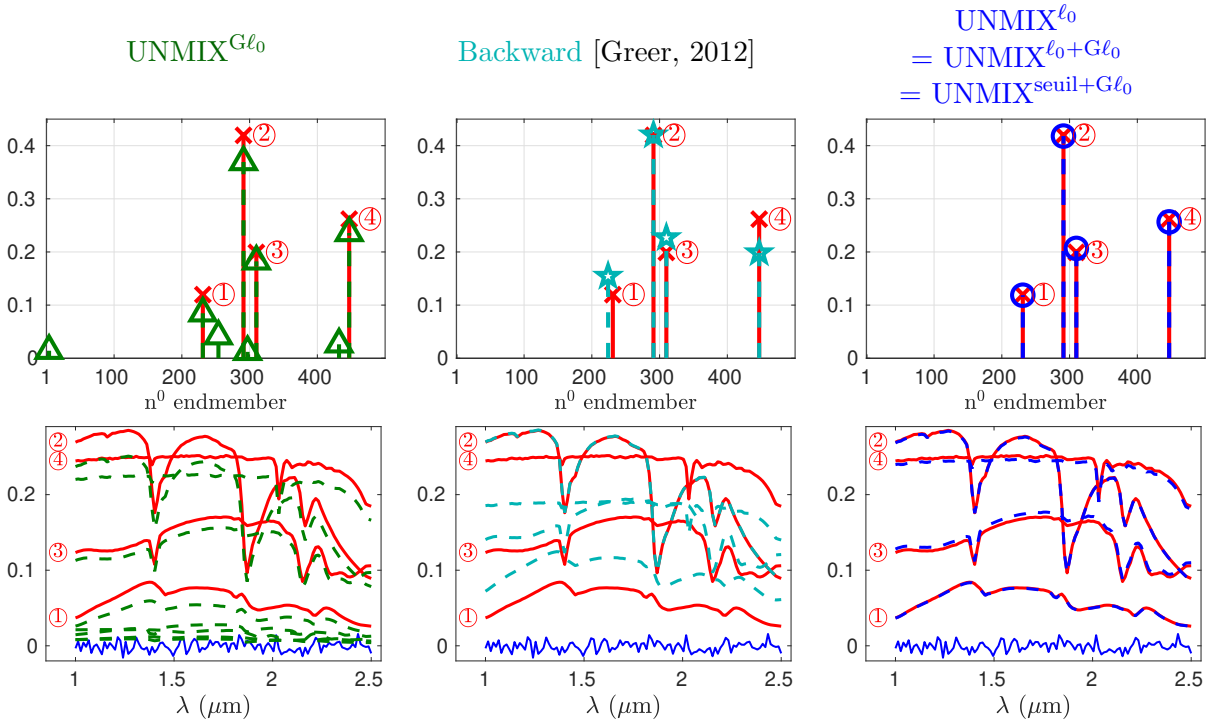


FIGURE 6.10 – Exemple de résultats obtenus par différentes méthodes de démélange, sur un problème simulé à K = 4 spectres, RSB = 40 dB. En haut, abondances estimées (cercles) et vraie séquence (croix). En bas, spectres estimés pondérés par leurs abondances (vrai en trait plein rouge, estimation en tireté). En bleu, le bruit.

spectres 2 et 3 avec des abondances élevées. Les abondances associées aux spectres 1 et 4 sont également non-nulles, mais avec des valeurs plus faibles que deux fausses détections. L'ajout des contraintes de groupe par UNMIX$^{\mathrm{G}\ell_0}$ améliore l'estimation, puisque les 4 abondances les plus élevées correspondant bien aux 4 spectres de la simulation ; quelques détections erronées subsistent cependant, et les abondances sont donc sous-estimées. La méthode *backward* détecte correctement les spectres 2, 3 et 4, mais le spectre 1 est remplacé par un spectre de forme proche et les abondances sont mal estimées. Les trois estimés UNMIX$^{\ell_0}$, UNMIX$^{\ell_0+\mathrm{G}\ell_0}$, UNMIX$^{\mathrm{seuil}+\mathrm{G}\ell_0}$ fournissent ici la même solution, détectant parfaitement les 4 composantes, les abondances étant alors correctement estimées.

La figure 6.11 montre des résultats moyennés sur 30 instances de chaque problème [11], pour un niveau de bruit faible (55 dB) et moyen (40 dB), en représentant l'erreur qua-

11. Pour les instances où l'optimisation n'a pas terminé en 1 000 s, nous gardons la dernière solution obtenue, non garantie donc.

dratique moyenne sur les abondances estimées (en haut) et le taux de détection correcte du support du vecteur d'abondances (en bas). UNMIX$^{\mathrm{G}\ell_0}$ améliore toujours FCLS et la méthode *backward* est meilleure que UNMIX$^{\mathrm{G}\ell_0}$ à fort rapport signal sur bruit ou lorsque le nombre de composantes est faible. Les trois estimés UNMIX$^{\ell_0}$, UNMIX$^{\ell_0+\mathrm{G}\ell_0}$, UNMIX$^{\mathrm{seuil}+\mathrm{G}\ell_0}$ donnent toujours les meilleurs résultats [12]. Sur les problèmes à fort rapport signal sur bruit, leurs détections sont parfaites jusqu'à $K = 7$, alors que FCLS et UNMIX$^{\mathrm{G}\ell_0}$ échouent sur une partie des problèmes dès que $K \geq 2$ (la méthode *backward* se dégrade à partir de $K = 4$). Sur les données plus bruitées, les résultats se dégradent évidemment : même si le minimiseur des problèmes MIP est calculé de manière garantie, il ne permet pas d'identifier correctement les abondances. De plus, comme nous le verrons plus bas, certaines des solutions correspondant aux estimateurs « ℓ_0 » ne sont pas garanties, l'algorithme n'ayant pas convergé en la limite de temps impartie.

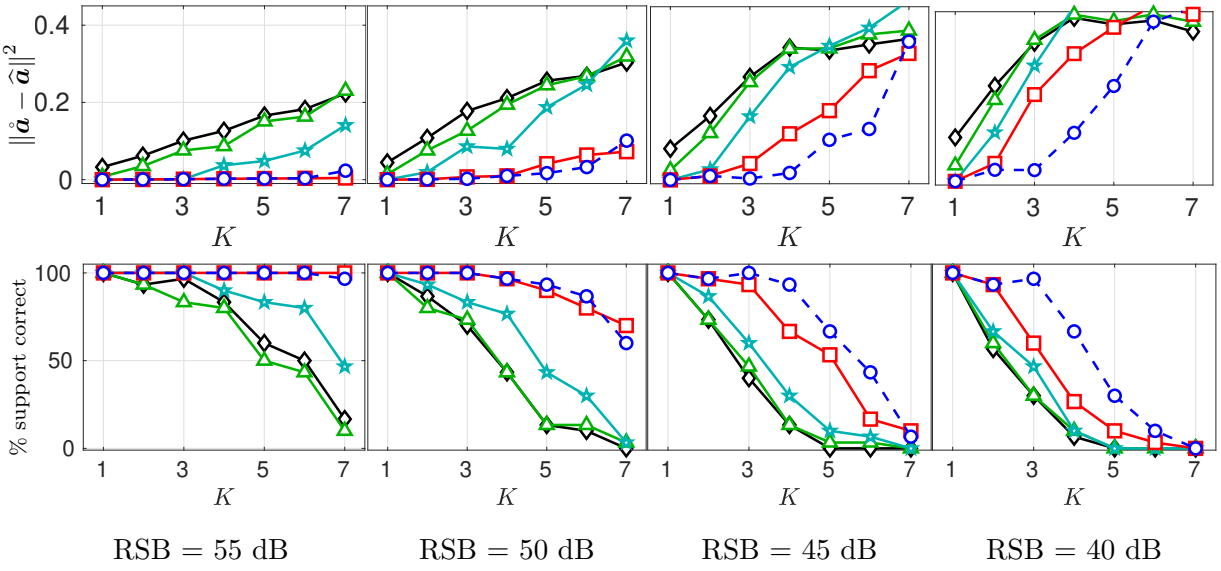


FIGURE 6.11 – Erreur quadratique moyenne entre les vraies abondances $\mathring{\boldsymbol{a}}$ et les abondances estimées $\widehat{\boldsymbol{a}}$ (haut) et taux de détection correcte du support (bas) pour différentes méthodes de démélange, en fonction du niveau de bruit et du nombre de composantes dans le mélange : FCLS ($\diamond$), UNMIX$^{\mathrm{G}\ell_0}$ ($\triangle$), backward ($\star$), UNMIX$^{\ell_0}$= UNMIX$^{\ell_0+\mathrm{G}\ell_0}$ ($\circ$), UNMIX$^{\mathrm{seuil}+\mathrm{G}\ell_0}$ ($\square$).

Le tableau 6.4 montre enfin les temps de calcul moyennés sur ces mêmes problèmes. Nous remarquons tout d'abord que, même s'il requiert la résolution d'un MIP, UNMIX$^{\mathrm{G}\ell_0}$ ne requiert jamais plus de quelques secondes. Les temps de calcul pour les trois approches ℓ_0 (UNMIX$^{\ell_0}$, UNMIX$^{\ell_0+\mathrm{G}\ell_0}$, UNMIX$^{\mathrm{seuil}+\mathrm{G}\ell_0}$) restent modérés tant que le bruit et le nombre de composantes sont faibles. En particulier, la solution produite par UNMIX$^{\ell_0+\mathrm{G}\ell_0}$ (la moins coûteuse des trois) est calculée en moins de 1 000 s sur l'ensemble des problèmes jusqu'à $K = 6$ à fort RSB et jusqu'à $K = 4$ à RSB moyen. Dans nos simulations, UNMIX$^{\ell_0}$et UNMIX$^{\ell_0+\mathrm{G}\ell_0}$ ont toujours fourni la même solution. Cependant, l'ajout des contraintes d'exclusivité de groupe permet de réduire le temps de calcul sur les problèmes les plus difficiles, ce qui semble logique, l'espace de recherche combinatoire étant alors plus contraint. Les problèmes imposant une valeur minimale aux abondances non nulles sont bien plus difficiles à résoudre, ce qui s'explique encore par le fait que l'espace de

12. UNMIX$^{\ell_0+\mathrm{G}\ell_0}$ a toujours fourni la même solution que UNMIX$^{\ell_0}$; les deux estimateurs sont donc associés à une seule courbe sur la figure 6.11

RSB	Méthode	K=3	K=4	K=5	K=6	K=7
55 dB	UNMIX$^{\mathrm{G}\ell_0}$	1.3	2.2	3.3	5.9	9.8
	UNMIX$^{\ell_0}$	1.5	3	11.8	$124^{(3)}$	$468^{(9)}$
	UNMIX$^{\ell_0+\mathrm{G}\ell_0}$	1.6	3.1	7.1	$82^{(1)}$	$320^{(7)}$
	UNMIX$^{\mathrm{seuil}+\mathrm{G}\ell_0}$	2.5	11.5	29	76	$264^{(5)}$
40 dB	UNMIX$^{\mathrm{G}\ell_0}$	1.3	2.2	2.1	3.2	2.9
	UNMIX$^{\ell_0}$	9.2	$107^{(1)}$	$574^{(13)}$	$963^{(27)}$	$1\,000^{(30)}$
	UNMIX$^{\ell_0+\mathrm{G}\ell_0}$	8.1	75	$515^{(11)}$	$929^{(25)}$	$982^{(28)}$
	UNMIX$^{\mathrm{seuil}+\mathrm{G}\ell_0}$	$282^{(6)}$	$426^{(8)}$	$746^{(19)}$	$936^{(26)}$	$997^{(29)}$

TABLE 6.4 – Temps de calcul (en secondes) pour l'optimisation des problèmes MIP, moyennés sur 30 réalisations de chaque problème (moyennes réalisées sur les instances ayant abouti en moins de 1 000 s). Entre parenthèses, nombre d'instances pour lesquelles l'optimisation n'a pas terminé en moins de 1 000 s. Repris de [Ben Mhenni *et al.*, 2018].

recherche est moins contraint que dans les formulations à contrainte ℓ_0 explicite (le nombre maximum de composantes imposé est ici de $1/\tau = 10$).

Notons pour terminer que la méthode *backward*, qui requiert la résolution successive d'un grand nombre (jusqu'à $P - K$) de problèmes de type FCLS, a un coût relativement élevé (une trentaine de secondes) mais fixe pour l'ensemble des problèmes.

6.4 Conclusions

Nous avons montré dans ce chapitre que l'optimisation de critères parcimonieux en norme ℓ_0 pouvait être réalisée de manière exacte, dans des limites de complexité dépendant du nombre de composantes recherchées et de la précision du modèle. L'optimisation du critère des moindres carrés contraint par la norme ℓ_0, la minimisation de la norme ℓ_0 sous contrainte d'une erreur quadratique limitée et celle de la version pénalisée minimisant la somme des deux objectifs ont toutes trois ont été reformulés comme des problèmes de programmation en nombres mixtes, pouvant être résolus par des méthodes d'optimisation globale reposant sur le principe de branch-and-bound. L'évaluation des performances de ces reformulations sur des problèmes simulés de déconvolution et de démélange spectral ont mis en évidence l'intérêt de la recherche du minimum global de critères en norme ℓ_0, fournissant une solution de meilleure qualité que les approches classiques d'estimation parcimonieuse, pour un coût de calcul certes bien plus élevé mais sans commune mesure avec celui d'une recherche combinatoire totale, rendant son utilisation possible pour l'analyse de données de taille modérée et ne requérant pas un résultat « instantané ».

L'essentiel de mes travaux sur l'optimisation en norme ℓ_0 a concerné le développement d'algorithmes de résolution dédiés de type branch-and-bound. Nous avons ainsi montré que l'exploitation des structures mathématiques des problèmes, non prises en compte par les logiciels génériques, permettait de dépasser, parfois très largement, les performances d'un solveur MIP reconnu comme l'un des plus puissants. Ainsi, notre stratégie d'exploration combinatoire exploite la parcimonie de la solution, privilégiant l'activation de variables non nulles dans le parcours de l'arbre de décision. Une telle stratégie n'est pas

implémentable sous CPLEX. Par ailleurs, en interprétant la relaxation continue mise en jeu à chaque nœud de l'arbre de recherche comme un problème d'optimisation en norme ℓ_1, nous avons pu proposer des méthodes de calcul dédiées. Nos algorithmes de branch-and-bound abordent ainsi les trois formulations pour un coût similaire, et au final se passent de l'utilisation de variables binaires et de la reformulation MIP !

En application au démélange spectral, la prise en compte de la parcimonie « au sens ℓ_0 » s'est révélée efficace pour des problèmes de complexité limitée par le nombre de composantes présentes et le niveau de bruit. Lorsqu'elle est calculable, la solution fournie améliore les performances de détection par rapport à la solution classique FCLS, sur des problèmes où la parcimonie en norme ℓ_1 et l'utilisation d'algorithmes gloutons sont peu utiles. Dans nos simulations, la résolution exacte de problèmes en norme ℓ_0 a été possible pour un dictionnaire d'environ 500 spectres de référence, impliquant jusqu'à 6 composantes avec un fort rapport signal sur bruit, et jusqu'à 4 composantes pour un rapport signal sur bruit moyen. Le développement d'algorithmes de résolution dédiés, potentiellement plus efficaces que le solveur CPLEX utilisé ici, est une piste de recherche prioritaire afin d'aborder des problèmes plus gros. Nous y reviendrons dans les perspectives du Chapitre 7. Enfin, la reformulation MIP permet également de traduire de manière exacte des contraintes réputées difficiles (parcimonie ℓ_0 simple ou structurée, seuils minimaux) et souvent prises en compte sous une forme relâchée non exacte. Qui plus est, la conjonction de ces contraintes est également un levier de réduction du temps de calcul, restreignant l'espace de recherche combinatoire.

6.5 Références

[Belotti *et al.*, 2013] Belotti, P., Kirches, C., Leyffer, S., Linderoth, J., Luedtke, J. et Mahajan, A. (2013). Mixed-integer nonlinear optimization. *Acta Numerica*, 22:1–131.

[Ben Mhenni *et al.*, 2019] Ben Mhenni, R., Bourguignon, S. et Ninin, J. (2019). Global optimization for sparse solution of least squares problems. Rapport technique, École Centrale de Nantes.

[Ben Mhenni *et al.*, 2017] Ben Mhenni, R., Bourguignon, S., Ninin, J. et Schmidt, F. (2017). Démélange parcimonieux exact dans une approche supervisée en imagerie hyperspectrale. *In Actes du 26e colloque GRETSI*, Juan-les-Pins.

[Ben Mhenni *et al.*, 2018] Ben Mhenni, R., Bourguignon, S., Ninin, J. et Schmidt, F. (2018). Spectral unmixing with sparsity and structuring constraints. *In Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing (WHISPERS)*, Amsterdam, The Netherlands.

[Bertsimas et Shioda, 2009] Bertsimas, D. et Shioda, R. (2009). Algorithm for cardinality-constrained quadratic optimization. *Computational Optimization and Applications*, 43(1):1–22.

[Bienstock, 1996] Bienstock, D. (1996). Computational study of a family of mixed-integer quadratic programming problems. *Mathematical Programming*, 74(2):121–140.

[Bioucas-Dias *et al.*, 2012] Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P. et Chanussot, J. (2012). Hyperspectral unmixing overview : Geometrical, statistical, and sparse regression-based approaches. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 5(2):354–379.

[Bixby, 2012] Bixby, R. (2012). A brief history of linear and mixed-integer programming computation. *Documenta Mathematica*, Extra Volume : Optimization Stories:107–121.

[Blumensath et Davies, 2008] Blumensath, T. et Davies, M. (2008). Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654.

[Bolte *et al.*, 2014] Bolte, J., Sabach, S. et Teboulle, M. (2014). Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1–2):459–494.

[Bourguignon *et al.*, 2016] Bourguignon, S., Ninin, J., Carfantan, H. et Mongeau, M. (2016). Exact sparse approximation problems via mixed-integer programming : Formulations and computational performance. *IEEE Transactions on Signal Processing*, 64(6):1405–1419.

[Burdakov *et al.*, 2016] Burdakov, O., Kanzow, C. et Schwartz, A. (2016). Mathematical programs with cardinality constraints : Reformulation by complementarity-type conditions and a regularization method. *SIAM Journal on Optimization*, 26(1):397–425.

[Chandrasekaran *et al.*, 2012] Chandrasekaran, V., Recht, B., Parrilo, P. A. et Willsky, A. S. (2012). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849.

[Chen *et al.*, 1998] Chen, S. S., Donoho, D. L. et Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61.

[Clark *et al.*, 2003] Clark, R. N., Swayze, G. A., Wise, R., Livo, K. E., Hoefen, T., Kokaly, R. F. et Sutley, S. J. (2003). USGS digital spectral library splib05a. *US Geological Survey, Digital Data Series*, 231.

[Clason, 2012] Clason, C. (2012). L^∞ fitting for inverse problems with uniform noise. *Inverse Problems*, 28(10):104007.

[Cui *et al.*, 2013] Cui, X. T., Zheng, X. J., Zhu, S. S. et Sun, X. L. (2013). Convex relaxations and MIQCQP reformulations for a class of cardinality-constrained portfolio selection problems. *Journal of Global Optimization*, 56(4):1409–1423.

[Donoho et Tsaig, 2008] Donoho, D. L. et Tsaig, Y. (2008). Fast solution of ℓ_1-norm minimization problems when the solution may be sparse. *IEEE Transactions on Information Theory*, 54:4789–4812.

[Drumetz *et al.*, 2019] Drumetz, L., Meyer, T. R., Chanussot, J., Bertozzi, A. L. et Jutten, C. (2019). Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms. *IEEE Transactions on Image Processing*, pages 1–1.

[Efron *et al.*, 2004] Efron, B., Hastie, T., Johnstone, I. et Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

[Greer, 2012] Greer, J. B. (2012). Sparse demixing of hyperspectral images. *IEEE Transactions on Image Processing*, 21(1):219–228.

[Heinz et Chein-I-Chang, 2001] Heinz, D. C. et Chein-I-Chang (2001). Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 39(3).

[Herrity *et al.*, 2006] Herrity, K., Gilbert, A. et Tropp, J. (2006). Sparse approximation via iterative thresholding. *In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 624–627.

[Iordache *et al.*, 2011] Iordache, M. D., Bioucas-Dias, J. M. et Plaza, A. (2011). Sparse unmixing of hyperspectral data. *IEEE Trans. Geosci. Remote Sens.*, 49(6).

[Iordache *et al.*, 2014] Iordache, M. D., Bioucas-Dias, J. M. et Plaza, A. (2014). Collaborative sparse regression for hyperspectral unmixing. *IEEE Transactions on Geoscience and Remote Sensing*, 52(1):341–354.

[Jokar et Pfetsch, 2008] Jokar, S. et Pfetsch, M. (2008). Exact and approximate sparse solutions of underdetermined linear equations. *SIAM Journal on Scientific Computing*, 31(1):23–44.

[Karahanoglu *et al.*, 2013] Karahanoglu, N. B., Erdogan, H. et Birbil, S. I. (2013). A mixed integer linear programming formulation for the sparse recovery problem in compressed sensing. *In IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5870–5874.

[Keshava et Mustard, 2002] Keshava, N. et Mustard, J. F. (2002). Spectral unmixing. *IEEE signal processing magazine*, 19(1).

[Kowalski et Torrésani, 2009] Kowalski, M. et Torrésani, B. (2009). Sparsity and persistence : mixed norms provide simple signal models with dependent coefficients. *Signal, Image and Video Processing*, 3(3):251–264. nombre de pages : 14.

[Le Thi *et al.*, 2015] Le Thi, H., Le, H. et Pham Dinh, T. (2015). Feature selection in machine learning : an exact penalty approach using a difference of convex function algorithm. *Machine Learning*, 101(1-3):163–186.

[Li *et al.*, 2006] Li, D., Sun, X. et Wang, J. (2006). Optimal lot solution to cardinality constrained mean–variance formulation for portfolio selection. *Mathematical Finance*, 16(1):83–101.

[Liang et Wang, 2017] Liang, X. et Wang, Y. (2017). Homotopy algorithm for box-constrained LASSO and its convergence. *International Journal of Pure and Applied Mathematics*, 112(02):333–340.

[Lu et Zhang, 2013] Lu, Z. et Zhang, Y. (2013). Sparse approximation via penalty decomposition methods. *SIAM Journal on Optimization*, 23(4):2448–2478.

[Marks *et al.*, 1978] Marks, R. J., Wise, G. L., Haldeman, D. S. et Whited, J. L. (1978). Detection in Laplace noise. *IEEE Transactions on Aerospace and Electronic Systems*, 14(6):866–872.

[Meyer *et al.*, 2016] Meyer, T. R., Drumetz, L., Chanussot, J., Bertozzi, A. L. et Jutten, C. (2016). Hyperspectral unmixing with material variability using social sparsity. *In 2016 IEEE International Conference on Image Processing (ICIP)*, pages 2187–2191.

[Miller, 2002] Miller, A. (2002). *Subset selection in regression.* Chapman and Hall/CRC.

[Mohimani *et al.*, 2009] Mohimani, H., Babaie-Zadeh, M. et Jutten, C. (2009). A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm. *IEEE Transactions on Signal Processing*, 57(1):289–301.

[O'Brien *et al.*, 1994] O'Brien, M. S., Sinclair, A. N. et Kramer, S. M. (1994). Recovery of a sparse spike time series by L1 norm deconvolution. *IEEE Transactions on Signal Processing*, 42:3353–3365.

[Osborne *et al.*, 2000] Osborne, M. R., Presnell, B. et Turlach, B. A. D. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*.

[Singer et McCord, 1979] Singer, R. B. et McCord, T. B. (1979). Mars - Large scale mixing of bright and dark surface materials and implications for analysis of spectral reflectance. *In Lunar and Planetary Science Conference Proceedings*, volume 10.

[Soubies *et al.*, 2015] Soubies, E., Blanc-Féraud, L. et Aubert, G. (2015). A continuous exact ℓ_0 penalty (cel0) for least squares regularized problem. *SIAM Journal on Imaging Sciences*, 8(3):1607–1639.

[Soussen *et al.*, 2011] Soussen, C., Idier, J., Brie, D. et Duan, J. (2011). From Bernoulli Gaussian deconvolution to sparse signal restoration. *IEEE Transactions on Signal Processing*, 59(10):4572–4584.

[Thouvenin *et al.*, 2018] Thouvenin, P.-A., Dobigeon, N. et Tourneret, J.-Y. (2018). A hierarchical Bayesian model accounting for endmember variability and abrupt spectral changes to unmix multitemporal hyperspectral images. *IEEE Transactions on Computational Imaging*, 4(1):32–45.

[Wolsey, 1998] Wolsey, L. A. (1998). *Integer Programming.* Wiley, New York, NY, USA.

[Yuan et Yin, 2006] Yuan, M. et Yin, L. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 68(1):49–67.

[Zare et Ho, 2014] Zare, A. et Ho, K. C. (2014). Endmember variability in hyperspectral analysis : Addressing spectral variability during spectral unmixing. *IEEE Signal Processing Magazine*, 31(1):95–104.

Troisième partie

Perspectives et projets de recherche

Chapitre 7

Quelques perspectives de travaux à venir

Ce dernier chapitre évoque quelques perspectives de travaux de recherche en lien avec les thèmes développés aux Chapitres 4 à 6. Chacune de ses sections évoque ainsi des pistes de recherches pouvant déboucher, *a minima*, sur un futur sujet de thèse de doctorat.

7.1 Imagerie hyperspectrale et radio-astronomie

Si mes travaux en imagerie hyperspectrale astronomique ont surtout été menés pendant mon séjour post-doctoral à l'Observatoire de la Côte d'Azur, je m'intéresse à nouveau, depuis l'été 2018, au **traitement de données hyperspectrales, cette fois en radio-astronomie**, dans le cadre du projet Orion-B. Ce projet, porté par des astronomes de l'Institut de Radioastronomie Millimétrique (IRAM, Grenoble) et du Laboratoire d'Études du Rayonnement et de la Matière en Astrophysique (LERMA, Observatoire de Paris), vise à comprendre le processus de formation d'étoiles, *via* des mesures spectroscopiques (dans le domaine des ondes radio) de nuages de gaz entourant les zones de formation d'étoiles situés dans le nuage moléculaire d'Orion. Cette collaboration, initiée par un Projet Exploratoire Premier Soutien (PEPS) du CNRS « Astro-informatique », s'est concrétisée au printemps 2019 par mon appartenance officielle au consortium Orion-B.

À la différence de spectro-imageurs comme MUSE, opérant essentiellement dans le domaine visible et proche infra-rouge du spectre électromagnétique , l'acquisition de spectres d'émission dans le domaine radio (longueurs d'onde millimétriques et centimétriques) s'effectue ici séquentiellement, le radiotélescope pointant successivement des points voisins de l'espace. Malgré ces différences dans le processus d'acquisition, un certain nombre de points communs ont pu être identifiés par rapport aux travaux présentés au Chapitre 4 :

- ici aussi, l'objectif scientifique principal réside dans la caractérisation *spectrale* des sources, où la position et la largeur des raies d'émission est inconnue mais peut être contrainte par des connaissances physiques ;
- les données, acquises à travers l'atmosphère, subissent des dégradations par l'étalement d'une réponse instrumentale spectrale et par un fort bruit de niveau variable en longueur d'onde. Par ailleurs, les spectres contenant jusqu'à 200 000 canaux spectraux, la plupart des points de mesure ne contiennent que du bruit.

J'envisage donc d'y développer des méthodes dans la continuité de mes recherches, basées

à nouveau sur une hypothèse de parcimonie des spectres dans un dictionnaire adapté, pour la détection de raies d'émission ou d'absorption.

Comme pour les données de type MUSE, les spectres mesurés sont contaminés par l'émission parasite du fond du ciel. Pour les données d'Orion-B, ce problème est actuellement abordé *via* la soustraction préalable de cette émission, par un ajustement polynomial ou polynomial par morceaux. Il est clair qu'une approche conjointe, visant à estimer simultanément cette émission et les composantes d'intérêt, s'avérerait plus robuste en termes de détection de raies spectrales. Dans mes travaux précédents, par simplicité et par homogénéité avec le reste du modèle, cette composante continue était modélisée par une représentation parcimonieuse, dans un dictionnaire composé d'oscillations de basse fréquence. Il me semble souhaitable ici d'introduire un modèle plus conforme aux pratiques des astronomes pour **l'estimation conjointe des sources et du fond du ciel**. En particulier, avec les notations du Chapitre 4 (voir le § 4.2.2), un modèle polynomial de degré faible pourrait être introduit : $\boldsymbol{s} = \mathbf{D}\boldsymbol{u} + \mathbf{P}\boldsymbol{c}$, où le dictionnaire $\mathbf{D}$ exclut la partie continue et $\mathbf{P}$ représente les variations polynomiales ; le vecteur parcimonieux $\boldsymbol{u}$ et les coefficients polynomiaux $\boldsymbol{c}$ (non parcimonieux) seraient alors estimés conjointement. Afin de limiter la complexité calculatoire (les spectres comportant ici plusieurs dizaines de milliers de longueurs d'onde), il serait aussi opportun d'envisager des méthodes de *screening* [Xiang *et al.*, 2011], visant à réduire la dimension du problème en enlevant des atomes du dictionnaire pour lesquels on peut garantir, par un calcul à faible coût, qu'ils ne seront pas activés dans la décomposition. L'exploitation de grandes quantités de données pose également des **questions statistiques de détection en grande dimension**, qui se posaient déjà à l'époque du projet ANR DAHLIA [Paris, 2013].

À plus long terme, et toujours sous l'hypothèse d'un modèle de décomposition parcimonieuse des spectres, il me semble intéressant de développer des méthodes visant à la **détection et caractérisation de sources astrophysiques prenant en compte leur étendue spatiale**.

Une première approche relève de la **classification / segmentation d'images**, où une mesure de similarité entre pixels proches est définie dans l'espace des coefficients de décomposition plutôt que dans l'espace des spectres bruts, dominés par le bruit. Les pixels voisins et pour lesquels les spectres admettent des décompositions proches sont alors agrégés pour former un objet, sans imposer de contrainte spatiale sur sa forme. Des travaux dans cette direction ont été initiés pendant mon post-doctorat [Bourguignon *et al.*, 2012] – avec un dictionnaire « générique » composé, suivant les travaux de Donoho et Huo [Donoho et Huo, 2001], d'une matrice d'impulsions pour les raies et d'une matrice de Transformée en Cosinus Discrète pour la partie continue. Des résultats intéressants ont été obtenus, permettant d'améliorer les performances de détection de raies d'émission de faible amplitude. Il serait utile à mon avis d'envisager une méthodologie similaire pour les données Orion-B, reposant sur une décomposition des spectres dans un dictionnaire plus approprié.

Une deuxième approche pourrait considérer **un modèle plus contraint, séparable**, où la contribution d'un objet astrophysique est représentée comme le produit d'une carte de distribution spatiale par un spectre d'émission : $o(\boldsymbol{r}, \lambda) = I(\boldsymbol{r})\, s(\lambda)$. Les travaux de thèse de Céline Meillier [Meillier, 2015], pour des données hyperspectrales de type MUSE, visent à reconstruire de tels objets, en imposant surtout un profil spatial elliptique des galaxies. À l'inverse, il me semblerait plus opportun ici de relâcher les contraintes de forme spatiale (les nuages de gaz pouvant être assez diffus) et de travailler à nouveau sur des modèles contraints selon la dimension spectrale.

Un binôme d'étudiants de l'École Centrale de Nantes a commencé à travailler sur des données Orion-B lors d'un projet étudiant début 2019. Si les avancées en résultant sont mineures, elles m'ont néanmoins permis d'avancer dans la compréhension des spécificités des données et d'échanger avec le responsable principal du projet. La définition d'un dictionnaire approprié et le développement de méthodes parcimonieuses associées, prenant en compte un modèle polynomial de la ligne de base, pourraient faire l'objet d'un stage de Master co-encadré avec Jérôme Pety de l'IRAM, pour un.e étudiant.e issu.e d'une formation en traitement du signal ou en astrophysique. Selon les avancées de ce stage et l'intérêt du consortium Orion-B pour ces approches et pour des méthodes prenant en compte une dimension spatiale des objets observés, le développement de méthodes « spatiales-spectrales » pourrait donner matière à une thèse de doctorat.

7.2 CND ultrasonore : modèles plus complexes et co-conception

Mes différents travaux relatifs au CND ultrasonore présentés au Chapitre 5 ont visé à raffiner différents modèles de signaux et à exploiter ces modèles dans des procédures de reconstruction de séquences monodimensionnelles, ou de cartographies bidimensionnelles, de la réflectivité ultrasonore. Différents développements ont été proposés, de manière séparée, qui pourraient avantageusement être mis en commun dans la construction de nouvelles méthodes d'imagerie ultrasonore.

Une première piste concerne **la prise en compte de modèles de propagation acoustique** (atténuation fréquentielle et dispersion associée), que nous avons étudiés dans le cas monodimensionnel, **dans les méthodes de reconstruction d'image** actuellement développées dans la thèse de Nans Laroche, afin de pouvoir imager précisément des milieux plus complexes comme les matériaux polymères. La prise en compte de l'atténuation et de la dispersion ne devrait pas poser de difficulté majeure sur le plan méthodologique : les modèles directs introduits pour décrire les données multi-statiques ou FMC *Full Matrix Capture* portent déjà sur les signaux temporels correspondant à un couple émetteur-récepteur. Elle risque cependant d'alourdir les algorithmes de calcul.

Toujours dans le cas de données multi-statiques, il me semble également intéressant de raffiner le modèle d'acquisition par la prise en compte d'une **réponse spécifique pour chaque couple émetteur-récepteur**. Les travaux de Holmes *et al.* [Holmes *et al.*, 2005] considèrent par exemple le diagramme de directivité de chaque transducteur pour pondérer la contribution de chaque signal dans une procédure linéaire de reconstruction d'image de type TFM. Une telle pondération n'est pas prise en compte dans les méthodes d'inversion actuellement développées, qui donnent probablement trop d'importance aux données de moins bonne qualité acquises entre deux transducteurs très éloignés. Il serait donc intéressant d'envisager de manière similaire une pondération des différentes contributions dans les données FMC. Pour aller plus loin, on pourrait chercher à introduire la **dépendance en fréquence du diagramme de directivité** des capteurs, laquelle peut être estimée *via* une étape de calibrage ne dépendant que de la géométrie d'inspection.

Un volet complémentaire concerne l'**estimation de la réponse des transducteurs**. Les performances des méthodes que nous construisons sont en effet fortement conditionnées par la connaissance de cette forme d'onde et, en pratique, une opération de calibrage n'est pas toujours possible. Sur le plan méthodologique, il serait intéressant d'aborder **conjointement** les problèmes de déconvolution d'échogrammes ou de reconstruction d'images et de l'estimation de cette PSF. Dans le cas particulier de la déconvolution d'échos régulièrement espacés, une modélisation paramétrique de la forme d'onde s'est

avérée efficace. On pourrait donc envisager le développement de **méthodes aveugles ou myopes**, optimisant conjointement les séquences de réflectivité (mono ou bidimensionnelles) et les paramètres de la forme d'onde, pour des problèmes de reconstruction plus généraux.

Ces perspectives s'inscrivent naturellement dans le programme de la thèse en cours de Nans Laroche (actuellement en fin de 2ème année). La partie non abordée pourrait constituer le point de départ d'une nouvelle thèse, laquelle s'inscrirait dans un périmètre plus large où je souhaiterais aborder l'imagerie ultrasonore sous l'angle de la **co-conception acquisition-traitement**. Tout comme dans le domaine biomédical, les appareils d'acquisition ultrasonore pour le CND évoluent rapidement vers des dispositifs configurables très souplement, permettant par exemple de focaliser successivement en différents points du matériau inspecté par la synthèse de lois de retard appropriées entre les différents éléments. Une réflexion portant conjointement sur l'optimisation des séquences d'acquisition (visant à diminuer leur nombre) et le développement de méthodes de reconstruction avancées me semble un levier très prometteur afin de réduire le temps d'acquisition et la complexité des algorithmes de calcul associés.

Enfin, si les méthodes proposées jusqu'à présent ont été testées sur des problèmes simulés et, pour certaines, sur des données expérimentales, celles-ci restent encore à l'échelle d'expériences de laboratoire. Il me semble important de garder en tête l'objectif de leur évaluation et leur utilisation dans un contexte industriel, même si ce point ne fait pas partie de mes perspectives directes en termes de recherche.

7.3 Démélange spectral et parcimonie

L'exploitation de reformulations en programmes mixtes en nombres entiers (MIP) pour le démélange spectral parcimonieux, introduisant des variables binaires encodant la présence de chaque spectre de référence dans le mélange, a permis de traduire de manière exacte la parcimonie du vecteur d'abondances, là où les approches basées sur une relaxation continue de la norme ℓ_0 s'avèrent délicates. Jusqu'à présent, la résolution numérique des différentes formes de MIP pour le démélange spectral a été réalisée avec le solveur commercial CPLEX. Le **développement d'algorithmes branch-and-bound spécifiques au problème de démélange** est bien entendu une priorité dans mes recherches à venir, devant permettre de réduire le coût calculatoire et d'aborder des problèmes plus difficiles. Les algorithmes proposés dans ce manuscrit ne s'appliquent pas directement à ce cadre, où les contraintes supplémentaires de positivité et de somme à 1 des abondances rendent inutile la relaxation continue de la norme ℓ_0 en norme ℓ_1 : avec les notations de la Section 6.3, la relaxation de la contrainte $\|\boldsymbol{a}\|_0 \leq K$ s'écrit $\|\boldsymbol{a}\|_1 \leq K$, laquelle est sans effet sur le problème de relaxation puisque $\|\boldsymbol{a}\|_1 = \sum_{p=1}^{P} a_p = 1$. Qui plus est, la prise en compte de contraintes structurantes, se manifestant typiquement par des contraintes linéaires d'égalité ou d'inégalité sur les variables binaires, produit des problèmes de relaxation continue impliquant des normes mixtes, des contraintes de borne et éventuellement des contraintes linéaires, pour lesquels il sera possible de développer des algorithmes adaptés.

Par ailleurs, le cadre de modélisation sous forme de MIP se révèle particulièrement propice à l'**intégration d'autres types de contraintes réputées difficiles**. Récemment, divers travaux ont cherché à prendre le phénomène de variabilité spectrale (signifiant que les spectres de référence peuvent varier entre les différents pixels d'une même scène observée) à base d'approches sous-optimales. Ainsi, dans [Drumetz *et al.*, 2019], une relaxation de type *group-LASSO* basée sur une formulation en norme mixte ℓ_{21} cherche

à imposer un faible nombre de groupes contenant les spectres actifs, tout en laissant la possibilité d'un mélange dense à l'intérieur de chaque groupe. Ce problème se prête tout à fait à une **formulation exacte** à partir de variables binaires. Au-delà de cet exemple, je suis convaincu que des reformulations MIP pourront permettre de traduire de manière exacte différents types de contraintes complexes sur l'espace des solutions, qui restent à formaliser à ce stade. J'envisage ainsi de développer un réseau de collaboration à l'échelle régionale entre Pays de la Loire et Bretagne autour du traitement de données hyperspectrales exploitant la parcimonie, incluant en particulier Lucas Drumetz, spécialiste du traitement de données d'imagerie hyperspectrale au Lab-STICC à Brest, Jordan Ninin, du même laboratoire et avec qui je travaille déjà sur les aspects algorithmiques de la résolution de MIP et Jérémy Cohen de l'IRISA à Rennes, dont les recherches concernent également la factorisation en matrices non-négatives, le démélange spectral et les algorithmes d'optimisation combinatoire.

La construction d'algorithmes d'optimisation exacte dans ce contexte et leur application à des données réelles d'imagerie hyperspectrale pourrait être à la base d'un nouveau sujet de thèse dans le cadre collaboratif mentionné ci-dessus, ou d'un sujet de post-doctorat consécutif à la thèse de Ramzi Ben Mhenni, dont la fin est prévue début 2020.

7.4 Optimisation globale pour la résolution de problèmes en norme ℓ_0

Au moment de l'écriture de ce manuscrit, le développement d'algorithmes d'optimisation de critères impliquant la norme ℓ_0 est le thème de recherche sur lequel je m'investis le plus au niveau méthodologique ; il est donc particulièrement fourni en perspectives.

7.4.1 Pistes d'amélioration des algorithmes branch-and-bound.

J'envisage d'aborder plusieurs pistes dans l'objectif de réduire davantage le temps de calcul de la résolution exacte de problèmes en norme ℓ_0 et de pouvoir aborder des problèmes de plus grande taille. Un premier axe concerne le développement d'autres algorithmes de **calcul des relaxations** à chaque nœud d'une procédure branch-and-bound (voir la Section 6.2 pour le fonctionnement des algorithmes branch-and-bound). Nous avons montré que la relaxation continue des variables binaires s'apparentait à un problème impliquant la norme ℓ_1 sur une partie des variables et des contraintes de borne. Le choix d'un algorithme homotopique a permis d'aborder avec la même efficacité les relaxations mises en jeu dans les trois formulations ($\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$), et s'est révélé d'autant plus efficace que la solution est parcimonieuse. En revanche, cet algorithme requiert une initialisation mettant à zéro toutes les composantes impliquées dans la norme ℓ_1, potentiellement loin de la solution. Nous avons commencé à construire des **algorithmes basés sur le principe des contraintes actives** ou des ensembles actifs (*active sets*) [Osborne *et al.*, 2000, Lee *et al.*, 2007], qui convergent également en un nombre fini d'itérations, ce qui est important pour garantir la validité et la qualité de la relaxation. À l'inverse des méthodes homotopiques, ces algorithmes peuvent être initialisés en tout point et peuvent donc bénéficier du démarrage à chaud. En particulier, la relaxation calculée à un nœud donné de l'arbre de recherche ne diffère que d'une variable par rapport à celle évaluée au nœud parent. L'exploitation des calculs réalisés au nœud parent doit alors permettre d'initialiser efficacement la relaxation courante et donc de réduire le nombre d'itérations nécessaires à son évaluation. Des premiers résultats ont été obtenus sur la forme pénalisée $\mathcal{P}_{2+0}$, qui

ont été récemment publiés à la conférence du Gretsi 2019 [Ben Mhenni *et al.*, 2019] et un post-doctorant travaille actuellement à leur utilisation pour les relaxations impliquées dans la résolution du problème $\mathcal{P}_{2/0}$. Par ailleurs, une alternative à la résolution exacte des problèmes en norme ℓ_1 pourrait consister en l'utilisation d'**algorithmes primaux-duaux** [Chambolle et Pock, 2011]. Ces algorithmes alternent entre des étapes de minimisation du problème initial, dit primal, et de maximisation du problème dual, fournissant un encadrement de la valeur optimale, vers laquelle les itérés convergent asymptotiquement. Les itérés calculés sur le problème dual fournissent donc également des bornes inférieures valides du problème de relaxation continue qui s'améliorent au fil des itérations, permettant de tronquer les itérations sous l'angle d'un compromis entre la qualité de la borne inférieure obtenue et son coût calculatoire. L'extension d'algorithmes primaux-duaux en norme ℓ_1 existants à notre contexte (incluant des variables libres en dehors de la norme ℓ_1 et des contraintes de borne) ne devrait pas poser de très fort verrou méthodologique, et pourrait s'avérer plus efficace que les approches exactes (homotopie, ensembles actifs) développées jusqu'ici.

Par ailleurs, sur des problèmes particuliers de type $\mathcal{P}_{2/0}$, il a été montré que la **relaxation lagrangienne** est meilleure que la relaxation continue du problème MIP [Shaw *et al.*, 2008], dans le sens où elle fournit une borne inférieure plus élevée. C'est donc un levier supplémentaire pour réduire le nombre de nœuds explorés dans l'algorithme branch-and-bound. Son coût calculatoire par des algorithmes génériques s'est cependant montré prohibitif lors d'une première étude menée dans le stage de Master de Ghandy Ajib que j'ai encadré en 2018. Le développement d'algorithmes spécifiques pour ce calcul est donc également une perspective d'intérêt.

Enfin, si les algorithmes développés s'avèrent compétitifs par rapport à un solveur générique de problèmes MIP, c'est parce que nous avons proposé des interprétations et des choix dédiés au problème ℓ_0 (relaxations continues, stratégies de choix de variable et de branchement). Les logiciels commerciaux incorporent également de nombreux éléments supplémentaires que nous n'avons pas exploités, sur lesquels il serait également intéressant de travailler, en recherchant ici aussi des formes spécifiques aux problèmes en norme ℓ_0. Je pense en particulier aux **méthodes de coupe** [Wolsey, 1998] qui raffinent les relaxations à chaque nœud par la résolution de problèmes plus contraints facilement calculables. La mise en œuvre d'**heuristiques d'exploration combinatoire locale** plus complexes, reprenant par exemple le principe des algorithmes gloutons, doit également permettre de trouver plus rapidement des solutions réalisables « candidates » à l'optimum global, afin de trouver plus rapidement l'optimum global et de diminuer le nombre de nœuds explorés. Enfin, la **parallélisation** de l'algorithme sur une architecture multi-cœurs est aussi probablement un vecteur d'amélioration significative des temps de calcul.

7.4.2 Résolution d'autres problèmes impliquant la norme ℓ_0

S'appuyant sur le savoir-faire développé pour la résolution de problèmes de type $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ et $\mathcal{P}_{2+0}$, je compte également construire des algorithmes abordant d'autres problèmes d'optimisation faisant intervenir la norme ℓ_0.

L'exploitation de reformulations MIP et le développement d'algorithmes dédiés pourraient être étudiés dans un contexte de **représentation parcimonieuse,** *i.e.*, **la minimisation de la norme ℓ_0 sous des contraintes égalité $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$**. Si le contexte est différent de celui des problèmes inverses faisant l'objet de ce manuscrit, dans lesquels la présence d'un terme de perturbation dans le modèle est fondamentale (bruit de mesure, erreur de modèle), ces problèmes sont rencontrés dans de nombreuses applica-

tions : échantillonnage compressif, *inpainting*, désécrêtage de signaux audio, ... (voir par exemple [Elad, 2010]). Les analyses de temps de calcul présentées au Chapitre 6 laissent penser que, pour un modèle exact, des problèmes de plus grande taille pourraient être abordés efficacement.

Par ailleurs, nous avons montré que les solutions intermédiaires obtenues en limitant le temps de résolution des MIP pouvaient s'avérer plus satisfaisantes que les solutions parcimonieuses standard (voir par exemple le § 6.1.2 et la figure 6.3). Je pense qu'il y a donc également des recherches intéressantes à mener dans la construction d'algorithmes sur le principe branch-and-bound, réalisant une **exploration partielle de l'arbre de recherche combinatoire**, obtenant alors un compromis intéressant entre la qualité de solution obtenue et le coût de calcul, à l'instar de l'approche de type $A^\star$-OMP proposée dans [Karahanoglu et Erdogan, 2012].

D'un point de vue pratique, pour de nombreux problèmes d'approximation parcimonieuse, la formulation $\mathcal{P}_{0/2}$, contrainte par l'erreur, nous semble la plus appropriée, le réglage d'une tolérance sur l'erreur d'approximation étant souvent plus facile que celui du nombre de composantes non nulles ou du paramètre de pénalisation. Cette formulation est cependant moins bien posée en ceci que le minimum, discret, peut être atteint sur différentes configurations du support, en particulier lorsque l'erreur d'approximation est élevée [Bourguignon *et al.*, 2016]. Un algorithme branch-and-bound dédié pourrait alors permettre d'obtenir **l'ensemble des solutions les plus parcimonieuses satisfaisant la contrainte d'erreur**, là où la version actuelle ne permet que d'obtenir une des solutions minimisant $\mathcal{P}_{0/2}$. Notons encore que, parmi les formulations reposant sur le réglage de l'erreur d'approximation, le **calcul de la « meilleure » solution parmi les plus parcimonieuses**, définie par le problème d'optimisation :

$$\min_{\boldsymbol{x}\in\mathcal{D}_{0/2}(\alpha)} \frac{1}{2}\left\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\right\|_2^2, \quad \text{où} \quad \mathcal{D}_{0/2}(\alpha) := \operatorname{Arg\,min}_{\boldsymbol{x}} \left\|\boldsymbol{x}\right\|_0 \quad \text{s. c.} \quad \frac{1}{2}\left\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\right\|_2^2 \leq \epsilon$$

pourrait également être réalisé par un algorithme branch-and-bound spécifique.

Enfin, aborder le problème d'estimation parcimonieuse sous l'angle de l'**optimisation multi-objectif** permettrait de contourner la problématique du réglage du paramètre de chaque formulation. Il s'agit alors de fournir l'ensemble des solutions définissant la frontière de Pareto du problème bi-objectif :

$$\min_{\boldsymbol{x}} \left\{ \frac{1}{2} \left\|\boldsymbol{y}-\mathbf{A}\boldsymbol{x}\right\|^2, \left\|\boldsymbol{x}\right\|_0 \right\},$$

i.e., telles que, pour chaque valeur possible d'un des deux objectifs, l'autre objectif soit minimal. Ce problème s'y prête bien puisque la norme ℓ_0 est discrète ; il est alors possible d'énumérer l'ensemble (fini) des valeurs prises par cette dernière. Partant de la structure de l'arbre de recherche parcouru lors de la résolution du problème $\mathcal{P}_{2/0}$ pour une valeur de la contrainte : $\left\|\boldsymbol{x}\right\|_0 \leq K_{\max}$, les problèmes $\mathcal{P}_{2/0}$ avec $\left\|\boldsymbol{x}\right\|_0 \leq K$ pour $K < K_{\max}$ doivent alors pouvoir être résolus en exploitant un sous-arbre extrait du précédent. L'ensemble des problèmes $\mathcal{P}_{2/0}$ pour $K \leq K_{\max}$ devrait alors être résolu pour un coût de calcul de l'ordre de la résolution du plus difficile, *i.e.,* $K = K_{\max}$.

7.5 Références

[Ben Mhenni *et al.*, 2019] Ben Mhenni, R., Bourguignon, S., Mongeau, M., Ninin, J. et Carfantan, H. (2019). Algorithme *branch-and-bound* pour l'optimisation exacte en norme ℓ_0. *In Actes du 27e colloque GRETSI.*

[Bourguignon *et al.*, 2012] Bourguignon, S., Mary, D. et Slezak, É. (2012). Processing MUSE hyperspectral data : Denoising, deconvolution and detection of astrophysical sources. *Statistical Methodology*, 9(1):32–43.

[Bourguignon *et al.*, 2016] Bourguignon, S., Ninin, J., Carfantan, H. et Mongeau, M. (2016). Exact sparse approximation problems via mixed-integer programming : Formulations and computational performance. *IEEE Transactions on Signal Processing*, 64(6):1405–1419.

[Chambolle et Pock, 2011] Chambolle, A. et Pock, T. (2011). A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145.

[Donoho et Huo, 2001] Donoho, D. L. et Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inf. Theory*, 47(7):2845–2862.

[Drumetz *et al.*, 2019] Drumetz, L., Meyer, T. R., Chanussot, J., Bertozzi, A. L. et Jutten, C. (2019). Hyperspectral image unmixing with endmember bundles and group sparsity inducing mixed norms. *IEEE Transactions on Image Processing*, pages 1–1.

[Elad, 2010] Elad, M. (2010). *Sparse and Redundant Representations. From Theory to Applications in Signal and Image Processing.* Springer-Verlag New York.

[Holmes *et al.*, 2005] Holmes, C., Drinkwater, B. W. et Wilcox, P. D. (2005). Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation. *NDT & E International*, 38(8):701 – 711.

[Karahanoglu et Erdogan, 2012] Karahanoglu, N. B. et Erdogan, H. (2012). A* orthogonal matching pursuit : Best-first search for compressed sensing signal recovery. *Digital Signal Processing*, 22(4):555 – 568.

[Lee *et al.*, 2007] Lee, H., Battle, A., Raina, R. et Ng, A. Y. (2007). Efficient sparse coding algorithms. *In Advances in Neural Information Processing Systems*, pages 801–808.

[Meillier, 2015] Meillier, C. (2015). *Détection de sources quasi-ponctuelles dans des champs de données massifs* . Thèse de doctorat, Université Grenoble Alpes.

[Osborne *et al.*, 2000] Osborne, M. R., Presnell, B. et Turlach, B. A. D. (2000). A new approach to variable selection in least squares problems. *IMA Journal of Numerical Analysis*.

[Paris, 2013] Paris, S. (2013). *Méthodes de détection parcimonieuses pour signaux faibles dans du bruit : application à des données hyperspectrales de type astrophysique.* Thèse de doctorat, Université Nice Sophia Antipolis.

[Shaw *et al.*, 2008] Shaw, D. X., Liu, S. et Kopman, L. (2008). Lagrangian relaxation procedure for cardinality-constrained portfolio optimization. *Optimization Methods and Software*, 23(3):411–420.

[Wolsey, 1998] Wolsey, L. A. (1998). *Integer Programming.* Wiley, New York, NY, USA.

[Xiang *et al.*, 2011] Xiang, Z., Xu, H. et Ramadge, P. (2011). Learning sparse representations of high dimensional data on large scale dictionaries. *In Advances in Neural Information Processing Systems (NIPS).*

Annexes : sélection de publications

Annexe A

Global Optimization for Sparse Solution of Least Squares Problems.

R. Ben Mhenni, S. Bourguignon et J. Ninin : Global optimization for sparse solution of least squares problems. *Soumis à Optimization Methods and Software*, avril 2019.

To appear in *Optimization Methods & Software*
Vol. 00, No. 00, Month 20XX, 1–20

Global Optimization for Sparse Solution of Least Squares Problems

Ramzi BEN MHENNI[a*] , Sebastien BOURGUIGNON[a] and Jordan NININ[b]

[a]*LS2N, CNRS UMR 6004, École Centrale de Nantes, Nantes, France;*
[b]*Lab-STICC, CNRS UMR 6285, ENSTA Bretagne, Brest, France*

(*Received 00 Month 20XX; final version received 00 Month 20XX*)

Finding solutions to least-squares problems with low cardinality has found many applications, including cardinality-constrained portfolio optimization, subset selection in statistics, and many sparsity-enhancing inverse problems in signal processing. This problem is NP-hard, and most works from a global optimization perspective consider a mixed integer programming (MIP) reformulation with binary variables, whose resolution is performed *via* branch-and-bound methods. We propose dedicated branch-and-bound algorithms for three possible formulations: cardinality-constrained and cardinality-penalized least-squares, and cardinality minimization under quadratic constraints. We show that the continuous relaxation problems involved in each node of the search tree are ℓ_1-norm-based optimization problems. A dedicated algorithm is built, based on the homotopy continuation principle, which efficiently computes the relaxed solutions for the three kinds of problems. The performance of the resulting global optimization procedure is then shown to compete with or improve over the CPLEX MIP solver, especially for problems involving quadratic constraints. The proposed strategies are able to solve problems involving 500 to 1 000 unknowns in less than 1 000 seconds, for which CPLEX generally fails.

Keywords: Sparse approximation; Subset selection; Cardinality constraint; Branch-and-bound; Continuous relaxation; Homotopy continuation.

1. Introduction

We are interested in solving optimization problems mixing a quadratic data adjustment term and a sparsity measure. Such problems arise in many application fields, among which portfolio optimization [2, 7, 17, 25], sparse regularization for inverse problems [6, 12, 19, 26, 29] or compressed sensing [5, 13], and variable or subset selection in statistics [10, 20, 23, 27]. In operations research, many works addressed the cardinality-constrained problem [2, 3, 7, 17, 25]:

$$\mathcal{P}_{2/0} : \min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \text{ subject to (s.t.) } \|\boldsymbol{x}\|_0 \leq K,$$

where $\boldsymbol{y} \in \mathbb{R}^m$ and $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with usually $n > m$,$\|\boldsymbol{x}\|_0 = \text{Card}\{i|x_i \neq 0\}$ (which will be called the ℓ_0-"norm" in this paper) and $K \in \mathbb{N}$ is a given cardinality that is fixed *a priori*. In some applications, however, one may prefer solving the error-constrained

*CONTACT R. BEN MHENNI. Author. Email: ramzi.benmhenni@ls2n.fr

problem [20, 21, 28]:

$$\mathcal{P}_{0/2} : \min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{x}\|_0 \text{ s.t. } \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \epsilon.$$

Such a formulation may be more relevant in signal processing and statistics, where parameter $\epsilon \geq 0$ controls the approximation level and can be tuned according to prior knowledge about the data. One is then interested in finding the sparsest approximation compatible with some given noise level (or some prediction accuracy). Finally, the penalized problem:

$$\mathcal{P}_{2+0} : \min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu\|\boldsymbol{x}\|_0,$$

where $\mu > 0$ trades off between approximation error and sparsity, is also encountered in the field of inverse problems, *e.g.,* for Geophysics [19] or ultrasonic non-destructive testing [22, 29]. In the Bayesian statistical framework, the ℓ_0-norm penalization term corresponds to a Bernoulli-Gaussian prior assumption about the unknown components in $\boldsymbol{x}$, and parameter μ then depends on both the noise level and the expected rate of non-zero values in $\boldsymbol{x}$ [26].

Optimization of the three problems $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ and $\mathcal{P}_{2+0}$, which are NP-hard due to the discrete nature of the cardinality term [3, 21], is therefore of interest. Let us remark that, due to their non-convexity, the three formulations are not equivalent. Global optimization algorithms have been proposed for solving $\mathcal{P}_{2/0}$, which is a cardinality-constrained quadratic program, in the context of sparse portfolio selections and subset selection problems. To our knowledge, Bienstock was the first to propose a specific branch-and-cut algorithm for such problems, with additional non-negativity constraints [3]. Continuous relaxation problems involved in each iteration were solved *via* a specific convex quadratic programming algorithm. In [2], Bertsimas and Shioda extended this work using Lemke's pivoting method to solve the continuous relaxation. When the matrix involved in the quadratic term is the sum of a diagonal positive matrix and a positive definite one, branch-and-bound techniques using perspective reformulation [14], Lagrangian relaxation [7, 17, 25] or geometric approaches [15], were shown to give tighter lower bounds than the continuous relaxation. In many problems of the form $\mathcal{P}_{2/0}$, however, variables involved in the columns of the matrix $\boldsymbol{A}$ (explaining variables in statistics, dictionary atoms in sparse approximation) are often highly correlated; then, the matrix $\boldsymbol{A}^T\boldsymbol{A}$ is ill-conditioned and such decomposition is not possible.

Up to our knowledge, problems $\mathcal{P}_{0/2}$ and $\mathcal{P}_{2+0}$ were never addressed with dedicated methods from a global optimization perspective. Many works in signal processing and statistics proposed local strategies, either based on greedy local search algorithms or on the substitution of the ℓ_0 norm by the ℓ_1 norm, which can be applied to the three problem formulations (see for example [28] and references therein). In very particular cases as those addressed within the compressed sensing theory, such approaches may solve the ℓ_0-norm-based problems [13], but optimality conditions are very restrictive.

In [4], the three problems were reformulated as mixed-integer programs (MIPs), and their resolution was performed with the CPLEX solver. It was experimentally shown that the exact resolution of difficult sparse approximation problems was possible in practice for moderate-size problems. In this paper, we propose dedicated branch-and-bound resolution strategies for such problems. Following the works in [2, 3], our motivation lies in the fact that sparsity-enhancing least-squares problems are very specific MIPs, that

could be advantageously solved by dedicated implementation. In particular, we show that all continuous relaxation problems involved in the resolution are particular forms of problems involving the ℓ_1 norm, for which we build a dedicated continuous optimization based on the homotopy principle [9, 10, 23]. Our algorithm is able to solve the relaxation problems involved in any of the three problems $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ and $\mathcal{P}_{2+0}$. Therefore, these three problems can be solved exactly by our procedure, with similar computational burden.

The paper is organized as follows. In Section 2, we study the structure of continuous relaxation problems involved in each node of a branch-and-bound algorithm solving $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ and $\mathcal{P}_{2+0}$, which are reformulated as ℓ_1-norm-based problems. A dedicated optimization strategy is then built in Section 3. In Section 4, the performance of our method is evaluated through numerical experiments on sparse deconvolution and subset selection problems. The discussion in Section 5 closes the paper.

2. Continuous relaxations within a branch-and-bound algorithm

In this section, we consider the MIP reformulations of problems $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$ and $\mathcal{P}_{2+0}$, and we reformulate the continuous relaxation problems involved in each node of the branch-and-bound strategy.

We introduce binary decision variables b_i, such that $b_i = 0 \Leftrightarrow x_i = 0$. We use the classical *bigM* formulation: assuming that solutions of interest satisfy $\forall i, |x_i| \leq M$ for some known value M, the former logical constraint reads $-Mb_i \leq x_i \leq Mb_i$. Trivial extensions of such assumption write $-M_i^{\inf} \leq x_i \leq M_i^{\sup}$, with $M_i^{\inf}, M_i^{\sup} \geq 0$, but in the following we keep $|x_i| \leq M$ simplifying notations. The three problems can then be reformulated as the standard MIPs given in Table 1 (see for example [4]).

Table 1. Initial problems (left) and their MIP reformulations (right).

Sparsity-enhancing problem		MIP Reformulation	
$\min_{\boldsymbol{x}\in\mathbb{R}^n}$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2$	$\min_{\boldsymbol{b}\in\{0,1\}^n,\boldsymbol{x}\in\mathbb{R}^n}$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2$
s.t.	$\|\boldsymbol{x}\|_0 \leq K$	s.t.	$\sum_{i=1}^n b_i \leq K$
	$\|\boldsymbol{x}\|_\infty \leq M$		$\|\boldsymbol{x}\| \leq M\boldsymbol{b}$
$\min_{\boldsymbol{x}\in\mathbb{R}^n}$	$\|\boldsymbol{x}\|_0$	$\min_{\boldsymbol{b}\in\{0,1\}^n,\boldsymbol{x}\in\mathbb{R}^n}$	$\sum_{i=1}^n b_i$
s.t.	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \epsilon$	s.t.	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \epsilon$
	$\|\boldsymbol{x}\|_\infty \leq M$		$\|\boldsymbol{x}\| \leq M\boldsymbol{b}$
$\min_{\boldsymbol{x}\in\mathbb{R}^n}$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu\|\boldsymbol{x}\|_0$	$\min_{\boldsymbol{b}\in\{0,1\}^n,\boldsymbol{x}\in\mathbb{R}^n}$	$\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu\sum_{i=1}^n b_i$
s.t.	$\|\boldsymbol{x}\|_\infty \leq M$	s.t.	$\|\boldsymbol{x}\| \leq M\boldsymbol{b}$

We consider a resolution strategy based on a branch-and-bound procedure, as adopted by most MIP solvers. The initial problem defines the root node. At each iteration of the algorithm, one node is selected from the list of subproblems that have not been processed yet, and a lower bound for the node is computed *via* the continuous relaxation of the binary variables. If this bound is greater than the current upper bound (defined as the value of the cost function at the best feasible solution found), then the subproblem is discarded. Otherwise, two children of this node are built through the addition of constraints fixing one of the relaxed variables b_i to 0 and 1. The two new nodes are then

added to the list. In the following, we focus on the node evaluation step by continuous relaxation.

2.1 *Continuous relaxation at the root node*

At the root node, no decision has been made concerning any binary variable. The continuous relaxation of binary variables in $\mathcal{P}_{2/0}$ then reads:

$$\mathcal{P}^R_{2/0} : \min_{\boldsymbol{b}\in[0,1]^n, \boldsymbol{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \text{ s.t. } \begin{cases} \sum_{i=1}^n b_i \leq K \\ |\boldsymbol{x}| \leq M\boldsymbol{b} \end{cases} .$$

Proposition 1 Let $\mathcal{P}_{2/1}$ be the following problem:

$$\mathcal{P}_{2/1} : \min_{\boldsymbol{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \text{ s.t. } \begin{cases} \|\boldsymbol{x}\|_1 \leq KM \\ \|\boldsymbol{x}\|_\infty \leq M \end{cases} .$$

Then, $\mathcal{P}^R_{2/0}$ and $\mathcal{P}_{2/1}$ have the same minimum value.

Proof. Let $(\boldsymbol{b}^R, \boldsymbol{x}^R)$ be a minimizer of $\mathcal{P}^R_{2/0}$ and let $\boldsymbol{x}^1$ be a minimizer of $\mathcal{P}_{2/1}$. Let $\boldsymbol{b}^1 := \frac{1}{M}\left|\boldsymbol{x}^1\right|$. Then, $(\boldsymbol{b}^1, \boldsymbol{x}^1)$ is clearly feasible for $\mathcal{P}^R_{2/0}$, therefore $\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}^R\|_2^2 \leq \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}^1\|_2^2$. Conversely, $\boldsymbol{x}^R$ is feasible for $\mathcal{P}_{2/1}$ because $\|\boldsymbol{x}^R\|_1 \leq M\|\boldsymbol{b}^R\|_1 = M\sum_{i=1}^N b_i^R \leq KM$ and $\|\boldsymbol{x}^R\|_\infty \leq M\|\boldsymbol{b}^R\|_\infty \leq M$. Consequently, $\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}^1\|_2^2 \leq \|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}^R\|_2^2$ and the proposition follows. ■

We note that this result was given by [2, 3] for problems with non-negativity constraints, and by [1] in our case.

A similar result holds for $\mathcal{P}^R_{0/2}$, the continuous relaxation of $\mathcal{P}_{0/2}$:

$$\mathcal{P}^R_{0/2} : \min_{\boldsymbol{b}\in[0,1]^n, \boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^n b_i \text{ s.t. } \begin{cases} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \epsilon \\ |\boldsymbol{x}| \leq M\boldsymbol{b} \end{cases}$$

Proposition 2 Let $\mathcal{P}_{1/2}$ be the following problem:

$$\mathcal{P}_{1/2} : \min_{\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{M}\|\boldsymbol{x}\|_1 \text{ s.t. } \begin{cases} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq \epsilon \\ \|\boldsymbol{x}\|_\infty \leq M \end{cases} .$$

Then, $\mathcal{P}^R_{0/2}$ and $\mathcal{P}_{1/2}$ have the same minimum value.

Proof. Let $(\boldsymbol{b}^R, \boldsymbol{x}^R)$ be a minimizer of $\mathcal{P}^R_{0/2}$. We show that $|\boldsymbol{x}^R| = M\boldsymbol{b}^R$, from which the proof is straightforward. Suppose that $|x_i^R| < Mb_i^R$ for some component i. Let $\boldsymbol{b}' := \frac{1}{M}|\boldsymbol{x}^R|$, such that $b'_i < b_i^R$. Then, $(\boldsymbol{b}', \boldsymbol{x}^R)$ is feasible for $\mathcal{P}^R_{0/2}$, with $\sum_{i=1}^n b'_i < \sum_{i=1}^n b_i^R$, which contradicts the definition of $(\boldsymbol{b}^R, \boldsymbol{x}^R)$. ■

Finally, consider the continuous relaxation of binary variables in the penalized problem $\mathcal{P}_{2+0}$ as follows:

$$\mathcal{P}_{2+0}^{R} : \min_{\boldsymbol{b}\in[0,1]^n, \boldsymbol{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu\sum_{i=1}^{n} b_i \text{ s.t. } |\boldsymbol{x}| \leq M\boldsymbol{b} \ .$$

Proposition 3 Let $\mathcal{P}_{2+1}$ be the following problem:

$$\mathcal{P}_{2+1} : \min_{\boldsymbol{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \tfrac{\mu}{M}\|\boldsymbol{x}\|_1 \text{ s.t. } \|\boldsymbol{x}\|_\infty \leq M.$$

Then, $\mathcal{P}_{2+0}^{R}$ and $\mathcal{P}_{2+1}$ have the same minimum value.

Proof. The proof is similar to that of Proposition 2. ■

2.2 *Continuous relaxation in the branch-and-bound algorithm*

We now consider a given node of the branch-and-bound algorithm and the corresponding continuous relaxation sub-problem. Let S^0 (respectively, S^1) denote the index set of binary variables that are set to 0 (respectively, to 1), and let $\bar{\mathrm{S}}$ index the remaining (undetermined) variables:

$$\begin{cases} \forall i \in \mathrm{S}^0, b_i = 0 \text{ and } x_i = 0, \\ \forall i \in \mathrm{S}^1, b_i = 1 \text{ and } |x_i| \leq M, \\ \forall i \in \bar{\mathrm{S}}, b_i \in [0,1] \text{ and } |x_i| \leq M b_i. \end{cases} \tag{1}$$

Let $\boldsymbol{A}_{\mathrm{S}}$ denote the sub-matrix formed by all columns of the matrix $\boldsymbol{A}$ indexed by S. Similarly, $\boldsymbol{z}_{\mathrm{S}}$ denotes the corresponding sub-vector of $\boldsymbol{z}$.

For problem $\mathcal{P}_{2/0}$, the continuous relaxation of variables $\boldsymbol{b}_{\bar{\mathrm{S}}}$ in the corresponding sub-problem reduces to:

$$\mathcal{Q}_{2/0}^{R} : \min_{\substack{\boldsymbol{x}_{\mathrm{S}^1}\in\mathbb{R}^{n_1} \\ \boldsymbol{b}_{\bar{\mathrm{S}}}\in[0,1]^{\bar{n}} \\ \boldsymbol{x}_{\bar{\mathrm{S}}}\in\mathbb{R}^{\bar{n}}}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}-\boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}\|_2^2 \text{ s.t. } \begin{cases} \sum_{i\in\bar{\mathrm{S}}} b_i \leq K - n_1 \\ |\boldsymbol{x}_{\bar{\mathrm{S}}}| \leq M\boldsymbol{b}_{\bar{\mathrm{S}}} \\ \|\boldsymbol{x}_{\mathrm{S}^1}\|_\infty \leq M \end{cases},$$

where n_1 and $\bar{n}$ denote the size of S^1 and $\bar{\mathrm{S}}$, respectively, and where the variables $\boldsymbol{x}_{\mathrm{S}^0}$ which are fixed to zeros have been removed. Then, similarly to the developments in Section 2.1, one can show that $\mathcal{Q}_{2/0}^{R}$ and $\mathcal{Q}_{2/1}$ have the same minimum value, with:

$$\mathcal{Q}_{2/1} : \min_{\substack{\boldsymbol{x}_{\mathrm{S}^1}\in\mathbb{R}^{n_1} \\ \boldsymbol{x}_{\bar{\mathrm{S}}}\in\mathbb{R}^{\bar{n}}}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}-\boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}\|_2^2 \text{ s.t. } \begin{cases} \|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_1 \leq M(K - n_1) \\ \|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_\infty \leq M \\ \|\boldsymbol{x}_{\mathrm{S}^1}\|_\infty \leq M \end{cases} .$$

Applying a similar reasoning to the two other formulations, we finally obtain the equivalent problems summarized in Table 2. These properties are of major interest for our work, with two main consequences:

- Whatever the formulation (constrained or penalized), all continuously relaxed subproblems involved in the evaluation of each node in the branch-and-bound algorithm can be reformulated without binary variables.
- They all reduce to optimization problems mixing a least-squares function, ℓ_1-norm terms involving only a part of the variables, and box constraints.

Table 2. Continuous relaxation problems at any node in the branch-and-bound procedure (left), and equivalent problems without binary variables involving the ℓ_1 norm (right), for the three considered formulations.

Continuous relaxation problem	Equivalent problem without binary variables
$\mathcal{Q}^R_{2/0}: \min_{\boldsymbol{b}\in[0,1]^n,\,\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2$ s.t. $\sum_{i=1}^n b_i \le K$, $\|\boldsymbol{x}\| \le M\boldsymbol{b}$, $\boldsymbol{b}_{\mathrm{S}^1}=1$, $\boldsymbol{b}_{\mathrm{S}^0}=0$	$\mathcal{Q}_{2/1}: \min_{\boldsymbol{x}_{\mathrm{S}^1}\in\mathbb{R}^{n_1},\boldsymbol{x}_{\bar{\mathrm{S}}}\in\mathbb{R}^{\bar{n}}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}-\boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}\|_2^2$ s.t. $\|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_1 \le M(K-n_1)$, $\|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_\infty \le M$, $\|\boldsymbol{x}_{\mathrm{S}^1}\|_\infty \le M$
$\mathcal{Q}^R_{0/2}: \min_{\boldsymbol{b}\in[0,1]^n,\,\boldsymbol{x}\in\mathbb{R}^n} \sum_{i=1}^n b_i$ s.t. $\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \le \epsilon$, $\|\boldsymbol{x}\| \le M\boldsymbol{b}$, $\boldsymbol{b}_{\mathrm{S}^1}=1$, $\boldsymbol{b}_{\mathrm{S}^0}=0$	$\mathcal{Q}_{1/2}: \min_{\boldsymbol{x}_{\mathrm{S}^1}\in\mathbb{R}^{n_1},\boldsymbol{x}_{\bar{\mathrm{S}}}\in\mathbb{R}^{\bar{n}}} \frac{1}{M}\|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_1 + n_1$ s.t. $\frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}-\boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}\|_2^2 \le \epsilon$, $\|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_\infty \le M$, $\|\boldsymbol{x}_{\mathrm{S}^1}\|_\infty \le M$
$\mathcal{Q}^R_{2+0}: \min_{\boldsymbol{b}\in[0,1]^n,\,\boldsymbol{x}\in\mathbb{R}^n} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \mu\sum_{i=1}^n b_i$ s.t. $\|\boldsymbol{x}\| \le M\boldsymbol{b}$, $\boldsymbol{b}_{\mathrm{S}^1}=1$, $\boldsymbol{b}_{\mathrm{S}^0}=0$	$\mathcal{Q}_{2+1}: \min_{\boldsymbol{x}_{\mathrm{S}^1}\in\mathbb{R}^{n_1},\boldsymbol{x}_{\bar{\mathrm{S}}}\in\mathbb{R}^{\bar{n}}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}-\boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}\|_2^2 + \frac{\mu}{M}\|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_1 + \mu\, n_1$ s.t. $\|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_\infty \le M$, $\|\boldsymbol{x}_{\mathrm{S}^1}\|_\infty \le M$

In Section 3, we build a dedicated algorithm which solves the three problems $\mathcal{Q}_{2/1}$, $\mathcal{Q}_{1/2}$ and $\mathcal{Q}_{2+1}$.

3. A dedicated homotopy continuation algorithm for relaxed problems

Optimization involving a quadratic misfit and the ℓ_1 norm has been a very active field of research in the past ten years. Many dedicated convex, non-smooth, optimization algorithms have been developed (see for example [13, 28] and references therein), for solving problems:

$$\mathcal{P}_1^\tau: \min_{\boldsymbol{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \text{ s.t. } \|\boldsymbol{x}\|_1 \le \tau;$$

$$\mathcal{P}_1^\epsilon: \min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{x}\|_1 \text{ s.t. } \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 \le \epsilon;$$

$$\mathcal{P}_1^\lambda: \min_{\boldsymbol{x}\in\mathbb{R}^n} \tfrac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1.$$

Most works addressed the penalized form $\mathcal{P}_1^\lambda$, which resorts to unconstrained optimization. Since the two objectives are convex, the three problems are equivalent, that is, for any $\tau \ge 0$, there exists $\lambda^{(\tau)} \ge 0$ such that $\mathcal{P}_1^\tau$ and $\mathcal{P}_1^{\lambda^{(\tau)}}$ have the same solution, and reciprocally. Similarly, for any $\epsilon \ge 0$, there exists $\lambda^{(\epsilon)} \ge 0$ such that solutions of $\mathcal{P}_1^\epsilon$ and $\mathcal{P}_1^{\lambda^{(\epsilon)}}$ are identical. However, in general, there is no explicit mapping between the three parameters. We propose to solve the three problems $\mathcal{Q}_{2/1}$, $\mathcal{Q}_{1/2}$ and $\mathcal{Q}_{2+1}$ by an homotopy continuation algorithm [9, 10, 23]. This choice is motivated by the following reasons:

- First and foremost, the same algorithm can solve the three problems with the same computational efficiency.
- It can naturally incorporate specificities such as box constraints and ℓ_1-norm terms involving only one part of the variables.
- It is an exact algorithm, for which the solution is obtained in a finite number of iterations.

The homotopy method considers the penalized form $\mathcal{P}_1^\lambda$, and exploits the fact that the solution path is piecewise linear as a function of λ [9, 10, 23]. Starting from $\lambda^{(0)} = \|\boldsymbol{A}^T\boldsymbol{y}\|_\infty$ (such that the solution is identically zero $\forall\lambda > \lambda^{(0)}$), it iteratively computes all solutions by continuously decreasing the parameter λ until the target value, say $\lambda^\star$, is reached. Therefore, it can similarly solve problems $\mathcal{P}_1^\tau$ or $\mathcal{P}_1^\epsilon$, by stopping when the corresponding value $\tau^\star$ or $\epsilon^\star$ is reached, respectively. Figure 1 shows a typical solution path (left part) and the corresponding evolution on the Pareto front (right part).

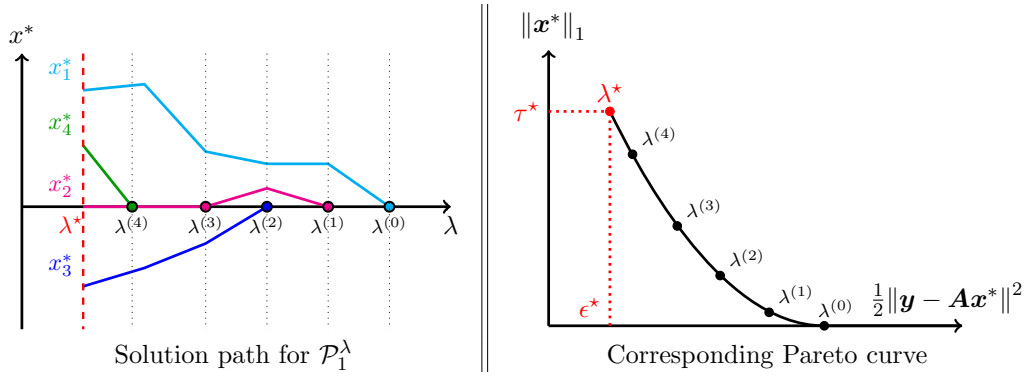


Figure 1. Homotopy method: an example of a solution path $\boldsymbol{x}^*(\lambda) = \arg\min_{\boldsymbol{x}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{x}\|_1$ as a function of λ (left), and corresponding set $\left(\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^*\|_2^2, \|\boldsymbol{x}^*\|_1\right)$ as a function of λ (right).

In the following, we generalize the homotopy method to the class of problems $\mathcal{Q}_{2/1}$, $\mathcal{Q}_{1/2}$ and $\mathcal{Q}_{2+1}$, where *free* variables (that is, variables that are not involved in the ℓ_1-norm term) and box constraints are included. Note that the homotopy method with box constraints was recently proposed in [18], which also established convergence proofs. Including free variables impacts initialization and requires additional tests to be performed at each iteration.

3.1 *Optimality conditions*

We first focus on problem $\mathcal{Q}_{2+1}$ in Table 2, that is, the relaxed problem involved in the cardinality-penalized form. We consider equivalently the optimization problem of the form:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} F(\boldsymbol{x}) := J(\boldsymbol{x}) + \lambda h(\boldsymbol{x}) \;\text{ s.t. }\; g_i(\boldsymbol{x}) \leq 0 \quad \forall i = 1, \ldots, n, \tag{2}$$

$$\begin{aligned} \text{with } \; J(\boldsymbol{x}) &:= \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}\|_2^2, \\ h(\boldsymbol{x}) &:= \|\boldsymbol{x}_{\bar{\mathrm{S}}}\|_1, \\ g_i(\boldsymbol{x}) &:= |x_i| - M, \end{aligned}$$

where index sets S^1 and $\bar{\mathrm{S}}$ are defined in equation (1). The function J is differentiable with $\nabla J(\boldsymbol{x}) = -\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})$. The subdifferentials of $h(\boldsymbol{x})$ and of $g_i(\boldsymbol{x})$ are respectively:

$$\partial h(\boldsymbol{x}) = \left\{ \boldsymbol{z} \in \mathbb{R}^n \;\middle|\; \begin{array}{ll} z_i = 0 & \text{if } i \in \mathrm{S}^1 \\ z_i = \operatorname{sgn}(x_i) & \text{if } i \in \bar{\mathrm{S}} \text{ and } x_i \neq 0 \\ z_i \in [-1, 1] & \text{if } i \in \bar{\mathrm{S}} \text{ and } x_i = 0 \end{array} \right\} \tag{3}$$

and

$$\partial g_i(\boldsymbol{x}) = \left\{ \boldsymbol{z} \in \mathbb{R}^n \;\middle|\; \begin{array}{ll} z_j = 0 & \text{for } j \neq i \\ z_i = \operatorname{sgn}(x_i) & \text{if } x_i \neq 0 \\ z_i \in [-1, 1] & \text{if } x_i = 0 \end{array} \right\}. \tag{4}$$

The vector $\boldsymbol{x}^*$ is a minimizer of (2) if and only if there exists $\boldsymbol{\pi} \in \mathbb{R}^n$ such that $(\boldsymbol{x}^*, \boldsymbol{\pi})$ satisfies the Karush-Kuhn-Tucker optimality conditions applied to continuous, convex, non-differentiable functions (see for example [24]):

$$\begin{cases} 0 \in -\boldsymbol{A}^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^*) + \lambda \partial h(\boldsymbol{x}^*) + \partial \displaystyle\sum_{i=1}^{n} \pi_i g_i(\boldsymbol{x}^*) & \text{(5a)} \\ g_i(\boldsymbol{x}^*) \leq 0 \;\; \forall i = 1, \ldots, n & \text{(5b)} \\ \pi_i \geq 0 \;\; \forall i = 1, \ldots, n & \text{(5c)} \\ \pi_i g_i(\boldsymbol{x}^*) = 0 \;\; \forall i = 1, \ldots, n. & \text{(5d)} \end{cases}$$

Particular points are those which activate the bound constraints ($x_i^* = \pm M$ for $i \in \bar{\mathrm{S}} \cup \mathrm{S}^1$) or non-differentiability points ($x_i^* = 0$ for $i \in \bar{\mathrm{S}}$). Therefore, we split the variable indices into the five possible cases.

- Case 1. Let $\bar{\mathrm{S}}_0 := \{i \in \bar{\mathrm{S}} \mid |x_i^*| = 0\}$. From equation (5d), $\boldsymbol{\pi}_{\bar{\mathrm{S}}_0} = 0$ and using equations (3) and (5a), optimality conditions of $\boldsymbol{x}^*_{\bar{\mathrm{S}}_0}$ become:

$$|\boldsymbol{A}^T_{\bar{\mathrm{S}}_0}(\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}^*_{\mathrm{S}^1} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}^*_{\bar{\mathrm{S}}})| < \lambda. \tag{6a}$$

- Case 2. Let $\bar{\mathrm{S}}_{in} := \{i \in \bar{\mathrm{S}} \mid 0 < |x_i^*| < M\}$. From equation (5d), $\boldsymbol{\pi}_{\bar{\mathrm{S}}_{in}} = 0$ and using equations (3) and (5a), optimality conditions of $\boldsymbol{x}^*_{\bar{\mathrm{S}}_{in}}$ become:

$$-\boldsymbol{A}^T_{\bar{\mathrm{S}}_{in}}(\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}^*_{\mathrm{S}^1} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}^*_{\bar{\mathrm{S}}}) + \lambda \operatorname{sgn}(\boldsymbol{x}^*_{\bar{\mathrm{S}}_{in}}) = 0. \tag{6b}$$

- Case 3. Let $\bar{\mathrm{S}}_{\square} := \{i \in \bar{\mathrm{S}} \mid |x_i^*| = M\}$. From equation (5c), $\boldsymbol{\pi}_{\bar{\mathrm{S}}_{\square}} \geq 0$ and using equations (4) and (5a), optimality conditions of $\boldsymbol{x}^*_{\bar{\mathrm{S}}_{\square}}$ become:

$$-\boldsymbol{A}^T_{\bar{\mathrm{S}}_{\square}}(\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}^*_{\mathrm{S}^1} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}^*_{\bar{\mathrm{S}}}) + \lambda \operatorname{sgn}(\boldsymbol{x}^*_{\bar{\mathrm{S}}_{\square}}) + \boldsymbol{\pi}_{\bar{\mathrm{S}}_{\square}} \odot \operatorname{sgn}(\boldsymbol{x}^*_{\bar{\mathrm{S}}_{\square}}) = 0, \quad \boldsymbol{\pi}_{\bar{\mathrm{S}}_{\square}} \geq 0, \tag{6c}$$

  where $\odot$ denotes the Hadamard (entrywise) product.
- Case 4. Let $\mathrm{S}^1_{in} := \{i \in \mathrm{S}^1 \mid 0 \leq |\boldsymbol{x}_i^*| < M\}$. From equation (5d), $\boldsymbol{\pi}_{\mathrm{S}^1_{in}} = 0$ and using equations (3) and (5a), optimality conditions of $\boldsymbol{x}^*_{\mathrm{S}^1_{in}}$ become:

$$-\boldsymbol{A}^T_{\mathrm{S}^1_{in}}(\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}^*_{\mathrm{S}^1} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}^*_{\bar{\mathrm{S}}}) = 0. \tag{6d}$$

- Case 5. Let $\mathrm{S}^1_\square := \{i \in \mathrm{S}^1 \mid |\boldsymbol{x}^*_i| = M\}$. From equation (5c), $\boldsymbol{\pi}_{\mathrm{S}^1_\square} \geq 0$ and using equations (4) and (5a), optimality conditions of $\boldsymbol{x}_{\mathrm{S}^1_\square{}^*}$ become:

$$-\boldsymbol{A}^T_{\mathrm{S}^1_\square}(\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}^*_{\mathrm{S}^1} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}^*_{\bar{\mathrm{S}}}) + \boldsymbol{\pi}_{\mathrm{S}^1_\square} \odot \operatorname{sgn}(\boldsymbol{x}^*_{\mathrm{S}^1_\square}) = 0, \quad \boldsymbol{\pi}_{\mathrm{S}^1_\square} \geq 0. \tag{6e}$$

Let us remark that equations (6b), (6c) and (6e) concern non-zero variables, therefore the corresponding sign function is well-defined.

3.2 *Homotopy continuation algorithm*

We now build the homotopy algorithm that solves the problem (2). Let $\boldsymbol{r} := \boldsymbol{y} - \boldsymbol{A}_{\bar{\mathrm{S}}_\square}\boldsymbol{x}^*_{\bar{\mathrm{S}}_\square} - \boldsymbol{A}_{\mathrm{S}^1_\square}\boldsymbol{x}^*_{\mathrm{S}^1_\square}$, where each component in $\boldsymbol{x}_{\bar{\mathrm{S}}_\square}$ and $\boldsymbol{x}_{\mathrm{S}^1_\square}$ equals $\pm M$. Equations (6b) and (6d) are linear systems in $\boldsymbol{x}^*_{\bar{\mathrm{S}}_{in}}$ and $\boldsymbol{x}^*_{\mathrm{S}^1_{in}}$, whose solution is:

$$\begin{cases} \boldsymbol{x}^*_{\bar{\mathrm{S}}_{in}} = \left(\boldsymbol{A}^T_{\bar{\mathrm{S}}_{in}}(\boldsymbol{I} - \boldsymbol{P}^{\mathrm{S}^1_{in}})\boldsymbol{A}_{\bar{\mathrm{S}}_{in}}\right)^{-1}\left(\boldsymbol{A}^T_{\bar{\mathrm{S}}_{in}}(\boldsymbol{I} - \boldsymbol{P}^{\mathrm{S}^1_{in}})\boldsymbol{r} - \lambda \operatorname{sgn}(\boldsymbol{x}^*_{\bar{\mathrm{S}}_{in}})\right), & \text{(7a)} \\ \boldsymbol{x}^*_{\mathrm{S}^1_{in}} = (\boldsymbol{A}^T_{\mathrm{S}^1_{in}}\boldsymbol{A}_{\mathrm{S}^1_{in}})^{-1}(\boldsymbol{A}^T_{\mathrm{S}^1_{in}}\boldsymbol{r} - \boldsymbol{A}^T_{\mathrm{S}^1_{in}}\boldsymbol{A}_{\bar{\mathrm{S}}_{in}}\boldsymbol{x}^*_{\bar{\mathrm{S}}_{in}}), & \text{(7b)} \end{cases}$$

where $\boldsymbol{P}^{\mathrm{S}^1_{in}} := \boldsymbol{A}_{\mathrm{S}^1_{in}}(\boldsymbol{A}^T_{\mathrm{S}^1_{in}}\boldsymbol{A}_{\mathrm{S}^1_{in}})^{-1}\boldsymbol{A}^T_{\mathrm{S}^1_{in}}$ and $\boldsymbol{I}$ is the identity matrix of appropriate size. These equations show that, in a given configuration of the index sets $\{\bar{\mathrm{S}}_{in}, \mathrm{S}^1_{in}, \bar{\mathrm{S}}_\square, \mathrm{S}^1_\square, \bar{\mathrm{S}}_0\}$, which we call the *support configuration*, the solution of the problem (2) is linear in λ (recall that variables in $\bar{\mathrm{S}}_\square$ and $\mathrm{S}^1_\square$ are fixed to $\pm M$, and variables in $\bar{\mathrm{S}}_0$ are zero). The homotopy method then constructs the solution path (the set of all solutions as a function of λ) by identifying iteratively the different *breakpoints* that lead to changes in the support configuration. These breakpoints will occur at specific values of λ, for which (at least) one of the conditions in equations (6a)–(6e) is violated. The algorithm works as follows:

(1) First, it is clear that as $\lambda \to +\infty$, ℓ_1-norm-penalized variables $\boldsymbol{x}_{\bar{\mathrm{S}}}$ are zero. In that case, other variables $\boldsymbol{x}_{\mathrm{S}^1}$ are found by solving the least-squares problem: $\min_{-M \leq \boldsymbol{x}_{\mathrm{S}^1} \leq M} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}\|_2^2$. We note $\boldsymbol{x}^{(0)}$ the vector defined by

$$\begin{cases} \boldsymbol{x}^{(0)}_{\mathrm{S}^1} := \operatorname*{argmin}_{-M \leq \boldsymbol{x}_{\mathrm{S}^1} \leq M} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}\|_2^2, & \text{(8a)} \\ \boldsymbol{x}^{(0)}_{\bar{\mathrm{S}}} := \boldsymbol{0}. & \text{(8b)} \end{cases}$$

Equation (6a) shows that $\boldsymbol{x}^{(0)}$ is the solution of the problem (2) as long as $\lambda \geq \lambda^{(0)}$, with:

$$\lambda^{(0)} := \|\boldsymbol{A}^T_{\bar{\mathrm{S}}}(\boldsymbol{y} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}^{(0)}_{\mathrm{S}^1})\|_\infty. \tag{8c}$$

(2) As λ decreases below $\lambda^{(0)}$, indices $j \in \bar{\mathrm{S}}$ such that $|\boldsymbol{a}^T_j(\boldsymbol{y} - \boldsymbol{A}^T_{\mathrm{S}^1}\boldsymbol{x}^{(0)}_{\mathrm{S}^1})| = \lambda^{(0)}$ leave $\bar{\mathrm{S}}_0$ to form the new subset $\bar{\mathrm{S}}_{in}$. This new support configuration remains valid for any $\lambda \in [\lambda^{(1)}, \lambda^{(0)}]$, where $\lambda^{(1)}$ defines the next breakpoint, *etc.* A monotonically decreasing sequence $\{\lambda^{(k)}\}_k$ is built iteratively, by testing all possible changes that can occur to the support configuration, and selecting the one(s) corresponding

to the smallest decrease in λ. Then, the support configuration is updated, and a new breakpoint in λ is searched. Since the solution path is piecewise linear as a function of λ, the solution $\boldsymbol{x}^{(k)}$ at the k-th breakpoint reads:

$$\begin{cases} \boldsymbol{x}^{(k)} = \boldsymbol{x}^{(k-1)} + \gamma^{(k)} \boldsymbol{d}^{(k)} & \text{(9a)} \\ \text{and } \lambda^{(k)} = \lambda^{(k-1)} - \gamma^{(k)}, & \text{(9b)} \end{cases}$$

where $\boldsymbol{d}^{(k)}$ represents the vector of slope changes and $\gamma^{(k)} > 0$ represents the length of interval $[\lambda^{(k)}, \lambda^{(k-1)}]$. From equations (7a) and (7b), the direction $\boldsymbol{d}^{(k)}$ is obtained by:

$$\begin{cases} \boldsymbol{d}^{(k)}_{\bar{\mathrm{S}}_{in}} = (\boldsymbol{A}^T_{\bar{\mathrm{S}}_{in}} (\boldsymbol{I} - \boldsymbol{P}^{\mathrm{S}^1_{in}}) \boldsymbol{A}_{\bar{\mathrm{S}}_{in}})^{-1} \mathrm{sgn}(\boldsymbol{x}^{(k-1)}_{\bar{\mathrm{S}}_{in}}) & \text{(10a)} \\ \boldsymbol{d}^{(k)}_{\mathrm{S}^1_{in}} = -(\boldsymbol{A}^T_{\mathrm{S}^1_{in}} \boldsymbol{A}_{\mathrm{S}^1_{in}})^{-1} \boldsymbol{A}^T_{\mathrm{S}^1_{in}} \boldsymbol{A}_{\bar{\mathrm{S}}_{in}} \boldsymbol{d}^{(k)}_{\bar{\mathrm{S}}_{in}} & \text{(10b)} \\ d^{(k)}_i = 0 \ \forall i \notin \{\bar{\mathrm{S}}_{in} \cup \mathrm{S}^1_{in}\}, & \text{(10c)} \end{cases}$$

where the last equality concerns variables that are fixed to zero or to $\pm M$. The step size $\gamma^{(k)}$ is obtained as the smallest positive value $\gamma > 0$ such that $\boldsymbol{x}^{(k-1)} + \gamma \boldsymbol{d}^{(k)}$ reaches a new breakpoint. Five different cases can occur, which are detailed hereafter. We introduce the following notations:

$$\begin{aligned} \boldsymbol{t}^{(k-1)} &:= \boldsymbol{y} - \boldsymbol{A}_{\bar{\mathrm{S}}} \boldsymbol{x}^{(k-1)}_{\bar{\mathrm{S}}} - \boldsymbol{A}_{\mathrm{S}^1} \boldsymbol{x}^{(k-1)}_{\mathrm{S}^1}, & \boldsymbol{u}^{(k)} &:= \boldsymbol{A}_{\bar{\mathrm{S}}_{in}} \boldsymbol{d}^{(k)}_{\bar{\mathrm{S}}_{in}} + \boldsymbol{A}_{\mathrm{S}^1} \boldsymbol{d}^{(k)}_{\mathrm{S}^1}, \\ \boldsymbol{v}^{(k-1)} &:= \boldsymbol{A}^T \boldsymbol{t}^{(k-1)}, & \boldsymbol{w}^{(k)} &:= \boldsymbol{A}^T \boldsymbol{u}^{(k)}. \end{aligned} \tag{11}$$

(a) A component with index $\ell \in \bar{\mathrm{S}}_0$ becomes nonzero when equality in equation (6a) is reached. Inserting equations (7a) and (7b) into equation (6a), one can show that it may become positive (respectively, negative) when:

$$\gamma = \frac{\lambda^{(k-1)} + v^{(k-1)}_\ell}{1 - w^{(k)}_\ell} \ \Big(\text{respectively, when } \gamma = \frac{-\lambda^{(k-1)} + v^{(k-1)}_\ell}{-1 - w^{(k)}_\ell}\Big). \tag{12a}$$

(b) A component with index $\ell \in \bar{\mathrm{S}}_{in}$ becomes zero. From equation (9a), this may occur when:

$$\gamma = \frac{-x^{(k-1)}_\ell}{d^{(k)}_\ell}. \tag{12b}$$

(c) A component with index $\ell \in \bar{\mathrm{S}}_{in}$ or S^1_{in} yields the bound M or $-M$, depending on its current sign. From (9a), this may occur when:

$$\gamma = \frac{M \mathrm{sgn}(x^{(k-1)}_\ell) - x^{(k-1)}_\ell}{d^{(k)}_\ell}. \tag{12c}$$

(d) The bound constraint for some component with index $\ell \in \bar{\mathrm{S}}_\square$ becomes inactive. This may occur when the corresponding Lagrange multiplier $\pi_\ell = 0$

in equation (6c), which yields:

$$\gamma = \frac{\operatorname{sgn}(x_\ell^{(k-1)})\lambda^{(k-1)} - v_\ell^{(k-1)}}{\operatorname{sgn}(x_\ell^{(k-1)}) - w_\ell^{(k)}}. \tag{12d}$$

(e) The bound constraint for some component with index $\ell \in \mathrm{S}_\square^1$ becomes inactive. This may occur when the corresponding Lagrange multiplier $\pi_\ell = 0$ in equation (6e), which yields:

$$\gamma = \frac{-v_\ell^{(k-1)}}{w_\ell^{(k)}}. \tag{12e}$$

The shortest step size $\gamma^{(k)}$ is then defined as the shortest positive step among all possible ones, defined by equations (12a)–(12e). In theory, $\gamma^{(k)}$ may be obtained by several conditions above simultaneously; should this happen, the support configuration is updated correspondingly.

(3) The algorithm stops when the target λ, say $\lambda^\star$, is reached, that is, after iteration k such that $\lambda^\star \in [\lambda^{(k)}, \lambda^{(k-1)}]$. Then, the optimal solution $\boldsymbol{x}^\star$ for $\lambda = \lambda^\star$ is found by:

$$\boldsymbol{x}^\star = \boldsymbol{x}^{(k-1)} + \gamma^\star \boldsymbol{d}^{(k)}, \tag{13}$$

with $\gamma^\star = \lambda^\star - \lambda^{(k)}$.

The homotopy algorithm is summarized in Algorithm 1. Figure 2 shows a typical solution path for a toy example with 5 variables: $\bar{\mathrm{S}} = \{1, 2, 3\}$ and $\mathrm{S}^1 = \{4, 5\}$.

Algorithm 1 Homotopy algorithm for solving the problem $\mathcal{Q}_{2+1}$ in Table 2, reformulated as the problem (2) with $\lambda = \lambda^\star$.

1: Set $k = 0$. Initialize $\lambda^{(0)}$, $\boldsymbol{x}^{(0)}$ by equations (8a)–(8c).
2: **while** $\lambda^{(k)} > \lambda^*$ **do**
3: $k \leftarrow k + 1$
4: Update $\boldsymbol{d}^{(k)}$ by equations (10a)–(10b).
5: Determine the step size $\gamma^{(k)}$ as the smallest positive value among all cases in equations (12a)–(12e).
6: Compute accordingly $(\boldsymbol{x}^{(k)}, \lambda^{(k)})$ by equations (9a)–(9b).
7: Update index sets $\{\bar{\mathrm{S}}_{in}, \mathrm{S}_{in}^1, \bar{\mathrm{S}}_\square, \mathrm{S}_\square^1, \bar{\mathrm{S}}_0\}$.
8: Compute $\boldsymbol{x}^\star$ by equation (13).

3.3 *Solutions to constrained problems $\mathcal{Q}_{2/1}$ and $\mathcal{Q}_{1/2}$*

As λ is continuously decreased, the ℓ_1 norm of the penalized variables $\|\boldsymbol{x}_{\bar{\mathrm{S}}}^*\|_1$ is continuously increased and the least-squares function $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}^* - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}^*\|_2^2$ is continuously decreased. Therefore, the homotopy method can also solve the constrained problems $\mathcal{Q}_{2/1}$ and $\mathcal{Q}_{1/2}$ in Table 2. More precisely:

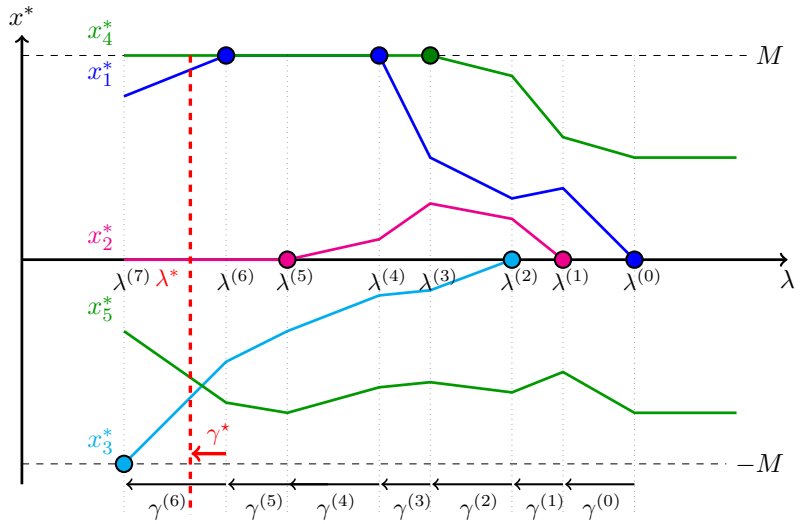


Figure 2. Example of solution path $\boldsymbol{x}^*$ giving the solutions of the problem (2) as a function of λ, with 5 variables: $\bar{\mathrm{S}} = \{1, 2, 3\}$ and $\mathrm{S}^1 = \{4, 5\}$. Circles represent the events that cause a change in the support configuration. Vertical dotted lines represent the breakpoints.

- For $\mathcal{Q}_{2/1}$, iterations stop at the first breakpoint such that the ℓ_1 norm of the penalized variables $\|\boldsymbol{x}_{\bar{\mathrm{S}}}^{(k)}\|_1$ exceeds the threshold value $\tau := M(K - n_1)$. Then, in the corresponding interval $[\lambda^{(k)}, \lambda^{(k-1)}]$, the solution is given by equation (13). By construction, there is no sign change between $\boldsymbol{x}^{(k-1)}$ and the optimal solution $\boldsymbol{x}^*$ such that $\|\boldsymbol{x}^*\|_1 = \tau$. One can then easily show that the value of γ such that $\|\boldsymbol{x}^\star\|_1 = \tau$ is:

$$\gamma^\star := \frac{\tau - \|\boldsymbol{x}^{(k-1)}\|_1}{\mathrm{sgn}(\boldsymbol{x}^{(k-1)})^T \boldsymbol{d}^{(k)}}.$$

- Similarly, for $\mathcal{Q}_{1/2}$, iterations stop at the first breakpoint such that $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}^{(k)} - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}^{(k)}\|_2^2 \leq \epsilon$. Substituting equation (13) in the least-squares expression, the value of γ such that $\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}_{\bar{\mathrm{S}}}\boldsymbol{x}_{\bar{\mathrm{S}}}^\star - \boldsymbol{A}_{\mathrm{S}^1}\boldsymbol{x}_{\mathrm{S}^1}^\star\|_2^2 = \epsilon$ can be found by solving a scalar quadratic equation, whose solution is:

$$\gamma^\star := \frac{\boldsymbol{t}^{(k-1)^T}\boldsymbol{u}^{(k)} - \sqrt{\left(\boldsymbol{t}^{(k-1)^T}\boldsymbol{u}^{(k)}\right)^2 - \boldsymbol{u}^{(k)^T}\boldsymbol{u}^{(k)}\left(\boldsymbol{t}^{(k-1)^T}\boldsymbol{t}^{(k-1)} - 2\epsilon\right)}}{\boldsymbol{u}^{(k)^T}\boldsymbol{u}^{(k)}},$$

where $\boldsymbol{t}^{(k-1)}$ and $\boldsymbol{u}^{(k)}$ are defined in equation (11).

3.4 *Implementation and practical issues*

Some practical remarks concerning the numerical implementation of the homotopy algorithm 1 are detailed in this section.

First, each iteration mostly consists of solving linear systems of equations (10a)–(10b), whose size respectively corresponds to the current number of variables in $\bar{\mathrm{S}}_{in}$ and S_{in}^1.

Since the support configuration only slightly changes between two breakpoints[1], the matrix inverses of equations (10a)–(10b) can be computed recursively by performing rank-one updates. In our simulations, using the *block matrix inversion formulas* appeared to be the most efficient strategy.

We also note that, for each non-zero component with index $\ell \in \bar{\mathrm{S}}_{in}$, only one of the two computations defined by equations (12b) and (12c) is necessary. In particular, if $x_\ell^{(k-1)} > 0$ and $d_\ell^{(k)} > 0$ (respectively, $d_\ell^{(k)} < 0$), then the only possible change in the support configuration is when x_ℓ hits the upper bound M (respectively, $x_\ell = 0$), corresponding to case (2)(c in Section 3.2 (respectively, to case (2)(b). A similar reasoning applies to negative components.

We conclude this section with an important remark concerning the resolution of the error-constrained problem $\mathcal{P}_{0/2}$. At each node, the continuous relaxation brought by the solution to problems such as $\mathcal{Q}_{1/2}$ in Table 2 provides a lower bound on the global optimum value of $\mathcal{P}_{0/2}$. This lower bound is compared to the global upper bound, say z^u, provided by the best known feasible solution. Since z^u is discrete, it is clear that the node can be pruned if the corresponding lower bound exceeds $z^u - 1$. With the homotopy method, the objective function in $\mathcal{Q}_{1/2}$ monotonically increases at each iteration. Therefore, the homotopy algorithm is stopped and the node is pruned as soon as the ℓ_1 norm of the current iterate exceeds $z^u - 1$, as shown in Figure 3.

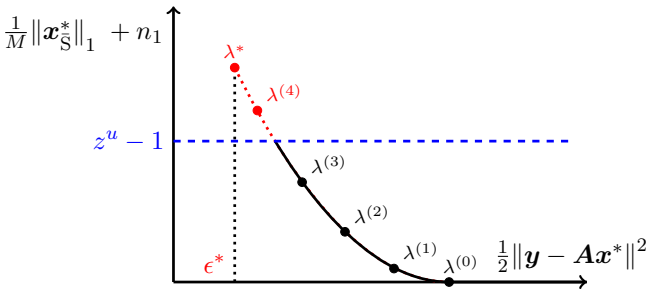


Figure 3. Set of optimal solutions $\left(\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^*\|_2^2, \frac{1}{M}\|\boldsymbol{x}^*_{\bar{\mathrm{S}}}\|_1 + n_1\right)$ as a function of λ, and illustration of a stopping criterion of the homotopy algorithm for $\mathcal{Q}_{1/2}$. The algorithm can be stopped (and the corresponding node is pruned) as soon as $\frac{1}{M}\|\boldsymbol{x}^*_{\bar{\mathrm{S}}}\|_1 + n_1 \geq z^u - 1$, where z^u is the cardinality of the best known feasible solution to $\mathcal{P}_{0/2}$ at a given iteration of the branch-and-bound procedure.

4. Experimental results

We now insert the continuous relaxation algorithms built in Section 3 into dedicated branch-and-bound procedures for the resolution of the three problems $\mathcal{P}_{2/0}$, $\mathcal{P}_{0/2}$, and $\mathcal{P}_{2+0}$. We use *depth-first search*, and our branching strategy is based on selecting the continuous variable, say x_i, with the maximum absolute value in the solution of the relaxed problem. In the standard MIP resolution, this corresponds to choosing the binary variable b_i whose value in the continuous relaxation problems is the closest to 1. We branch *up* first, that is, we first explore the branch corresponding to the decision $x_i \neq 0$ (*i.e.*, . $b_i = 1$). This strategy is well adapted for problem $\mathcal{P}_{2/0}$, in which the depth limit is imposed by the cardinality constraint. It also allows one to quickly find feasible solutions

[1]Although several support configuration changes may happen simultaneously, in practice, $\bar{\mathrm{S}}_{in}$ and S^1_{in} are usually modified by at most one element at each breakpoint, corresponding to the activation of one condition among equations (12a)–(12e).

for problems $\mathcal{P}_{2+0}$ and $\mathcal{P}_{0/2}$ with limited depth search, since we know in advance that their solutions are sparse. For problem $\mathcal{P}_{2+0}$, we also exploit the fact that the initialization step in the homotopy algorithm at each node evaluation (see Section 3.2) produces a feasible solution, which is used to improve the best solution found.

In this section, we compare our homotopy-based branch-and-bound implementation, named B&B$_{\text{R-HOM}}$, to:

- the same branch-and-bound exploration strategy, where the continuous relaxation problems $\mathcal{Q}^R_{2/0}$, $\mathcal{Q}^R_{0/2}$ and $\mathcal{Q}^R_{2+0}$ in Table 2 are solved with the CPLEX 12.8 quadratic programming solver (named B&B$_{\text{R-CPLEX}}$);
- the resolution of the MIP problems in Table 1 with the CPLEX 12.8 MIP solver (named MIP$_{\text{CPLEX}}$).

All methods are implemented in C++ and executed on a UNIX machine equipped with 31.1 Go RAM and four Intel Core i7-6600U central processing units (CPUs) clocked at 2.60 GHz. Computations are restricted to one core in order to focus on the algorithm performance, disabling parallelization capacities. The maximum time allowed is set to 1 000 s and all computations with CPLEX are run with its default settings.

In Section 4.1, we study the performance of our algorithm on simulated sparse deconvolution problems typically encountered in signal processing. Then, simulated subset selection problems with random entries are considered in Section 4.2.

4.1 *Sparse deconvolution problems*

Sparse deconvolution is a classical signal processing problem [16, Chapter 5], which aims to estimate a sparse sequence $\boldsymbol{x}$ from filtered and noisy observations. Collecting sampled data in vector $\boldsymbol{y}$ and the unknown sparse sequence in vector $\boldsymbol{x}$, we obtain a model of the form $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\xi}$, where $\boldsymbol{A}$ is a convolution matrix and $\boldsymbol{\xi}$ is a random term representing noise and model errors. Deconvolution is an ill-posed inverse problem, whose resolution with sparsity-enhancing regularization is often addressed through suboptimal ℓ_1-norm-based or greedy methods. It was shown in [4] that exact resolution of ℓ_0-norm problems (with the CPLEX MIP solver) achieves better solutions, obviously with higher computation times. We consider the problem instances[2] proposed by [4], where $\boldsymbol{y} \in \mathbb{R}^{120}$, $\boldsymbol{A} \in \mathbb{R}^{100\times 120}$, noise samples in $\boldsymbol{\xi}$ are independent, identically and normally distributed with signal-to-noise ratio SNR = 10 dB and $M = 1.1\|\boldsymbol{A}^T\boldsymbol{y}\|_\infty$. Columns of $\boldsymbol{A}$ have unit norm. The cardinality varies from $K = 5$ to $K = 9$ for $\mathcal{P}_{2/0}$. For problems $\mathcal{P}_{0/2}$ and $\mathcal{P}_{2+0}$, the respective parameters ϵ and λ are tuned from statistical rules accounting for the noise level and the sparsity degree (see [4] for details): ϵ is tuned such that the probability $P(\|\boldsymbol{\xi}\|_2^2 \leq \epsilon) = 95$ %, and $\lambda = 2\sigma^2 \log(1/\rho - 1)$, where σ^2 is the noise variance and $\rho = K/n$.

Computational results are reported in Table 3. For all problems, B&B$_{\text{R-HOM}}$ and B&B$_{\text{R-CPLEX}}$ explore the same number of nodes, since they use the same branch-and-bound strategy (slight differences are observed, though, that may be due to numerical issues). However, B&B$_{\text{R-HOM}}$ requires significantly less execution time, which is reduced by a factor between 8 and 25 for $\mathcal{P}_{2/0}$ and $\mathcal{P}_{2+0}$, and by a factor between 60 and 120 for $\mathcal{P}_{0/2}$. It also successfully solves more instances in 1 000 s.

For problems $\mathcal{P}_{2/0}$ and $\mathcal{P}_{2+0}$, B&B$_{\text{R-HOM}}$ performs better than MIP$_{\text{CPLEX}}$ on instances with lower complexity ($K = 5$ and $K = 7$), with a reduction of the computation time by a factor of up to 4 on the simplest problems ($\mathcal{P}_{2/0}$, $K = 5$). For $K = 9$, MIP$_{\text{CPLEX}}$

[2]Data are available online at `pagesperso.ls2n.fr/~bourguignon-s/download_MIP.html`

Table 3. Computational results for simulated sparse deconvolution problems. For each algorithm, columns "Time", "Nodes", and "F" respectively give the average computing time in seconds, the average number of explored nodes, and the number of problems that could not be solved in 1 000 s. Fifty instances of each problem were considered. Averages are performed only over the instances which could be solved in less than 1 000 s by the three algorithms.

Problem		Branch-and-bound						MIP solver CPLEX		
		$B\&B_{\text{R-HOM}}$			$B\&B_{\text{R-CPLEX}}$			CPLEX 12.8		
		Time (s)	Nodes (10^3)	F	Time (s)	Nodes (10^3)	F	Time (s)	Nodes (10^3)	F
$\mathcal{P}_{2/0}$	$K=5$	**0.7**	1.28	0	7.7	1.28	0	3.0	1.71	0
	$K=7$	**11.6**	17.89	0	141.9	17.89	2	16.6	21.51	0
	$K=9$	**43.5**	57.37	9	448.1	57.46	30	53.8	72.04	6
$\mathcal{P}_{2+0}$	$K=5$	**1.8**	2.01	0	32.6	2.02	0	3.2	1.98	0
	$K=7$	**7.3**	10.20	0	187.3	10.22	7	7.4	9.61	0
	$K=9$	25.6	31.80	5	470.7	31.87	28	**17.3**	23.74	2
$\mathcal{P}_{0/2}$	$K=5$	**0.1**	0.21	0	6.0	0.21	0	25.7	6.71	0
	$K=7$	**0.9**	2.32	0	85.2	2.32	0	114.8	49.54	2
	$K=9$	**2.5**	5.22	0	296.9	5.22	18	328.2	101.07	17

solves more instances in less than 1 000 s, and is up to 30% faster (for $\mathcal{P}_{2+0}$) on the solved instances. This is not surprising, since the CPLEX quadratic MIP solver benefits from many additional developments that are not implemented in our branch-and-bound algorithm, such as cutting planes, heuristics, *etc.*

Results are much more contrasted with problem $\mathcal{P}_{0/2}$, where $B\&B_{\text{R-HOM}}$ strongly outperforms $\text{MIP}_{\text{CPLEX}}$ on all instances, both in terms of number of explored nodes and of computation time. In particular, the computation time per node is reduced by more than 8 with the homotopy algorithm. As a result, $B\&B_{\text{R-HOM}}$ runs 120 to 250 times faster than $\text{MIP}_{\text{CPLEX}}$ on these instances, even with a rather simple branch-and-bound exploration strategy.

The performance profile, defined by Dolan and Moré in [8], is a visual tool to benchmark algorithms. The fraction of solved problems is represented as a function of the performance ratio: for each instance, the ratio between the computing time of each algorithm and the smaller one is computed. Then, the performance profile draws the cumulative distribution functions for such ratios. The performance profile obtained over the 450 considered instances (50 instances for each of the 9 problems defined in Table 3) is shown in Figure 4 left, which confirms the results in Table 3. In particular, due the strong superiority of $B\&B_{\text{R-HOM}}$ over $\text{MIP}_{\text{CPLEX}}$ in all instances of problems $\mathcal{P}_{0/2}$, the performance of $B\&B_{\text{R-HOM}}$ is far above that of $\text{MIP}_{\text{CPLEX}}$ in all parts of the profile. For example, in twice the best algorithm time, $B\&B_{\text{R-HOM}}$ solves 90% of problems, whereas $\text{MIP}_{\text{CPLEX}}$ only solves 40% of them (see the vertical line in Figure 4).

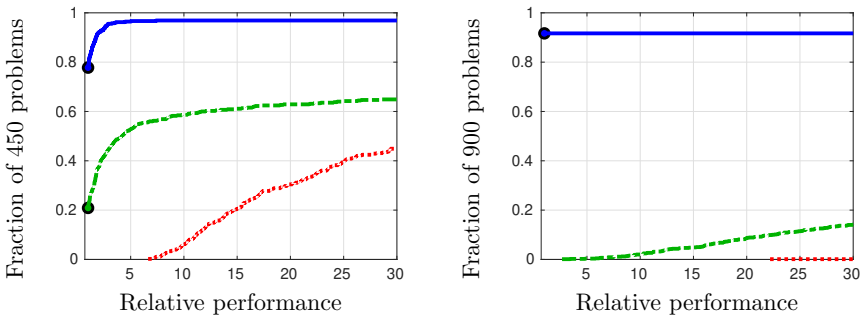


Figure 4. Performance profiles obtained on 450 sparse deconvolution problems (left, see Section 4.1) and on 900 random subset selection problems (right, see Section 4.2), for the three algorithms: B&B$_{\text{R-HOM}}$ (full line), MIP$_{\text{CPLEX}}$ (dotted line) and B&B$_{\text{R-CPLEX}}$ (dashed line).

4.2 *Simulated Subset selection problems*

We now study algorithmic performance on artificial subset selection problems with random entries (see [2, 30] for similar simulations). We consider a dictionary $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, with $n = 2 \times m$, composed of independent and identically distributed random entries, which are drawn from a centered, unit-variance, normal distribution (columns have been normalized). Sparse vectors are generated with randomly chosen non-zero coordinates, and corresponding amplitudes are drawn as $u + \text{sgn}(u)$, where u follows a centered, unit variance, Gaussian distribution. Gaussian noise is added such that SNR = 10 dB. The parameter M and formulation-specific parameters K, ϵ and μ are tuned as in Section 4.1. Note that the random nature of the dictionary makes problems easier to solve, which allows us to increase the problem dimension: we consider here $n = 500$ and $n = 1\,000$, and the cardinality varies from $K = 5$ to $K = 15$. Indeed, the compressed sensing theory (see for example [13]) shows that, if the correlation between columns in $\boldsymbol{A}$ is sufficiently low, then the ℓ_1-norm relaxation may solve the original ℓ_0-norm problem with high probability. Consequently, fewer nodes are expected to be explored in the branch-and-bound algorithm than with deconvolution problems of the same size.

Computational results are reported in Table 4. As in Section 4.1, B&B$_{\text{R-HOM}}$ explores the same number of nodes as B&B$_{\text{R-CPLEX}}$, but B&B$_{\text{R-HOM}}$ requires significantly less execution time: about a factor of 50 to 500 for $\mathcal{P}_{2/0}$ and $\mathcal{P}_{2+0}$, and about a factor of 200 to 500 for $\mathcal{P}_{0/2}$. It also successfully solves more instances in less than 1 000 s.

B&B$_{\text{R-HOM}}$ performs dramatically better than MIP$_{\text{CPLEX}}$ on all problems in terms of execution time: it generally runs more than 20 times faster (up to 1 800 times faster on problems $\mathcal{P}_{0/2}$ with $K = 5$ and $n = 1\,000$). For all problems, the average number of explored nodes is lower with B&B$_{\text{R-HOM}}$, with a factor up to 5 for $\mathcal{P}_{2/0}$ and $\mathcal{P}_{2+0}$; for $\mathcal{P}_{0/2}$, MIP$_{\text{CPLEX}}$ could only solve a few instances, for which the number of nodes was 30 to 40 times greater than for B&B$_{\text{R-HOM}}$, which successfully solved most instances in less than 1 000 s. Surprisingly, for this problem, MIP$_{\text{CPLEX}}$ also performs worse than B&B$_{\text{R-CPLEX}}$, probably due to inefficient exploration strategies in the CPLEX quadratically constrained MIP solver. Most of all, we observe that the most important improvement achieved by B&B$_{\text{R-HOM}}$ is due to the efficiency of the continuous relaxation algorithm: the computing time per node with the proposed homotopy method is 5 times smaller (problem $\mathcal{P}_{2+0}$, $n = 500$, $K = 15$) to almost 100 times smaller (problem $\mathcal{P}_{0/2}$, $n = 1\,000$, $K = 10$) than that of MIP$_{\text{CPLEX}}$. The performance profile obtained over the 900 considered instances (50 instances for each of the 18 problems defined in Table 3), shown in Figure 4 right, confirms such superiority.

Table 4. Computational results for simulated subset selection problems. For each algorithm, columns "Time", "Nodes", and "F" respectively give the average computing time in seconds, the average number of explored nodes, and the number of problems that could not be solved in 1 000 s. Fifty instances of each problem were considered. Averages are performed only over the instances which could be solved in less than 1 000 s by the three algorithms, except on the last row, where only B&B$_{\text{R-HOM}}$ could solve the 50 instances

Problem			Branch-and-bound						MIP solver CPLEX		
			B&B$_{\text{R-HOM}}$			B&B$_{\text{R-CPLEX}}$			CPLEX 12.8		
			Time (s)	Nodes (10^3)	F	Time (s)	Nodes (10^3)	F	Time (s)	Nodes (10^3)	F
$n = 500$	$\mathcal{P}_{2/0}$	$K = 5$	**0.4**	0.03	0	23.5	0.03	0	16.8	0.04	0
		$K = 10$	**3.8**	0.26	0	249.4	0.26	12	107.0	0.51	5
		$K = 15$	**7.6**	0.59	24	538.7	0.59	48	161.4	1.09	45
	$\mathcal{P}_{2+0}$	$K = 5$	**0.4**	0.13	0	205.0	0.13	1	18.8	0.12	0
		$K = 10$	**1.5**	0.29	0	442.0	0.29	21	53.7	0.76	3
		$K = 15$	**7.3**	0.63	26	626.9	0.63	49	179.4	3.28	47
	$\mathcal{P}_{0/2}$	$K = 5$	**0.03**	0.01	0	85.8	0.01	2	191.1	4.28	17
		$K = 10$	**0.5**	0.07	0	552.6	0.07	20	-	-	50
		$K = 15$	**1.5**	0.17	7	809.1	0.17	49	-	-	50
$n = 1\,000$	$\mathcal{P}_{2/0}$	$K = 5$	**1.1**	0.02	0	109.2	0.02	0	70.7	0.03	0
		$K = 10$	**6.2**	0.06	0	437.7	0.06	13	371.7	0.31	5
		$K = 15$	**33.9**	0.38	7	-	-	50	665.3	0.72	42
	$\mathcal{P}_{2+0}$	$K = 5$	**0.6**	0.06	0	275.5	0.06	10	118.1	0.11	0
		$K = 10$	**2.5**	0.13	0	508.2	0.13	32	189.4	0.42	3
		$K = 15$	**19.5**	0.52	11	-	-	50	665.3	2.15	40
	$\mathcal{P}_{0/2}$	$K = 5$	**0.2**	0.01	0	99.3	0.01	0	360.9	0.28	28
		$K = 10$	**2.6**	0.04	0	462.9	0.04	6	-	-	50
		$K = 15$	**113**	1.02	0	-	-	50	-	-	50

5. Conclusion

We have investigated exact optimization algorithms for least squares problems with low cardinality. A dedicated branch-and-bound procedure was proposed, which removes the binary variables that are usually introduced for the MIP reformulation. An algorithm was built for solving the continuous relaxation problems involved in any node of the search tree, which were recast as convex non-smooth optimization problems, involving the ℓ_1

norm. Inspired by the homotopy principle, this algorithm can be applied with similar efficiency for the three addressed formulations: cardinality-constrained and cardinality-penalized least-squares, and cardinality minimization under quadratic constraints. On all conducted experiments, the resulting algorithms at least equal and often outperform the state-of-the art MIP solver CPLEX. Our algorithms were able to tackle problems of higher dimension (number of variables, cardinality), which become intractable with CPLEX. In particular, the proposed strategy was shown to solve much more efficiently the quadratically constrained formulation, which is of major interest in many applications.

Such improvements were achieved by exploiting mathematical properties of the problem, which are not considered by a generic solver. Further works may be developed following the same guideline. Building tighter relaxations, *e.g.* based on Lagrangian relaxation, developing more refined exploration strategies, for example based on ad-hoc greedy algorithms [26] or building dedicated cutting plane methods are some possibilities. The three formulations could also be tackled jointly from a multi-objective optimization perspective, involving dedicated branch-and-bound techniques [11].

References

[1] D. Bertsimas, A. King, and R. Mazumder, *Best subset selection via a modern optimization lens*, The Annals of Statistics 44 (2016), pp. 813–852.

[2] D. Bertsimas and R. Shioda, *Algorithm for cardinality-constrained quadratic optimization*, Computational Optimization and Applications 43 (2009), pp. 1–22.

[3] D. Bienstock, *Computational study of a family of mixed-integer quadratic programming problems*, Mathematical Programming 74 (1996), pp. 121–140.

[4] S. Bourguignon, J. Ninin, H. Carfantan, and M. Mongeau, *Exact sparse approximation problems via mixed-integer programming: Formulations and computational performance*, IEEE Transactions on Signal Processing 64 (2016), pp. 1405–1419.

[5] A.M. Bruckstein, D.L. Donoho, and M. Elad, *From sparse solutions of systems of equations to sparse modeling of signals and images*, SIAM Review 51 (2009), pp. 34–81.

[6] S.S. Chen, D.L. Donoho, and M.A. Saunders, *Atomic decomposition by basis pursuit*, SIAM Journal on Scientific Computing 20 (1998), pp. 33–61.

[7] X. Cui, X. Zheng, S. Zhu, and X. Sun, *Convex relaxations and MIQCQP reformulations for a class of cardinality-constrained portfolio selection problems*, Journal of Global Optimization 56 (2013), pp. 1409–1423.

[8] E.D. Dolan and J.J. Moré, *Benchmarking optimization software with performance profiles*, Mathematical Programming 91 (2002), pp. 201–213.

[9] D.L. Donoho and Y. Tsaig, *Fast solution of ℓ_1-norm minimization problems when the solution may be sparse*, IEEE Transactions on Information Theory 54 (2008), pp. 4789–4812.

[10] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, The Annals of Statistics 32 (2004), pp. 407–499.

[11] M. Ehrgott, *Multicriteria Optimization*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[12] M. Elad, *Sparse and Redundant Representations. From Theory to Applications in Signal and Image Processing*, Springer-Verlag New York, 2010.

[13] Y. Eldar and G. Kutyniok, *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012.

[14] A. Frangioni and C. Gentile, *Perspective cuts for a class of convex 0–1 mixed integer programs*, Mathematical Programming 106 (2006), pp. 225–236.

[15] J. Gao and D. Li, *Optimal cardinality constrained portfolio selection*, Operations Research 61 (2013), pp. 745–761.

[16] J. Idier, *Bayesian Approach to Inverse Problems*, ISTE Ltd and John Wiley & Sons Inc, 2018.

[17] D. Li, X. Sun, and J. Wang, *Optimal lot solution to cardinality constrained mean–variance formulation for portfolio selection*, Mathematical Finance 16 (2006), pp. 83–101.
[18] X. Liang and Y. Wang, *Homotopy algorithm for box-constrained LASSO and its convergence*, International Journal of Pure and Applied Mathematics 112 (2017), pp. 333–340.
[19] J.M. Mendel, *Optimal Seismic Deconvolution*, Academic Press, 1983.
[20] A. Miller, *Subset selection in regression*, Chapman and Hall/CRC, 2002.
[21] B.K. Natarajan, *Sparse approximate solutions to linear systems*, SIAM Journal on Computing 24 (1995), pp. 227–234.
[22] M.S. O'Brien, A.N. Sinclair, and S.M. Kramer, *Recovery of a sparse spike time series by L1 norm deconvolution*, IEEE Transactions on Signal Processing 42 (1994), pp. 3353–3365.
[23] M. Osborne, B.P. B, and B.T. BAD, *A new approach to variable selection in least squares problems*, IMA Journal of Numerical Analysis (2000).
[24] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.
[25] D.X. Shaw, S. Liu, and L. Kopman, *Lagrangian relaxation procedure for cardinality-constrained portfolio optimization*, Optimization Methods and Software 23 (2008), pp. 411–420.
[26] C. Soussen, J. Idier, D. Brie, and J. Duan, *From Bernoulli Gaussian deconvolution to sparse signal restoration*, IEEE Transactions on Signal Processing 59 (2011), pp. 4572–4584.
[27] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B 58 (1996), pp. 267–288.
[28] J.A. Tropp and S.J. Wright, *Computational methods for sparse solution of linear inverse problems*, Proceedings of the IEEE 98 (2010), pp. 948–958.
[29] C. Zala, *High-resolution inversion of ultrasonic traces*, IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control 39 (1992), pp. 458–463.
[30] X. Zheng, X. Sun, D. Li, and J. Sun, *Successive convex approximations to cardinality-constrained convex programs: a piecewise-linear DC approach*, Computational Optimization and Applications 59 (2014), pp. 379–397.

Annexe B

Exact sparse approximation problems via mixed-integer programming : Formulations and computational performance.

S. Bourguignon, J. Ninin, H. Carfantan et M. Mongeau : Exact sparse approximation problems via mixed-integer programming : Formulations and computational performance. *IEEE Transactions on Signal Processing, 64(6) :1405–1419*, mars 2016.

IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 64, NO. 6, MARCH 15, 2016

Exact Sparse Approximation Problems *via* Mixed-Integer Programming: Formulations and Computational Performance

Sébastien Bourguignon, Jordan Ninin, Hervé Carfantan, and Marcel Mongeau, *Member, IEEE*

***Abstract*—Sparse approximation addresses the problem of approximately fitting a linear model with a solution having as few non-zero components as possible. While most sparse estimation algorithms rely on suboptimal formulations, this work studies the performance of exact optimization of ℓ_0-norm-based problems through Mixed-Integer Programs (MIPs). Nine different sparse optimization problems are formulated based on ℓ_1, ℓ_2 or ℓ_∞ data misfit measures, and involving whether constrained or penalized formulations. For each problem, MIP reformulations allow exact optimization, with optimality proof, for moderate-size yet difficult sparse estimation problems. Algorithmic efficiency of all formulations is evaluated on sparse deconvolution problems. This study promotes error-constrained minimization of the ℓ_0 norm as the most efficient choice when associated with ℓ_1 and ℓ_∞ misfits, while the ℓ_2 misfit is more efficiently optimized with sparsity-constrained and sparsity-penalized problems. Exact ℓ_0-norm optimization is shown to outperform classical methods in terms of solution quality, both for over- and underdetermined problems. Numerical simulations emphasize the relevance of the different ℓ_p fitting possibilities as a function of the noise statistical distribution. Such exact approaches are shown to be an efficient alternative, in moderate dimension, to classical (suboptimal) sparse approximation algorithms with ℓ_2 data misfit. They also provide an algorithmic solution to less common sparse optimization problems based on ℓ_1 and ℓ_∞ misfits. For each formulation, simulated test problems are proposed where optima have been successfully computed. Data and optimal solutions are made available as potential benchmarks for evaluating other sparse approximation methods.**

***Index Terms*—Deconvolution, mixed-integer programming, ℓ_0-norm-based problems, optimization, sparse approximation.**

Manuscript received March 02, 2015; revised July 29, 2015 and September 20, 2015; accepted October 12, 2015. Date of publication October 30, 2015; date of current version February 09, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ami Wiesel. This work was partially supported by the French Groupement de Recherche ISIS (Information, Signal, Image, viSion) through its "young researchers" program. One of the authors of this work has been supported by French National Research Agency (ANR) through JCJC program (ANR project ATOMIC 12-JS02-009-01).

S. Bourguignon is with IRCCyN, CNRS UMR 6597, Ecole Centrale de Nantes, 44321, Nantes Cedex 3, France (e-mail: Sebastien.Bourguignon@ec-nantes.fr).

J. Ninin is with Lab-STICC, CNRS UMR 6285, ENSTA Bretagne, 29806 Brest Cedex, France (e-mail: Jordan.Ninin@ensta-bretagne.fr).

H. Carfantan is with IRAP, CNRS UMR 5277, Université Toulouse III – Paul Sabatier, 31400 Toulouse, France (e-mail: Herve.Carfantan@irap.omp.eu).

M. Mongeau is with MAIAA, ENAC, Université de Toulouse, 31055 Toulouse cedex 4, France (e-mail: Marcel.Mongeau@enac.fr).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org provided by the authors. This includes simulated data and optimization results obtained by MIP approaches, corresponding to the sparse deconvolution problems addressed in Sections V and VI. This material is 3.5 MB in size.

Digital Object Identifier 10.1109/TSP.2015.2496367

I. Introduction

A. Sparse Estimation for Inverse Problems

THE problem of *sparse representation* of data $\boldsymbol{y} \in \mathbb{R}^N$ in a *dictionary* $\mathbf{H} \in \mathbb{R}^N \times \mathbb{R}^Q$ consists in finding a solution $\boldsymbol{x} \in \mathbb{R}^Q$ to the system $\boldsymbol{y} = \mathbf{H}\boldsymbol{x}$ with the fewest non-zero components, *i.e.*, with the lowest *sparsity level*. In *sparse approximation*, in order to account for noise and model errors, the equality constraint is relaxed through the minimization of the data misfit measure $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|$, where $\|\cdot\|$ generally stands for the standard Euclidean norm in $\mathbb{R}^N$. Such sparsest representation and approximation problems are essentially combinatorial. Finding the best K-sparse solution (the solution with K non-zero components) is usually considered too difficult in practical large-scale instances. Indeed, the brute-force approach that amounts to exploring all the $\binom{Q}{K}$ possible combinations, is computationally prohibitive. In the abundant literature on sparse approximation, much work has been dedicated to the relaxation approach that replaces the ℓ_0-"norm" sparsity measure, $\|\boldsymbol{x}\|_0 := \mathrm{Card}\{i \,|\, x_i \neq 0\}$, with the ℓ_1 norm $\|\boldsymbol{x}\|_1 := \sum_i |x_i|$. Many specific convex optimization algorithms have been proposed in the past decade, see for example [1], [2]. In addition, conditions were established for which both the ℓ_0 and the relaxed ℓ_1 problems yield the same *solution support* (the set of non-zero components). These mostly rely on a low sparsity level assumption and on structural hypotheses on the matrix $\mathbf{H}$, such as low correlation of its columns (see [1] and references therein). Alternatively, greedy algorithms build a sparse solution by iteratively adding non-zero components to an initially empty-support solution [3]–[5]. More complex forward-backward methods [6], [7] may show better performance in practice but with higher computation time. Tree-search-based methods also try to improve the classical greedy algorithms using heuristics to reduce the complexity of exhaustive combinatorial exploration (see *e.g.,* [8] and references therein). Other support exploration strategies maintain the desired sparsity level at each iteration, and perform local combinatorial exploration steps [9]. Optimality proofs for all such strategies also rely on very restrictive hypotheses [10], [11]. More "ℓ_0-oriented" approaches were proposed, *e.g.*, by successive continuous approximations of the ℓ_0 norm [12], by descent-based Iterative Hard Thresholding (IHT) [13], [14] and by penalty decomposition methods [15]. However, without additional assumptions on $\mathbf{H}$, one can only prove that the solution

found is a local minimum of the optimization problem. Moreover, for IHT, optimality conditions suffer from the same restrictions as the aforementioned greedy methods [13], [14].

In many inverse problems, the model $\boldsymbol{y} \simeq \mathbf{H}\boldsymbol{x}$ results from the discretization of an intrinsically continuous physical model. A typical example is sparse deconvolution, where $\mathbf{H}\boldsymbol{x}$ models the convolution of a spike train (in one-dimensional signals) or of point sources (in imaging problems) by the impulse response of the acquisition device [7], [16], [17]. A similar problem concerns nonlinear parameter identification, where parameters are discretized on arbitrarily thin grids [7], [18], [19] and estimation amounts to finding a sparse solution to a linear problem of high dimension. In such cases, the columns of $\mathbf{H}$ can be highly correlated, so no optimality guaranty can be obtained for greedy and ℓ_1-norm-based methods. Similar problems also arise for variable selection in machine learning and statistics [6], where the set of features (the columns of $\mathbf{H}$) is not designed to satisfy any recovery property.

The aforementioned problems essentially focus on the correct estimation of the support of $\boldsymbol{x}$. In deconvolution, for example, support identification corresponds to detection and localization of the sources. Since the true sparsity measure is indeed the ℓ_0 norm, a global optimum of ℓ_0-based formulations is more likely to yield exact support identification than approximate solutions. Consequently, our interest focuses on optimization methods for ℓ_0-norm-based criteria providing global optimality guarantees. Such *exact* approaches are usually discarded, based on the argument that sparse optimization problems are NP hard [20]. It is also commonly considered that exact optimization amounts to combinatorial exploration of all possible supports, which is nothing but a worst-case-scenario argument. Note however that, in order to reduce the number of explored supports, Tropp and Wright [1] mentioned the possible use of *cutting-plane methods*, which are one of the basic elements of resolution of the mixed-integer programs explored in the present paper.

Here, we focus on sparse optimization occurring in certain inverse problems with moderate size yet with a complexity sufficient to make the usual methods fail in estimating the sparsest solutions. Examples include spike train deconvolution in ultrasonic nondestructive testing (NDT) [16] or Geophysics [17], sparse hyperspectral unmixing [21], and spectral analysis with short data sets [22]. A first objective of this contribution is to show the viability of exact resolution approaches for such problems.

B. Global Optimization via Mixed-Integer Programming

We focus on Mixed-Integer Programs (MIP), that is, optimization problems involving both continuous and integer variables. Such problems are well suited to ℓ_0-norm-based optimization, since the ℓ_0 norm naturally introduces a binary decision variable for each component (zero or non-zero?). In this paper, MIP refers to the minimization of linear or quadratic criteria subject to linear or quadratic inequality constraints. It is commonly claimed that, in the past fifteen years, a factor 10^9 was gained in the required computing time for solving such problems. This gain is due in (roughly) equal parts to (i) hardware improvement, (ii) progress in the resolution of linear programs, and (iii) implementation efficiency of advanced mathematical techniques [23]. Therefore, as it will be shown in this paper, some exact approaches can now be advantageously used to address the moderate-size, yet difficult, applications enumerated at the end of Section I.A.

To our knowledge, the first MIP reformulation of a sparse optimization problem is proposed in [24]. However, the authors argue that the assumption that $|\boldsymbol{x}|$ is upper bounded, which is required for the MIP reformulation, leads to computational inefficiency. Therefore, they choose to consider only a related problem: the maximum feasible subsystem problem, for which exact solutions could be found only for very small instances ($N = 16$, $Q = 32$) and no result is given concerning the MIP approach. A similar formulation with binary variables appears in [25], but binary variables are replaced by continuous variables in [0,1] in order to yield a convex problem, which is obviously not equivalent to the original one. In [26], some exact and approximate reformulations of ℓ_0-based problems are surveyed. The authors deplore the inherent combinatorial difficulty of such MIP problems but no practical result is provided. Finally, in [27], noise-free sparse *representation* problems are formulated as MIP. Here, we address the noisy case, which opens perspectives to different possible formulations of the optimization problem. Establishing MIP reformulations for such problems, studying their computational efficiency, investigating properties of optimal solutions and comparing them with the results of standard methods are the core of this paper.

C. Objectives of the Paper

This paper shows that different possible reformulations of sparse approximation problems can be tackled by MIP solvers. Sparse approximation is intrinsically a bi-objective optimization problem, where both the sparsity measure and the data misfit measure are optimized. In inverse problems, it is usually formulated through the optimization of a weighted sum of the two terms. However, constrained formulations (involving one criterion to be minimized and the other subject to a constraint) may also be well suited to MIP reformulations, which are constrained programs by nature. Therefore, we study the efficiency of MIP solving techniques applied to the following three formulations:

- minimize the ℓ_0 norm under a bounded-data-misfit constraint,
- minimize the data misfit under an ℓ_0-boundedness constraint,
- minimize a weighted sum of the two criteria.

Additionally, we consider non-quadratic data misfit measures, which may be appropriate if the error term $\boldsymbol{y} - \mathbf{H}\boldsymbol{x}$ is non-Gaussian. Moreover, piecewise-linear alternatives to the ℓ_2 norm $\|\boldsymbol{x}\|_2 := \sqrt{\sum_i |x_i|^2}$, may also prove to be more attractive computationally, because MIP solvers essentially rely on the resolution of *linear* subproblems. In particular, the ℓ_1 and ℓ_∞($\|\boldsymbol{x}\|_\infty := \max_i |x_i|$) norms are easily linearized; formulations based on those norms naturally boil down to optimization problems involving linear inequality constraints or linear objective functions. Therefore, we also consider formulations involving ℓ_1 and ℓ_∞ misfits, for which much fewer algorithms have been proposed.

Our work establishes MIP reformulations of nine different sparse approximation problems, which are all evaluated in terms of computational efficiency, depending on the sparsity level and on the noise level. Then, these formulations are compared in their ability to identify the exact support in the presence of noise, depending on the noise statistical distribution. Our experimental results additionally show that the classical methods are far from reaching acceptable results in such cases, whereas solving the ℓ_0-norm formulations do yield more satisfactory solutions—but with much higher computing time. Note that experiments in both [24] and [27] involve random matrices, which are more likely to satisfy the conditions ensuring the optimality of ℓ_1-norm-based and greedy approaches. In most ill-posed inverse problems, the columns of $\mathbf{H}$ are highly correlated, so that such conditions certainly do not hold.

The remainder of the paper is organized as follows. Section II introduces nine optimization formulations of the sparse approximation problem, and discusses their statistical interpretation and the structure of the solution sets. MIP reformulations are established in Section III. Then, basic elements concerning the resolution of MIP problems are given in Section IV. Experimental results in Section V are dedicated to the evaluation of computational costs. Section VI compares the solutions obtained through MIP optimization with those of classical sparse approximation algorithms, on both overdetermined and underdetermined sparse deconvolution problems. Simulations in Section VII evaluate the support identification performance of ℓ_p-misfit-constrained formulations as a function of p and of the noise statistical distribution. Finally, a discussion is given in Section VIII.

II. Sparse Optimization Problems

In sparse approximation, both the sparsity of the solution and the fidelity of its corresponding data approximation are optimized. Therefore, the generic sparse approximation problem, which we are interested in, is the following unconstrained bi-objective optimization problem:

$$\min_{\boldsymbol{x}}(\|\boldsymbol{x}\|_0, \|\boldsymbol{y}-\mathbf{H}\boldsymbol{x}\|_p), \tag{1}$$

where p is either 1, 2 or ∞. This section presents nine formulations of this problem and discusses their statistical interpretation and the structure of the sets of solutions.

A. Taxonomy

Various mono-objective optimization problems can be formulated to address the bi-criterion problem (1). For $p \in \{1, 2, \infty\}$, the *bounded-error* problems read

$$\mathcal{P}_{0/p} : \min_{\boldsymbol{x}} \|\boldsymbol{x}\|_0 \text{ s.t. } \|\boldsymbol{y}-\mathbf{H}\boldsymbol{x}\|_p \leq \alpha_p,$$

and the *sparsity-constrained* problems read

$$\mathcal{P}_{p/0} : \min_{\boldsymbol{x}} \|\boldsymbol{y}-\mathbf{H}\boldsymbol{x}\|_p \text{ s.t. } \|\boldsymbol{x}\|_0 \leq K_p,$$

where α_p and K_p are user-defined threshold parameters. Finally, the *penalized* problems read

$$\mathcal{P}_{0+p} : \min_{\boldsymbol{x}} \ \mu_p\|\boldsymbol{x}\|_0 + \|\boldsymbol{y}-\mathbf{H}\boldsymbol{x}\|_p^p, \text{ for } p \in \{1, 2\},$$
$$\text{and } \mathcal{P}_{0+\infty} : \min_{\boldsymbol{x}} \ \mu_\infty\|\boldsymbol{x}\|_0 + \|\boldsymbol{y}-\mathbf{H}\boldsymbol{x}\|_\infty,$$

where μ_p and μ_∞ are user-defined penalty parameters.

In this paper, we propose a reformulation of each of these problems as MIPs. To the best of our knowledge, $\mathcal{P}_{0/\infty}$ is the only sparse approximation problem for which a MIP reformulation is mentioned [24]. Remark that the sparse *representation* case (noise-free data), which was recently tackled *via* MIPs in [27] with equality constraint $\boldsymbol{y} = \mathbf{H}\boldsymbol{x}$, is the special case of $\mathcal{P}_{0/p}$ with α_p set to 0, for which $\mathcal{P}_{0/1}, \mathcal{P}_{0/2}$, and $\mathcal{P}_{0/\infty}$ are obviously equivalent. Recall that the sparsity-based inverse problems considered here are sparse *approximation* problems: data are always contaminated by measurement noise and the model may be inexact, so that $\alpha_p \neq 0$.

Choosing one of the nine formulations and the value of the parameter (α_p, K_p or μ_p) amounts to selecting some particular solution among the wide variety of Pareto-optimal solutions of problem (1). Note that, for a given ℓ_p misfit, no equivalence between the three problems $\mathcal{P}_{0/p}$, $\mathcal{P}_{p/0}$ and $\mathcal{P}_{0+p}$ can be obtained because the ℓ_0 norm is not convex. In particular, solutions in the non-convex part of the Pareto frontier cannot be reached by solving the penalized formulation [28].

B. Statistical Interpretations and Parameter Tuning

In practice, one has to choose one among the nine optimization problems and must set a value for the corresponding parameter. Such choices can be based on statistical arguments.

The ℓ_p data-misfit measures, with $p \in \{1, 2, \infty\}$, can be interpreted in terms of likelihood functions. Let $p_{\boldsymbol{\epsilon}}$ be the statistical distribution of the additive noise term $\boldsymbol{\epsilon} = \boldsymbol{y} - \mathbf{H}\boldsymbol{x}$. The likelihood function is defined as: $\mathcal{L}(\boldsymbol{y};\boldsymbol{x}) := p_{\boldsymbol{\epsilon}}(\boldsymbol{y} - \mathbf{H}\boldsymbol{x})$. If noise samples ϵ_n, $n = 1, \ldots, N$ are independent and identically distributed (i.i.d.) according to a centered Gaussian distribution, then $-\log \mathcal{L}(\boldsymbol{y};\boldsymbol{x})$ is proportional to $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_2^2$ up to an additive constant. Similarly, $-\log \mathcal{L}(\boldsymbol{y};\boldsymbol{x})$ is proportional to $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_1$ (up to an additive constant) if ϵ_n are i.i.d. according to a centered Laplace distribution. Such a heavy-tailed distribution assumption may be appropriate in the presence of impulsive noise [16], [29]. The ℓ_∞ misfit is connected to an i.i.d. uniform noise distribution assumption. Suppose that $\boldsymbol{\epsilon}$ is uniformly distributed on $[-a, a]^N$, for some given $a > 0$. Then, the likelihood function is constant for any $\boldsymbol{x}$ such that $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_\infty \leq a$, otherwise it is zero. Consequently, $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_\infty$ is a maximum-likelihood estimator if noise samples are uniformly distributed on $[-a, a]$, for any $a > \|\boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{x}}\|_\infty$ [30, Ch. 7.1]. In this case, which arises for example when accounting for quantization noise, ℓ_∞ data fitting may be a relevant choice—see [31] for both theoretical and numerical arguments.

Consequently, $\mathcal{P}_{p/0}$ is a sparsity-constrained maximum likelihood estimation problem with the aforementioned corresponding noise distribution assumption. In a Bayesian setting, $\mathcal{P}_{0+p}$ defines a Maximum *A Posteriori* (MAP) estimate, where the ℓ_0 term results from a Bernoulli-Gaussian prior model with infinite variance Gaussian distribution [7]. Note that, within such a MAP interpretation, solving $\mathcal{P}_{0+\infty}$ reduces to solving $\mathcal{P}_{0/\infty}$. Indeed, since $-\log \mathcal{L}(\boldsymbol{y};\boldsymbol{x})$ is constant if $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_\infty \leq a$ and equals $+\infty$ otherwise, minimizing $\mu_\infty\|\boldsymbol{x}\|_0 - \log \mathcal{L}(\boldsymbol{y};\boldsymbol{x})$ amounts to minimizing $\|\boldsymbol{x}\|_0$ subject to $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_\infty \leq a$. Finally, $\mathcal{P}_{0/p}$ considers a maximal tolerance on the approximation error and cannot be interpreted as a maximum likelihood or MAP estimation problem.

The choice of the formulation also depends on available prior information about the considered practical problem. If reasonable bounds on the acceptable approximation error can be inferred, *e.g.*, from the knowledge of the signal-to-noise ratio or from a desired approximation quality, then $\mathcal{P}_{0/p}$ may be preferred. In particular, the parameter α_p can be fixed according to the statistics of $\|\boldsymbol{\epsilon}\|_p$, which can be obtained for any noise distribution (analytically or numerically). If the sparsity level is fixed or can be upper bounded, *e.g.*, in a compression context, then $\mathcal{P}_{p/0}$ may be appropriate. In $\mathcal{P}_{0+p}$, the parameter μ_p trades off between the noise level and the sparsity level. With the previous MAP interpretation, for $p \in \{1, 2\}$, it is an explicit function of the noise variance and of the rate of non-zero values in the Bernoulli process. Therefore, tuning μ_p requires more information than tuning the parameters of the two other formulations. When too little prior information is available, a practical solution consists in computing optimal solutions corresponding to different parameter tunings—whatever the considered formulation—and then selecting *a posteriori* the most appropriate one, according to some expert supervision or to model order selection criteria [32].

C. Structure of the Solution Sets

We investigate hereafter the structure of the solution sets of the different problems, for fixed values of the corresponding parameters α_p, K_p and μ_p. In the following, an *optimal support* refers to a set of indices which supports at least one optimal solution. The ℓ_0 norm is a piecewise constant function, where each value is attained on a finite number of supports. Hence, for any problem $\mathcal{P}$ defined in Section II.A, the set of minimizers can be defined as the finite union of sets of minimizers on each optimal support: if $\mathcal{S}$ denotes the set of optimal supports, then:

$$\operatorname{Arg} \mathcal{P} = \bigcup_{s \in \mathcal{S}} \operatorname{Arg} \mathcal{P}^s,$$

where $\mathcal{P}^s$ denotes the restriction of problem $\mathcal{P}$ to the support s.

Let us characterize the solution set of $\mathcal{P}^s$. We assume that the sparsity level of all solutions is lower than N and that the matrix $\mathbf{H}$ satisfies the *Unique Representation Property* (URP) [33], that is, any N columns of $\mathbf{H}$ are linearly independent. For any $\boldsymbol{x}$ supported by s, $\|\boldsymbol{x}\|_0$ is constant, hence $\mathcal{P}^s_{p/0}$ and $\mathcal{P}^s_{0+p}$ are solved by minimizing $\|\boldsymbol{y} - \mathbf{H}_s\boldsymbol{x}_s\|_p$, where $\boldsymbol{x}_s$ (respectively, $\mathbf{H}_s$) collects the non-zero components in $\boldsymbol{x}$ (respectively, the corresponding columns of $\mathbf{H}$). Thanks to the URP, $\mathbf{H}_s$ has full column rank and, for $p = 2$, $\mathcal{P}^s$ admits a unique solution (the least-squares solution). Consequently, the solution sets of $\mathcal{P}_{2/0}$ and $\mathcal{P}_{0+2}$ are both (finite) unions of singletons. The ℓ_p norms for $p = 1$ and $p = \infty$ are not strictly convex, therefore one can only claim that $\min_{\boldsymbol{x}_s} \|\boldsymbol{y} - \mathbf{H}_s\boldsymbol{x}_s\|_p$ is attained on a convex set, such that $\mathbf{H}_s\boldsymbol{x}_s$ lies in an ℓ_p-sphere in dimension $K_s - 1$, centered at $\boldsymbol{y}$. Consequently, for $p = 1$ and $p = \infty$, the solution sets of $\mathcal{P}_{p/0}$ and $\mathcal{P}_{0+p}$ are (finite) unions of convex sets of the form $\{\|\boldsymbol{y} - \mathbf{H}_s\boldsymbol{x}_s\|_p = \text{constant}\}$. Now, consider the solution set of $\mathcal{P}^s_{0/p}$, which is formed by all vectors $\boldsymbol{x}_s$ such that $\|\boldsymbol{y} - \mathbf{H}_s\boldsymbol{x}_s\|_p \leq \alpha_p$. In the particular case where $\min_{\boldsymbol{x}_s} \|\boldsymbol{y} - \mathbf{H}_s\boldsymbol{x}_s\|_p = \alpha_p$, it comes from the previous arguments that the solution set is a singleton for $p = 2$, and a convex set for $p = 1$ and $p = \infty$. But, in the most frequent case where $\min_{\boldsymbol{x}_s} \|\boldsymbol{y} - \mathbf{H}_s\boldsymbol{x}_s\|_p < \alpha_p$, the solution set $\{\boldsymbol{x}_s \mid \|\boldsymbol{y} - \mathbf{H}_s\boldsymbol{x}_s\|_p \leq \alpha_p\}$ is such that $\mathbf{H}_s\boldsymbol{x}_s$ lies in an ℓ_p-ball of dimension K_s, centered at $\boldsymbol{y}$, and the solution set of $\mathcal{P}_{0/p}$ is a finite union of such sets. Consequently, for a given p, the solution set of $\mathcal{P}_{0/p}$ is generally "larger" than the solution sets of $\mathcal{P}_{p/0}$ and $\mathcal{P}_{0+p}$. In particular, the minimizers of these last two problems may be unique. For example, with some additional assumptions on the data $\boldsymbol{y}$ and on the matrix $\mathbf{H}$, the solution of $\mathcal{P}_{0+2}$ is unique [34]. On the contrary, the minimizer of $\mathcal{P}_{0/p}$ is certainly not unique, except in very specific cases.

III. Mixed-Integer Reformulations

In this section, we establish the reformulations of optimization problems $\mathcal{P}_{0/p}$, $\mathcal{P}_{p/0}$ and $\mathcal{P}_{0+p}$ for $p \in \{1, 2, \infty\}$, as Mixed-Integer Linear Programs (MILPs), Mixed-Integer Quadratic Programs (MIQPs) or Mixed-Integer Quadratically Constrained (linear) Programs (MIQCPs).

A. Definitions of MILP, MIQP and MIQCP

The general form of an MILP is

$$\min_{\boldsymbol{v}} \boldsymbol{c}^\intercal \boldsymbol{v}, \text{ subject to (s.t.)} \begin{cases} \mathbf{A}_{\text{in}}\boldsymbol{v} \leq \boldsymbol{b}_{\text{in}}, \\ \mathbf{A}_{\text{eq}}\boldsymbol{v} = \boldsymbol{b}_{\text{eq}}, \\ \boldsymbol{l}_b \leq \boldsymbol{v} \leq \boldsymbol{u}_b, \\ v_j \in \mathbb{Z}, \forall j \in \mathcal{I}, \end{cases}$$

where $\boldsymbol{v} \in \mathbb{R}^J$ is the vector of optimization variables; $\boldsymbol{c} \in \mathbb{R}^J$ defines the linear objective function; $\boldsymbol{b}_{\text{in}} \in \mathbb{R}^{P_{\text{in}}}$, $\boldsymbol{b}_{\text{eq}} \in \mathbb{R}^{P_{\text{eq}}}$, $\mathbf{A}_{\text{in}} \in \mathbb{R}^{P_{\text{in}}} \times \mathbb{R}^J$ and $\mathbf{A}_{\text{eq}} \in \mathbb{R}^{P_{\text{eq}}} \times \mathbb{R}^J$ define the inequality and equality constraints; $\boldsymbol{l}_b$ and $\boldsymbol{u}_b \in \mathbb{R}^J$ are respectively the vectors of lower and upper bounds of the optimization variables; $\mathcal{I}$ is the index set corresponding to the components of $\boldsymbol{v}$ that are constrained to be integer-valued.

An MIQP has the general form:

$$\min_{\boldsymbol{v}} \frac{1}{2}\boldsymbol{v}^\intercal \mathbf{F} \boldsymbol{v} + \boldsymbol{c}^\intercal \boldsymbol{v}, \text{ s.t.} \begin{cases} \mathbf{A}_{\text{in}}\boldsymbol{v} \leq \boldsymbol{b}_{\text{in}}, \\ \mathbf{A}_{\text{eq}}\boldsymbol{v} = \boldsymbol{b}_{\text{eq}}, \\ \boldsymbol{l}_b \leq \boldsymbol{v} \leq \boldsymbol{u}_b, \\ v_j \in \mathbb{Z}, \forall j \in \mathcal{I}, \end{cases}$$

where $\mathbf{F}$ is a $J \times J$ matrix.

Finally, the form of an MIQCP that is of interest in this paper is:

$$\min_{\boldsymbol{v}} \boldsymbol{c}^\intercal \boldsymbol{v}, \text{ s.t.} \begin{cases} \frac{1}{2}\boldsymbol{v}^\intercal \mathbf{B} \boldsymbol{v} + \mathbf{d}^\intercal \boldsymbol{v} \leq e, \\ \mathbf{A}_{\text{in}}\boldsymbol{v} \leq \boldsymbol{b}_{\text{in}}, \\ \mathbf{A}_{\text{eq}}\boldsymbol{v} = \boldsymbol{b}_{\text{eq}}, \\ \boldsymbol{l}_b \leq \boldsymbol{v} \leq \boldsymbol{u}_b, \\ v_j \in \mathbb{Z}, \forall j \in \mathcal{I}, \end{cases}$$

where $\mathbf{B}$ is a $J \times J$ matrix, $\mathbf{d} \in \mathbb{R}^J$ and $e \in \mathbb{R}$.

B. Equivalent Reformulation Techniques

We now present standard reformulation techniques that enables to express each of the nine optimization problems introduced in Section II.A as an MILP, an MIQP or an MIQCP, without any approximation.

1) Boundedness Assumption and "Big-M" Reformulation of the ℓ_0 Norm: For each $q = 1, 2, \ldots, Q$, let us introduce an additional binary optimization variable b_q, such that

$$b_q = 0 \Leftrightarrow x_q = 0. \tag{2}$$

Then, the non-linear sparsity measure $\|\boldsymbol{x}\|_0$ is equal to the linear term $\mathbf{1}_Q^\intercal \boldsymbol{b}$ $(= \sum_q b_q)$, where $\mathbf{1}_Q$ is the Q-dimensional all-ones column vector. The logical constraint (2) must however be translated into (in)equality constraints compatible with MIP, ideally through *linear* constraints. One standard way to achieve this (see *e.g.* [24], [27]) is to assume that a solution $\boldsymbol{x}$ of the problem under consideration satisfies the following constraints for some sufficiently large pre-defined value $M > 0$:

$$-M\mathbf{1}_Q < \boldsymbol{x} < M\mathbf{1}_Q. \tag{3}$$

This assumption supposes that the problem admits bounded optimal solutions, which is not restrictive in our practical applications. The parameter M has to be large enough so that $\|\hat{\boldsymbol{x}}\|_\infty < M$ at any desirable optimal solution $\hat{\boldsymbol{x}}$. On the other hand, the bound M must be as tight as possible in order to improve computational efficiency; therefore tuning the value of parameter M may be a critical issue [24]. In the problems addressed in this paper, satisfactory results are obtained with a rather simple empirical rule discussed in Section V.B. Note that specific lower and upper bounds for each component x_q could also be advantageously considered [27] if corresponding prior information is available.

The reformulations of ℓ_0-norm-based constraints and objective functions are obtained through the two following respective lemmas.

Lemma 1: Considering the boundedness assumption (3),

$$\|\boldsymbol{x}\|_0 \leq K \Leftrightarrow \begin{cases} \exists \boldsymbol{b} \in \{0;1\}^Q \text{ such that} \\ \sum_{q=1}^{Q} b_q \leq K, \quad (i) \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b}. \quad (ii) \end{cases}$$

Proof: The $\Rightarrow$ implication is straightforward by considering $\boldsymbol{b}$ defined by (2). Now, let $\boldsymbol{b} \in \{0;1\}^Q$ satisfy (i) and (ii), and suppose $\|\boldsymbol{x}\|_0 > K$. From (ii), one has $(b_q = 0) \Rightarrow (x_q = 0)$, that is, $(x_q \neq 0) \Rightarrow (b_q = 1)$. Hence $b_q = 1$ for at least $K + 1$ indices q, which contradicts (i). Consequently, $\|\boldsymbol{x}\|_0 \leq K$. ∎

Lemma 2: Considering the boundedness assumption (3),

$$\min_{\boldsymbol{x}\in\mathcal{F}} \|\boldsymbol{x}\|_0 \Leftrightarrow \min_{\substack{\boldsymbol{x}\in\mathcal{F} \\ \mathbf{b}\in\{0,1\}^Q}} \sum_{q=1}^{Q} b_q \text{ s.t. } -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b},$$

where $\mathcal{F}$ represents the feasible domain of the problem under consideration.

Proof: Similar to that of Lemma 1. ∎

Such a reformulation technique is commonly referred to as "big-M" reformulation. Remark finally that another reformulation of the cumbersome logical constraint (2) consists in introducing the equality constraint $x_q(b_q - 1) = 0$. However, the latter is a bi-linear constraint, typically less interesting in terms of computation time for off-the-self MIP solvers than linear constraints [35].

2) Reformulation of the ℓ_1 Data Misfit Measure: The ℓ_1 misfit term can be written linearly as $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_1 = \sum_n w_n$, with additional constraints $w_n = |y_n - \boldsymbol{h}_n^r \boldsymbol{x}|, n = 1, \ldots, N$, where $\boldsymbol{h}_n^r$ denotes the nth row of $\mathbf{H}$. Then, these constraints can be relaxed (exactly) by the linear inequalities: $-\boldsymbol{w} \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq \boldsymbol{w}$, with column vector $\boldsymbol{w} = [w_1, \ldots, w_N]^\intercal$, thanks to the two following lemmas.

Lemma 3:

$$\min_{\boldsymbol{x}\in\mathcal{F}} \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_1 \Leftrightarrow \min_{\boldsymbol{x}\in\mathcal{F}, \boldsymbol{w}\in\mathbf{R}^N} \sum_n w_n \text{ s.t. } -\boldsymbol{w} \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq \boldsymbol{w}.$$

Proof: Let $f_n(\boldsymbol{x}) = y_n - \boldsymbol{h}_n^r \boldsymbol{x}$. The following optimization problems are equivalent:

$$\mathcal{P}_a : \min_{\boldsymbol{x}\in\mathcal{F}} \sum_n |f_n(\boldsymbol{x})|$$
$$\mathcal{P}_b : \min_{\boldsymbol{x}\in\mathcal{F}, \boldsymbol{w}\in\mathbb{R}^N} \sum_n w_n \text{ s.t.} |f_n(\boldsymbol{x})| = w_n, \forall n$$
$$\mathcal{P}_c : \min_{\boldsymbol{x}\in\mathcal{F}, \boldsymbol{w}\in\mathbb{R}^N} \sum_n w_n \text{ s.t. } |f_n(\boldsymbol{x})| \leq w_n, \forall n.$$

Indeed, $\mathcal{P}_a \Leftrightarrow \mathcal{P}_b$ is trivial. In order to show that $\mathcal{P}_b \Leftrightarrow \mathcal{P}_c$, one can simply remark that if an optimal solution $(\boldsymbol{x}^\star, \boldsymbol{w}^\star)$ of $\mathcal{P}_c$ is such that $|f_{n_0}(\boldsymbol{x}^\star)| < w^\star_{n_0}$ for some index n_0, then one can straightforwardly construct a better feasible solution for $\mathcal{P}_c$, which yields a contradiction. ∎

Lemma 3 will be used to obtain a MIP reformulation of $\mathcal{P}_{1/0}$ and $\mathcal{P}_{0+1}$, which involve the ℓ_1-misfit term in the objective function. For $\mathcal{P}_{0/1}$, which involves the ℓ_1-misfit term as a constraint, we use the following lemma:

Lemma 4: Let $(\boldsymbol{x}^\star, \boldsymbol{w}^\star)$ solve the optimization problem:

$$\mathcal{P}_d : \min_{\boldsymbol{x}\in\mathbb{R}^Q, \boldsymbol{w}\in\mathbb{R}^N} \|\boldsymbol{x}\|_0 \text{ s.t. } \begin{cases} \sum_n w_n \leq \alpha_1 \\ -\boldsymbol{w} \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq \boldsymbol{w}. \end{cases}$$

Then, $\boldsymbol{x}^\star$ is a solution of $\mathcal{P}_{0/1}$.

Proof: Suppose that $(\boldsymbol{x}^\star, \boldsymbol{w}^\star)$ solves $(\mathcal{P}_d)$ and let $w'_n := |\boldsymbol{y} - \boldsymbol{h}_n^r \boldsymbol{x}^\star|$, $\forall n$. Then, $\sum_n w'_n \leq \sum_n w^\star_n \leq \alpha_1$ and $(\boldsymbol{x}^\star, \boldsymbol{w}')$ is a solution of:

$$\mathcal{P}_e : \min_{\boldsymbol{x}\in\mathbb{R}^Q, \boldsymbol{w}\in\mathbb{R}^N} \|\boldsymbol{x}\|_0 \text{ s.t. } \begin{cases} \sum_n w_n \leq \alpha_1 \\ |y_n - \boldsymbol{h}_n^r \boldsymbol{x}| = w_n, \forall n. \end{cases}$$

Indeed, since $\mathcal{P}_d$ is a relaxation of $\mathcal{P}_e$—the feasible set of $\mathcal{P}_e$ is a subset of that of $\mathcal{P}_d$—, its optimal value, $\|\boldsymbol{x}^\star\|_0$, is a lower bound for the optimal value of $\mathcal{P}_e$. The solution $(\boldsymbol{x}^\star, \boldsymbol{w}')$ is clearly feasible for $\mathcal{P}_e$ and it attains the lower bound $\|\boldsymbol{x}^\star\|_0$. Hence, it is optimal for $\mathcal{P}_e$. Finally, $\mathcal{P}_e$ is clearly equivalent to $\mathcal{P}_{0/1}$. ∎

Let us remark finally that problems $\mathcal{P}_d$ and $\mathcal{P}_e$ are not strictly equivalent because they are not minimized by the same couple of vectors $(\boldsymbol{x}, \boldsymbol{w})$, but they share the same solution set for $\boldsymbol{x}$.

TABLE I
MIXED-INTEGER PROGRAMMING REFORMULATIONS OF NINE SPARSE APPROXIMATION PROBLEMS

Problem	Equivalent MIP reformulation	MIP class
Bounded-misfit formulations (parameters: α_2, α_1 and α_∞)		
$\mathcal{P}_{0/2}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q} \mathbf{1}_Q^\intercal \boldsymbol{b}$ s.t. $\begin{cases} \boldsymbol{x}^\intercal\mathbf{H}^\intercal\mathbf{H}\boldsymbol{x} - 2\boldsymbol{y}^\intercal\mathbf{H}\boldsymbol{x} \leq \alpha_2^2 - \boldsymbol{y}^\intercal\boldsymbol{y} \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$	MIQCP
$\mathcal{P}_{0/1}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q,\boldsymbol{w}\in\mathbb{R}^N} \mathbf{1}_Q^\intercal \boldsymbol{b}$ s.t. $\begin{cases} \mathbf{1}_N^\intercal\boldsymbol{w} \leq \alpha_1 \\ -\boldsymbol{w} \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq \boldsymbol{w} \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$	MILP
$\mathcal{P}_{0/\infty}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q} \mathbf{1}_Q^\intercal \boldsymbol{b}$ s.t. $\begin{cases} -\alpha_\infty\mathbf{1}_N \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq \alpha_\infty\mathbf{1}_N \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$	MILP
Sparsity-constrained formulations (parameters: K_2, K_1 and K_∞)		
$\mathcal{P}_{2/0}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q} \boldsymbol{x}^\intercal\mathbf{H}^\intercal\mathbf{H}\boldsymbol{x} - 2\boldsymbol{y}^\intercal\mathbf{H}\boldsymbol{x}$ s.t. $\begin{cases} \mathbf{1}_Q^\intercal\boldsymbol{b} \leq K_2 \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$	MIQP
$\mathcal{P}_{1/0}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q,\boldsymbol{w}\in\mathbb{R}^N} \mathbf{1}_N^\intercal \boldsymbol{w}$ s.t. $\begin{cases} \mathbf{1}_Q^\intercal\boldsymbol{b} \leq K_1 \\ -\boldsymbol{w} \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq \boldsymbol{w} \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$	MILP
$\mathcal{P}_{\infty/0}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q,z\in\mathbb{R}} z$ s.t. $\begin{cases} \mathbf{1}_Q^\intercal\boldsymbol{b} \leq K_\infty \\ -z\mathbf{1}_N \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq z\mathbf{1}_N \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$	MILP
Penalized formulations (parameters: μ_2, μ_1 and μ_∞)		
$\mathcal{P}_{0+2}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q} \mu_2\mathbf{1}_Q^\intercal\boldsymbol{b} + \boldsymbol{x}^\intercal\mathbf{H}^\intercal\mathbf{H}\boldsymbol{x} - 2\boldsymbol{y}^\intercal\mathbf{H}\boldsymbol{x}$ s.t. $\{ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b}$	MIQP
$\mathcal{P}_{0+1}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q,\boldsymbol{w}\in\mathbb{R}^N} \mu_1\mathbf{1}_Q^\intercal\boldsymbol{b} + \mathbf{1}_N{}^\intercal\boldsymbol{w}$ s.t. $\begin{cases} -\boldsymbol{w} \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq \boldsymbol{w} \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$	MILP
$\mathcal{P}_{0+\infty}$	$\min_{\boldsymbol{x}\in\mathbb{R}^Q,\boldsymbol{b}\in\{0,1\}^Q,z\in\mathbb{R}} \mu_\infty\mathbf{1}_Q^\intercal\boldsymbol{b} + z$ s.t. $\begin{cases} -z\mathbf{1}_N \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq z\mathbf{1}_N \\ -M\boldsymbol{b} \leq \boldsymbol{x} \leq M\boldsymbol{b} \end{cases}$	MILP

3) Reformulation of the ℓ_∞ Data Misfit Measure: $\mathcal{P}_{0/\infty}$ naturally brings linear inequality constraints as

$$\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_\infty \leq \alpha_\infty \Leftrightarrow -\alpha_\infty\mathbf{1}_N \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq \alpha_\infty\mathbf{1}_N.$$

For both $\mathcal{P}_{\infty/0}$ and $\mathcal{P}_{0+\infty}$, involving the ℓ_∞ norm in the objective function, one can simply introduce an additional scalar variable z such that minimizing $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_\infty$ amounts to minimizing z under the constraints

$$-z\mathbf{1}_N \leq \boldsymbol{y} - \mathbf{H}\boldsymbol{x} \leq z\mathbf{1}_N.$$

Here again, one can easily show that an optimal solution of the original problem will necessarily satisfy $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|_\infty = z$.

C. Mixed-Integer Programming Reformulations

Given the reformulation techniques of Section II.B, the nine problem reformulations proposed in Table I are straightforward.

IV. MIP Resolution: Basic Elements

Mixed-integer programming problems are not easy problems. MILP problems are already NP-hard [36]. As a consequence, reducing sparse approximation problems to MILP, MIQP or MIQCP problems does not *per se* reduce the complexity. Nevertheless, such MIP reformulations not only open up possibilities to prove the optimality (or quantify the sub-optimality) of solutions, but also allows one to benefit from decades of exponential progress in terms of required computing time to solve a given MIP problem. This progress does not simply reflect the doubling of computing power every 18 months, it is also a consequence of fast progress in both the theory and practice of linear programming and discrete optimization (duality, numerical linear algebra, interior-point methods, semi-definite positive relaxations, branch and bound/cut/price methods, decomposition approaches, global optimization, *etc.*) Once the sparse approximation problem is recast as a MIP problem, then state-of-the-art off-the-shelf software can be used, such as BARON, COUENNE, CPLEX, GloMIQO,

GUROBI, MOSEK or Xpress-MP—see for example [35] and references therein. We chose to use CPLEX [23] because it is unanimously considered among the best MIP solvers. CPLEX has been developed over the last thirty years and includes the best strategies developed by the MIP community. Moreover, it is freely available for research and teaching purposes.

The main method behind the CPLEX MIP solver is a *branch-and-cut* algorithm. Globally, it implements a *branch-and-bound* strategy (*i.e.*, a tree-structured implicit enumeration algorithm) based on successive continuous relaxations of the integer variables [37]. Each *branch* generates a subproblem by fixing some integer (in our case, binary) variables, and a set of branches is constructed corresponding to the different possible configurations of such variables. Then, the aim is to discard (huge) parts of the remaining combinatorial tree by *lower bounding* the objective function on the corresponding subproblems. To obtain such lower bounds, a continuous relaxation of each subproblem is formed by relaxing the integer variables. Linear constraints, such as Gomory *cutting planes* [38], are added to each generated (continuous relaxation) subproblem, so that the continuous solution converges to an integer solution. Such cutting planes remove parts of the feasible domain of the subproblem that does not contain any integer solution. This approach amounts to attempting to construct the convex hull of the set of integer feasible solutions of each subproblem. CPLEX incorporates several techniques in order to improve performance, such as constraint propagation techniques [39], linear algebra techniques [23] and heuristic techniques to find rapidly a good integer solution. Doing so, parts of the research space are eliminated only if it is proved that they do not contain the global minimum.

The best current integer solution provides an upper bound of the global minimum of the entire problem. The solution of the current relaxed (continuous) subproblem gives a lower bound of the global minimum of the current subproblem with integer constraints under consideration. The worst solution of all relaxed subproblems—the one that achieves the lowest lower bound—gives a certified lower bound of the global minimum of the entire problem. If such a lower bound is attained by the best current integer solution, then a global minimizer is found and optimality is proved. Otherwise, the entire process is iterated by creating new *branches*. The algorithm converges towards such a certified optimum in a finite number of steps. If the solver reaches the time limit, the *duality gap* (the difference between the best solution found and the certified lower bound of the global minimum) provides a measure of sub-optimality of the solution found. Note that tree-search based greedy algorithms such as in [8] also rely on tree-based (local) exploration and on lower-bounding the objective function on sets of solutions, which is used inside a greedy procedure. Therefore, they do not come with any optimality guarantee.

Remark that in most sparsity-inspired signal processing problems, the matrix $\mathbf{H}$ satisfies specific properties that are exploited for efficient computations. In particular, if $\mathbf{H}$ represents a redundant set of transforms based on multi-scale representations such as wavelets [40], matrix-vector products $\mathbf{H}\boldsymbol{v}$ and $\mathbf{H}^\intercal\boldsymbol{v}$ (where $\boldsymbol{v}$ is some given vector of an appropriate dimension) can be computed by fast algorithms using Fast Fourier Transforms (FFT). In 1D (respectively, 2D) deconvolution problems, $\mathbf{H}$ is a Toeplitz (respectively, Toeplitz-block-Toeplitz) matrix, so that matrix-vector products can also be computed by FFT. No such matrix structure is exploited within the standard mixed-integer optimization algorithms, as they are implemented through general-purpose MIP solvers. The only matrix property that is exploited here is the possible sparsity of the matrix $\mathbf{H}$ to compute fast vector products. This is particularly the case for the finite impulse response (FIR) deconvolution problems that are considered hereafter.

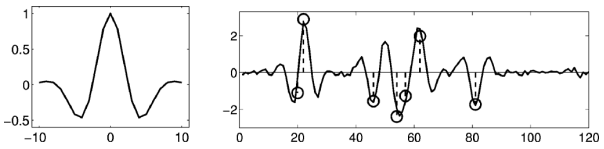


Fig. 1. Example of sparse deconvolution data. Left: impulse response. Right: 7-sparse sequence $\boldsymbol{x}$ (circles) and noisy convolved data $\boldsymbol{y}$, with $\mathrm{SNR} = 20$ dB.

V. Experimental Results: Optimization Performance

This section presents the test problems and the computational framework. Then, the computational efficiency of the nine MIP reformulations is studied.

A. Definition of Test Problems

The different MIP reformulations are evaluated on one-dimensional sparse deconvolution problems. Such problems are typically encountered for example in ultrasonic NDT [16] and in seismic reflection in Geophysics [17]. In the following, $\boldsymbol{x}$ is a K-sparse sequence in $\mathbb{R}^{100}$ (*i.e.*, $Q = 100$), with uniformly distributed spike locations, where the sparsity level, K, is varying. In order to avoid arbitrary small spike values, each non-zero amplitude is drawn as $\mathrm{sign}(u) + u$, where u is a centered Gaussian sample with unit variance. The matrix $\mathbf{H}$ is the discrete convolution matrix corresponding to the 21-sample impulse response shown in Fig. 1 (left). With the boundary assumption that $\boldsymbol{x}$ is zero outside its domain, $\boldsymbol{y}$ is a 120-sample signal (*i.e.*, $N = 120$). White noise is added with variable signal-to-noise ratio (SNR). In this section, noise samples are generated according to a centered normal distribution $\mathcal{N}(0, \sigma^2)$, with σ such that $\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}(\|\mathbf{H}\boldsymbol{x}\|_2^2/(N\sigma^2))$. We name such problems $\mathrm{SA}_K^{\mathrm{SNR}}$. Note that they are slightly overdetermined ($Q < N$), whereas typical sparse approximation problems deal with largely under-determined systems. However, trivial inversion is not satisfactory here, because of the presence of noise and of the ill-conditioned nature of $\mathbf{H}$ ($\mathrm{cond}(\mathbf{H}) \simeq 10^3$).

We consider problems $\mathrm{SA}_K^{\mathrm{SNR}}$ with K varying between 5 and 11, and SNR varying from $+\infty$ (noise-free data) downto 10 dB. One example of data is given in Fig. 1 (right) for $K = 7$ and $\mathrm{SNR} = 20$ dB. It illustrates the difficulty of sparse deconvolution problems arising for example in ultrasonic NDT [16]. The oscillating impulse response and the proximity of the spikes produce overlapping echoes in the available data. As the echoes cannot be distinguished visually, numerical techniques are required. All data and optimization results of this section are available online as supplementary multimedia material.

TABLE II
CPU TIMES (IN SECONDS) OBTAINED BY THE NINE MIP REFORMULATIONS, AS A FUNCTION OF THE SIGNAL-TO-NOISE RATIO (SNR) AND OF THE SPARSITY LEVEL (K), AVERAGED OVER 50 INSTANCES OF EACH PROBLEM. THE FIRST NUMBER INDICATES THE CPU TIME FOR ESTABLISHING OPTIMALITY OF THE SOLUTION. THE SECOND NUMBER GIVES THE CPU TIME AT WHICH THE CORRESPONDING OPTIMUM WAS FOUND. NUMBERS IN PARENTHESES COUNT THE INSTANCES FOR WHICH OPTIMALITY WAS NOT PROVED IN LESS THAN 1000 SECONDS. FOR REFERENCE, THE COMPUTATION TIME OF CLASSICAL SPARSE APPROXIMATION ALGORITHMS ON OUR HARDWARE CAN BE FOUND IN FIG. 2

	Error-constrained problems			Sparsity-constrained problems			Penalized problems		
SNR = ∞	$\mathcal{P}_{0/2}$	$\mathcal{P}_{0/1}$	$\mathcal{P}_{0/\infty}$	$\mathcal{P}_{2/0}$	$\mathcal{P}_{1/0}$	$\mathcal{P}_{\infty/0}$	$\mathcal{P}_{0+2}$	$\mathcal{P}_{0+1}$	$\mathcal{P}_{0+\infty}$
K = 5	53 | 50 (2)	0.07 | 0.03	0.06 | 0.03						
K = 7	3.8 | 3.6	0.09 | 0.03	0.08 | 0.03						
K = 9	4.4 | 4.2	0.12 | 0.03	0.10 | 0.03						
K = 11	5.2 | 5.0	0.14 | 0.03	0.13 | 0.03						
SNR = 30 dB									
K = 5	13 | 7.2	0.3 | 0.2	0.1 | 0.1	0.3 | 0.2	0.5 | 0.4	0.5 | 0.4	0.2 | 0.1	0.4 | 0.2	0.4 | 0.2
K = 7	17 | 6.7	1.4 | 0.4	0.3 | 0.1	0.5 | 0.3	1.4 | 0.8	1.4 | 1.0	0.4 | 0.2	1.0 | 0.6	0.8 | 0.4
K = 9	132 | 37	3.1 | 1.0	1.7 | 0.5	3.0 | 1.0	9.3 | 3.0	25 | 7.5	1.9 | 0.8 (1)	19 | 2.7	17 | 2.7
K = 11	399 | 96 (18)	18 | 4.1	19 | 2.1	29 | 4.0 (1)	72 | 11 (2)	78 | 17 (9)	11 | 2.7 (1)	60 | 13 (1)	48 | 14 (4)
SNR = 20 dB									
K = 5	13 | 8.6	0.3 | 0.2	0.1 | 0.1	0.3 | 0.2	0.6 | 0.3	0.5 | 0.4	0.3 | 0.2	0.6 | 0.3	0.4 | 0.2
K = 7	48 | 11 (1)	1.0 | 0.4	0.5 | 0.1	1.3 | 0.6	4.0 | 1.8	4.2 | 3.5	6.6 | 0.9	14 | 1.5 (1)	29 | 14
K = 9	285 | 82 (9)	6.6 | 2.7	3.7 | 0.4	18 | 3.4 (1)	30 | 6.4 (1)	72 | 64 (2)	28 | 13 (2)	11 | 2.8 (2)	87 | 18 (2)
K = 11	- | - (50)	67 | 16	65 | 11 (3)	166 | 21 (6)	244 | 22 (11)	355 | 260 (20)	104 | 16 (5)	179 | 58 (13)	299 | 38 (17)
SNR = 10 dB									
K = 5	20 | 4.9	0.5 | 0.2	0.2 | 0.1	0.4 | 0.2	1.3 | 0.6	1.7 | 1.4	1.9 | 1.7	2.0 | 0.5 (2)	15 | 0.8
K = 7	195 | 27 (2)	2.3 | 0.6	1.4 | 0.2	3.5 | 0.7	11 | 3.0	40 | 39	42 | 2.4	73 | 4.0 (3)	66 | 15 (6)
K = 9	485 | 63 (34)	24 | 3.9	26 | 1.2	54 | 12 (1)	189 | 28 (4)	220 | 218 (11)	96 | 25 (11)	230 | 11 (18)	215 | 32 (16)
K = 11	- | - (50)	287 | 44 (12)	211 | 9.0 (9)	358 | 26 (24)	670 | 75 (42)	600 | 598 (39)	340 | 34 (32)	300 | 13 (43)	508 | 101 (44)

B. Machine Configuration and Implementation Details

Optimization is run with IBM ILOG CPLEX V12.6.0 from a Matlab interface on a computer with eight Intel Xeon X5472 processors with Central Processing Units (CPU) clocked at 3 GHz. The maximum time allowed for each resolution is set to $T_{\max} = 1000$s. The other CPLEX parameters are set to their default value. For each problem, the "big-M" constant is set to $M = 1.1x^1_{\max}$, where $x^1_{\max} := \|\mathbf{H}^\intercal \boldsymbol{y}\|_\infty / \|\boldsymbol{h}\|_2^2$ corresponds to the maximum amplitude of 1-sparse solutions estimated by least-squares. If the boundedness assumption (3) is saturated—*i.e.*, one component in the solution $\hat{\boldsymbol{x}}$ reaches the value $-M$ or M—then the optimization is successively run again with M replaced with $1.1M$, until the obtained solution satisfies $\|\hat{\boldsymbol{x}}\|_\infty < M$. The CPU times given below include such restarts. With this heuristic strategy, in our simulations, only few cases led to such saturation: no restart was necessary in 90% of the cases, and the average number of restarts in the other cases was approximately 1.6.

C. Evaluation of Computational Costs

Each of the nine MIP reformulations of Table I is run for fifty random instances (for both spike distributions and noise realizations) of each $\mathrm{SA}_K^{\mathrm{SNR}}$ problem. In this section, in order to ensure a fair comparison (in terms of computational efficiency), the parameters K_p are set to the true sparsity level K, and parameters α_p and μ_p are set by trial and error, until the sparsity level of each solution equals K. Note that in our case, the matrix $\mathbf{H}$ has full column rank, hence ℓ_0 inequality constraints in $\mathcal{P}_{p/0}$ will yield the same results as if they were equality constraints. That is, when imposing $\|\boldsymbol{x}\|_0 \leq K$, all the K columns of matrix $\mathbf{H}$ contribute to the reduction of the data misfit. However, formulations with inequality constraints yielded lower computing times. Sparse representation problems (noise-free data, SNR $= \infty$) are addressed through $\mathcal{P}_{0/p}$ for $p \in \{1, 2, \infty\}$, with threshold $\alpha_p = 10^{-8}$. Remark that in the noise-free case, no sparsity-enhancing algorithm is indeed necessary, since the solution can simply be computed by least squares.

Average CPU times obtained for MIP reformulations are given in Table II. The figures on the left-hand side of each column is the time required to *prove* the global optimality of the solution found. The figures on the right-hand side indicate the time after which the support of the solution was *found*, which is generally much lower. The figures in parentheses indicate the number of instances for which optimality was not proved within 1000s.

All CPU times increase with the sparsity level K, but also with the noise level. In particular, for SNR $= 10$ dB and $K = 11$, for each formulation, optimality of the solutions was obtained in less than 1000s only on a fraction of the fifty instances. In order to explain such behavior, let us remark that the sparsity level (respectively, the noise level) increases the size of the feasible domain of $\mathcal{P}_{p/0}$ (respectively, of $\mathcal{P}_{0/p}$). More generally, for all problems, if either the sparsity level or the noise level increases, then the branch-and-bound strategy becomes less efficient in discriminating between concurrent candidates and in eliminating some of them.

With SNR $= 30$ dB, the lowest CPU times are achieved by solving $\mathcal{P}_{0+2}$. When the noise level increases, solving $\mathcal{P}_{0/1}$ and $\mathcal{P}_{0/\infty}$ problems becomes more efficient computationally, and their superiority over other problems increases with both the sparsity level and the noise level. In particular, $\mathcal{P}_{0/1}$ problems were always solved exactly in less than 1000s, except

for (SNR $=$ 10 dB, $K = 11$). Results are slightly not as good for $\mathcal{P}_{0/\infty}$, where optimization did not terminate within 1000s for three instances with SNR $=$ 20 dB and $K = 11$. Note that even a small amount of noise severely degrades the resolution performance of problems $\mathcal{P}_{0/p}$. Optimization of $\mathcal{P}_{0/2}$, the only problem with a quadratic constraint, is the most time-consuming among all proposed formulations. Non-linear constraints are known to make the MIP really difficult to solve [35]. An element of explanation can be found by comparing the Lagrangians of formulations involving ℓ_2 misfits. Indeed, the Lagrangian of $\mathcal{P}_{0/2}$ contains trilinear terms. On the contrary, the Lagrangians of $\mathcal{P}_{2/0}$ and of $\mathcal{P}_{0+2}$ are quadratic functions. Therefore, optimizing a linear function under quadratic constraints is more complex than optimizing a quadratic function under linear constraints.

For any $p \in \{1, 2, \infty\}$, solving problems $\mathcal{P}_{0+p}$ generally performs better than solving problems $\mathcal{P}_{p/0}$ at high SNR. On the contrary, as the noise level increases, sparsity-constrained formulations outperform penalized versions. For both formulations, which involve the data misfit in the objective function, using an ℓ_2 misfit measure is the most efficient choice, and both ℓ_1- and ℓ_∞-misfit optimizations behave similarly.

We also note a high dispersion of the required CPU times among the fifty realizations of each problem. For example, the average time for the resolution of $\mathcal{P}_{0+1}$ on SA_9^{20} problems was approximately 11s on forty-eight instances, whereas optimization did not terminate after 1000s on the two other instances. We also remark that, for $\mathcal{P}_{0/2}$, optimality was not proved within 1000s for two instances of the simplest test problem $\mathrm{SA}_5^{+\infty}$. Two other instances of $\mathrm{SA}_5^{+\infty}$ also required a much higher CPU time than the others, which leads to an atypically high average time of 53s, reflecting once again the difficulty of the quadratically-constrained problem $\mathcal{P}_{0/2}$.

Finally, let us evaluate the CPU time of exhaustive combinatorial exploration for $\mathcal{P}_{2/0}$. Using notations introduced in Section II.C, for a given support s with K components, the minimizer of $\mathcal{P}_{2/0}^s$ has the closed-form expression $\hat{\boldsymbol{x}}_s = (\mathbf{H}_s{}^\intercal \mathbf{H}_s)^{-1} \mathbf{H}_s{}^\intercal \boldsymbol{y}$. Then, the least-squares misfit value is computed by: $\|\boldsymbol{y} - \mathbf{H}_s \hat{\boldsymbol{x}}_s\|^2 = \|\boldsymbol{y}\|^2 - \|\mathbf{H}_s \hat{\boldsymbol{x}}_s\|^2$. In practice, $\hat{\boldsymbol{x}}_s$ can be computed by performing Cholesky decomposition of $\mathbf{H}_s{}^\intercal \mathbf{H}_s$, so that one computation of the objective function mainly amounts to two $K \times K$ triangular inversions. The CPU time, denoted c_K, of one such computation is estimated by averaging over 10^5 inversions. Then, neglecting the cost of Cholesky factorizations, the cost for testing all K-sparse solutions is extrapolated as $\binom{Q}{K} c_K$. It yields approximately 1500s for $K = 5$, 4 days for $K = 7$, and more than one year for $K = 9$. Problems based on both ℓ_1 and ℓ_∞ misfits require the resolution of a linear program for each support, therefore the corresponding exhaustive search yields still higher computational costs than in the ℓ_2 case. Consequently, exhaustive search cannot be considered as a practical solution, even for such moderate-size problems. In order to emphasize the ability of the MIP solver to remove important parts of the full combinatorial tree search, we give a last indicator. For $\mathcal{P}_{2/0}$ with SNR $=$ 30 dB and $K = 9$, for which all instances were successfully solved, the average number of combinations that were explicitly considered in the branch-and-bound procedure is about $4.5\,10^5$, to be compared with the total number of $\binom{100}{9} \simeq 1.9\,10^{12}$ combinations.

VI. Experimental Results: Evaluation of Solutions

We now compare the solutions obtained *via* MIP optimization with those of classical sparse approximation methods: Orthogonal Matching Pursuit (OMP) [5], Single Best Replacement (SBR) [7], Iterative Hard Thresholding (IHT)[1] [13], [14] and the minimization of the ℓ_1-norm-penalized least-squares criterion (ℓ_1-relax) using homotopy continuation [41]. All algorithms are tuned so that all solutions have the correct number of spikes. Therefore, all methods are put in a favorable setting for evaluating their ability to retrieve the correct support of the solution.

Recall that for sparse deconvolution, none of the classical methods are theoretically guaranteed to solve the ℓ_0-norm problem. As seen in the previous section, the MIP approach can compute an exact solution, but requires a larger computation time. In order to evaluate intermediate solutions found by the MIP solver, we consider several values of the maximum time allowed for each MIP resolution: $T_{\max} = 1\text{s}, 10\text{s}, 100\text{s}$ and 1000s. If the maximum time is reached, then the current solution is considered—which is the best solution found, without any optimality guarantee. The parameter M is tuned as explained in Section V.B when running optimization with $T_{\max} = 1000\text{s}$. Then, this value is also used for lower values of $T_{\max}$.

In this section, we consider two types of sparse deconvolution problems. The first problems are similar to those of Section V.A, with $N = 120$ and $Q = 100$, and are therefore slightly overdetermined. The second ones are underdetermined problems, the most frequent case in sparse approximation. Such problems may arise in *high resolution* sparse deconvolution. Indeed, the true spike locations are generally continuous-valued (*e.g.*, representing times of flight of reflected waves), and the model $\mathbf{H}\boldsymbol{x}$ is a discrete approximation of a continuous integral equation. The discretization step is usually chosen equal to the sampling period of the data, thus $\mathbf{H}$ is a discrete convolution matrix. Such a choice may be too rough and can lead to erroneous spike locations [42]. In order to improve the model, one may consider an upsampled convolution model, where both the impulse response and the sparse sequence are upsampled by an integer factor, UF. Then, $\mathbf{H}$ can be viewed as the concatenation of UF discrete convolution matrices. Detail about the corresponding matrix structure can be found in [42]. In the following, we consider UF $= 2$, so that $N = 120$ and $Q = 200$. Recall however that the intrinsic difficulty of sparse deconvolution is mostly due to the ill-conditioned nature of matrix $\mathbf{H}$ and to the presence of noise, even in the overdetermined case.

Fifty random instances are run with UF $= 1$ and UF $=$ 2, SNR $= \infty$ and SNR $=$ 20 dB, and sparsity levels K varying between 5 and 11. In the noise-free case, for the MIP

[1]We used T. Blumensath's implementation of IHT (program AIHT.m) available at: http://www.personal.soton.ac.uk/tb1m08/sparsify/sparsify.html

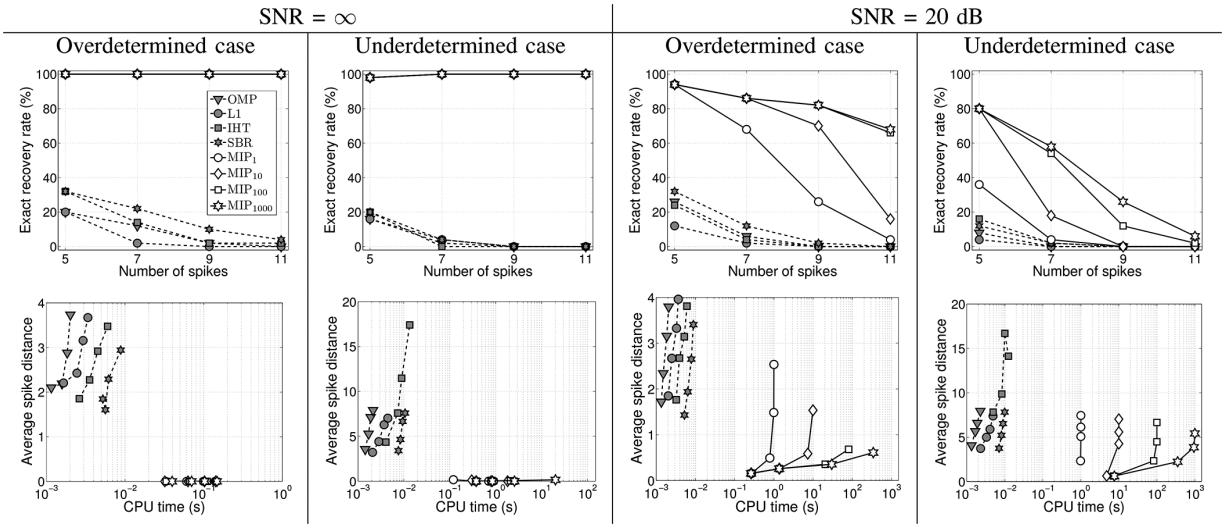


Fig. 2. Performance of classical sparse approximation algorithms and MIP optimization on sparse deconvolution problems, as a function of the signal-to-noise ratio and of the size of the unknown vector. Overdetermined cases correspond to the standard deconvolution problem ($N = 120, Q = 100$). Underdetermined cases correspond to an upsampling factor $\mathrm{UF} = 2$ such that $N = 120$ and $Q = 200$. Top row: *exact recovery rate* as a function of the sparsity level K. Bottom row: *average spike distance* to the true sequence as a function of the computation time. For each algorithm, the four marks correspond to $K = 5, 7, 9$ and 11. Note the log-scale and the different scalings of the time axes. The notation $\mathrm{MIP}_{T_{\max}}$ corresponds to the MIP optimization strategy running in a maximum of $T_{\max}$ seconds. All results are averaged over 50 random realizations. Note that the four marks of MIP algorithms superimpose on the two first panels of the top row.

approach, only $\mathcal{P}_{0/p}$ problems are relevant. Therefore, solutions are computed *via* the optimization of $\mathcal{P}_{0/\infty}$ (with threshold $\alpha_\infty = 10^{-8}$), which was the most efficient computationally (see Section V.C). All data and MIP optimization results of this section are available online as supplementary multimedia material.

In the noisy case, we consider $\mathcal{P}_{2/0}$, since the classical methods rely on ℓ_2 data misfit measures. Two quality indices are computed for each simulated data set:

- the *exact recovery rate*, which is the average number of simulations for which a given algorithm correctly locates the support of the true sequence;
- the *average spike distance* between the estimated and the true sequences. To this aim, a distance similar to that used in neuroscience [43] is defined: both estimated and true spike trains are convolved with a continuous-domain Gaussian kernel (with standard deviation chosen equal to the discretization step of the spikes). Then, the ℓ_2-norm error between both signals is computed. Such a criterion is less severe than the exact recovery rate if the estimated spikes are slightly shifted with respect to the true ones, and also gives more importance to spikes with high amplitudes.

Note that with $\mathrm{SNR} = \infty$, the true sequence is a minimizer of $\mathcal{P}_{0/p}$, with $p = 1, 2, \infty$. Hence the expected exact recovery rate for the MIP approach is 100%. On the other hand, in the noisy case, the minimizer of $\mathcal{P}_{2/0}$ may not be obtained on the true support. Therefore, successful global optimization does not always produce exact recovery.

Results are summarized in Fig. 2. The top row shows the average exact recovery rate as a function of the number of spikes, and the bottom row plots the average spike distance to the true sequence as a function of the CPU time.

Let us first focus on the noise-free case (left columns). Recall that, in the overdetermined noise-free case, the solution can simply be computed by least squares. Simulations are still of interest, however, in order to compare the algorithms in an ideal and simple context. For all classical algorithms, the exact recovery rate is lower than 40% in the overdetermined case, and decreases as the sparsity level increases (top row). Their performance is still worse in the underdetermined case, where the exact recovery rate is close to zero, except for the simplest problems ($\mathrm{K} = 5$). Their average spike distance to the true sequence (bottom row) is logically lower for algorithms requiring more computation time. We note in particular the bad results obtained by IHT in the underdetermined case. Such bad performance of IHT was already attested when theoretical optimality conditions do not hold [14]. This is particularly true in the underdetermined case, where the columns of $\mathbf{H}$ are strongly correlated. On the contrary, the MIP strategy correctly retrieves the support in nearly 100% of the noise-free instances, even with the computation time limited to 1s. Actually, only one instance led to erroneous support identification (for $\mathrm{UF} = 2$ and $K = 5$), meaning that the solution was not found, even within 1000s. The MIP approach also gives an average spike distance close to zero, which means that both the supports *and* the amplitudes of the solutions have been correctly recovered, even in the underdetermined case, but with a larger computation time (from 0.03s to 0.2s for $\mathrm{UF} = 1$, and from 0.1s to 20s for $\mathrm{UF} = 2$). Note however that all classical algorithms are still much faster on such relatively small problems (between 10^{-3}s and 10^{-2}s).

In the more realistic noisy case ($\mathrm{SNR} = 20$ dB), the results of the classical algorithms are very similar to those obtained in the noise-free case, both in terms of exact recovery rate, average spike distance and CPU time. In contrast, the MIP performance deteriorates, and the exact recovery rate quickly de-

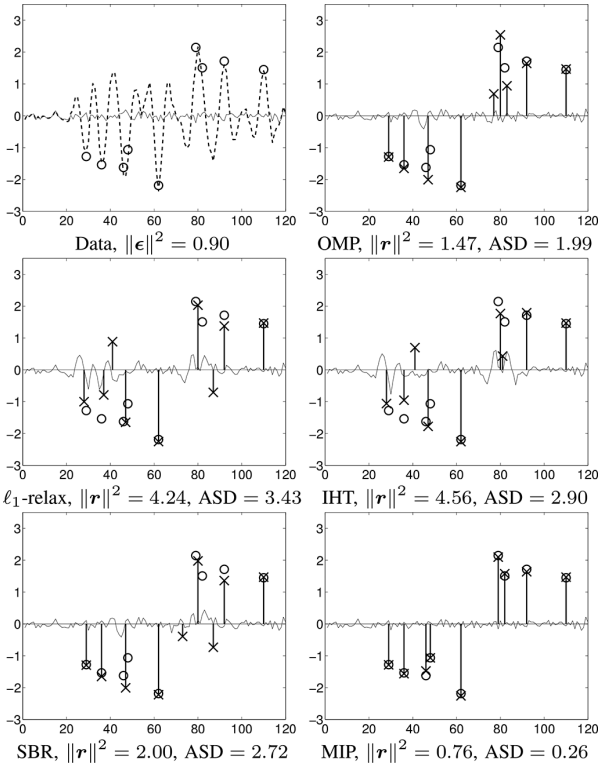


Fig. 3. Data (top-left) and deconvolution results obtained by different sparse approximation algorithms in the overdetermined case ($N = 120, Q = 100$), with SNR $= 20$ dB and $K = 9$ spikes. Circles locate the true spikes. On the top-left panel, the dashed (respectively, solid) line represents the data $\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{\epsilon}$ (respectively, the noise $\boldsymbol{\epsilon}$). On the other panels, crosses show the obtained minimizer $\hat{\boldsymbol{x}}$ and the solid line represents the residual $\boldsymbol{r} = \boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{x}}$. For each algorithm, the residual norm and the average spike distance to the true sequence (ASD) are given.

creases as the number of spikes increases. Recall however that, in the presence of noise, the minimizer of $\mathcal{P}_{2/0}$ may *not* retrieve the correct support: in the overdetermined case, for example, the MIP solver returns the optimal solution in less than 1000s in 94% of the instances, whereas the average exact recovery rate is much lower. However, it is still better than that of the classical methods with $T_{\max} = 1$s, and even much better if $T_{\max}$ is increased to 100s or 1000s. The MIP approach also outperforms classical methods in terms of average spike distance, in particular if $T_{\max}$ is high enough. In the overdetermined case, the average computing time ranges from 0.25s (for $K = 5$) to 350s (for $K = 11$) and, as mentioned earlier, global optimality was obtained in less than 1000s for most simulations. In the underdetermined case, however, an optimum was not proved to be found within 1000s in 51 % of the instances, that mostly correspond to the cases where $K = 9$ and $K = 11$. This analysis corroborates the results in Section V.C: the presence of noise strongly impacts the computing time of the MIP solver, and therefore it impacts the quality of the solutions obtained by early stopping.

A typical data set and estimated sparse sequences are shown in Fig. 3. It corresponds to the overdetermined case, the true sequence is 9-sparse and SNR $= 20$ dB. In this example, the MIP approach is the only algorithm that correctly identifies the support. Note that the resulting ℓ_2 misfit at the MIP solution is lower than the ℓ_2 norm of the noise.

VII. Experimental Results: Relevance of ℓ_1-, ℓ_2- and ℓ_∞-Norm Data Misfits

In this section, the impact of the data misfit measure (through ℓ_1, ℓ_2 and ℓ_∞ norms) on the quality of the solution is studied, as a function of the noise distribution—as motivated by the discussion in Section II.B. To this aim, data are simulated in a manner similar to Section V.A and Fig. 1. The 7-sparse spike sequence $\boldsymbol{x}^{\text{true}}$ is fixed, and 200 noise realizations are drawn, where noise samples are i.i.d. according to Gaussian, Laplacian and uniform distributions, successively. The SNR here is set to 15 dB. The three error-constrained problems $\mathcal{P}_{0/p}$ are considered here. We focus on these formulations because, in practical cases, tuning the parameter α_p requires less prior information than tuning the parameter K_p for $\mathcal{P}_{p/0}$ or the parameter μ_p for $\mathcal{P}_{0+p}$. Indeed, for any given noise distribution $p_{\boldsymbol{\epsilon}}$, the parameters α_p, $p \in \{1, 2, \infty\}$, can be naturally estimated from a common statistical rule. More precisely, setting α_p to a value satisfying $\Pr(\|\boldsymbol{\epsilon}\|_p \leq \alpha_p | \boldsymbol{\epsilon} \sim p_{\boldsymbol{\epsilon}}) = 95\%$ amounts to considering that the approximation error cannot be distinguished from noise with probability 95%. Doing so, fair comparisons can be performed between the solutions of the three problems $\mathcal{P}_{0/p}$.

These three problems share the same objective function, hence they can be compared through their minimum value, that is, the estimated number of spikes. In order to evaluate the quality of the solutions, we also consider the *support error* $e^{\text{supp}}(\hat{\boldsymbol{x}}, \boldsymbol{x}^{\text{true}}) := ||\hat{\boldsymbol{b}} - \boldsymbol{b}^{\text{true}}||_0$, where $\hat{b}_q$ (respectively, b_q^{true}) equals 1 if $\hat{x}_q \neq 0$ (respectively, $x_q^{\text{true}} \neq 0$), and 0 otherwise. Optimization terminated in less than 1000s for all simulations. Fig. 4 shows the distribution of the ℓ_0 norms (left column) and of the support errors (right column) for the three noise distributions, and for the three ℓ_p data misfits.

We first note that the correct value of the ℓ_0 norm is the most frequently reached in all cases, except if an ℓ_∞ misfit is used with Laplacian noise. When the estimated ℓ_0 norm is wrong, it is generally lower than the true value. Indeed, in most cases (statistically, in 95% of the cases from the definition of α_p), the noise realization satisfies $\|\boldsymbol{\epsilon}\|_p < \alpha_p$. That is, the true sequence $\boldsymbol{x}^\star$ satisfies the misfit-bound constraint. Thus, the threshold α_p allows a higher approximation error than the noise in the data, which enables the possibility of solutions with fewer spikes. Consequently, in such cases, $\|\hat{\boldsymbol{x}}\|_0 \leq \|\boldsymbol{x}^\star\|_0$. This is particularly true if the noise distribution is heavy-tailed, where the ℓ_p-norm of the noise samples may be much below the threshold α_p, as show the ℓ_0 norm estimation statistics in the case of Laplacian noise. On the other hand, if $\|\boldsymbol{\epsilon}\|_p \geq \alpha_p$, then the true sequence does not correspond to a solution of the optimization problem, and $\|\boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{x}}\|_p < \|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}^\star\|_p$. In such cases, one may have $\|\hat{\boldsymbol{x}}\|_0 > \|\boldsymbol{x}^\star\|_0$. In our simulations, only very few instances led to such an overestimation of the ℓ_0 norm.

As it could be expected, the lowest support errors are achieved by using the ℓ_2 (respectively, ℓ_1 and ℓ_∞) misfit in the case of Gaussian (respectively, Laplacian and uniform) noise. For each noise distribution, the corresponding misfit yields the smallest average support error, and more frequently achieves

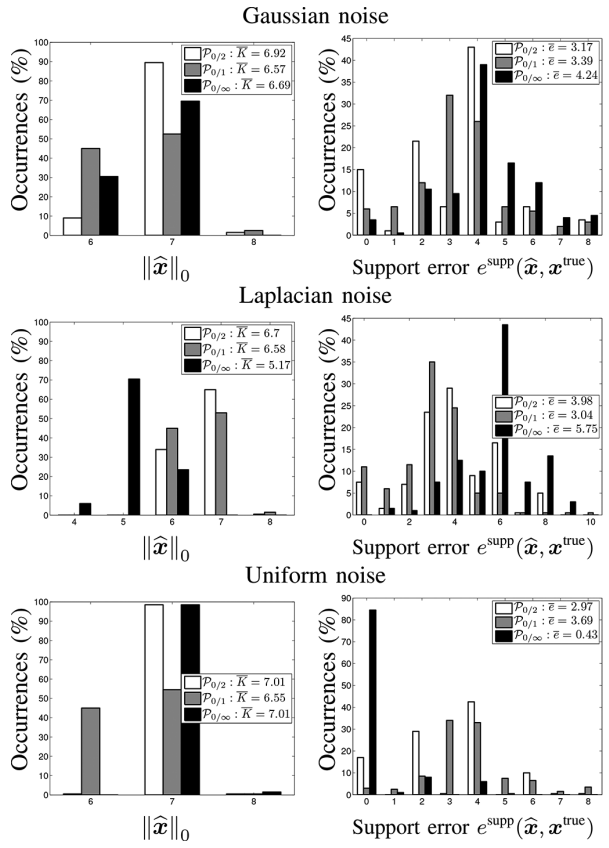


Fig. 4. Estimation results obtained for the deconvolution of a 7-sparse sequence with SNR $= 15$ dB, averaged over 200 noise realizations, in the case of Gaussian (top), Laplacian (center), and uniform (bottom) noise distributions. Error-constrained problems $\mathcal{P}_{0/p}$ are considered, with $p = 2$ (white), $p = 1$ (gray) and $p = \infty$ (black). Left: distributions of the ℓ_0 norm of the solutions for the three ℓ_p misfits. For each problem, $\overline{K}$ indicates the average estimated ℓ_0 norm value. Right: distributions of the support errors. For each problem, $\overline{e}$ indicates the average support error between the estimated and the true sequences.

correct support identification—even if, for Gaussian or Laplacian noise, it is only obtained in a few cases (respectively, 15% and 11%). We also remark that switching from ℓ_2 to ℓ_1 misfits with Gaussian noise only slightly degrades the support identification performance, whereas optimization is computationally more efficient in the ℓ_1 case—see Section V.C. Much better support identification is achieved with uniform noise combined with the ℓ_∞ misfit, which yields exact identification in 90 % of the cases, whereas ℓ_1 and ℓ_2 data fitting achieve much worse results. Note that with both ℓ_∞ and ℓ_2 misfits, the ℓ_0 norm is correctly estimated in 98% of the cases. However, support recovery performance is much worse in the ℓ_2 case, as some spikes are misplaced. Such superiority of ℓ_∞ data fitting for uniform noise was already attested in [31] in a non-sparse context.

Fig. 5 displays typical results obtained for one particular realization of the noise process. For Gaussian (respectively, Laplacian and uniform) noise distributions, one example is shown such that ℓ_2 (respectively, ℓ_1 and ℓ_∞) data fitting yields a solution with the most frequent support error obtained among the 200 realizations. Note that, in each case, the solution shown corresponds to *one* solution of the considered optimization problem, that is, with the lowest ℓ_0 norm that satisfies the bounded-ℓ_p-misfit constraint. Recall indeed that, for most values of threshold parameters α_p, problems $\mathcal{P}_{0/p}$ feature an infinite number of solutions—see Section II.C. Consequently, the presented solution is almost certainly not the solution with minimal ℓ_p misfit. With Gaussian noise, the minimizer of $\mathcal{P}_{0/2}$ has the correct ℓ_0 norm, but with two misplaced spikes, that leads to a support error equal to 4. With the minimizer of $\mathcal{P}_{0/1}$, two spikes are slightly misplaced, and a third one is not detected. The minimizer of $\mathcal{P}_{0/\infty}$ also has the correct ℓ_0 norm, but its spikes are very badly located. For Laplacian noise, the most frequent support error for the minimizer of $\mathcal{P}_{0/1}$ is 3, which corresponds to one misplaced spike and the non-detection of one spike. Note that on the presented example, both minimizers of $\mathcal{P}_{0/1}$ and of $\mathcal{P}_{0/2}$ identify the same support, whereas the solution of $\mathcal{P}_{0/\infty}$ features only four spikes (among which one is erroneous). In the case of uniform noise, the solution of $\mathcal{P}_{0/\infty}$ correctly locates all spikes. The solution of $\mathcal{P}_{0/1}$ misplaces one spike and misses another one, and the solution of $\mathcal{P}_{0/2}$ is still worse, with three misplaced spikes—although with the correct sparsity level.

VIII. Discussion

In this paper, nine sparse approximation problems involving the ℓ_0 norm were considered and reformulated as mixed-integer programs (MIP). Bounded-error, sparsity-constrained and penalized formulations were studied, involving ℓ_p-norm data misfit measures, for $p \in \{1, 2, \infty\}$. Thanks to efficient dedicated MIP solvers, we demonstrated that moderate-size sparse approximation problems can be solved exactly, whereas exhaustive search remains computationally prohibitive for such instances. In particular, the use of a branch-and-bound strategy, coupled with efficient cutting planes methods, allows most combinations to be discarded without being evaluated.

Computational costs were evaluated on simulated difficult sparse deconvolution problems. Simulated data and corresponding optimal solutions are made available as potential benchmarks for evaluating other (potentially suboptimal) sparse approximation algorithms[2]. Our experiments show that misfit-constrained minimization of the ℓ_0 norm is the most efficient optimization formulation when associated with ℓ_1 and ℓ_∞ misfit measure. Conversely, the ℓ_2 misfit measure is advantageously used as an objective function, not as a constraint. All CPU times increase with the number of non-zero components in the true solution, and also with the amount of noise in the data. Our encouraging numerical results tend to indicate however that such optimization formulations may be appropriate for tackling sparse approximation problems with several hundreds of unknowns, as long as the solution is highly sparse and/or the noise level is low. In particular, they do represent an alternative to ℓ_1-norm-based and greedy methods for difficult estimation problems with highly correlated dictionaries, both of which are likely to fail. Simulations revealed, in particular, that exact

[2]Matlab implementations of the nine formulations are available at http://www.irccyn.ec-nantes.fr/~bourguignon

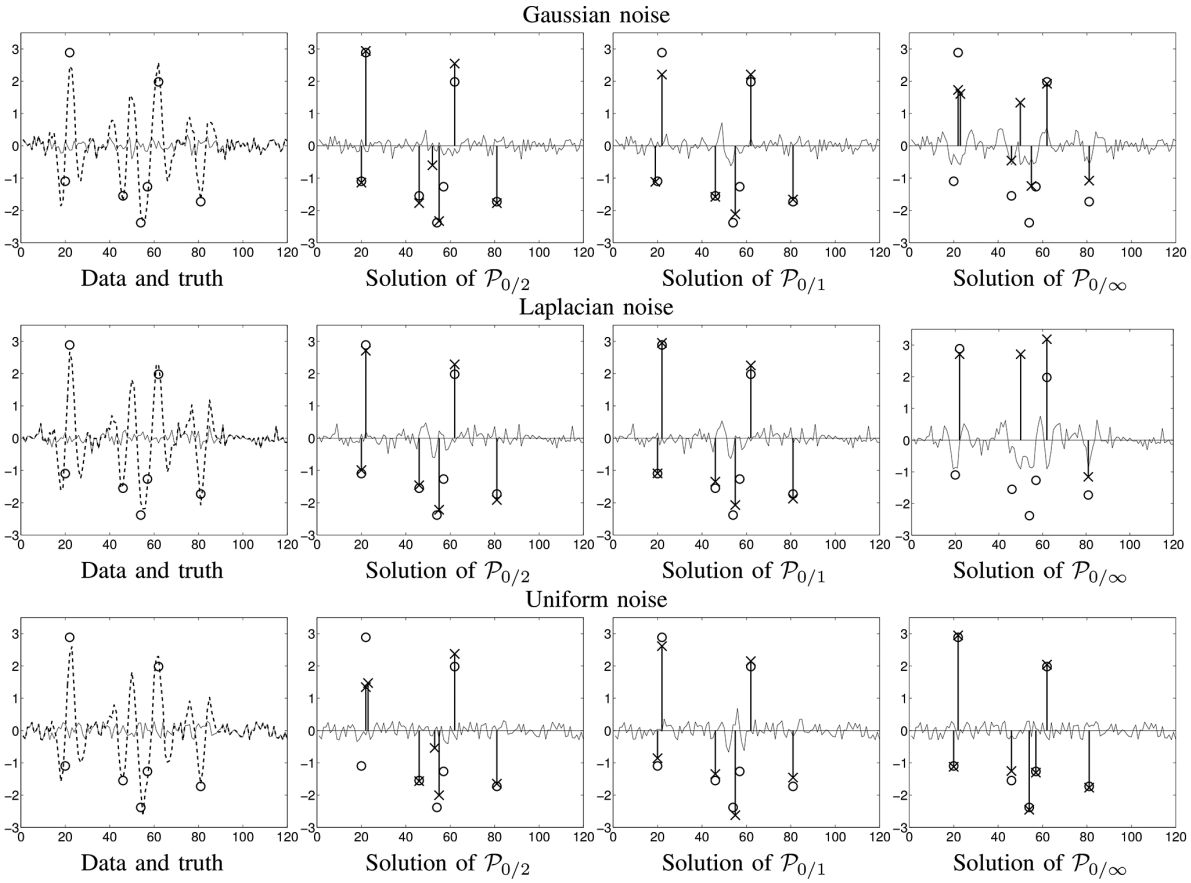


Fig. 5. Solutions of deconvolution problems $\mathcal{P}_{0/p}$ with Gaussian (top), Laplacian (center) and uniform (bottom) noises, for one particular noise realization, with SNR $= 15$ dB. Circles locate the true spikes. On the left column, the dashed (respectively, solid) line represents the data $\boldsymbol{y}$ (respectively, the noise $\boldsymbol{\epsilon}$). On the three other columns, crosses show the obtained minimizer $\hat{\boldsymbol{x}}$ and the solid line represents the residual $\boldsymbol{y} - \mathbf{H}\hat{\boldsymbol{x}}$.

solutions of ℓ_0 problems almost always achieve perfect support recovery for underdetermined noise-free problems, whereas classical methods perform relatively badly. In the presence of noise, the MIP solutions still outperform that of classical methods (in both over- and underdetermined cases), although the required computing time for obtaining exact solutions dramatically increases.

The ℓ_0 sparse approximation problem with ℓ_2 data misfit measure has been used in a huge quantity of works in signal processing, statistics, machine learning, *etc*. To the best of our knowledge, the methods presented in this paper are the only guaranteed-optimality alternatives to exhaustive search that do not rely on any strong assumption on the dictionary structure. With the introduced MIP reformulations, we also proposed to solve exactly less common sparse optimization problems based on ℓ_1 and ℓ_∞ misfits. Such problems may be of interest from an informational point of view. Simulations illustrated this point: choosing an ℓ_p misfit with $p = 2$ (respectively, $p = 1$ and $p = \infty$) is relevant if the noise distribution is Gaussian (respectively, Laplacian and uniform) *as far as support identification is concerned*. In particular, with uniformly distributed noise, introducing an ℓ_∞ misfit constraint frequently achieves correct support identification, which is not the case for any other combination of data misfit and noise distribution.

Several points in the MIP reformulations could be considered in order to improve computational efficiency. First, as acknowledged in previous works on MIP reformulations of sparsity-based problems [24], [27], tuning the value, M, in the "big-M" reformulation impacts algorithmic performance. For a given problem, statistical rules may be used in order to infer reasonable M values. Then, new constraints in the optimization formulations may be added in order to reduce the feasible domain. For example, in [27], an upper bound on the ℓ_0 norm of the solution sought is considered. Furthermore, many signal processing problems naturally involve linear constraints such as positivity or sum-to-one requirements. The proposed MIP-based approaches can easily be adapted to such cases, for which exact solutions can still be obtained. Adding such extra constraints may also contribute to reducing the computational time, whereas it generally penalizes the efficiency of classical (convex or greedy) sparse approximation algorithms. One may also consider directly the bi-objective optimization problem with multi-criterion optimization methods [44] in order to propose a whole range of trade-off (sparsity vs. data fitting) solutions.

Global optimization of criteria involving *structured sparsity* would also be worth being studied, where (possibly overlapping) subsets of coefficients are jointly zero or non-zero.

Such problems are generally tackled by convex optimization approaches involving mixed norms [45] or by extensions of greedy algorithms [46]. Both suffer from similar limitations than their scalar ℓ_1-norm relaxation and greedy counterparts, as far as optimality with respect to the ℓ_0-based problem is concerned. We believe that exact optimization of such problems through MIP should also be possible for moderate-size problems. For example, MIP-like formulations of some structured sparsity problems are shown in [25]—although the authors finally resort to (inexact) continuous relaxation of the binary variables—and in [47], where specific structured sparsity problems defined through totally unimodular systems allow exact optimization in polynomial time.

ACKNOWLEDGMENT

The authors thank the anonymous referees for their insightful and constructive comments, and Sonia Cafieri for helpful discussions on Mixed-Integer Programming.

REFERENCES

[1] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE*, vol. 98, no. 6, pp. 948–958, June 2010.

[2] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Rev.*, vol. 51, no. 1, pp. 34–81, Feb. 2009.

[3] S. Chen, S. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, Nov. 1989.

[4] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[5] Y. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 1993, vol. 1, pp. 40–44.

[6] A. J. Miller, *Subset Selection in Regression*, 2nd ed. London, UK: Chapman and Hall, 2002.

[7] C. Soussen, J. Idier, D. Brie, and J. Duan, "From Bernoulli-Gaussian deconvolution to sparse signal restoration," *IEEE Trans. Signal Process.*, vol. 59, no. 10, pp. 4572–4584, Oct. 2011.

[8] N. Karahanoglu and H. Erdogan, "A* orthogonal matching pursuit: Best-first search for compressed sensing signal recovery," *Digit. Signal. Process.*, vol. 22, no. 4, pp. 555–568, July 2012.

[9] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Appl. Comput. Harmon. A.*, vol. 26, no. 3, pp. 301–321, May 2009.

[10] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[11] C. Soussen, R. Gribonval, J. Idier, and C. Herzet, "Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares," *IEEE Trans. Inf. Theory*, vol. 59, pp. 3158–3174, May 2013.

[12] H. Mohimani, M. Babaie-Zadeh, and C. Jutten, "A fast approach for overcomplete sparse decomposition based on smoothed ℓ^0 norm," *IEEE Trans. Signal Process.*, vol. 57, no. 1, pp. 289–301, Jan. 2009.

[13] K. Herrity, A. Gilbert, and J. Tropp, "Sparse approximation via iterative thresholding," in *Proc. IEEE ICASSP*, 2006, vol. 3, pp. 624–627.

[14] T. Blumensath and M. Davies, "Normalized iterative hard thresholding: Guaranteed stability and performance," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 298–309, Apr. 2010.

[15] Z. Lu and Y. Zhang, "Sparse approximation via penalty decomposition methods," *SIAM J. Optimization*, vol. 23, no. 4, pp. 2448–2478, 2013.

[16] M. S. O'Brien, A. N. Sinclair, and S. M. Kramer, "Recovery of a sparse spike time series by ℓ_1 norm deconvolution," *IEEE Trans. Signal Process.*, vol. 42, no. 12, pp. 3353–3365, Dec. 1994.

[17] J. Mendel, "Some modeling problems in reflection seismology," *ASSP Magazine, IEEE*, vol. 3, no. 2, pp. 4–17, Apr. 1986.

[18] S. Bourguignon, H. Carfantan, and J. Idier, "A sparsity-based method for the estimation of spectral lines from irregularly sampled data," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 575–585, Dec. 2007.

[19] G. Tang, B. Bhaskar, and B. Recht, "Sparse recovery over continuous dictionaries - just discretize," in *Proc. Asilomar Conf. Signals, Syst., Comput.*, 2013, pp. 1043–1047.

[20] B. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comp.*, vol. 2, no. 24, pp. 227–234, Apr. 1995.

[21] M.-D. Iordache, J. Bioucas-Dias, and A. Plaza, "Sparse unmixing of hyperspectral data," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 6, pp. 2014–2039, June 2011.

[22] A. Klein, H. Carfantan, D. Testa, A. Fasoli, and J. Snipes, JET EFDA Contributors, "A sparsity-based method for the analysis of magnetic fluctuations in unevenly-spaced Mirnov coils," *Plasma Phys. Contr. F.*, vol. 50, no. 12, p. 125005, 2008.

[23] R. Bixby, "A brief history of linear and mixed-integer programming computation," *Doc. Math.*, Optimization Stories, pp. 107–121, 2012.

[24] S. Jokar and M. Pfetsch, "Exact and approximate sparse solutions of underdetermined linear equations," *SIAM J. Sci. Comp.*, vol. 31, no. 1, pp. 23–44, 2008.

[25] I. Tosic and S. Drewes, "Learning joint intensity-depth sparse representations," *IEEE Trans. Image Process.*, vol. 23, no. 5, May 2014.

[26] X. Sun, X. Zheng, and D. Li, "Recent advances in mathematical programming with semi-continuous variables and cardinality constraint," *J. Oper. Res. Soc. China*, vol. 1, no. 1, pp. 55–77, Mar. 2013.

[27] N. B. Karahanoglu, H. Erdogan, and S. I. Birbil, "A mixed integer linear programming formulation for the sparse recovery problem in compressed sensing," in *Proc. IEEE ICASSP*, 2013, pp. 5870–5874.

[28] I. Das and J. Dennis, "A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems," *Struct. Optimization*, vol. 14, no. 1, pp. 63–69, Aug. 1997.

[29] R. J. Marks, G. L. Wise, D. S. Haldeman, and J. L. Whited, "Detection in Laplace noise," *IEEE Trans. Aerospace Electronic Syst.*, vol. 14, no. 6, pp. 866–872, Nov. 1978.

[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, USA: Cambridge University Press, 2004.

[31] C. Clason, "L^∞ fitting for inverse problems with uniform noise," *Inverse Probl.*, vol. 28, no. 10, p. 104007, Oct. 2012.

[32] P. Stoica and Y. Selen, "Model-order selection: A review of information criterion rules," *IEEE Sig. Proc. Mag.*, vol. 21, no. 4, pp. 36–47, July 2004.

[33] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted minimum norm algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.

[34] M. Nikolova, "Description of the minimizers of least squares regularized with ℓ_0-norm. Uniqueness of the global minimizer," *SIAM J. Imaging Sciences*, vol. 6, no. 2, pp. 904–937, 2013.

[35] P. Belotti, C. Kirches, S. Leyffer, J. Linderoth, J. Luedtke, and A. Mahajan, "Mixed-integer nonlinear optimization," *Acta Numer.*, vol. 22, pp. 1–131, May 2013.

[36] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. New York, USA: W. H. Freeman & Co., 1979.

[37] A. Lodi, , M. Jünger, Ed. *et al.*, "Mixed integer programming computation," in *50 Years of Integer Programming 1958-2008*. Berlin, Germany: Springer, 2010, pp. 619–645.

[38] R. Gomory, "Outline of an algorithm for integer solutions to linear programs," *Bull. Amer. Math. Soc.*, vol. 64, no. 5, pp. 275–278, 1958.

[39] F. Benhamou, F. Goualard, L. Granvilliers, and J.-F. Puget, "Revising hull and box consistency," in *Proc. ICLP*, 1999, pp. 230–244.

[40] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. New York, NY, USA: Academic Press, 2009.

[41] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Homotopy continuation for sparse signal representation," in *Proc. IEEE ICASSP*, Philadelphia, USA, Mar. 2005, vol. 5, pp. 733–736.

[42] E. Carcreff, S. Bourguignon, J. Idier, and L. Simon, "Resolution enhancement of ultrasonic signals by up-sampled sparse deconvolution," in *Proc. IEEE ICASSP*, Vancouver, Canada, May 2013, pp. 6511–6515.

[43] M. C. W. Van Rossum, "A novel spike distance," *Neural Computation*, vol. 13, no. 4, pp. 751–763, Apr. 2001.

[44] M. Ehrgott, *Multicriteria Optimization*, 2nd ed. Berlin, Germany: Springer, 2005.

[45] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, Nov. 2011.

[46] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," in *Proc. ICML*, New York, USA, 2009, pp. 417–424.

[47] L. Baldassarre, N. Bhan, V. Cevher, and A. Kyrillidis, "Group-sparse model selection: Hardness and relaxation," *IEEE Trans. Inf. Theory*, arXiv preprint arXiv:1303.3207v4, to be published.

Sébastien Bourguignon was born in Dijon, France, in 1977. He received the diploma degree in electrical engineering from École Supérieure d'Électricité, Gif-sur-Yvette, France; the engineer degree from ETSIT, Universidad Politécnica de Madrid, Spain, in 2001; and the Ph.D. degree in signal processing from the University of Toulouse, France, in 2005. From 2002 to 2007, he was with the Institut de Recherche en Astrophysique et Planétologie, Toulouse, France. From 2007 to 2008, he was with IFREMER, the French research institute for exploitation of the sea, Brest, France. Then, he worked as a postdoctoral fellow at the Côte d'Azur Observatory, Nice, France. Since 2011, he is an associate professor at École Centrale de Nantes and at the IRCCyN laboratory, Nantes, France. His research interests include inverse problems in signal and image processing, statistical inference, sparse approximation, optimization algorithms, and applications to different kinds of experimental data.

Jordan Ninin was born in Villers-Semeuse, France in 1983. He received his Master degree from the engineering school ENSEEIHT in France in 2007 and his Ph.D. degree in applied mathematics and computer science in 2010 from the University of Toulouse. Since 2012, he is an associate professor at ENSTA Bretagne, Brest, France. He is a member of the team IHSEV of the Lab-STICC laboratory in France and an associate member of the research group GERAD in Canada. His research deals with exact global optimization algorithms with applications to a wide variety of problems in robotics, synthesis of robust control laws, optimal design of polygons and signal processing.

Hervé Carfantan was born in France in 1968. He received the Engineer degree in control and signal processing from the École Supérieure d'Informatique Électronique Automatique, Paris, France, in 1992, and the M.S. degree in control and signal processing and the Ph.D. degree in physics from the Université de Paris-Sud, Orsay, France, in 1992 and 1996, respectively. He is currently an Assistant Professor with the Université Toulouse 3 Paul Sabatier, Toulouse, France, and a Researcher with the Signal, Image en Sciences de l'Univers Team, Institut de Recherche en Astrophysique et Planétologie, Toulouse, which is part of the Université de Toulouse, Toulouse, and the Centre National de la Recherche Scientifique, Paris. His main interests are in inverse problems, estimation, spectral analysis (in particular, the case of irregularly sampled data), sparse approximation and optimisation. His application fields essentially concern astronomical and remote sensing data analysis.

Marcel Mongeau (M'11) received his B.Sc. (1985) and M.Sc. (1987) degrees in mathematics from Universite de Montreal, and his Ph.D. (1991) in Combinatorics & Optimization from the University of Waterloo (Canada). He was then a post-doctoral researcher at CRM (Universite de Montreal), at INRIA (France) and at the University of Edinburgh. From 1994 to 2011, he was at IMT, Université Paul Sabatier (France), where he received a Habilitation à Diriger des Recherches in 2003. He is currently Professor in Operations Research at ENAC in Toulouse (France). His research interests include Global Optimization, Numerical Optimization, and Operations Research with applications to aeronautics.

Annexe C

A linear model approach for ultrasonic inverse problems with attenuation and dispersion.

E. Carcreff, S. Bourguignon, J. Idier et L. Simon. : A linear model approach for ultrasonic inverse problems with attenuation and dispersion. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control, 61(7) :1191–1203*, juillet 2014.

IEEE TRANSACTIONS ON ULTRASONICS, FERROELECTRICS, AND FREQUENCY CONTROL, VOL. 61, NO. 7, JULY 2014

A Linear Model Approach for Ultrasonic Inverse Problems With Attenuation and Dispersion

Ewen Carcreff, Sébastien Bourguignon, Jérôme Idier, *Member, IEEE*, and Laurent Simon, *Member, IEEE*

***Abstract*—Ultrasonic inverse problems such as spike train deconvolution, synthetic aperture focusing, or tomography attempt to reconstruct spatial properties of an object (discontinuities, delaminations, flaws, etc.) from noisy and incomplete measurements. They require an accurate description of the data acquisition process. Dealing with frequency-dependent attenuation and dispersion is therefore crucial because both phenomena modify the wave shape as the travel distance increases. In an inversion context, this paper proposes to exploit a linear model of ultrasonic data taking into account attenuation and dispersion. The propagation distance is discretized to build a finite set of radiation impulse responses. Attenuation is modeled with a frequency power law and then dispersion is computed to yield physically consistent responses. Using experimental data acquired from attenuative materials, this model outperforms the standard attenuation-free model and other models of the literature. Because of model linearity, robust estimation methods can be implemented. When matched filtering is employed for single echo detection, the model that we propose yields precise estimation of the attenuation coefficient and of the sound velocity. A thickness estimation problem is also addressed through spike deconvolution, for which the proposed model also achieves accurate results.**

I. Introduction

Ultrasonic waves are widely used for nondestructive testing (NDT) of materials [1], [2], tissue characterization [3] and biomedical imaging [4], [5]. Many applications in these fields can be formulated as inverse problems such as spike train deconvolution [6]–[8], biomedical image restoration [4], [9], time-of-flight tomography [10]–[12] and synthetic aperture focusing techniques (SAFT) [13]–[15]. Such problems rely on both an accurate direct model describing the acquisition process and appropriate prior information constraining the solution [8].

Attenuation and dispersion can arise in the aforementioned applications. Attenuation is due to two basic causes; namely, scattering and absorption. Scattering results from the fact that the material is not strictly homogeneous, implying multiple direction propagation. Absorption is caused by the excitation of the particles that converts sound energy into heat. Consequently, attenuation increases with frequency. Most models in the literature consider a frequency power law [3], [16]. This frequency-dependent loss has a low-pass filtering effect on the transmitted waves and grows as propagation distance increases. It generates a shape broadening of the echoes that degrades the resolution [17]. Dispersion—a by-product of attenuation—means that the phase velocity depends on frequency, producing a frequency-dependent phase variation of the echoes [5], [17], [18].

Attenuation and dispersion are often overlooked in ultrasonic direct models for inversion purposes. Most formulations consider a linear model that is invariant with respect to the propagation distance. However, if the ultrasonic propagation characteristics are not considered, the performance of the algorithms can be degraded because the model accuracy is too weak. Several approaches have been developed to include attenuation in acoustical inverse problems, mostly in geophysics [19]–[21]. Indeed, similar effects impact the propagation of seismic waves that is modeled as the Q-filter [19], [22]. In the field of ultrasonic NDT, some methods overcome the issue of wave distortion by proposing a greater flexibility in the direct model. On the one hand, parametric methods associate a specific shape to each echo, which is usually modeled as a modulated Gaussian pulse. The parameters of each echo are then estimated by nonlinear least-squares fitting [23] or by greedy procedures such as matching pursuit [24]–[26]. On the other hand, nonparametric approaches employ a blind strategy in which weaker constraints are imposed on the echo shape—as, for example, slow variations between neighboring intervals of the propagation distance [27]. Even if such models allow for some shape variation of the echoes with respect to the propagation distance, they do not introduce any physical knowledge about the propagation properties.

The goal of the present paper is to contribute to the solving of ultrasonic inverse problems by including attenuation and dispersion in the direct model. In particular, we propose to account for physical attenuation profiles defined in the frequency domain like power law attenuation models. Our objective is threefold. First, we improve the ultrasound model accuracy compared with the standard attenuation-free model. Second, in contrast with the aforementioned methods [23]–[27], we yield a more constrained description of the data. Consequently, better performance of the inversion procedure is expected. In particular, a more accurate model aims at improving echo detection for

Manuscript received January 18, 2014; accepted April 18, 2014. This work was partially supported by the French Région Pays de la Loire as part of the scientific program Non-Destructive Testing and Evaluation–Pays de la Loire (ECND-PdL).

E. Carcreff, S. Bourguignon, and J. Idier are with the Institut de Recherche en Communications et Cybernétique de Nantes (IRCCyN), Nantes, France (e-mail: ewen.carcreff@irccyn.ec-nantes.fr).

E. Carcreff and L. Simon are with the Laboratoire d'Acoustique de l'Université du Maine (LAUM), Le Mans, France.

DOI http://dx.doi.org/10.1109/TUFFC.2014.3018

long propagation distances where the signal-to-noise ratio is low. Last, our framework yields a linear direct model which enables the use of many acknowledged inversion methods [8]. Related works [7], [28] proposed similar approaches but with an empirical description of attenuation within a time-domain signal model. The model that we propose is derived from the physics of wave propagation and is described in the frequency domain [3], [16].

In this paper, the signal model is formulated as a set of transfer functions in the Fourier domain [29]. The propagation in the medium is modeled by a so-called radiation transfer function [30] depending on the wavenumber $k(f) = \beta(f) - j\alpha(f)$. The term $\beta(f)$ is related to the phase velocity and $\alpha(f)$ is the attenuation. To ensure the consistency of the corresponding radiation impulse response, $\alpha(f)$ and $\beta(f)$ are analytically linked by conditions derived from the Kramers–Kronig relations [31]. Several models have been developed, both for continuous-time and discrete-time signals, and validated for materials having linear and nonlinear attenuation [31]–[34]. In this paper, we will use the formulation proposed by Kuc [35], [36]. Because attenuation depends on propagation distance, we build a set of radiation impulse responses by an appropriate discretization of the space domain. We then obtain a discrete linear model of data $\mathbf{y} = \mathbf{H}_\mathrm{e}\mathbf{H}_\mathrm{a}\mathbf{x}$. The matrix $\mathbf{H}_\mathrm{e}$ represents the instrumental response, which is invariant with respect to the propagation distance. The matrix $\mathbf{H}_\mathrm{a}$ stands for the set of radiation impulse responses. The vector $\mathbf{x}$ represents the unknown spatial distribution of targets. Note that standard direct models consider an invariant model with respect to the propagation distance; that is, $\mathbf{y} = \mathbf{H}_\mathrm{e}\mathbf{x}$.

A major advantage of the obtained model is in its generality for a large variety of ultrasonic inverse problems. The purpose of inversion is then to estimate the object $\mathbf{x}$ based on the knowledge of $\mathbf{y}$, $\mathbf{H}_\mathrm{e}$, and $\mathbf{H}_\mathrm{a}$, and on some prior information on $\mathbf{x}$. In NDT for example, $\mathbf{x}$ is expected to have a few nonzero elements, corresponding to the positions of impedance discontinuities, including flaws [6]. The description and the validation of sophisticated inversion algorithms are out of the scope of this paper. Consequently, experiments are dedicated to rather simple estimation problems in homogeneous plates. We use two basic inverse methods, namely, a matched filtering procedure and an ℓ_1-norm-based spike train deconvolution method [6] applied to non-overlapping echoes.

The paper is organized as follows. Section II describes the model of ultrasonic signals, defined in the frequency domain. The relation proposed by Kuc [35] between the phase and the magnitude of the radiation transfer function is detailed. The model is validated on experimental data acquired in a polymethyl methacrylate (PMMA) plate. Then, in Section III, a linear direct model is built, based on the discretization of the unknown spatial source distribution. Such a frequency-based approach is compared with Olofsson's time-domain model [7]. In Section IV, experimental results are shown through the nondestructive evaluation of a polycarbonate plate to estimate the attenuation coefficient and the velocity. Section V considers a typical plate thickness estimation problem in which a spike train deconvolution method is considered using our model. The paper ends with a discussion in Section VI.

II. Propagation Model of Ultrasonic Signals

A. Model for a Single Target

Here, the ultrasonic measurement is made in a homogeneous and isotropic medium containing a single point-like target placed at spatial location $\mathbf{r}_\mathrm{T}$. The configuration can either use two transducers, a transmitter and a receiver (T/R), or a single transducer in pulse–echo (P/E) mode—see Fig. 1 for the description of both configurations. The received signal can be defined in the Fourier domain by a set of transfer functions [29], [37]:

$$Y(f, \mathbf{r}_\mathrm{T}) = U(f) H_\mathrm{ea}(f) H_\mathrm{r}(f, \mathbf{r}_\mathrm{T}) H_\mathrm{ae}(f), \tag{1}$$

as illustrated in Fig. 2. $U(f)$ is the electrical excitation pulse sent to the emitting transducer. The functions $H_\mathrm{ea}(f)$ and $H_\mathrm{ae}(f)$ are the electro-acoustical and acousto-electrical frequency responses of the transducers, respectively. The global instrumental function $H_\mathrm{e}(f) = U(f)H_\mathrm{ea}(f)H_\mathrm{ae}(f)$ can be defined by collecting the functions that do not depend on the propagation in the material. The received spectrum is therefore

$$Y(f, \mathbf{r}_\mathrm{T}) = H_\mathrm{e}(f) H_\mathrm{r}(f, \mathbf{r}_\mathrm{T}). \tag{2}$$

The radiation transfer function $H_\mathrm{r}(f, \mathbf{r}_\mathrm{T})$ represents the transfer function related to the propagation path [29], [30]. As an example, let us consider the one-way path from the emitting transducer with surface S to the target. The radiation transfer function[1] in $\mathbf{r}_\mathrm{T}$ is the sum over elementary contributions of sources over the surface S [30]:

$$H_\mathrm{r}^1(f, \mathbf{r}_\mathrm{T}) = \int_{\mathbf{r}_0 \in S} \frac{e^{-jk(f)\|\mathbf{r}_\mathrm{T} - \mathbf{r}_0\|}}{2\pi \|\mathbf{r}_\mathrm{T} - \mathbf{r}_0\|} \mathrm{d}S. \tag{3}$$

In T/R mode, as represented in Fig. 1, the overall radiation transfer function is the product of two different radiation functions (transmitter to target and target to receiver), whereas in P/E mode, the two transfer functions are equal [38]. Note that one could consider specific target surfaces, leading to different reflector signatures [37]. For example, Lhémery has developed a model with small oriented targets, including the diffraction of the transducers [38]. In the current paper, we consider targets with identical signatures.

The complex-valued frequency-dependent wavenumber $k(f)$ can be written

[1] By convention, for a frequency f, we consider a plane wave $e^{j(2\pi ft - k(f)x)}$ propagating in the positive x direction, where $k(f)$ is the wavenumber.

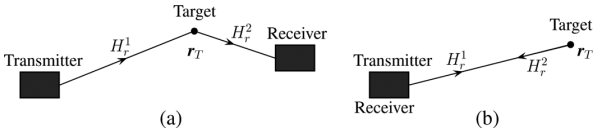


Fig. 1. Two possible configurations for ultrasonic data aquisition: (a) transmit–receive mode and (b) pulse–echo mode.

$$k(f) = \beta(f) - j\alpha(f). \tag{4}$$

The term $\beta(f)$ describes the propagation of the wave such that $\beta(f) = 2\pi f/c(f)$, where $c(f)$ is the phase velocity. The term $\alpha(f)$ represents the attenuation in the material. By neglecting the diffraction of the transducers, we assume that the distance of propagation is roughly constant for all points of the radiating and receiving surfaces [29]. Let z represent the distance of the travel path from the center of the emitter to the center of the receiver through the target. From (3) and (4), one can write the radiation transfer function:

$$\begin{aligned} H_{\mathrm{r}}(f,z) &= b(z)e^{-\alpha(f)z}e^{-j\beta(f)z} \\ &= b(z)e^{-\alpha(f)z}e^{-j2\pi fz/c(f)}. \end{aligned} \tag{5}$$

The frequency-independent term $b(z)$ depends on the propagation distance and on several factors such as the transducer and target surfaces and their relative positions. For each z, the radiation impulse response $h_{\mathrm{r}}(t,z)$ is given by the inverse Fourier transform of $H_{\mathrm{r}}(f,z)$. Similarly, $h_{\mathrm{e}}(t)$ is the impulse response corresponding to $H_{\mathrm{e}}(f)$. From (2), the time-domain signal received by the transducer for a single target is

$$y(t,z) = h_{\mathrm{e}}(t) * h_{\mathrm{r}}(t,z), \tag{6}$$

which is the convolution between the instrumental impulse response $h_{\mathrm{e}}(t)$ and the radiation impulse response $h_{\mathrm{r}}(t,z)$ depending on the propagation distance z.

Most ultrasound propagation models in tissue characterization and NDT consider a frequency power law attenuation model [16]:

$$\alpha(f) = \alpha_0|f|^{\gamma}, \tag{7}$$

where α_0 and γ are real positive parameters characterizing a given material. Generally, the frequency power parameter satisfies $1 \leq \gamma \leq 2$ [16], [34]. For tissues, one typically has $1 \leq \gamma \leq 1.5$ [39]. Many materials have linear attenuation, that is, $\gamma = 1$ [3], [32]. Numerous methods have been proposed to measure the parameter α_0 in such a case [3], [35], [36], [40]. In NDT, $\gamma > 1$ corresponds to nonlinear attenuation and is related to more complex material structures, for instance, polyethylene ($\gamma \approx 1.13$), synthetic rubber ($\gamma \approx 1.38$), or castor oil ($\gamma \approx 1.67$) [34], [40].

Note that the attenuation model (7) is similar to the constant-Q model employed in geophysics for seismic waves [22]. Q is a quality factor and can be defined as

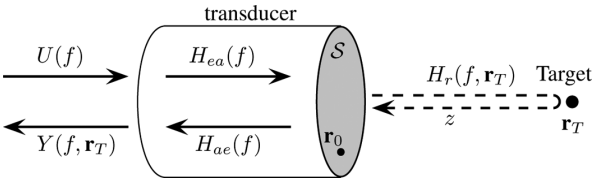


Fig. 2. Pulse–echo measurement for a target located at $\mathbf{r}_{\mathrm{T}}$ using a radiating surface $\mathcal{S}$. The received signal $Y(f,\mathbf{r}_{\mathrm{T}})$ is modeled through a set of transfer functions: $Y(f,\mathbf{r}_{\mathrm{T}}) = U(f)H_{\mathrm{ea}}(f)H_{\mathrm{r}}(f,\mathbf{r}_{\mathrm{T}})H_{\mathrm{ae}}(f)$. z is the overall propagation distance.

$Q = \pi/\alpha_0 c_0$, if c_0 is the constant wave velocity [21]. The parameter Q is also inversely proportional to α_0, meaning that infinite Q corresponds to a lossless medium ($\alpha_0 = 0$). Similarly, it has been shown that propagation in soils is adequately modeled with $\gamma \in [1,2]$ [21].

From (5) and (7), attenuation has a low-pass effect, causing a downshift of the center frequency of the echoes as the propagation distance increases [16], [17], [41] and limiting the resolution. Such an effect is represented in Fig. 3(a), which shows a typical radiation transfer function $|H_{\mathrm{r}}(f,z)|$ as a function of both frequency f and propagation distance z. This implies the use of relatively low-frequency transducers and a strong amplitude loss for long distances. In the time domain, this effect causes a broadening of the echoes that lessens the resolution as the distance increases.

B. Causality of the Radiation Impulse Response

For physical reality purpose, constraints are applied on the radiation impulse response $h_{\mathrm{r}}(t,z)$. First, the response is real-valued and therefore implies the Hermitian symmetry property $H_{\mathrm{r}}(-f,z)^* = H_{\mathrm{r}}(f,z)$, where the asterisk stands for complex conjugation. As a consequence, according to (4) and (5), the wavenumber has the anti-Hermitian symmetry property, leading to $\alpha(f)$ even and $\beta(f)$ odd. Second, in acoustics, the phase velocity $c(f)$ increases as a function of frequency, which is called the anomalous dispersion [34]. Therefore, there exists a maximum velocity for $f = \infty$ [42], say, c_0. For a given distance z, this maximum velocity is directly linked to a minimum time of flight $t_0 = z/c_0$ such that $h_{\mathrm{r}}(t,z) = 0$ for $t < t_0$. In other words, an ultrasonic wave emitted at $t = 0$ should not appear before t_0 for a target located at distance z. Note that c_0 is larger than the group velocity, which is the velocity of the envelope of the waveform. Such causality principle implies specific relations between the phase and the magnitude in (5) [34], [43].

Kak and Dines [3] proposed a linear phase model, under a linear attenuation assumption $\alpha(f) = \alpha_0|f|$, by considering a constant phase velocity $c(f) = c_0$:

$$H_{\mathrm{r}}(f,z) = b(z)e^{-\alpha(f)z}e^{-j2\pi fz/c_0}. \tag{8}$$

In this case, c_0 also corresponds to the group velocity. The inverse Fourier transform of $e^{-j2\pi fz/c_0}$ is a delta function

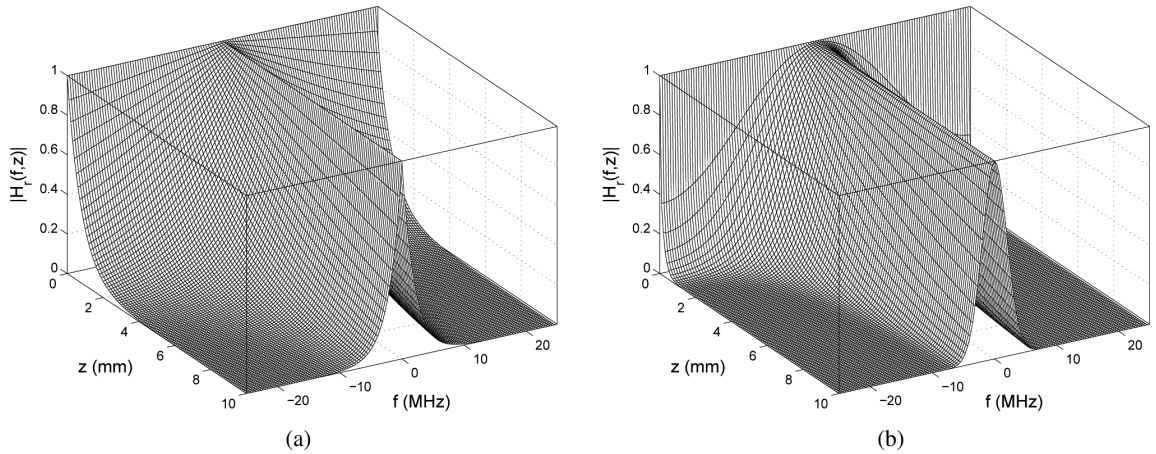


Fig. 3. Examples of attenuation as a function of frequency f and of propagation distance z. The larger the distance, the more important is the low-pass filtering effect of attenuation. (a) Linear attenuation model: $|H_\mathrm{r}(f,z)| = e^{-\alpha_0|f|z}$, with $\alpha_0 = 50$ Np/MHz/m. (b) Olofsson's model [7] with $\alpha = 0.2$ (see Section III-C).

$\delta(t - z/c_0)$ corresponding to a pure delay of $t_0 = z/c_0$. However, under this linear phase assumption, the impulse response $h_\mathrm{r}(t,z)$ is symmetric with respect to t_0 and hence is not causal [3].

Indeed, the Paley–Wiener condition states that $h_\mathrm{r}(t,z)$ is causal if and only if [44]

$$\int_{-\infty}^{+\infty} \frac{|\ln|H_\mathrm{r}(f,z)||}{1+f^2}\mathrm{d}f < \infty. \tag{9}$$

In such a case, the corresponding phase term is derived from the Kramers–Kronig relations [31], [44]. For the power law attenuation model (7), equation (9) is verified only for $\gamma < 1$ [40], [42], [43]; that is, $h_\mathrm{r}(t,z)$ is not causal for $\gamma \geq 1$.

Gurumurthy and Arthur [32] considered a minimum-phase model [44], [45] accounting for dispersion in the case $\gamma = 1$. They considered that attenuation grows only sub-linearly at high frequencies to compute the dispersion from the Kramers–Kronig relations [32]. Nevertheless, this model is not strictly causal because the Paley–Wiener condition is not respected. A dispersion term $\varepsilon(f)$ is added to the linear phase term in (8), which yields

$$H_\mathrm{r}(f,z) = b(z)e^{-\alpha(f)z}e^{-j2\pi fz/c_0}e^{-j\varepsilon(f)z}. \tag{10}$$

$H_\mathrm{r}(f,z)$ can then be separated into the linear-phase function $b(z)e^{-j2\pi fz/c_0}$ and the attenuation function:

$$H_\mathrm{a}(f,z) = e^{-\alpha(f)z}e^{-j\varepsilon(f)z}, \tag{11}$$

that is, in the time domain:

$$\begin{aligned} h_r(t,z) &= b(z)h_\mathrm{a}(t,z) * \delta(t - z/c_0) \\ &= b(z)h_\mathrm{a}(t - z/c_0, z), \end{aligned} \tag{12}$$

with $h_\mathrm{a}(t,z)$ the attenuation impulse response corresponding to $H_\mathrm{a}(f,z)$.

Analogously, Kuc derived a minimum-phase model for discrete-time signals for $\gamma = 1$ [35], [36], that can be easily extended to nonlinear attenuation. In this paper, we will use this formulation to describe $H_\mathrm{a}(f,z)$, which is detailed hereafter.

C. Causality Constraint on the Discrete-Time Impulse Response

For continuous-time signals, the causality of the system defined by the transfer function $H(f) = H_\mathrm{R}(f) + jH_\mathrm{I}(f)$ imposes that H_I is the Hilbert transform of H_R [44]. The equivalent characterization for discrete-time signals reads [45]

$$H_\mathrm{I}(f) = -\frac{1}{f_\mathrm{S}}\mathcal{P}\int_{-f_\mathrm{S}/2}^{f_\mathrm{S}/2} H_\mathrm{R}(g)\cot\left(\frac{\pi}{f_\mathrm{S}}(f-g)\right)\mathrm{d}g, \tag{13}$$

where f_S is the sampling frequency and $\mathcal{P}$ denotes the Cauchy principal value of the integral. Taking the logarithm of $H_\mathrm{a}(f,z)$ in (11), $\ln H_\mathrm{a}(f,z) = -\alpha(f)z - j\varepsilon(f)z$, and assuming that the corresponding impulse response is causal [45] leads to

$$\varepsilon(f) = \mathcal{H}(\alpha(f)), \tag{14}$$

where $\mathcal{H}(\alpha(f))$ is defined by

$$\mathcal{H}(\alpha(f)) = -\frac{1}{f_\mathrm{S}}\mathcal{P}\int_{-f_\mathrm{S}/2}^{f_\mathrm{S}/2} \alpha(g)\cot\left(\frac{\pi}{f_\mathrm{S}}(f-g)\right)\mathrm{d}g. \tag{15}$$

Note that this expression is well-defined for a large class of attenuation models $\alpha(f)$. Moreover, $\mathcal{H}(\alpha(f))$ has an analytic expression for linear attenuation as established in the appendix, which is useful for fast and precise computations. For power-law attenuation models with $\gamma \neq 1$, (15) can be computed by numerical integration.

The dispersion term (15) has been used in the case of linear attenuation to model attenuated signals in PMMA plates [35], [36]. Several studies precisely compared phase velocities from measured signals with the model in (14) [40], [43]. They concluded to a satisfactory agreement from materials having linear and nonlinear attenuation (PMMA, rubber, castor oil).

A synthetic example of impulse responses $h_\mathrm{r}(t,z)$ is presented in Fig. 4. A single target is located at distance z = 1 mm in a homogeneous and isotropic material with propagation parameters c_0 = 2000 m/s, γ = 1, α_0 = 50 Np/MHz/m, and $b(z)$ = 1. Three propagation models are used: without attenuation, with linear phase (8), and with dispersive phase (14). The response without attenuation is a delta function with a shift of $t_0 = z/c_0 = 0.5$ µs. The impulse response of the linear phase model has a symmetric shape on both sides of t_0, which violates the causality condition. As expected, the impulse response of the dispersive model appears to be causal because the values are very close to zero—less than 10^{-10}—before t_0.

D. Validation With Experimental Data

We now assess the accuracy of the dispersive model and compare it with other available models. The measurements are performed using a flat circular transducer of diameter 12.7 mm and center frequency 2.25 MHz. The object under test is a 25-mm-thick PMMA plate, immersed in a water tank as illustrated in Fig. 5. The plate is in the far field of the transducer and with normal incidence. Data are acquired at sampling frequency f_S = 100 MHz and averaged over 100 realizations to reduce the noise level. We extract the two significant echoes from the data: the front wall echo $y_\mathrm{f}(t)$ [Fig. 6(a)] and the back wall echo $y_\mathrm{b}(t)$ [Fig. 6(b)]. Their respective Fourier transforms, $Y_\mathrm{f}(f)$ and $Y_\mathrm{b}(f)$, are plotted in moduli in Fig. 6(c), showing a strong amplitude loss and a frequency downshift between the front wall and the back wall echoes.

From (2) and (5), one has

$$Y_\mathrm{f}(f) = b(2D)H_\mathrm{e}(f)e^{-(\alpha_\mathrm{w}(f)+j\beta_\mathrm{w}(f))2D}, \tag{16}$$

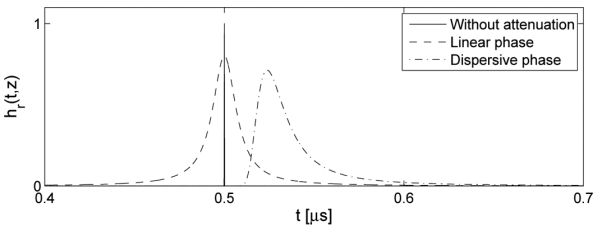


Fig. 4. Simulated impulse responses $h_\mathrm{r}(t,z)$ using three propagation models: without attenuation, with linear phase, and with dispersive phase. Parameters: c_0 = 2000 m/s, γ = 1, α_0 = 50 Np/MHz/m, z = 1 mm, and $b(z)$ = 1. The minimum time of flight is then $t_0 = z/c_0 = 0.5$ µs. This value corresponds to: 1) the position of the delta function $\delta(t - t_0)$, 2) the center of the linear-phase impulse response, and 3) the starting time of the dispersive-phase impulse response.

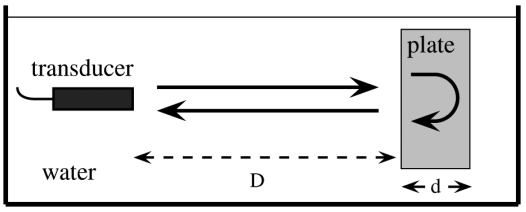


Fig. 5. Simplified data measurement setup. The plate of thickness d is placed normally in the far field of the transducer.

where $b(2D)$ is the reflection coefficient at the front face of the plate [46]. The terms $\alpha_\mathrm{w}(f)$ and $\beta_\mathrm{w}(f)$ stand for the propagation parameters in water [see (4)]. The back wall echo is modeled as

$$Y_\mathrm{b}(f) = b(2D+2d)H_\mathrm{e}(f)e^{-(\alpha_\mathrm{w}(f)+j\beta_\mathrm{w}(f))2D-(\alpha(f)+j\beta(f))2d}, \tag{17}$$

with $b(2D + 2d)$ the resulting amplitude for the whole wave travel. Dividing $Y_\mathrm{b}(f)/Y_\mathrm{f}(f)$ enables the cancellation of $H_\mathrm{e}(f)$ and of the terms related to the propagation in water:

$$\frac{1}{2d}\ln\left(\frac{|Y_\mathrm{b}(f)|}{|Y_\mathrm{f}(f)|}\right) = \alpha(f) + \frac{1}{2d}\ln\left|\frac{b(2D+2d)}{b(2D)}\right|. \tag{18}$$

Fig. 6(d) shows that such a function is approximately linear with respect to frequency, say $\alpha(f) \simeq \alpha_0|f|$. A linear regression leads to $\alpha_0 \simeq 11.55$ Np/MHz/m, which is in

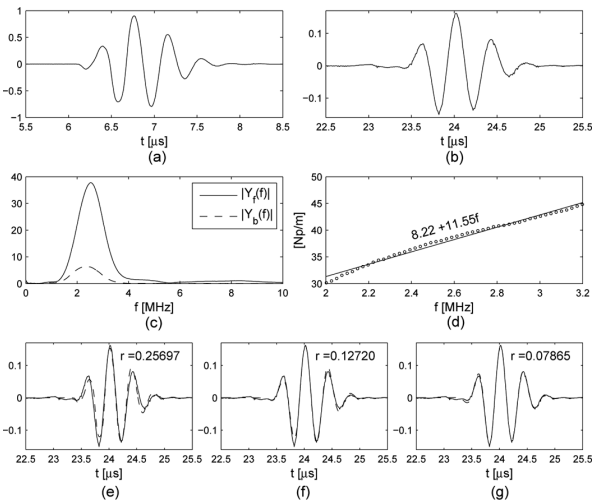


Fig. 6. Attenuation and dispersion models for data acquired from a PMMA plate with thickness d = 25 mm using a 2.25 MHz probe. (a) front wall and (b) back wall echoes. (c) Magnitude spectra $|Y_\mathrm{f}(f)|$ (solid line) and $|Y_\mathrm{b}(f)|$ (dashed line). (d) Attenuation measurement: spectral magnitude ratio $1/(2d)\ln(|Y_\mathrm{b}(f)|/|Y_\mathrm{f}(f)|)$ (circles) and linear regression (solid line). (e)–(g) Measured back wall echo (solid line) and three different models (dashed line), with r the quadratic error between data and model; (e) without attenuation, (f) linear phase, and (g) dispersive phase. Parameters: c_0 = 2802 m/s, α_0 = 11.55 Np/MHz/m, γ = 1 (linear attenuation assumption).

agreement with the values proposed in the literature [40], [47].

By neglecting attenuation and dispersion in water, which is usually assumed in the literature [40], one has $\alpha_{\mathrm{w}}(f) = 0$ and $\beta_{\mathrm{w}} = 2\pi f/c_{\mathrm{w}}$, where c_{w} is the constant speed of sound in water. Therefore, the front wall echo is

$$Y_{\mathrm{f}}(f) = b(2D)H_{\mathrm{e}}(f)e^{-j2\pi f 2D/c_{\mathrm{w}}} \tag{19}$$

in the frequency domain, that is, in the time domain: $y_{\mathrm{f}}(t) = b(2D)h_{\mathrm{e}}(t - 2D/c_{\mathrm{w}})$. According to (17), we then fit the back wall echo $y_{\mathrm{b}}(t)$ with

$$\hat{y}_{\mathrm{b}}(t) = b(2D + 2d)h_{\mathrm{e}}(t - 2D/c_{\mathrm{w}}) * h_{\mathrm{r}}(t, 2d), \tag{20}$$

where $h_{\mathrm{e}}(t - 2D/c_{\mathrm{w}})$ is obtained from $y_{\mathrm{f}}(t)$ and $h_{\mathrm{r}}(t, 2d)$ is set from the three models introduced in Section II-B: without attenuation, with linear phase as in (8), and with dispersive phase as in (14). In the two last cases, attenuation is supposed linear, with α_0 at the previously estimated value. The velocity c_0 is calculated from $2d/(t_{\mathrm{b}} - t_{\mathrm{f}})$ where t_{f} and t_{b} stand for the front wall and back wall times of flight, respectively, yielding $c_0 \approx 2802$ m/s. For each model, the amplitudes and the times of flight of the echoes are optimized to achieve the best least-squares fit. The results are plotted in Figs. 6(e)–6(g). The quadratic error between the data $y_{\mathrm{b}}(t)$ and the model $\hat{y}_{\mathrm{b}}(t)$ is also computed and displayed on the corresponding subfigures. As expected, the dispersive model gives the best results, followed by the linear phase model.

III. A Discrete-Time Linear Model for Inverse Problems

A. Inversion Framework

Let us consider a propagation medium composed of an unknown distribution of point-like targets. From (6) and (12), the received signal is then the sum over all the target contributions:

$$\begin{aligned} y(t) &= \int_z h_{\mathrm{e}}(t) * h_{\mathrm{r}}(t, z)\mathrm{d}z \\ &= \int_z b(z)h_{\mathrm{e}}(t) * h_{\mathrm{a}}(t - z/c_0, z)\mathrm{d}z \\ &= \int_z b(z)\Big(\int_u h_e(u)h_{\mathrm{a}}(t - u - z/c_0, z)\mathrm{d}u\Big)\mathrm{d}z, \end{aligned} \tag{21}$$

where $b(z)$ describes the material spatial content independently of the ultrasonic wave propagation effects. In this context, inversion aims at reconstructing such a function from a finite number of noisy samples of $y(t)$. The reconstruction procedure might incorporate some prior knowledge on $b(z)$. For example, from an acquired A-scan in NDT, $b(z)$ may be a spike train containing the spatial positions of the acoustical impedance changes [7]:

$$b(z) = \sum_{k=1}^{K} b_k \delta(z - z_k), \tag{22}$$

where unknown parameters z_k and b_k represent the position of the kth impedance discontinuity and the associated amplitude, respectively. In the case in which no attenuation is taken into account, that is, $h_{\mathrm{a}}(t,z) = \delta(t)$, (21) and (22) formulate the classical spike train deconvolution problem that has been widely addressed in the literature [6]–[8].

Note that such a formulation can be extended to bidimensional data such as B-scan images. It can also be adapted to two-dimensional data in ultrasonic image restoration [4], [9] and SAFT [13]–[15].

B. A Linear Model Including Attenuation and Dispersion

Let us consider the discrete-time signal $y_n = y(n\Delta_t)$, $n = 0, \ldots, N_y - 1$, where $\Delta_t = 1/f_{\mathrm{S}}$ is the sampling period. From (21), one has

$$y_n = \int_z b(z)\int_u h_{\mathrm{e}}(u)h_{\mathrm{a}}(n\Delta_t - u - z/c_0, z)\mathrm{d}u\mathrm{d}z. \tag{23}$$

Discretizing the time integral at rate Δ_t yields:

$$\begin{aligned} y_n &\simeq \int_z b(z)\Delta_t \sum_m h_{\mathrm{e}}(m\Delta_t)h_{\mathrm{a}}\Big((n - m)\Delta_t - \frac{z}{c_0}, z\Big)\mathrm{d}z \\ &= \Delta_t \sum_m h_{\mathrm{e}}(m\Delta_t)\int_z b(z)h_{\mathrm{a}}\Big((n - m)\Delta_t - \frac{z}{c_0}, z\Big)\mathrm{d}z. \end{aligned} \tag{24}$$

Note that using the data sampling period for the discretization rate is a practical choice that is commonly made in inverse problems. In particular, it yields a Toeplitz matrix structure that can be exploited for fast computations [48]. In [49], we have recently proposed a model with higher discretization rate, showing better estimation performance in some spike train deconvolution problems—but under the usual framework $h_{\mathrm{a}}(t, z) = \delta(t)$. In this paper, we restrict the description to the discretization at rate Δ_t for the sake of clarity.

Similarly, the spatial integral is discretized at a given step size Δ_z:

$$y_n \simeq \Delta_t\Delta_z \sum_m h_{\mathrm{e}}(m\Delta_t)\sum_i b(i\Delta_z)h_{\mathrm{a}}\Big((n - m)\Delta_t - \frac{i\Delta_z}{c_0}, i\Delta_z\Big). \tag{25}$$

A natural choice is to consider Δ_z corresponding to the data time sampling $\Delta_z = c_0\Delta_t$. We then obtain

$$y_n \simeq \Delta_t\Delta_z \sum_m h_{\mathrm{e}}(m\Delta_t)\sum_i b(i\Delta_z)h_{\mathrm{a}}((n - m - i)\Delta_t, i\Delta_z). \tag{26}$$

Let us now denote $x_i = \Delta_t\Delta_z b(i\Delta_z)$ and let us consider the column vectors $\mathbf{y} = [y_0, \ldots, y_{N_y - 1}]^{\mathrm{T}}$ and $\mathbf{x} = [x_0, \ldots, x_{N_x - 1}]^{\mathrm{T}}$,

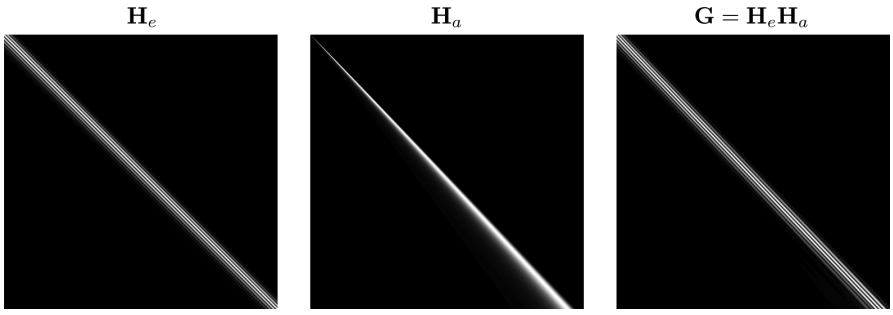


Fig. 7. Example of matrices $\mathbf{H}_\mathrm{e}$, $\mathbf{H}_\mathrm{a}$, and $\mathbf{G}$, where each column is normalized for representation clarity. $\mathbf{H}_\mathrm{e}$ is the instrumental convolution matrix obtained from the front wall echo in Fig. 6. In $\mathbf{H}_\mathrm{a}$, each line corresponds to an instant $t_n = n\Delta_t$ and each column corresponds to a spatial distance $z_i = i\Delta_z$. $\mathbf{H}_\mathrm{a}$ depends on both the attenuation and dispersion models. In this example, $\alpha(f) = \alpha_0|f|$ with $\alpha_0 = 50$ Np/MHz/m. The dispersion is set using the model in (14). The matrix $\mathbf{G}$ combines both instrumental and attenuative effects.

where the superscript T denotes matrix transposition. Note that N_x and N_y are not necessarily equal, depending on the boundary assumptions of the convolution [8]. We finally obtain the matrix-vector model:

$$\mathbf{y} = \mathbf{H}_\mathrm{e}\mathbf{H}_\mathrm{a}\mathbf{x} + \mathbf{e} = \mathbf{G}\mathbf{x} + \mathbf{e}, \tag{27}$$

where

- $\mathbf{H}_\mathrm{e}$ is the convolution matrix corresponding to the instrumental response with elements $\{h_\mathrm{e}((p-q)\Delta_t)\}_{p=0,\ldots,N_y-1,q=0,\ldots,N_x-1}$, where p and q, respectively, are the line and column indices. This matrix has a Toeplitz structure.
- $\mathbf{H}_\mathrm{a}$ is the attenuation matrix with elements $\{h_\mathrm{a}((n-i)\Delta_t, i\Delta_z)\}_{n=0,\ldots,N_x-1,i=0,\ldots,N_x-1}$. That is, the ith column of $\mathbf{H}_\mathrm{a}$ corresponds to the radiation impulse response at distance $z_i = i\Delta_z$. In our approach, it is computed from $H_\mathrm{a}(f, i\Delta_z)$ in (11), with the power-law attenuation model in (7) and the corresponding dispersive phase model defined by (14). In practice, $H_\mathrm{a}(f, i\Delta z)$ is evaluated on a frequency grid in $[-f_\mathrm{S}/2, f_\mathrm{S}/2]$ with thin spacing such that temporal aliasing can be neglected. Then, the impulse response $h_\mathrm{a}((n-i)\Delta t, i\Delta z), n = 0, \ldots, N_x - 1$ is obtained by inverse discrete Fourier transform. Causality is imposed by setting $h_\mathrm{a}((n-i)\Delta t, i\Delta z) = 0$ for $n < i$. Consequently, $\mathbf{H}_\mathrm{a}$ is lower triangular.
- $\mathbf{G} = \mathbf{H}_\mathrm{e}\mathbf{H}_\mathrm{a}$ combines the effects of the instrumental impulse responses and of the radiation impulse responses.
- $\mathbf{x}$ is the unknown sequence describing the target distribution.
- $\mathbf{e}$ is a perturbation term accounting for noise and model errors.

An example of matrices $\mathbf{H}_\mathrm{e}$, $\mathbf{H}_\mathrm{a}$, and $\mathbf{G}$ is given in Fig. 7. The matrix $\mathbf{H}_\mathrm{e}$ is built from the front wall echo in Fig. 6(a). The matrix $\mathbf{H}_\mathrm{a}$ is generated from the linear attenuation model $\alpha(f) = \alpha_0|f|$ and the dispersive relation defined in (14), with $\alpha_0 = 50$ Np/MHz/m. An example of data generated from the columns 200 and 1000 of the previous matrix $\mathbf{G}$ is also plotted in Fig. 8. Note that attenuation provokes the widening of the second echo. Dispersion causes a phase distortion that creates a time shift of the echo envelope. Indeed, in $\mathbf{H}_\mathrm{a}$, the maximum of each column is down-shifted away from the diagonal.

The model (27) states that data $\mathbf{y}$ are a noisy linear combination of columns of $\mathbf{G}$, and $\mathbf{x}$ collects the associated weights. Estimating $\mathbf{x}$ from $\mathbf{y}$ is an inverse problem that cannot be satisfactorily inverted in a least-squares sense: the matrix $\mathbf{G}$ is ill-conditioned and the generalized inverse $(\mathbf{G}^\mathrm{T}\mathbf{G})^{-1}\mathbf{G}^\mathrm{T}\mathbf{y}$ therefore suffers from uncontrolled noise amplification [8]. In the preceding example, the condition number of matrix $\mathbf{G}$ (which is a 1500×1500 matrix) is approximately $9 \cdot 10^{19}$.

C. Comparison With Olofsson's Model

From (5), the radiation transfer function at $z_i = i\Delta_z$ reads

$$\begin{aligned} H_\mathrm{r}(f, z_i) &= b(z_i)e^{-[\alpha(f)+j\beta(f)]i\Delta_z} \\ &= b(z_i)[e^{-(\alpha(f)+j\beta(f))\Delta z}]^i \\ &= b(z_i)P(f)^i. \end{aligned} \tag{28}$$

It corresponds, up to a multiplicative constant, to the ith power of the frequency kernel $P(f) = e^{-(\alpha(f)+j\beta(f))\Delta z}$. Let ρ_n, $n = 0, \ldots, N-1$ represent the discrete-time sequence with Fourier transform $P(f)$ for $f \in [-f_\mathrm{S}/2, f_\mathrm{S}/2]$. It represents the radiation impulse response

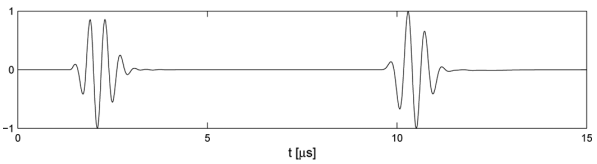


Fig. 8. Example of signal generated from the matrix $\mathbf{G}$ in Fig. 7. The data $\mathbf{y}$ are computed from the columns 200 and 1000 of matrix $\mathbf{G}$ such that: $\mathbf{y} = \mathbf{g}_{200} + \mathbf{g}_{1000}$ where $\mathbf{g}_i$ is the ith column of $\mathbf{G}$. Both echoes are normalized in amplitude for visualization clarity.

between two elementary spatial layers, separated by Δ_z. Eq. (28) states that the radiation impulse response at z_i is equal to $i-1$ self-convolutions of the time kernel ρ.

In a time-domain setting, Olofsson and Stepinski [7] proposed an empirical choice for the kernel ρ, defined by

$$\rho_0 = 0, \quad \rho_1 = 1-a, \quad \rho_2 = a, \text{ and } \rho_n = 0 \quad \forall\ n \geq 3, \tag{29}$$

with $a > 0$ and close to zero. Such a kernel implies a small distortion between two elementary layers spaced by Δ_z and is hence close to a delayed Kronecker delta function. Because of its first zero, this kernel generates causal impulse responses in $\mathbf{H}_\mathrm{a}$. The transfer function between two elementary layers consequently behaves as a low-pass filter:

$$|P(f)|^2 = 1 - 4a(1-a)\sin^2(\pi\Delta_t f). \tag{30}$$

The corresponding attenuation function can be found by

$$\alpha(f) = -\frac{\ln|P(f)|}{\Delta_z}. \tag{31}$$

Such a model has shown satisfactory results in a deconvolution context [7]. However, it seems less accurate in terms of physical reality, because $\alpha(f)$ is more adequately modeled by a frequency power law. In addition, the tuning of parameter a looks somehow arbitrary, whereas frequency-based models can be tuned according to physical models, or even set from a material catalog [47]. An example of radiation transfer functions $|H_\mathrm{r}(f,z)|$ for Olofsson's model is plotted in Fig. 3(b) for $a = 0.2$. As will be shown in Section IV-C, this value yields a model which can be compared with the linear attenuation model with $\alpha_0 = 50$ Np/MHz/m at propagation distance $z = 10$ mm. However, the two attenuation models yield quite different radiation transfer functions for other distances, as the comparison between Figs. 3(a) and 3(b) shows. The performances of frequency-based models and of Olofsson's model are compared for basic experimental data in the following section.

IV. Model Exploitation in the Case of a Single Echo: Estimation of Material Parameters

A. Single Echo Detection by Matched Filtering

We consider data made up of two well-separated echoes in a configuration similar to the one of Figs. 5 and 6. The instrumental response h_e is identified from the front wall echo, from which the matrix $\mathbf{H}_\mathrm{e}$ in (27) is built. The matrix $\mathbf{H}_\mathrm{a}$ depends on the considered attenuation model. Identifying the back wall echo then amounts to selecting the column of matrix $\mathbf{G}$, say $\mathbf{g}_i$, that best fits the data.[2]

[2]In practice, the contribution of the front wall echo is previously removed from the data.

That is, $\mathbf{x}$ should have only one nonzero coefficient at index $\hat{i}$ that corresponds to the spatial position $z_{\hat{i}} = \hat{i}\Delta_z$ of the back surface. Consider the minimization of the least-squares misfit criterion between the data and the model:

$$(\hat{i}, \hat{x}_{\hat{i}}) = \underset{i=0,\ldots,N_x-1, x_i\in\mathbb{R}}{\arg\min} \|\mathbf{y} - \mathbf{g}_i x_i\|^2, \tag{32}$$

which statistically corresponds to the maximum likelihood estimation of the one-column model $\mathbf{y} = \mathbf{g}_i x_i + \mathbf{e}$ under the assumption that the noise samples in $\mathbf{e}$ are zero-mean, independently, identically, and normally distributed [8]. Because this problem is linear in x_i, the best $\hat{x}_i$ can be found for a given $\mathbf{g}_i$ by

$$\hat{x}_i = \frac{\mathbf{g}_i^\mathrm{T}\mathbf{y}}{\|\mathbf{g}_i\|^2}. \tag{33}$$

By inserting (33) into (32), after simple manipulations, the optimal position $\hat{i}$ is

$$\hat{i} = \underset{i=0,\ldots,N_x-1}{\arg\max} \frac{|\mathbf{g}_i^\mathrm{T}\mathbf{y}|}{\|\mathbf{g}_i\|^2}. \tag{34}$$

Finally, the estimated echo is obtained by $\hat{\mathbf{y}} = \mathbf{g}_{\hat{i}}\hat{x}_{\hat{i}}$. Such a procedure follows a matched filtering approach in that it selects the column in $\mathbf{G}$ that yields the maximum correlation with data $\mathbf{y}$—up to a normalization term. However, it is not strictly speaking a matched filter; $\mathbf{G}$ is not a convolution matrix because of the spatially variant nature of the attenuation matrix $\mathbf{H}_\mathrm{a}$.

B. Joint Estimation of the Attenuation Parameter and Sound Velocity

For a frequency power-law attenuation model $\alpha(f) = \alpha_0|f|^\gamma$ with given γ, the previous echo detection procedure enables the joint estimation of the attenuation parameter α_0 and the reference velocity c_0, which are physical quantities of interest for a given material. Indeed, the attenuation transfer function in (11) at distance $z_i = i\Delta_z$ is

$$\begin{aligned} H_\mathrm{a}(f, z_i) &= e^{-[\alpha(f) + j\mathcal{H}(\alpha(f))]i\Delta_z} \\ &= e^{-[|f|^\gamma + j\mathcal{H}(|f|^\gamma)]\alpha_0 c_0 i\Delta_t}. \end{aligned} \tag{35}$$

Hence, for a given γ, the model (11) only depends on the single parameter $\chi_0 = \alpha_0 c_0$. For each χ_0, the velocity c_0 can be deduced from the optimal position $\hat{i}$, found by the matched filter as in (34), and from the thickness of the plate d by

$$c_0 = \frac{2d}{\hat{i}\Delta_t}. \tag{36}$$

The associated attenuation parameter is then given by $\alpha_0 = \chi_0/c_0$. As mentioned in Section II-B, c_0 is not the group velocity but the phase velocity for $f = \infty$. Note that

standard nondestructive evaluation (NDE) methods generally rely on the estimation of the group velocity because dispersion is not considered.

In practice, we can apply the matched filtering procedure for different values of χ_0 in an arbitrarily thin grid, and select the best value $\hat{\chi}_0$ that minimizes the residue between the data and the model $r = \|\mathbf{y} - \hat{\mathbf{y}}\|/\|\mathbf{y}\|$. This value consequently leads to $\hat{c}_0$ and $\hat{\alpha}_0$ that are the best estimations in a least-squares sense of those material quantities for a given propagation model. This procedure can be used for NDE which attempts to estimate the material properties. However, it can only be applied if the front wall and the back wall echoes are well separated.

Note that if the mechanical properties of the material are known, c_0 can be obtained from the analytic formula of the sound speed of longitudinal waves [1]:

$$c_0 = \sqrt{\frac{E(1-\nu)}{\rho(1+\nu)(1-2\nu)}}, \tag{37}$$

where E is the Young's modulus, ν is the Poisson's ratio, and ρ is the density.

C. Application to Polycarbonate

We apply this method to estimate the attenuation and velocity parameters of a homogeneous material. The same measurement configuration as in Section II-D is designed with a 2.25-MHz center frequency transducer used in pulse–echo mode. We use a clear polycarbonate plate of thickness $e = 10.2$ mm, known to have linear attenuation ($\gamma = 1$) and to be highly attenuative [5], [47]. The thickness is measured with a digital caliper, with precision ± 0.1 mm, accounting for both instrument imprecision and irregularities of the plate.

We compare the estimations obtained with the following propagation models:

1) without attenuation,
2) with linear attenuation and linear phase,
3) with linear attenuation and dispersive phase, and
4) Olofsson's model.

For Olofsson's model, we select parameter a in (29) that best fits the echo. The corresponding matched filtering procedure also yields an estimation of c_0—see (36)—but it does not provide any estimation of α_0. The residue values of the estimated back wall echoes are displayed in Figs. 9(a) and 9(b), as a function of χ_0 for linear attenuation models and of a for Olofsson's model, respectively.

One can clearly see that the dispersive model produces the lowest residue ($r \sim 0.08$). The linear phase model and Olofsson's model lead to approximately the same residue, $r \sim 0.27$ and $r \sim 0.25$, respectively. The estimated waveform without attenuation shows the greatest discrepancy with the data ($r \sim 0.45$). Because of the poor adequacy of the model, the matched filter returns a positive amplitude $\hat{x}_i$ whereas a negative value is expected. Indeed the reflection coefficient between polycarbonate and water is negative [46]. As a consequence, a better estimated waveform should be left-shifted by half the wave cycle. The corresponding values of $\hat{c}_0$ and $\hat{\alpha}_0$ are listed in Table I. The attenuation parameter given by the dispersive model is 54.1 Np/MHz/m, and that obtained from the linear phase model is 52.1 Np/MHz/m. Both are in the range of the values reported in the literature (for example, 50 to 57 Np/MHz/m in [47]). We consider the value given by the dispersive model as more likely as this model yields the lowest estimation residual. The optimal attenuation parameter of Olofsson's model is 0.210, but it cannot be linked to any reference value.

As expected, the linear phase model that considers a constant phase velocity $c(f)$ leads to $\hat{c}_0 = 2274$ m/s, very close to the reference group velocity given in [47]: 2270 m/s. The value returned by the dispersive model, 2380 m/s, is logically larger than the group velocity, because it corresponds to the phase velocity at infinite frequency (see Section II-B). Olofsson's model leads to 2749 m/s, which is difficult to interpret. The inaccurate modeling obtained with the nonattenuated waveform leads to an estimated velocity of 2329 m/s, whereas it should return the same value as the linear phase model because both models consider constant phase velocity.

In this study, the dispersive model gives accurate results for polycarbonate, for which frequency-dependent attenuation can be well represented by a linear model. Such an approach could also be applied to nonlinear attenuation models with given γ, e.g., for synthetic rubber and castor oil [40].

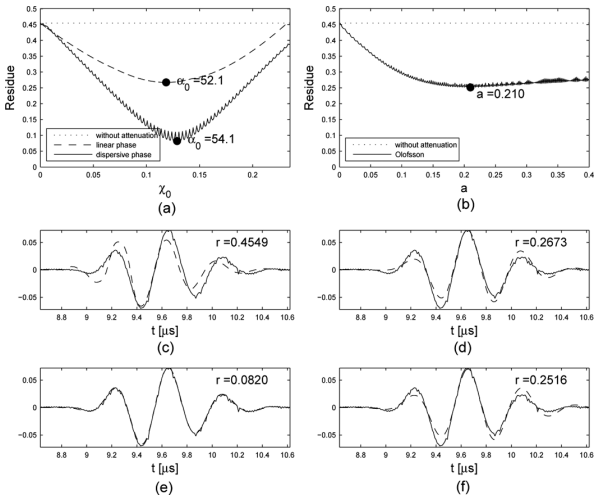


Fig. 9. Matched-filtering-based back wall echo estimation from a polycarbonate plate of thickness $d = 10.2$ mm using a 2.25-MHz transducer. (a) Residue of estimation as a function of the parameter χ_0, for several models of linear attenuation (without attenuation, linear phase, dispersive phase). (b) Residue of estimation as a function of the parameter a for Olofsson's model. (c)–(f) Data $\mathbf{y}$ (solid line) and estimation $\mathbf{y}$ (dashed line) with the optimal attenuation parameters, (c) without attenuation, (d) with linear phase, (e) with dispersive phase, and (f) with Olofsson's model (r is the residue).

TABLE I. ESTIMATED $\hat{\chi}_0$, TIMES OF FLIGHT, REFERENCE VELOCITIES, AND ATTENUATION COEFFICIENTS FOR A 10.2-MM-THICK PLATE OF POLYCARBONATE USING A 2.25-MHZ PROBE.

	$\hat{\chi}_0$	$\hat{i}\Delta_t$ (µs)	$\hat{c}_0$ (m/s)	$\hat{\alpha}_0$ (Np/MHz/m)
1 Without attenuation		8.76	2329	
2 Linear phase	0.118	8.97	2274	52.1
3 Dispersive phase	0.129	8.57	2380	54.1
4 Olofsson's model		7.42	2749	

For models 2 and 3, $\gamma = 1$.

V. APPLICATION TO THE DECONVOLUTION OF NDT DATA: ESTIMATION OF THE THICKNESS OF A PLATE

In this section, an inverse problem of spike train deconvolution is considered. From a pulse–echo measurement, our goal is to estimate the thickness of the polycarbonate plate that was already used in Section IV-C. More precisely, we consider the problem of estimating a spike train $\mathbf{x}$ from data $\mathbf{y} = \mathbf{G}\mathbf{x} + \mathbf{e}$, where the locations of the spikes—the nonzero elements in vector $\mathbf{x}$—correspond to the positions of the echoes. The distance between consecutive spikes then corresponds to twice the plate thickness.

Such a framework can be applied to a large variety of practical NDT problems [1] to measure the thickness of a layer or a wall. Applications occur for the manufacture of pipes, plates, strips, etc., and for the control of walls in severe environments (power plants, chemical industry), when the back area is out of reach. In such cases, a pulse–echo acquisition may be an appropriate solution. Note that, in the presented example, the echoes are well separated; hence the deconvolution problem is not complex from an informational point of view. In the case of thinner layers, the reflected echoes overlap and advanced processing methods such as deconvolution are appropriate tools to estimate the positions of the echoes.

Here, we use a transducer with 5 MHz center-frequency. Compared with the 2.25-MHz transducer used in the previous experiment, better resolution is expected but the attenuation effect is stronger. The back wall echo is strongly distorted and highly attenuated in amplitude, with a very low signal-to-noise ratio [see Figs. 10(a) and 10(b)]. A direct time-of-flight identification by visual inspection might lead to inaccurate thickness estimation because of the low signal-to-noise ratio and of the phase shift.

Deconvolution is performed through the minimization of the least-squares data misfit function, penalized by the ℓ_1-norm $||\mathbf{x}||_1 = \sum_i |x_i|$:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{G}\mathbf{x}\|^2 + \lambda \|\mathbf{x}\|_1. \tag{38}$$

Such a sparsity-inducing penalization produces few nonzero elements in the solution for appropriate λ. Criterion (38) is a convex function and admits a unique minimum for a given λ. The regularization parameter λ introduces a trade-off between the least-squares fit and the penalization. The value $\lambda = 0$ corresponds to the least-squares solution, which is not acceptable because of the ill-posedness of the problem [8]. Then, as λ increases, the number of spikes in $\mathbf{x}$ tends to decrease. This kind of formulation has attracted a great interest in inverse problems in the past decades [50], and particularly in ultrasonic deconvolution problems [6], [8]. Optimization is performed using the homotopy continuation method described in [51]. In this part, λ is set manually to retrieve one spike in the area of 10 mm.

We consider deconvolution using the attenuation-free model and the model with attenuation and dispersion. The first model is the approach commonly adopted in deconvolution problems. The matrix $\mathbf{G}$ is equal to $\mathbf{H}_\mathrm{e}$, which is built from the front wall echo shown in Fig. 10(a). The second model includes $\mathbf{G} = \mathbf{H}_\mathrm{e}\mathbf{H}_\mathrm{a}$ where $\mathbf{H}_\mathrm{a}$ is built from the dispersive model (11). Its parameters are set from the experimental results in Table I. We have seen in Section IV-C that the estimation of c_0 for the model without attenuation was erroneous. Hence, for this model, we set c_0 to the most plausible estimated group velocity; that is, 2274 m/s. Note that the sampling frequency is 100 MHz, leading to the spatial precision $\Delta_z = c_0\Delta_t = 0.0274$ mm and 0.0238 mm for the two models, respectively. Accord-

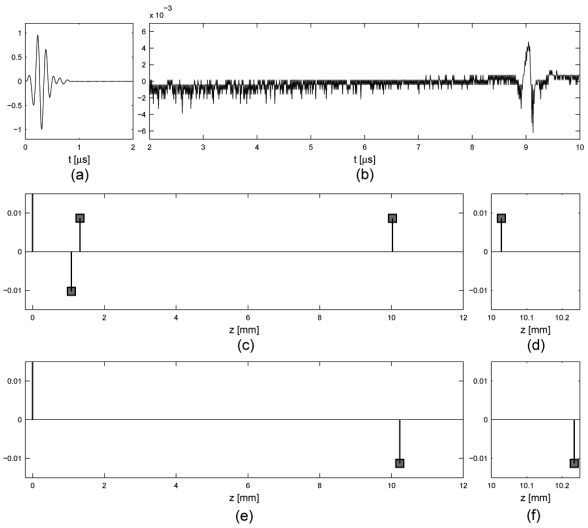


Fig. 10. Deconvolution of data acquired from a polycarbonate plate of thickness $d = 10.2$ mm with a 5-MHz transducer. (a) Data for $t \in [0, 2]$ µs corresponding to the front wall echo, (b) data for $t \in [2, 10]$ µs where a zoom in amplitude is performed. (c), (d) Deconvolution without attenuation model ($c_0 = 2274$ m/s). (e), (f) Deconvolution with linear attenuation and dispersive model ($c_0 = 2380$ m/s, $\gamma = 1$, $\alpha_0 = 54.1$ Np/MHz/m).

ing to [46], the true sequence is composed of a positive spike at 0 mm and of a negative spike at 10.2 mm.

Results are shown in Figs. 10(c)–10(f). Deconvolution with the attenuation-free model estimates four spikes. They correspond to the front wall position, two false detections and the back wall position. The last spike has a positive amplitude, which is inconsistent with the truth (as in Section IV-C). The corresponding thickness estimation is 10.5 mm. As shown in Figs. 10(e) and 10(f), the result using attenuation and dispersion in the model shows two spikes that correspond to the front wall and the back wall positions. They have a positive and a negative sign respectively, which is coherent with the expected object. The estimated distance is 10.23 mm, which is closer to the measured value. Using the model with attenuation and dispersion in criterion (38) achieves more accurate spike detection than the standard convolution model with $\mathbf{G} = \mathbf{H}_{\mathrm{e}}$, and hence leads to more satisfactory thickness estimation.

VI. Discussion and Future Works

We have presented a discrete model of ultrasonic signals that considers attenuation and dispersion. The model is built in the frequency domain, where specific radiation transfer functions are computed for a finite set of distances. We obtain a linear model between the data and the unknown spatial distribution of targets. Compared with the usual formulation, it amounts to introducing an attenuation matrix depending on acoustic propagation parameters. Experimental results from attenuative materials reveal the accuracy of such a formulation for the modeling of the backscattered echoes. Better results are obtained compared with the non-causal model that considers constant phase velocity [3] and to the causal empirical model proposed in [7]. The proposed model also yields the best results on the problem of detecting and locating the back wall echo in a polycarbonate plate. Finally, we consider a spike train deconvolution problem based on ℓ_1-norm regularization to estimate the thickness of a plate. With the proposed model, the back wall echo blurred into noise is successfully located, whereas the solution based on the standard model shows false detections and an imprecise echo location.

Future works could concern the application of the developed model to more complex data, in particular to the deconvolution of A-scans with overlapping echoes. In such cases, the better adequacy of the proposed model should improve the spike detection performance compared with standard deconvolution approaches using the generic, stationary, convolution model. We also expect that algorithms based on such a model can yield better performance than parametric models [23], [24], which also allow some flexibility in the echo shapes but do not integrate any constraint resulting from the ultrasound propagation properties.

Similar models for ultrasonic data could also be developed for two- and three-dimensional acquisitions. In particular, our approach could be extended for modeling SAFT [13], [14] and full matrix capture [15] data in ultrasonic imaging.

We have seen that our model depends on a reference velocity, which corresponds to the phase velocity at infinite frequency. The reference velocity can be obtained from the mechanical properties of the material. We have also shown that it can be estimated from a material evaluation process. In addition, complementary works could link such reference velocity to the group velocity, inspired by Gurumurthy and Arthur [32] who proposed an empirical relation between the two quantities.

Finally, our model considers a single geometrical signature of the acoustical targets, which is particularly appropriate if the targets have the same shape, as for plane surfaces in the presented experiments. Future works could include the diffraction of typical reflectors (flat-bottom holes for instance) by considering different possible signatures. In such an approach, the attenuation matrix would be replaced by a set of matrices, each one characterizing a specific diffraction signature. We believe that such dictionary-based model, coupled with efficient sparsity-aware algorithms [50], may be appropriate to address complex NDT problems for the detection and the characterization of flaws. A similar approach could be used to detect cracks or delaminations in multilayered materials. Indeed, a set of several attenuation matrices can also describe the different paths produced by multiple reflections.

Appendix
Expression of the Dispersive Phase With Linear Attenuation

With linear attenuation $\alpha(f) = \alpha_0|f|$, the phase term (15) reads $H(\alpha(f)) = -(\alpha_0 f_S)/(4\pi^2)\mathcal{J}((2\pi f)/f_S)$, with

$$\mathcal{J}(\omega) = \mathcal{P}\int_{-\pi}^{\pi} |v| \cot\left(\frac{\omega - v}{2}\right) \mathrm{d}v, \quad \omega \in [-\pi, \pi]. \tag{39}$$

Function $J(\omega)$ is odd with $J(\pi) = 0$. Consider $\omega \in]0, \pi[$. With $u = \omega - v$, one has

$$\mathcal{J}(\omega) = \mathcal{P}\int_{\omega-\pi}^{\omega} (\omega - u) \cot\frac{u}{2}\,\mathrm{d}u - \int_{\omega}^{\omega+\pi} (\omega - u) \cot\frac{u}{2}\,\mathrm{d}u, \tag{40}$$

where the Cauchy principal value of the first integral excludes 0 from its domain. Let $F_\omega(u)$ be an antiderivative of $(\omega - u)\cot(u/2)$. An antiderivative of $\cot(u/2)$ is $2\ln|\sin(u/2)|$. Hence, integration by parts yields for any $u \in]-2\pi, 2\pi[$, $u \neq 0$:

$$F_\omega(u) = 2(\omega - u)\ln\left|\sin\frac{u}{2}\right| + 2\int_0^u \ln\left|\sin\frac{\varphi}{2}\right| \mathrm{d}\varphi. \tag{41}$$

The last integral also reads

$$\int_0^u \ln\left|\sin\frac{\varphi}{2}\right| \mathrm{d}\varphi = -\mathrm{Cl}_2(u) - u\ln(2), \tag{42}$$

where $\mathrm{Cl}_2(u) = -\int_0^u \ln|2\sin(\varphi/2)|\mathrm{d}\varphi$ is the Clausen function of order 2, i.e., the imaginary part of the dilogarithm of e^{ju} [52]. From (40)–(42), one can show that

$$\begin{aligned} J(\omega) &= 2\mathrm{Cl}_2(\omega+\pi) + 2\mathrm{Cl}_2(\omega-\pi) - 4\mathrm{Cl}_2(\omega) \\ &= 4(\mathrm{Cl}_2(\omega+\pi) - \mathrm{Cl}_2(\omega)). \end{aligned} \tag{43}$$

Finally, it can be shown that (43) also holds for $w \in [-\pi, 0]$.

Acknowledgments

The authors thank A. Duclos from the University of Maine in Le Mans, France, for the help provided to perform the ultrasonic measurements and for helpful discussions about ultrasonic propagation. We also acknowledge V. Baltazart from IFSTTAR in Nantes, France, for valuable discussions about Q-filtering.

We are also grateful to the reviewers for their critical reading and valuable comments.

References

[1] J. Krautkramer and H. Krautkramer, *Ultrasonic Testing of Materials*. Berlin, Germany: Springer-Verlag, 1990.

[2] S.-K. Sin and C.-H. Chen, "A comparison of deconvolution techniques for the ultrasonic nondestructive evaluation of materials," *IEEE Trans. Image Process.*, vol. 1, pp. 3–10, Jan. 1992.

[3] A. C. Kak and K. A. Dines, "Signal processing of broadband pulsed ultrasound: Measurement of attenuation of soft biological tissues," *IEEE Trans. Biomed. Eng.*, vol. 25, pp. 321–344, Jul. 1978.

[4] M. Fatemi and A. C. Kak, "Ultrasonic B-scan imaging: Theory of image formation and a technique for restoration," *Ultrason. Imaging*, vol. 2, pp. 1–47, Jan. 1980.

[5] P. Droin, G. Berger, and P. Laugier, "Velocity dispersion of acoustic waves in cancellous bone," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 45, pp. 581–592, May 1998.

[6] M. S. O'Brien, A. N. Sinclair, and S. M. Kramer, "Recovery of a sparse spike time series by L1 norm deconvolution," *IEEE Trans. Signal Process.*, vol. 42, pp. 3353–3365, Dec. 1994.

[7] T. Olofsson and T. Stepinski, "Minimum entropy deconvolution of pulse-echo signals acquired from attenuative layered media," *J. Acoust. Soc. Am.*, vol. 109, pp. 2831–2839, Jun. 2001.

[8] J. Idier, *Bayesian Approach to Inverse Problems*, London, UK: ISTE Ltd. and Wiley, 2008.

[9] J. A. Jensen, "Estimation of pulses in ultrasound B-scan images," *IEEE Trans. Med. Imaging*, vol. 10, pp. 164–172, Jun. 1991.

[10] J. Greenleaf, S. Johnson, S. Lee, G. Hermant, and E. Woo, "Algebraic reconstruction of spatial distributions of acoustic absorption within tissue from their two-dimensional acoustic projections," in *Acoustical Holography*, P. Green, Ed., New York, NY: Springer, 1974, pp. 591–603.

[11] G. Glover and J. C. Sharp, "Reconstruction of ultrasound propagation speed distributions in soft tissue: Time-of-flight tomography," *IEEE Trans. Sonics Ultrason.*, vol. 24, pp. 229–234, Jul. 1977.

[12] J. Greenleaf and R. Bahn, "Clinical imaging with transmissive ultrasonic computerized tomography," *IEEE Trans. Biomed. Eng.*, vol. 28, pp. 177–185, Feb. 1981.

[13] M. Karaman, L. Pai-Chi, and M. O'donnell, "Synthetic aperture imaging for small scale systems," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 42, pp. 429–442, May 1995.

[14] F. Lingvall, T. Olofsson, and T. Stepinski, "Synthetic aperture imaging using sources with finite aperture: Deconvolution of the spatial impulse response," *J. Acoust. Soc. Am.*, vol. 114, pp. 225–234, Jul. 2003.

[15] C. Holmes, B. W. Drinkwater, and P. Wilcox, "Post-processing of the full matrix of ultrasonic transmit-receive array data for non-destructive evaluation," *NDT Int.*, vol. 38, pp. 701–711, Dec. 2005.

[16] P. A. Narayana and J. Ophir, "A closed form method for the measurement of attenuation in nonlinearly dispersive media," *Ultrason. Imaging*, vol. 5, pp. 17–21, Jan. 1983.

[17] K. Wear, "The effects of frequency-dependent attenuation and dispersion on sound speed measurements: Applications in human trabecular bone," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 47, pp. 265–273, Jan. 2000.

[18] G. Haiat, F. Padilla, R. Cleveland, and P. Laugier, "Effects of frequency-dependent attenuation and velocity dispersion on in vitro ultrasound velocity measurements in intact human femur specimens," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 53, pp. 39–51, Jan. 2006.

[19] D. Hale, "An inverse Q-filter," Stanford Exploration Project, vol. 26, pp. 231–244, 1981.

[20] J. M. Mendel, *Optimal Seismic Deconvolution: An Estimation Based Approach*. New York, NY: Academic, 1983.

[21] Y. Wang, *Seismic Inverse Q-filtering*. London, UK: ISTE Ltd. and Wiley, 2009.

[22] E. Kjartansson, "Constant Q-wave propagation and attenuation," *J. Geophys. Res.*, vol. 84, pp. 4737–4748, Aug. 1979.

[23] R. Demirli and J. Saniie, "Model-based estimation of ultrasonic echoes. Part I: Analysis and algorithms," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 48, pp. 787–802, May 2001.

[24] Y. Lu and J. Michaels, "Numerical implementation of matching pursuit for the analysis of complex ultrasonic signals," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 55, pp. 173–182, Jan. 2008.

[25] G.-M. Zhang, D. M. Harvey, and D. R. Braden, "Signal denoising and ultrasonic flaw detection via overcomplete and sparse representations," *J. Acoust. Soc. Am.*, vol. 124, pp. 2963–2972, Nov. 2008.

[26] E. Mor, A. Azoulay, and M. Aladjem, "A matching pursuit method for approximating overlapping ultrasonic echoes," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 57, pp. 1996–2004, Sep. 2010.

[27] K. Kaaresen and E. Bolviken, "Blind deconvolution of ultrasonic traces accounting for pulse variance," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 46, pp. 564–573, May 1999.

[28] K. Rasmussen, "Maximum likelihood estimation of the attenuated ultrasound pulse," *IEEE Trans. Signal Process.*, vol. 42, pp. 220–222, Jan. 1994.

[29] M. Fink and J.-F. Cardoso, "Diffraction effects in pulse-echo measurement," *IEEE Trans. Sonics Ultrason.*, vol. 31, pp. 313–329, Jul. 1984.

[30] P. R. Stephanishen, "Transient radiation from pistons in an infinite baffle," *J. Acoust. Soc. Am.*, vol. 49, no. 5, pp. 1629–1638, 1971.

[31] M. O'Donnell, E. T. Jaynes, and J. G. Miller, "General relationships between ultrasonic attenuation and dispersion," *J. Acoust. Soc. Am.*, vol. 63, pp. 1935–1937, Jun. 1978.

[32] K. Gurumurthy and R. Arthur, "A dispersive model for the propagation of ultrasound in soft tissue," *Ultrason. Imaging*, vol. 4, pp. 355–377, Oct. 1982.

[33] C. Lee, M. Lahham, and B. Martin, "Experimental verification of the Kramers–Kronig relationship for acoustic waves," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 37, pp. 286–294, Jul. 1990.

[34] T. Szabo, "Causal theories and data for acoustic attenuation obeying a frequency power law," *J. Acoust. Soc. Am.*, vol. 97, pp. 14–24, Jan. 1995.

[35] R. Kuc, "Generating a minimum-phase digital filter model for the acoustic attenuation of soft tissue," in *Ultrasonics Symp.*, 1983, pp. 794–796.

[36] R. Kuc, "Modeling acoustic attenuation of soft tissue with a minimum-phase filter," *Ultrason. Imaging*, vol. 6, pp. 24–36, Jan. 1984.

[37] C.-H. Chen, W.-L. Hsu, and S.-K. Sin, "A comparison of wavelet deconvolution techniques for ultrasonic NDT," in *Int. Conf. Acoustics, Speech, and Signal Processing*, 1988, vol. 2, pp. 867–870.

[38] A. Lhémery, "Impulse-response method to predict echo-responses from targets of complex geometry. Part I: Theory," *J. Acoust. Soc. Am.*, vol. 90, pp. 2799–2807, Nov. 1991.

[39] F. A. Duck, *Physical Properties of Tissue*. London, UK: Academic Press, 1990.
[40] P. He, "Experimental verification of models for determining dispersion from attenuation," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 46, pp. 706–714, May 1999.
[41] J. Ophir and P. Jaeger, "Spectral shifts of ultrasonic propagation through media with nonlinear dispersive attenuation," *Ultrason. Imaging*, vol. 4, pp. 282–289, Jul. 1982.
[42] J. F. Kelly, R. J. McGough, and M. M. Meerschaert, "Analytical time-domain Green's functions for power-law media," *J. Acoust. Soc. Am.*, vol. 124, pp. 2861–2872, Nov. 2008.
[43] T. Szabo, "Time domain wave equations for lossy media obeying a frequency power law," *J. Acoust. Soc. Am.*, vol. 96, pp. 491–500, Jul. 1994.
[44] A. Papoulis, *The Fourier Integral and its Applications* (McGraw-Hill electronic sciences series). New York, NY: McGraw-Hill, 1962.
[45] A. Oppenheim and R. Schafer, *Discrete-Time Signal Processing* (Prentice-Hall signal processing series). Upper Saddle River, NJ: Prentice Hall, 1989.
[46] J. Saniie and D. Nagle, "Pattern recognition in the ultrasonic imaging of reverberant multilayered structures," *IEEE Trans. Ultrason. Ferroelectr. Freq. Control*, vol. 36, pp. 80–92, Jan. 1989.
[47] A. Selfridge, "Approximate material properties in isotropic materials," *IEEE Trans. Sonics Ultrason.*, vol. 32, pp. 381–394, May 1985.
[48] G. Golub and C. Van Loan, *Matrix Computations*, vol. 3. Baltimore, MD: Johns Hopkins University Press, 1996.
[49] E. Carcreff, S. Bourguignon, J. Idier, and L. Simon, "Resolution enhancement of ultrasonic signals by up-sampled sparse deconvolution," in *Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing*, 2013, pp. 6511–6515.
[50] J. Tropp and S. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. IEEE*, vol. 98, pp. 948–958, Jun. 2010.
[51] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Homotopy continuation for sparse signal representation," in *IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2005, vol. 5, pp. 733–736.
[52] L. C. Maximon, "The dilogarithm function for complex argument," *Proc. R. Soc. Lond. A*, vol. 459, pp. 2807–2819, Nov. 2003.

Ewen Carcreff was born in Quimperlé, France, in 1984. He received the engineering degree from the ESEO school of engineering, Angers, France, in 2007 and the M.S. degree in acoustics from the University of Maine, Le Mans, France, in 2009. Mr. Carcreff worked for a few years as a Research and Development engineer in France and Japan, in the areas of audio signal processing and nondestructive testing. Since 2011, he has been pursuing a Ph.D. degree at the University of Maine in the IRCCyN laboratory, Nantes, France. His thesis interest focuses on inverse problems applied to ultrasonic nondestructive testing.

Sébastien Bourguignon was born in Dijon, France, in 1977. He received the diploma degree in electrical engineering from École Supérieure d'Électricité, Gif-sur-Yvette, France; the engineer degree from ETSIT, Universidad Politécnica de Madrid, Spain, in 2001; and the Ph.D. degree in signal processing from the University of Toulouse, France, in 2005. From 2002 to 2007, he was with the Institut de Recherche en Astrophysique et Planétologie, Toulouse, France. From 2007 to 2008, he was with IFREMER, the French research institute for exploitation of the sea, Brest, France. Then, he worked as a postdoctoral fellow at the Côte d'Azur Observatory, Nice, France. He is currently an associate professor at École Centrale de Nantes and at the Institut de Recherche en Communications et Cybernétique, Nantes, France.

His research interests include statistical inference and estimation, sparse approximation, optimization and MCMC algorithms, and applications to different kinds of observational data.

Jérome Idier is a Research Director at the Centre National de la Recherche Scientifique (CNRS). He was born in France in 1966. He received the diploma degree in electrical engineering from the Ecole Superieure d'Electricité, Gif-sur-Yvette, France, in 1988; the Ph.D. degree in physics from the Universite de Paris-Sud, Orsay, France, in 1991; and the HDR (Habilitation a diriger des recherches) from the same university in 2001. Since 1991, he has been a full-time researcher at CNRS. He was with the Laboratoire des Signaux et Systemes, Gif-sur-Yvette, France, from 1991 to 2002, and with the Institut de Recherches en Cybernetique de Nantes, France (IRCCyN) since September 2002. Since 2012, he has also been an elected member of the French National Committee of Scientific Research (CoNRS).

His major scientific interest is in statistical approaches to inverse problems for signal and image processing. More specifically, he studies probabilistic modeling, inference, and optimization issues yielded by data processing problems such as denoising, deconvolution, spectral analysis, and reconstruction from projections. The investigated applications are mainly nondestructive testing, astronomical imaging, and biomedical signal processing, and also radar imaging and geophysics. Dr. Idier has been involved in joint research programs with several specialized research centers: Electricite de France, Research and Development Division (EDF); Commissariat a l'Energie Atomique (CEA); Centre National d'Etudes Spatiales (CNES); Office National d'Etudes et de Recherches Aerospatiales (ONERA); Loreal; Thales; and Schlumberger.

Dr. Idier is the coauthor of two international patents. He has edited two books on inverse problems in the field of signal and image processing (published by Hermes Science, ISTE Ltd. and John Wiley & Sons Inc.) and he has authored 35 peer-reviewed papers and more than 110 conference papers.

He has served as an Associate Editor for the *IEEE Transactions on Signal Processing* (2010–2012) and for the *Journal of Electronic Imaging*, co-published by SPIE and IS&T, since 2009.

Laurent Simon was born in France in 1965. He received the Ph.D. degree in acoustics from the Université du Maine, Le Mans, in 1994. He is currently a professor with the Laboratoire d'Acoustique de l'Université du Maine, UMR-CNRS 6613, Le Mans. His research interests mainly concern signal processing for vibration and acoustics, including nonlinear system identification, inverse problems for NDT, and spectral estimation of missing data.

Annexe D

Resolution enhancement of ultrasonic signals by up-sampled sparse deconvolution

E. Carcreff, S. Bourguignon, J. Idier et L. Simon : Resolution enhancement of ultrasonic signals by up-sampled sparse deconvolution. *Proceedings of the IEEE International Conference on Acoustic, Speech and Signal Processing*, Vancouver, Canada, mai 2013.

RESOLUTION ENHANCEMENT OF ULTRASONIC SIGNALS BY UP-SAMPLED SPARSE DECONVOLUTION

Ewen Carcreff[1,2]*, Sébastien Bourguignon*[1]*, Jérôme Idier*[1]*, Laurent Simon*[2]

[1] LUNAM Université, École Centrale de Nantes / IRCCyN UMR CNRS 6597, 1 rue de la Noë, 44321 Nantes, France
[2] LUNAM Université, Université du Maine / LAUM UMR CNRS 6613, avenue Olivier Messiaen, 72085, Le Mans, France

ABSTRACT

This paper deals with the estimation of the arrival times of overlapping ultrasonic echoes. We focus on approaches based on discrete sparse deconvolution. Such methods are limited by the time resolution imposed by the model discretization, which is usually considered at the data sampling rate. In order to get closer to the continuous-time model, we propose to increase the time precision by introducing an up-sampling factor in the discrete model. The problem is then recast as a Multiple Input Single Output (MISO) deconvolution problem. Then, we propose to revisit standard sparse deconvolution algorithms for MISO systems. Specific and efficient algorithmic implementation is derived in such setting. Algorithms are evaluated on synthetic data, showing improvements in robustness toward discretization errors and competitive computational time compared to the standard approaches.

Index Terms— deconvolution, sparse approximation, MISO systems, ultrasonic data.

1. INTRODUCTION

Estimation of arrival times and amplitudes of superimposed echoes from noisy observations arises in many applications such as RADAR, seismic exploration, ultrasonic nondestructive testing (NDT) or medical imaging. In NDT for example, a known waveform is sent through a material, and reflection occurs at each impedance change. The precise estimation of the echo parameters then leads to the localization and the characterization of the geometrical properties (including flaws) of the inspected object. Consider the signal model:

$$y(t) = \sum_i a_i h(t - t_i) + e(t), \tag{1}$$

where $h(t)$ is the *a priori* known waveform and $e(t)$ stands for additive noise. The purpose is then to estimate the parameters t_i and a_i from sampled data $\boldsymbol{y} = [y(nT_S)]_{n=1,\dots,N}$, where T_S is the sampling period. This can be a hard task when the echoes overlap, creating constructive or destructive interferences. It is particularly critical for ultrasonic data, where the waveform $h(t)$ has generally a strongly oscillating shape.

Many approaches aim to identify a_i and t_i in Eq. (1) as continuous parameters. Cross-correlation methods [1] are computationally simple but show poor performance when echoes overlap. The finite rate of innovation theory [2] offers exact reconstruction provided that the sampling rate is high enough, although in a rather different context than ours – in particular, specific *sampling kernels* $h(t)$ are considered. Related subspace-based methods have also been widely used in this context, even though they are better suited to multiple snapshot data in order to yield robust covariance estimates [3]. For ultrasonic NDT, parametric methods were proposed where (a_i, t_i) are jointly estimated together with shape parameters for each echo, by minimizing a least-squares distance [4]. This is hence a nonlinear approach that can be very sensitive to model errors and local minima.

On the other hand, extensive research has been carried out on deconvolution methods. Indeed, Eq. (1) formulates a continuous convolution :

$$y(t) = (h*x)(t)+e(t) = \int_{-\infty}^{+\infty} h(\tau)\, x(t-\tau)\mathrm{d}\tau+e(t), \tag{2}$$

where $x(t)$ is a spike train with time positions t_i and amplitudes a_i. Many deconvolution methods then consider a discretized version of the right-hand term in Eq. (2), which yields a discrete and linear inverse problem $\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{e}$. Regularization is then addressed by introducing a sparsity constraint on the sequence $\boldsymbol{x}$. By exploiting linearity with appropriate regularization techniques, such approaches have shown satisfactory results in the presence of strong overlapping and noise [5]. However, the time resolution is obviously limited by the discretization precision, that usually corresponds to the data sampling frequency.

In this paper, we show that it is practically possible to increase the time resolution in sparse deconvolution algorithms by an up-sampling approach. This is relevant to estimate times of flight, which are continuous values. We propose to revisit well-known sparse approximation algorithms in this context, based on greedy strategies [6, 7, 8], and on ℓ_0 [9] and ℓ_1-norm penalization [10, 11]. Section 2 establishes the

This work has been partially supported by Région Pays de la Loire as part of the scientific program "Non-Destructive Testing and Evaluation-Pays de la Loire" (ECND-PdL).

discrete up-sampled convolution model, which is recast as a MISO system. Based on the resulting structure of matrix $\mathbf{H}$, Section 3 studies implementation issues of sparse approximation for up-sampled deconvolution and, more generally, for the estimation of sparse inputs in MISO systems. From synthetic data, a comparison between standard and up-sampled deconvolution is conducted in Section 4. Algorithms are also compared in terms of computational efficiency and performance through Monte-Carlo simulations. Conclusions are finally given in Section 5.

2. UP-SAMPLED CONVOLUTION AS A MISO SYSTEM

Consider the continuous-time convolution model (2), where available data is sampled at period T_S: we note $y_n = y(nT_S)$. Up to our knowledge, all the works in the field of deconvolution consider a *discrete convolution* model, that reads:

$$y_n = \sum_{m=0}^{M-1} h_m \, x_{n-m} + e_n. \tag{3}$$

That is, the right-hand term in Eq. (2) is sampled at the data sampling rate : $h_n = h(nT_S)$ and $x_n = x(nT_S)$. Note that the error term e_n should now also include model errors due to inexact discretization. Let column vectors $\boldsymbol{y}$, $\boldsymbol{h}$, $\boldsymbol{x}$ and $\boldsymbol{e}$ collect the samples of y_n, h_n, x_n and e_n, respectively. Eq. (3) then reads $\boldsymbol{y} = \mathbf{H}\boldsymbol{x} + \boldsymbol{e}$ where $\mathbf{H}$ is a convolution matrix, whose n-th line is a delayed version of the reversed sequence $[h_{M-1}, \ldots, h_0]$ with $n-1$ zeros inserted at the beginning. The Toeplitz structure of $\mathbf{H}$ can be exploited to perform efficient computations with Fast Fourier Transform (FFT) algorithms [12]. Note that such definition of $\mathbf{H}$ corresponds to the post-windowing boundary assumption, for which $\boldsymbol{x} = [x_{-M+2}, \ldots, x_0, x_1, \ldots, x_N]^T$, where superscript ^{T} denotes the transposition.

However, in many applications, the data sampling rate is limited and such discretization may not be appropriate. This is particularly true for sparse deconvolution, since the searched sequence is not band-limited. Hence, it may be of interest to consider that h and x in Eq. (2) are discretized at rate T_S/K with K integer. The discrete model becomes:

$$y_n = \sum_{p=0}^{P-1} \overline{h}_p \overline{x}_{nK-p} + e_n \tag{4}$$

with $\overline{h}_p = h(pT_S/K)$, $\overline{x}_p = x(pT_S/K)$ and $P = KM$. Let column vectors $\overline{\boldsymbol{h}}$ and $\overline{\boldsymbol{x}}$ collect the samples $\overline{h}_n$ and $\overline{x}_n$. In matrix form, model (4) reads $\boldsymbol{y} = \overline{\mathbf{H}}\,\overline{\boldsymbol{x}} + \boldsymbol{e}$ where each line of $\overline{\mathbf{H}}$ is formed by the reversed sequence $\left[\overline{h}_{P-1} \ldots \overline{h}_0\right]$, with $(n-1)K$ zeros inserted at the beginning of line n. $\overline{\mathbf{H}}$ is now an $N \times KN$ matrix and is no more Toeplitz. One can show, however, that model (4) also reads as the sum of K discrete convolutions:

$$y_n = \sum_{k=1}^{K} \left(\sum_{m=0}^{M-1} h_m^k \, x_{n-m}^k \right) + e_n, \tag{5}$$

where $\boldsymbol{h}^k$, $k = 1 \ldots K$ are K *sub-waveforms* with sampling period T_S, such that $h_m^k = h\big((k-1)T_S/K + mT_S\big)$. Similarly, $\boldsymbol{x}^k$ are the corresponding sparse sub-sequences with N points. The matrix form hence reads $\boldsymbol{y} = \sum_{k=1}^{K} \mathbf{H}^k \, \boldsymbol{x}^k + \boldsymbol{e}$, where $\mathbf{H}^k$ are Toeplitz *sub-matrices* obtained by taking every K columns of $\overline{\mathbf{H}}$. In other words, it can be seen as a specific MISO system, as illustrated in Figure 1, where the K filters are obtained by sampling the continuous-time impulse response $h(t)$ at period T_S, with K subsample time shifts $(k-1)T_S/K$, $k = 1 \ldots K$. In the following section, we describe the algorithmic implementations of sparse deconvolution for generic MISO systems.

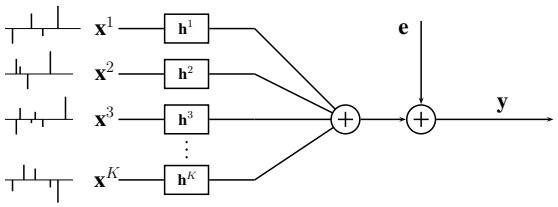


Fig. 1. Diagram of a MISO system.

3. MISO DECONVOLUTION WITH SPARSE APPROXIMATION METHODS

Sparse approximation has become an important field of research in the past fifteen years [11]. It aims at approximating the data $\boldsymbol{y}$ with $\overline{\mathbf{H}}\overline{\boldsymbol{x}}$ where $\overline{\boldsymbol{x}}$ is a sparse sequence[1], that is, $\overline{\boldsymbol{x}}$ has only few non-zero components. We focus on implementation issues for five acknowledged sparse approximation methods applied to MISO deconvolution:

- Three *greedy* algorithms are implemented, namely, by increasing complexity: Matching Pursuit (MP) [6], Orthogonal Matching Pursuit (OMP) [7] and Orthogonal Least Squares (OLS) [8]. Each iteration of such procedures comprises the selection of one component improving the data approximation, and the update of the solution as a combination of the selected components.

- The Single Best Replacement algorithm was recently introduced in [9] and performs local minimization of the penalized least-squares criterion:

$$\left\| \boldsymbol{y} - \overline{\mathbf{H}}\overline{\boldsymbol{x}} \right\|^2 + \mu \|\overline{\boldsymbol{x}}\|_0 \tag{6}$$

  where $\|\overline{\boldsymbol{x}}\|_0$ is the number of non-zero components in $\overline{\boldsymbol{x}}$. This is a combinatorial problem, and local exploration is performed by moves affecting only one component. Each iteration either adds or removes one element in the current support, and the replacement is selected which most decreases criterion (6).

[1] Notation $\overline{\mathbf{H}}\overline{\boldsymbol{x}}$ is used here for homogeneity, although most works on sparse approximation do not consider convolution-based operators.

- Last, ℓ_1-norm penalization is considered by minimizing

$$\left\|\boldsymbol{y} - \overline{\mathbf{H}}\overline{\boldsymbol{x}}\right\|^2 + \mu \sum_{\ell} |\overline{x}_\ell|. \tag{7}$$

  Optimization is performed with the homotopy continuation principle described in [10, 13], which shows formal similarities with greedy methods and even more with SBR since it performs removal moves as well. This algorithm will be referred to as ℓ_1-HC.

The reader is referred to the corresponding references for detailed descriptions of the algorithms. Note that such algorithms were compared for deconvolution purposes in [14], but only within the standard convolution setting of Eq. (3).

The selection steps for MP and OMP mostly amount to computing matrix products $\mathbf{H}^T\cdot$. For MISO systems, using notations of Section 2, such a product is decomposed into K products $\mathbf{H}_k^T\cdot$ with Toeplitz matrices $\mathbf{H}_k$. These are actually cross-correlations, which can be implemented in the frequency domain using two FFTs and one inverse FFT (in dimension N) [12]. In practice, the Discrete Fourier Transforms of all $\boldsymbol{h}_k$ are computed before the algorithm starts. One selection step is then executed by $K+1$ FFTs. Hence, the corresponding cost increases linearly with K. For MP and OMP, the update of the residue is identical to the standard versions. For OMP, it requires the inversion of matrix $\mathbf{H}_\star^T\mathbf{H}_\star$, where subscript $_\star$ indexes the active columns of $\mathbf{H}$. In the proposed implementation, the Cholesky factorization of $\mathbf{H}_\star^T\mathbf{H}_\star$ is updated at each iteration at low cost, since one iteration only performs rank-one modifications to such matrix. Doing so, system inversions amount to two triangular system inversions of complexity $\mathcal{O}(i^2)$ where i is the number of active elements.

Efficient implementations of OLS and SBR require extensive access to elements of the Gram matrix $\mathbf{H}^T\mathbf{H}$ – more precisely, to $\mathbf{H}_\star^T\mathbf{H}_\circ$, where subscript $_\circ$ indexes the non-active columns of $\mathbf{H}$ at a given iteration, see for example implementation details given in [9]. For MISO systems, $\mathbf{H}^T\mathbf{H}$ is composed of blocks $\mathbf{H}_k^T\mathbf{H}_\ell$, that are Toeplitz matrices with elements corresponding to the cross-correlation between $\boldsymbol{h}^k$ and $\boldsymbol{h}^\ell$. Hence, the pre-computation of the $K(K+1)/2$ distinct cross-correlation sequences – also in the Fourier domain – between the K impulse responses gives all useful information about the Gram matrix.

ℓ_1-HC minimizes criterion (7) by gradually decreasing the value of parameter μ [10, 13]. At iteration j, all possible values of μ producing a change in the sign of the current solution are computed, among which the next value $\mu^{(j)}$ is selected as the maximal one that satisfies $\mu^{(j)} < \mu^{(j-1)}$. Such computations are indeed similar to those of previously described greedy methods (see for example [13] for explicit equations). More precisely, addition tests require the computation of two matrix products $\mathbf{H}^T\cdot$, that is, $2(K+1)$ FFTs. Removal tests amount to two system inversions with matrix $\mathbf{H}_\star^T\mathbf{H}_\star$, performed by Cholesky factorization as previously explained.

All algorithms require the pre-computation of the products $\mathbf{H}_k^T\boldsymbol{y}$, which are performed using FFT. To sum up, the complexity of each iteration of MP, OMP and ℓ_1-HC is proportional to K. On the contrary, OLS and SBR require the pre-computation of $(K+1)^2$ FFTs, but their core computations remain roughly constant as K increases.

4. SIMULATION RESULTS

4.1. Deconvolution of a complex signal

We consider the data shown in Figure 2, generated from Eq. (1) with 8 echoes, randomly distributed on the continuous time axis. Consequently, none of them falls exactly on any restoration grid. The waveform is a 5 MHz sine wave with a Gaussian envelope [4]. The data is sampled at 25 MHz and corrupted by 10 dB SNR Gaussian noise. Three overlapping problems occur at approximately 1, 4.5 and 8.5 μs.

Deconvolution is performed with standard algorithms (*i.e.,* $K = 1$) and using up-sampling with $K = 6$. Greedy methods are stopped when the norm of the approximation error $\|\boldsymbol{y} - \mathbf{H}\boldsymbol{x}\|^2$ becomes lower than a given threshold, depending on the noise power (see for example [14]). Similarly, the regularization parameter for SBR and ℓ_1-HC is tuned[2] in order to get solutions with similar approximation errors. All the standard methods fail to correctly locate the echoes on any of the three problems and many spike locations and signs are badly estimated. The up-sampled deconvolution leads to more satisfactory results. In particular, the erroneous behavior of all algorithms at 8.5 μs has been corrected for $K = 6$. OLS and SBR with $K = 6$ also solve the overlapping problem at 1 μs. However, they still fail to correctly locate the two close echoes at 4.5 μs, where estimation results are even slightly worse than with $K = 1$. This is due to the suboptimal nature of the greedy algorithms, that reached a local minimum of the data misfit criterion. The ℓ_1-norm deconvolution achieves correct location of the two close echoes at 4.5 μs and at 8.5 μs. On the other hand, it produces spurious small spikes and double spikes, which are typical artifacts of ℓ_1-norm-based sparse approximation.

4.2. Monte-Carlo simulations

We now compare algorithmic performances with Monte-Carlo simulations. The deconvolution algorithms are run for 2000 synthetic data sets, containing 15 echoes, with the same waveform, SNR and total duration as the data used in Figure 2. The signals therefore contain strongly overlapping echoes. Algorithms are tuned as explained in § 4.1.

Since true spikes do not belong to any of the discrete reconstruction grids, estimation errors are computed using a distance between two spike trains inspired by the work

[2] Note that the ℓ_1-norm introduces bias on amplitude estimation. Thus, for ℓ_1-HC, amplitudes are corrected before computing the approximation error.

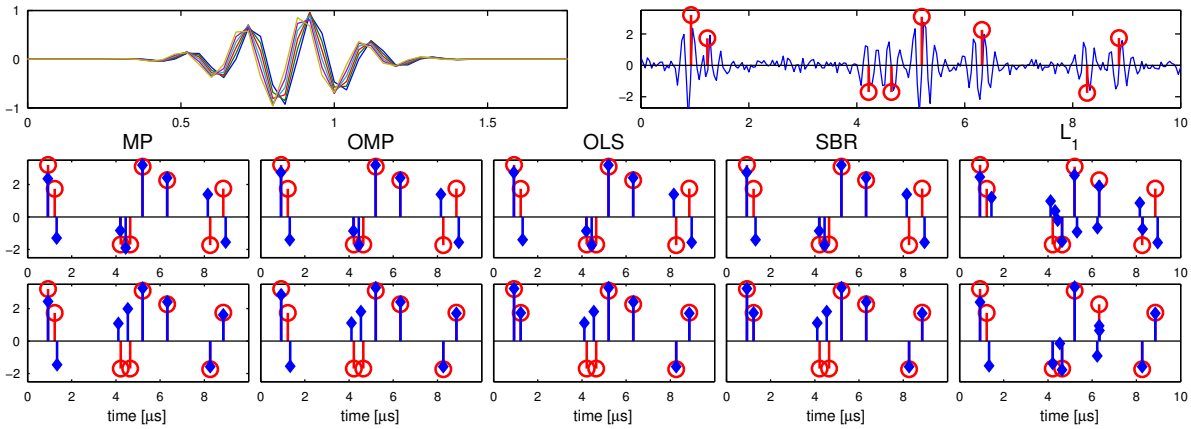


Fig. 2. Deconvolution example from synthetic data. Top left: waveforms used for deconvolution with $K = 6$. Top right: data (–) and true spikes (○). Middle: results with $K = 1$. Bottom: results with $K = 6$. True spikes (○) and estimated spikes (◇).

in [15]: amplitudes are first binarized to ± 1 in order to give the same importance to all detections. Then, the spike trains are convolved with a double-side exponential kernel $e^{-|t|/\tau}$, with $\tau = T_S$, producing slight spike spreading. Finally, the ℓ_2-norm between the two convolved spike trains is computed. Figure 3 (top) shows such estimation errors obtained by the implemented algorithms for different values of the up-sampling factor K. As can be expected, errors decrease with K. For example, using $K = 4$ yields an error reduction of about 25% with respect to $K = 1$, except for ℓ_1-HC. The relative performances of the different algorithms are also in accordance with their complexity, that is, MP has the greater error, followed by OMP, OLS and SBR. The results for ℓ_1-HC appear to be less sensitive to up-sampling, and show the worst performance among all methods for $K \geq 2$. Most of this behavior can be explained by the nature of the spike distance, which strongly penalizes the false detections of small amplitude spikes, inherent to ℓ_1-norm penalization. On the contrary, other simulation showed that ℓ_1-HC yields the smallest errors using a distance without amplitude binarization. Note also that for $K \geq 6$, error reduction becomes negligible. This can be explained by the intrinsic variance on the time delay estimation due to the presence of noise [1].

Central Processing Unit (CPU) times are evaluated with Matlab running on a personal laptop computer with 4 Go RAM and double-Core CPUs clocked at 2.5 GHz. Results are plotted in Figure 3 (bottom). MP and OMP are the fastest and their cost increase linearly with K, which is coherent with the analysis in Section 3. OLS, followed by SBR, are more costly, which is in accordance with their increased complexity. We note that most CPU time required by these algorithms is due to pre-computations, whose cost increase roughly quadratically with K (see Section 3). Note that the cost of ℓ_1-norm deconvolution is the highest one, which is also in accordance with the results in Figures 2 and 3 (top): ℓ_1-HC estimates show more spikes than other algorithms, hence their computations require more iterations.

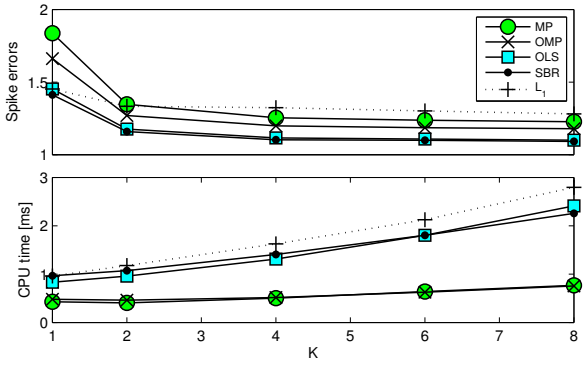


Fig. 3. Estimation errors (top) and computational costs (bottom) for different algorithms versus the up-sampling factor K. Results are averaged on 2000 random realizations.

5. CONCLUSION

An up-sampling approach for sparse deconvolution has been proposed for the estimation of time delays in typical ultrasonic NDT data. A model was introduced based on a finer time discretization of the convolution model than usual approaches. The model was recast has a MISO system, for which well-known sparse approximation methods were studied and computationally optimized for deconvolution. Synthetic simulations revealed the efficiency of up-sampled deconvolution to estimate times of arrival in presence of noise, even for strong overlapping. Computation costs were evaluated that confirmed the implementation efficiency. In particular, increasing the up-sampling factor only produces a reasonable increase of the CPU time.

6. REFERENCES

[1] A. Quazi, "An overview on the time delay estimate in active and passive systems for target localization," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 3, pp. 527–533, June 1981.

[2] M. Vetterli, P. Marziliano, and T. Blu, "Sampling signals with finite rate of innovation," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1417–1428, June 2002.

[3] A. Bruckstein, Tie-Jun Shan, and T. Kailath, "The resolution of overlapping echoes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 6, pp. 1357–1367, December 1985.

[4] R. Demirli and J. Saniie, "Model-based estimation of ultrasonic echoes. Part I: Analysis and algorithms," *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, vol. 48, no. 3, pp. 787–802, May 2001.

[5] J. Idier, *Bayesian Approach to Inverse Problems*, ISTE Ltd and John Wiley & Sons Inc, April 2008.

[6] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[7] Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, November 1993, vol. 1, pp. 40–44.

[8] S. Chen, S.A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of Control*, vol. 50, no. 5, pp. 1873–1896, 1989.

[9] C. Soussen, J. Idier, D. Brie, and J. Duan, "From Bernoulli Gaussian Deconvolution to Sparse Signal Restoration," *IEEE Transactions on Signal Processing*, vol. 59, no. 10, pp. 4572–4584, October 2011.

[10] D. M. Malioutov, M. Cetin, and A. S. Willsky, "Homotopy continuation for sparse signal representation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Philadelphia, USA, March 2005, vol. 5, pp. 733–736.

[11] M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*, Springer, August 2010.

[12] G.H. Golub and C.F. Van Loan, *Matrix Computations*, vol. 3, Johns Hopkins University Press, Baltimore and London, October 1996.

[13] S. Maria and J.-J. Fuchs, "Application of the global matched filter to stap data: an efficient algorithmic approach," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006, vol. 4.

[14] S. Bourguignon, C. Soussen, H. Carfantan, and J. Idier, "Sparse deconvolution: Comparison of statistical and deterministic approaches," in *IEEE Statistical Signal Processing Workshop*, June 2011, pp. 317–320.

[15] M. C. W. Van Rossum, "A novel spike distance," *Neural Computation*, vol. 13, no. 4, pp. 751–763, April 2001.

Annexe E

Restoration of astrophysical spectra with sparsity constraints : Models and algorithms.

S. Bourguignon, D. Mary et É. Slezak : Restoration of astrophysical spectra with sparsity constraints : Models and algorithms. *IEEE J. Sel. Topics Signal Processing, 5(5) : 1002–1013*, septembre 2011.

Restoration of Astrophysical Spectra With Sparsity Constraints: Models and Algorithms

Sébastien Bourguignon, David Mary, and Éric Slezak

***Abstract*—We address the problem of joint signal restoration and parameter estimation in the context of the forthcoming MUSE instrument, which will provide spectroscopic measurements of light emitted by very distant galaxies. Restoration of spectra is formulated as a linear inverse problem, accounting for the instrument response and the noise spectral variability. Estimation is considered in the setting of sparse approximation, where restoration is performed jointly with the detection of relevant patterns in the spectra. To this aim, a dictionary of elementary spectral features is designed according to astrophysical spectroscopy. Sparse estimation is considered through the minimization of a quadratic data misfit criterion with an ℓ^1 -norm penalization, where nonzero components are associated to the detected features. An efficient optimization strategy is proposed, based on the Iterative Coordinate Descent (ICD) principle, with accelerations that dramatically reduce the computational cost. The algorithm does not rely on fast transforms and can be applied to a wide variety of criteria if the sparsity constraint is separable. Results on simulated MUSE-like data reveal satisfactory performance in terms of denoising and detection of physically relevant spectral features. On such data, the proposed algorithm is shown to outperform both state-of-the-art gradient-based and homotopy continuation methods. Simulations with a compressed sensing-like random matrix also reveal better performance compared with usual algorithms, showing that ICD can be a powerful strategy for sparse optimization.**

***Index Terms*—ℓ^1 -norm penalization, deconvolution, denoising, iterative coordinate descent, sparse approximation, sparse optimization.**

I. Introduction

MUSE (Multi-Unit Spectroscopic Explorer) is a second-generation instrument under construction for the European Southern Observatory, which will be installed at the Very Large Telescope in Chile in 2012. It is a very powerful integral field spectrograph [1], which will provide massive hyperspectral data cubes with images of 300×300 pixels, with 0.2 arcsec angular resolution, at up to 4000 wavelengths, covering the visible and the near infrared parts of the electromagnetic spectrum. One of the main science cases of MUSE, which is of interest here, concerns the detection of very distant galaxies and their characterization by their spectra.

Because of their large distance to the observer and of other disturbances, data will be collected in a very noisy environment, with spectrally variable characteristics. In particular, noise is not expected to be identically distributed along the wavelength axis. The Line Spread Function (LSF) of MUSE—the instrument impulse response in the spectral domain—is also spectrally variable, producing more degradations as wavelength increases. In this paper, we consider the restoration of MUSE-like spectra within the setting of inverse problems regularization [2], which provides a robust framework for taking into account the latter observational specificities. Regularization is also an efficient approach for adding prior information on the solution. This is absolutely necessary in our case where, given the very high level of noise contaminating the data, estimation has to be constrained with prior assumptions.

Spectra of galaxies have been studied for a long time [3]. Several spectral components can be identified in the wavelength range covered by MUSE, which depend mainly on the star formation history of each object, on its chemical composition and dust content. In particular, such spectra contain emission and absorption lines and a continuum, which can exhibit a blueward break due to the absorption of photons by intergalactic hydrogen clouds along the line of sight. Hence, we consider that a galaxy spectrum can be modeled as the superposition of these three basic components. Since the characteristics of the features of each component are unknown, we build a catalog containing a high number of possible ones with discretized parameters, some of which will be selected in order to fit the spectrum. In the works that we present next, restoration is consequently combined with the detection of physically relevant parameters.

Such an approach can be viewed as a sparse approximation problem, where only *a few* components are selected in the catalog. Sparsity-based methods have been widely used for signal and image denoising problems in the past 20 years [4]. They rely on the assumption that, for a given class of signals, most information concentrates into a small number of significant coefficients, expressed in some appropriate space. Many contributions initiated by the work of Donoho and Johnstone [5] considered transforms based on multi-scale representations such as wavelets, curvelets, or other XX-let transforms that have been introduced for specific problems [4]. Additionally to the fact that most information in natural signals (and especially, images) can be efficiently represented by a few number of decomposition coefficients—wavelets are good generic *sparsifying* transforms—two "technical" reasons also drove the design of such transforms.

Manuscript received October 17, 2010; revised February 22, 2011; accepted April 11, 2011. Date of publication April 25, 2011; date of current version August 17, 2011. This work was supported in part by ANR project 08-BLAN-0253-01 DAHLIA—Dedicated Algorithms for HyperspectraL Imaging in Astronomy and in part by PPF-ISSO, University of Nice Sophia Antipolis. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jean-Luc Starck.

S. Bourguignon and É. Slezak are with the Cassiopée Laboratory, University of Nice Sophia Antipolis, CNRS, Côte d'Azur Observatory, F-06304 Nice, France (e-mail: sebastien.bourguignon@oca.eu; eric.slezak@oca.eu).

D. Mary is with the Fizeau Laboratory, University of Nice Sophia Antipolis, CNRS, Côte d'Azur Observatory, F-06103 Nice, France (e-mail: david.mary@unice.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2011.2147278

Orthogonality: If the transform is orthogonal, then sparse estimation amounts to thresholding the transform coefficients, under the assumption of white independent and identically distributed (i.i.d.) noise contaminating the data [5]. This provides a simple framework for both computation and analysis of estimator properties [6], [7]. In contrast, numerous recent works explored the use of *redundant* transforms, such as overcomplete families of homogeneous functions [8], [9], or unions of structurally different bases [8], [10]—a concept sometimes called morphological diversity [11]. In these cases, near-orthogonality is also a crucial property in the transform choice, in order to guarantee the ability of algorithms to retrieve the sparsest solution [10], [12]. This property is also a cornerstone of the powerful emergent theory of compressed sensing (CS) [13].

Fast Transforms: if fast algorithms are available, then large data sets can be processed efficiently. In the redundant case especially, where simple thresholding generally does not provide the best sparse estimates, efficient optimization is needed. Many recent convex optimization algorithms in this field rely on intensive computations of gradient-like functionals which exploit fast transforms [14]–[18].

Although the methodology considered in this paper is based on a sparse decomposition of data, fundamental differences exist with such usual setting. Considered data are one-dimensional and with "reasonable" size, so that we can afford the construction of a specific dictionary (with no associated fast operator), which better models sparse prior information than generic transforms. We suppose that galaxy spectra can be synthesized as the superposition of astrophysically meaningful features [3]. Doing so, a physical interpretation can be associated to the active coefficients in the decomposition, indicating for example the *detection* of spectral lines or breaks. In other words, we are interested in both restoring the spectrum and estimating associated parameters by means of the synthesis coefficients, whereas usual denoising mainly focuses on reconstruction in data space. Moreover, sparsity should be better expressed in such an adapted dictionary than in generic ones; hence better denoising performance is expected. This approach has a double price to pay, however. First, dictionary atoms are highly correlated. Hence, theoretical properties about the resulting sparse approximation cannot be obtained. Nevertheless, as shown below, satisfactory results are achieved in practice. A second disadvantage concerns optimization, since no fast transform can be used to compute matrix-vector products. Note that such sparse representation problem with "constrained" dictionary arises in many applications, e.g., sparse linear regression in statistics, source separation, or when the dictionary is learned from the data. Hence, the optimization context of this paper is a rather generic one.

The methodological objectives of this paper are mainly twofold. First, attention is given to precisely modeling both data formation – in particular, noise variability and instrumental characteristics – and prior information. A dictionary of elementary spectral features is built based on prior physical knowledge, so that sparse estimation is combined with the detection of relevant spectral information in the data. Estimation is set under the usual sparsity-promoting ℓ^1-penalization framework [8], [19]. Observational specificities are shown to modify the equivalent dictionary, whose properties are studied.

The second part of our paper concerns sparse optimization when no fast transform can be used. Alternatives to gradient-based methods are considered which exploit sparsity, namely Homotopy Continuation (HC) and Iterative Coordinate Descent (ICD). HC or Least Angle Regression [20]–[23] is specifically designed for sparse solutions, where most computations are concentrated towards identifying the support of the solution. ICD [24], [25] performs successive componentwise optimization steps and can be interpreted similarly. Whereas HC is often considered as the most efficient strategy in this context [21], [23], [26], the efficiency of ICD for sparse estimation was recently exhibited, e.g., in [27]–[29]. In particular, Friedman *et al.* [27] showed that ICD can outperform HC in large size problems, and conclude their work by stating that "coordinate-wise descent algorithms deserve more attention in convex optimization." In this paper, we build an ICD-based strategy, with accelerations specifically designed for sparse problems. It can be efficiently applied to any penalized least-squares criterion with separable penalization, provided that coordinatewise optimizations can be performed at low cost and that the solution is sparse. Hence, we also consider a "compressed sensing like" scenario with high-dimensional random matrix, similarly to an example proposed in [17].

The paper is organized as follows. Section II introduces a model for data observation, with variable LSF and non identically distributed noise. In Section III, a prior model is proposed, based on a sparse representation of the data, and a specific dictionary is built. Estimation is formulated in Section IV as an ℓ^1-norm penalized optimization problem. Hyperparameter tuning is addressed and the resulting equivalent dictionary is studied. Section IV ends with the description of a posterior amplitude re-estimation step. Optimization is studied in Section V. The use of HC and ICD strategies is motivated, and improvements on standard ICD are introduced by exploiting the sparsity of the solution. Simulation results are given in Section VI. An application to deconvolution and noise reduction for a MUSE-like simulated spectrum is presented. Then, the behavior and performance of several optimization strategies are compared, revealing the efficiency of the proposed algorithm, both on MUSE-like data and on a CS-like example with random and noisy measurements of a sparse process. Conclusions and directions for further work are finally given in Section VII.

II. Data Formation Model

Let $\boldsymbol{y} = [y_1, \ldots, y_N]^T \in \mathbb{R}^N$ denote a spectrum as observed by MUSE, discretized at equispaced wavelengths $\lambda_1, \ldots, \lambda_N$. The spectral sampling step here is $\Delta\lambda = 0.13$ nm, ranging from $\lambda_1 = 450$ nm to $\lambda_N = 900$ nm. This yields $\mathrm{N} = 3\,463$. We consider the following linear observational model:

$$\boldsymbol{y} = \mathbf{H}\boldsymbol{s} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{s} \in \mathbb{R}^N$ is the spectrum to be restored, $\mathbf{H}$ is the $N \times N$ matrix form of the LSF, and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_N]^T$ is an additive perturbation term. Model (1) supposes that $\boldsymbol{s}$ is reconstructed at the

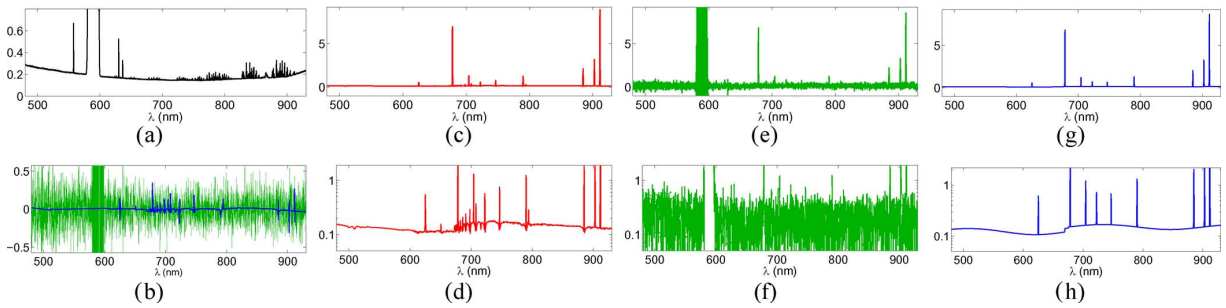


Fig. 1. An example of MUSE-like spectrum and corresponding estimation results. (a) Standard deviation of the noise affecting each wavelength. (c)–(h) Noise-free (red), noisy (green), and restored (blue) spectra at two amplitude scales (bottom panels (d)-(h) are in log scale). (b) Reconstruction error (blue) and original noise (green). Amplitudes are light fluxes, in $\mathrm{erg} \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-2}$ ($\times 10^{-20}$).

resolution of the data $\boldsymbol{y}$. Actually, the true spectrum is a continuous function of wavelength, which may have much higher *frequencies*[1] than the Nyquist limit $1/2\Delta\lambda$, hence model (1) can be viewed as a band-limited approximation of the true reconstruction problem. Restoration at a higher resolution could be addressed by considering a smaller sampling step for $\boldsymbol{s}$. However, this would increase the numerical complexity and the ill-posedness of problem (1)[2, Ch. 2].

A. Variable Line Spread Function

The LSF represents the instrumental spreading affecting every point in the spectrum: each column $\boldsymbol{h}_p = [h_{1,p}, \ldots, h_{N,p}]^T$ of $\mathbf{H}$ is the instrument's impulse response at wavelength λ_n, such that

$$y_n = \sum_{p=1}^{N} h_{n,p} s_p + \epsilon_n.$$

The LSF of MUSE is variable: as wavelength increases, it becomes more spiky and the spectral blurring effect decreases. Consequently, $h_{n,p}$ cannot be written under the typical form h_{n-p}, $\mathbf{H}$ is not a Toeplitz matrix and product $\mathbf{H}\boldsymbol{s}$ cannot be written as a convolution. Note that, at the instrument's spectral resolution, each $\boldsymbol{h}_p$ has a support of 11 points (that is, a spectral extension of 1.43 nm), which is small compared with the size of the spectra, so that matrix $\mathbf{H}$ is very sparse.

B. Non-Identically Distributed Noise

Noise affecting MUSE-like observations is also expected to strongly vary with wavelength, as shown in Fig. 1(a). Three main factors contribute to such variability. First, as in all ground-based astronomical observations, data will suffer from a parasite atmospheric emission. Indeed, some chemical components in the atmosphere emit light at specific wavelengths. In practice, such emission can be supposed spatially constant across the 1 arcmin $\times$ 1 arcmin field of view of the instrument. Hence, it can be estimated from the related 300×300 pixels at each wavelength, and then subtracted from the data. Note that current MUSE simulations suppose perfect estimation of such background emission, whereas estimation residuals will for sure affect observational data. Even so, however, fluctuations remain in the background-subtracted spectrum, which are proportional to the emission level because light emission is a Poisson process: the higher the emission, the higher the associated noise variance. Consequently, in model (1), $\boldsymbol{\epsilon}$ contains a noise source whose power varies with wavelength. In particular, one can see in Fig. 1(a) the signature of powerful parasite emission lines at 528 nm and around 600 nm, which correspond to [OI] emission, and of line packets at higher wavelengths, caused by the presence of water in the atmosphere [3]. MUSE quantum efficiency—the number of produced electrons over the number of photons hitting the detector—is also variable with wavelength, with decreasing performance at each extremity of the spectral range, where consequently the noise influence is stronger. Least, a laser reference star system will be implemented for adaptive optics, generating a very powerful parasite emission in a 20-nm-wide spectral band around the sodium line $\lambda_{\mathrm{Na}} = 589.2$ nm, where no signal can be detected. These three effects are visible in Fig. 1(a), which shows a typical model for the variation of the noise level, jointly to the noise-free and noisy simulated spectra of a moderately bright galaxy in MUSE-like data (panels c)-d) and e)-f), respectively). MUSE data for a deep extragalactic exposure will contain thousands of such spectra, and for the vast majority of them the noise level is expected to be as high, or even higher, as that of Fig. 1.

Note that MUSE is still under construction. Hence, only simulated data generated by the MUSE consortium are available at this time: astronomical scenes are first computed from high-complexity astrophysical simulations, and a mock observation is then generated by applying the MUSE Instrument Numerical Model [30]. In the following, we will suppose that both LSF and noise variance are known at each wavelength.

III. Prior Model With Sparse Representations

A. Motivation for Sparsity Priors

Estimating spectrum $\boldsymbol{s}$ from data $\boldsymbol{y}$ in (1) can be viewed as a denoising and deconvolution inverse problem. Given the high level of noise contamination in the data, only poor results can be reached without any additional assumption on the spectrum. To improve the restoration quality, a typical approach incorporates prior information on the searched solution [2]. Widespread sparsity-based methods can be viewed within this setting. The unknown spectrum $\boldsymbol{s}$ is supposed to have a sparse representation in an appropriate dictionary, say $\boldsymbol{s} = \mathbf{W}\boldsymbol{x}$ with $\boldsymbol{x}$ *sparse*, that

[1] *Frequency* here denotes the inverse (dual) dimension of wavelength.

is, only a few coefficients in $\boldsymbol{x}$ are significant. Since the signal energy is concentrated in a few points, nonzero coefficients in $\boldsymbol{x}$ take higher values than coefficients in $\boldsymbol{s}$, so that coefficients are relatively less affected by noise in the first case. *Denoising* then consists in estimating such sparse vector, say $\widehat{\boldsymbol{x}}$, and restoring the signal by $\widehat{\boldsymbol{s}} = \mathbf{W}\widehat{\boldsymbol{x}}$.

As far as astrophysical spectra are concerned, we consider sparsity in a twofold objective. Denoising and deconvolution is a first one, but we also aim at associating a physical interpretation to the detected atoms. The latter aspect relies on a dictionary design specially adapted to such spectra, which is addressed hereafter.

B. Modeling Galaxies' Spectra: The Dictionary

In most signal and image denoising applications with sparsity assumptions, the dictionary is generally chosen according to the following guidelines:

- the considered data must, of course, contain information that is sparse in some transform domain(s);
- fast transforms must be available to compute products $\mathbf{W}\cdot$ (*synthesis* operator) and $\mathbf{W}^T\cdot$ (*analysis* operator), enabling algorithms to handle efficiently large-size data;
- orthogonality (or near-orthogonality for redundant dictionaries) is another property of interest, which is crucial for theoretical analysis such as uniqueness of the sparse solution and ℓ^0-ℓ^1 equivalence.

In this paper, we adopt a rather different approach, where we prefer using a specific, highly redundant and correlated dictionary, which is more adapted to the morphological features of the considered spectra than generic transforms. We consider that a spectrum is made of three components, each of which is supposed to have a sparse decomposition in an appropriate dictionary: a line spectrum, a step-like spectrum and a continuous spectrum.

1) Complex Line Spectrum: Most relevant information in astrophysical spectroscopy is contained in emission and absorption lines. Each line is characterized by its central wavelength, amplitude and width—we do not consider here the line spreading produced by the LSF, already included in the model (1). Lines can be either *resolved* if their profile spans several points in the wavelength axis, or *unresolved* if the linewidth is smaller than the spectral sampling step. Hence, a dictionary $\mathbf{W}_\ell$ is built, composed of spectral lines with variable widths and central wavelengths. Positive and negative coefficients in the decomposition then characterize, respectively, emission and absorption lines. Unresolved lines are modeled by N delta functions[2] at wavelengths $\lambda_{n,\,n=1\ldots N}$. The shapes of resolved lines are modeled with spline functions of variable size, corresponding to different linewidths. Because resolved lines cannot have arbitrarily large widths, we consider spline widths ranging from 0.39 nm to 18 nm, corresponding to supports varying from 3 to 138 points. In order to limit the dictionary size and to avoid too high correlation between spline atoms, the spacing between adjacent splines with same support size S was set to $S/8$ (up to rounding error). Empirical studies led us to the following selection of parameters, written under the form (support size, translation step) and given in number of points: (3, 1), (5, 1), (9, 1), (11, 1), (17, 2), (25, 3), (35, 4), (49, 6), (69, 9), (97, 12), (138, 17). By doing so, the number of spline atoms is approximately reduced by half compared with the "full" version with unitary translation steps, while the loss in precision is small.

Note that although the rest-frame spectrum of almost all chemical components are well characterized by laboratory measurements, the locations in frequency of these lines in the observed spectrum are unknown, due to the cosmological *redshift* [3] of the data: for an object measured at a redshift $z > 0$ (the larger z, the farther the object), the observed wavelengths λ_{obs} are shifted with respect to the known rest-frame wavelengths λ_{lab} according to the famous equation $1 + z = \lambda_{\mathrm{obs}}/\lambda_{\mathrm{lab}}$. MUSE should be sensitive to galaxies with redshifts as high as 6 or more, that is to objects as far away as 10 billion light-years.

2) Step Spectrum: One major science case of MUSE is the detection of distant galaxies exhibiting a strong discontinuity in their observed spectra [1], [3]. This discontinuity can be modeled by a break in the spectrum. We consider for this purpose the dictionary $\mathbf{W}_s$ made of step functions, which we initially centered at wavelengths $\lambda_{n,\,n=1\ldots N}$. Since the step functions centered at the very first wavelengths possess very high correlations with the "continuous component" (the mean value, which is always nonzero), they may be selected by approximation algorithms although no significant break is present in the data. In order to avoid this effect, only steps centered at wavelengths $\lambda_{n,\,n=50\ldots N}$ were considered.

3) Continuous Spectrum: Several works in the field of sparse representations [8], [10] (see also [31] for a previous work of the authors on astrophysical spectra) have considered a sparse decomposition in the discrete cosine transform (DCT) basis (or, similarly, in the discrete Fourier transform basis) for smoothly varying signals. An N-point signal is implicitly modeled as the sum of a few sinusoids, taken in a dictionary of N atoms with distinct *frequencies* between 0 and the Nyquist limit. In our case, the continuous spectrum is supposed to show very smooth variations, so that high frequencies are unnecessary. On the other hand, more accuracy on the low frequency model is desired. Hence, dictionary $\mathbf{W}_c$ is built by considering all sinusoids with:

- reduced frequencies $f_k = k/N$ with $k = 0 \ldots 8$;
- for $f_k \neq 0$, 8 equispaced phases $\varphi_\ell = \ell\pi/8$, $\ell = 0 \ldots 7$.

The whole dictionary $\mathbf{W} = [\mathbf{W}_\ell \;\; \mathbf{W}_s \;\; \mathbf{W}_c]$ has $N = 3\,463$ lines (number of data) and $M = 26\,015$ columns (number of atoms). Note that $\mathbf{W}$ shows obvious redundancies: a delta function in $\mathbf{W}_c$ is the difference of two adjacent step functions in $\mathbf{W}_s$, and three sine functions with same frequency and different phases are linearly dependent. In both cases, however, the sparsity constraint should remove the ambiguity by favoring a combination with the fewest atoms. Estimation accounting for MUSE LSF and noise statistics is discussed in Section IV. This is shown to lead to an *equivalent dictionary*, whose structure is studied in Section IV-D.

Note that alternatives exist for designing dictionaries adapted to specific data. In particular, approaches based on dictionary learning (e.g., [32], [33]) build a dictionary which sparsifies a

[2]Note that locating unresolved lines with higher precision than that imposed by the instrument would require a high-resolution formulation of problem (1).

given set of training data. The objective of this paper is different, where attention is given in priority to the physical meaning of the dictionary elements. An interesting and more related recent work [34] considers parametric dictionary design, where parameters of elementary components are discretized in order to satisfy a minimal coherence criterion. In our case, we prefer building specific sets of atoms according to precise prior information, even if the resulting dictionary is very coherent.

Of course, our dictionary design only performs an *approximation* of galaxies spectra with simple atoms, and true data do not exactly correspond to a superposition a such atoms. In particular, emission and absorption line profiles are rarely symmetric. Neither does the parametrization of the continuous spectrum with sine functions correspond to physical reality—except for the imposed low-frequency trend. Indeed, such dictionary results from a compromise between sufficiently rich and physically meaningful modeling and limited complexity.

IV. Estimation Setting

A. ℓ^1-Norm Penalization

Let $\mathbf{W}$ denote the dictionary built in Section III-B. One wants to find a sparse approximation of spectrum $\boldsymbol{s}$ using dictionary $\mathbf{W}$ under the observation model (1), that is, a sparse vector $\boldsymbol{x} \in \mathbb{R}^M$ to problem

$$\boldsymbol{y} = \mathbf{H}\mathbf{W}\boldsymbol{x} + \boldsymbol{\epsilon}. \tag{2}$$

In the literature of sparse representations, two well-known approaches coexist. The greedy approach was formalized as Matching Pursuit (MP) by Mallat and Zhang [35]—even if the principle can be traced back to much earlier works such as [36]. MP is an iterative procedure, which removes the most correlated component between the data and the dictionary, and repeats the process on the residual until some stopping condition is met. Although it benefits from a low computational cost, it is known to propagate erroneous atom selections in cases where atoms of $\mathbf{W}$ are too much correlated ([37], see also examples in [8]). Improvements such as Orthogonal Matching Pursuit [38] and Orthogonal Least Squares [39] try to overcome this problem by performing more complex iterations, at the price of a higher computational cost. They remain, however, sensitive to error propagation and are discarded in the rest of this paper, where the dictionary is highly correlated—see Section IV-D.

We consider the usual alternative to greedy approaches, which turns the problem into convex optimization, by defining estimate $\widehat{\boldsymbol{x}}$ as the minimizer of the following criterion [19]:

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \frac{1}{2} \|\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}\|^2 + \gamma \|\boldsymbol{x}\|_1 \text{ with } \gamma > 0$$

which is a compromise between data fidelity and the sparsity measure $\|\boldsymbol{x}\|_1 = \sum_m |x_m|$. Such an approach, named *Basis Pursuit De-Noising* after the work of Chen *et al.* [8], is known to yield a sparse solution for well chosen values of γ [10], [12]. Compared with greedy methods, convexity of the optimization problem brings less sensitivity of the estimate toward high correlations between atoms of $\mathbf{W}$ [8], [23], [40].

Note that other penalizations can be used to enforce sparsity. Strictly convex approximations of the ℓ^1-norm (e.g., [41]) yield a strictly convex criterion, for which uniqueness of the minimizer is ensured and optimization can be tackled by a wide range of algorithms. However, corresponding estimates are not strictly sparse, which is not in accordance with our detection purpose. On the other hand, other sparsity-promoting functions than the ℓ^1-norm are non-convex, so that optimization can be trapped in local minima. Hence, we select the ℓ^1-norm as the limiting case of a convex (but not strictly) penalization function that yields strict sparsity, which is also a key point for the efficiency of the optimization procedure proposed in Section V.

The former generic ℓ^1-based criterion has to be adapted, however, in order to integrate noise statistics and the necessary dictionary normalization. These topics are addressed in Sections IV-B and IV-C, respectively, that yield the final optimization criterion given in (5).

B. Adjustment for Non-i.i.d. Noise: Equivalent Dictionary

Data-misfit term $(1/2)\|\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}\|^2$ can be viewed as the neg-log-likelihood of model (2) under the assumption that perturbations ϵ_n are i.i.d. centered Gaussian [2]. The specific structure of the noise affecting MUSE spectra was described in Section II-B. We suppose in the following that data are contaminated by zero-mean Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_N^2)$ and σ_n^2 is the variance of the noise contaminating the spectrum at wavelength λ_n. Hence, the neg-log-likelihood of model (2) reads (up to an additive constant):

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}\|_{\boldsymbol{\Sigma}}^2 &\triangleq \frac{1}{2}(\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \mathbf{H}\mathbf{W}\boldsymbol{x}) \\ &= \frac{1}{2}\left\|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{y} - \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{W}\boldsymbol{x}\right\|^2\end{aligned}$$

with $\boldsymbol{\Sigma}^{-1/2} = \operatorname{diag}(\sigma_1^{-1}, \ldots, \sigma_N^{-1})$. This expresses the correct data-misfit measurement to be considered in order to account for MUSE noise statistics. Note that non-diagonal noise covariance could also be taken into account similarly. The optimization criterion then reads

$$\frac{1}{2}\|\boldsymbol{z} - \mathbf{A}\boldsymbol{x}\|^2 + \gamma\|\boldsymbol{x}\|_1 \tag{3}$$

where

- $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{y}$ are weighted data;
- $\mathbf{A} = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{W}$ is the weighted and convolved dictionary.

C. Hyperparameter Tuning and Dictionary Normalization

Tuning the weight γ of the ℓ^1-norm penalization term in (3) is a crucial issue. For $\gamma \geq \left\|\mathbf{A}^T\boldsymbol{z}\right\|_\infty$, the minimizer is identically zero [12]. Conversely, for too small γ, the solution may not be sparse. In this section, a statistical interpretation of γ is given and the need for dictionary normalization is evidenced, which yields a practical rule for hyperparameter tuning.

Consider a slightly more general penalization function of the form $\sum_m \gamma_m|x_m|$. Karush–Kuhn–Tucker (KKT) conditions stipulate [19] that $\widehat{\boldsymbol{x}}$ minimizes $\|\boldsymbol{z} - \mathbf{A}\boldsymbol{x}\|^2/2 + \sum_m \gamma_m|x_m|$ if and only if

$$\begin{cases} \text{for } \widehat{x}_m = 0: & |\widehat{e}_m| < \gamma_m \\ \text{for } \widehat{x}_m \neq 0: & \widehat{e}_m = \gamma_m \operatorname{sign}(\widehat{x}_m) \end{cases} \tag{4}$$

where $\widehat{e}_m$ is the mth component of projected residual $\widehat{\boldsymbol{e}} = \mathbf{A}^T(\boldsymbol{z} - \mathbf{A}\widehat{\boldsymbol{x}})$ and the sign function equals 1, -1 and 0 for positive, negative and zero arguments, respectively. Thus, γ_m can be viewed as a threshold on $\widehat{e}_m$ under which the mth component is not detected.

Suppose that data contain only noise. One then wants $\widehat{\boldsymbol{x}} = \mathbf{0}$ so that residual $\boldsymbol{z} - \mathbf{A}\widehat{\boldsymbol{x}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}_N)$ and $\widehat{e}_m \sim \mathcal{N}(0, ||\boldsymbol{a}_m||)$, where $\boldsymbol{a}_m$ is the mth column of $\mathbf{A}$. From (4), a detection test with false alarm rate τ_{FA} is achieved by choosing the threshold on $\widehat{e}_m$ at $\gamma_m = q\,||\boldsymbol{a}_m||$, so that $\tau_{\text{FA}} = Pr(\widehat{x}_m \neq 0) = 1 - \text{erf}(q/\sqrt{2})$, where erf is the Gaussian error function. That is, the weight γ_m on $|x_m|$ should be proportional to $||\boldsymbol{a}_m||$. Equivalently, $\mathbf{A}$ should have normalized columns so that a unique hyperparameter $\gamma_m = \gamma$ yields a uniform false detection rate on each component. A practical illustration of such a dependence between hyperparameter values, dictionary normalization and false alarms can be found in [31].

Let $\mathbf{N}_A$ denote the diagonal matrix with elements $\{||\boldsymbol{a}_m||\}_{m=1\ldots M}$, so that the columns of $\mathbf{B} \triangleq \mathbf{A}\mathbf{N}_A^{-1}$ have unit norm. The problem we want to solve reads finally

$$\begin{aligned} \widehat{\boldsymbol{x}} &= \arg\min_{\boldsymbol{x}} \frac{1}{2}||\boldsymbol{z} - \mathbf{A}\boldsymbol{x}||^2 + q\sum_m ||\boldsymbol{a}_m||\,|x_m| \\ \Leftrightarrow \widehat{\boldsymbol{u}} &= \arg\min_{\boldsymbol{u}} J(\boldsymbol{u}), \text{ where} \\ J(\boldsymbol{u}) &= \frac{1}{2}||\boldsymbol{z} - \mathbf{B}\boldsymbol{u}||^2 + q\sum_m |u_m| \end{aligned} \tag{5}$$

with $\boldsymbol{u} = \mathbf{N}_A\boldsymbol{x}$, that is $u_m = ||\boldsymbol{a}_m||\,x_m$. Typical values of q are 3 or 4 for which one has, respectively, $\tau_{\text{FA}} \simeq 2.7\ 10^{-3}$ and $\tau_{\text{FA}} \simeq 6.3\ 10^{-5}$. The latter formulation (5) is used in the rest of the paper.

D. Equivalent Dictionary Analysis and Solution Uniqueness

Equation (5) defines a classical *Basis Pursuit De-Noising* problem, where relevant atoms are searched in the equivalent dictionary $\mathbf{B} = \boldsymbol{\Sigma}^{-1/2}\mathbf{H}\mathbf{W}\mathbf{N}_A^{-1}$. Such a dictionary is imposed by observational constraints, prior information on the spectra of galaxies, and a desired uniform false alarm rate. It is certainly not designed in order to satisfy usual recovery conditions such as low mutual coherence [10], [12], positive Exact Recovery Coefficient [42] or the Restricted Isometry Property [43], that can be used to prove uniqueness of the minimizer of (5). The Unique Representation Property (URP, [44]) is another sufficient condition for uniqueness, and supposes that any N columns of the dictionary are independent, where N is the number of data. In our case, even small sets of columns in $\mathbf{W}$ are obviously linearly dependent—see Section III-B. Thus, the same columns in $\mathbf{B}$ are also linearly dependent. Hence, $\mathbf{B}$ does not satisfy the URP. Addressing uniqueness in our case thus seems intractable. Consequently, we can only claim that the minimizer is a convex set of solutions, possibly not reduced to a single element, but where all members are equivalent solutions: in the Bayesian maximum *a posteriori* estimation framework [2], all of them maximize the posterior distribution under a Laplacian prior assumption on each coefficient: $p(u_m) \propto \exp(-q|u_m|)$. Let us remark that, in our experiments, all tested optimization algorithms always converged to the same solution, suggesting that uniqueness may be obtained in practice.

Let us note here that in the literature of compressed sensing, the first theoretical results regarding dictionaries with high correlations were, to the best of our knowledge, established only very recently in [45]. These results cannot be used in the present case, however, because they concern a sparse analysis approach (for which $\mathbf{W}^T\boldsymbol{s}$ is sparse) rather than the synthesis approach considered here (for which $\boldsymbol{s} = \mathbf{W}\boldsymbol{x}$ with $\boldsymbol{x}$ sparse).

E. Amplitude Re-Estimation

The ℓ^1-norm penalization is known to introduce bias on the amplitudes of the nonzero components in $\widehat{\boldsymbol{u}}$ [12]. In practice, once such components have been identified by solving (5), their best fit to the data can be obtained in the least-squares sense. Let $\widehat{\boldsymbol{u}}_\star$ collect only the nonzero components in $\widehat{\boldsymbol{u}}$, and $\mathbf{B}_\star$ collect the corresponding columns of $\mathbf{B}$. Amplitude re-estimation is performed by computing

$$\widehat{\boldsymbol{u}}_\star^{\text{reest}} = \arg\min_{\boldsymbol{u}_\star} ||\boldsymbol{z} - \mathbf{B}_\star\boldsymbol{u}_\star||^2 = \left(\mathbf{B}_\star^T\mathbf{B}_\star\right)^{-1}\mathbf{B}_\star^T\boldsymbol{z} \tag{6}$$

where the matrix inversion in the latter equation is properly defined if $\mathbf{B}_\star$ is full-rank, that is, if optimal atoms are linearly independent. Despite the high number of linear dependencies in our dictionary and its high coherence, we expect the sparsity constraint to be strong enough in order to select linearly independent atoms. In practice, given the very low signal-to-noise ratio, no more than a few tens of active atoms are selected, so that in all our experiments, we never had to face non-invertible or ill-conditioned matrices $\mathbf{B}_\star^T\mathbf{B}_\star$. Equation (6) can be efficiently computed by conjugate gradients [16], [17], or by direct matrix inversion if the number of nonzero components is small enough. Note that such re-estimation is applied to the output of the normalized problem (5). Spectral components are then finally retrieved by multiplying the coefficients of $\widehat{\boldsymbol{u}}_\star^{\text{reest}}$ by the corresponding atoms of $\mathbf{W}\mathbf{N}_A^{-1}$.

V. Optimization Without Fast Transforms

The growing interest for ℓ^1-based regularization in signal and image processing gave birth to an abundant literature about optimization of criteria such as (5). In many applications, priority is given to the ability of processing large-size data—in particular, images—and thus to the computational efficiency of the transforms. In practice, dictionaries based on DCT or wavelets are not built explicitly, and optimization efficiency relies on fast transforms to compute products $\mathbf{B}\cdot$ and $\mathbf{B}^T\cdot$ (or $\mathbf{B}^T\mathbf{B}\cdot$). In this category fall all gradient-based methods such as Iterative Thresholding [14], [15], Fast Iterative Shrinkage-Thresholding (FISTA, [18]), Gradient Projection for Sparse Reconstruction (GPSR, [16]) or Sparse Reconstruction by Separable Approximation (SpaRSA, [17]).

The problem tackled in this paper is different, where priority is given to the design of an adapted dictionary. The price to pay is the impossibility to compute high-dimensional operations at low cost, causing the inefficiency of gradient-based methods. Algorithmic solutions exist, though, that are suited to this problem. In particular, homotopy continuation (HC) methods [20], [21], introduced for sparse regression in statistics, do not

rely on the use of fast transforms. HC is known to perform very efficiently, in particular it outperforms usual quadratic programming methods, especially when the solution is highly sparse [21], [22]. Iterative Coordinate Descent (ICD) is another alternative, whose efficiency was recently exhibited for large-scale sparse regression problems [27], [29] and for sparse spectral analysis [28].

A. Support Exploration Algorithms

Both HC and ICD can be viewed as *support exploration* algorithms, because they are particularly efficient at quickly finding the correct support of the solution if it is sparse enough. Identifying the *signed support* of the solution, that is, the location of the nonzero components and their corresponding signs, is the hardest task in ℓ^1 minimization. Indeed, once the signed support is found, amplitude estimation is straightforward. Let $\widehat{\boldsymbol{u}}$ denote a minimizer of (5), let $\widehat{\boldsymbol{u}}_\star$ collect its nonzero components and $\mathbf{B}_\star$ collect the corresponding columns of $\mathbf{B}$. One has $\nabla \|\widehat{\boldsymbol{u}}_\star\|_1 = \operatorname{sign}(\widehat{\boldsymbol{u}}_\star)$ and minimization of (5) in $\widehat{\boldsymbol{u}}_\star$—other components of $\widehat{\boldsymbol{u}}$ are zero—can be written as the gradient cancellation

$$-\mathbf{B}_\star^T(\boldsymbol{z} - \mathbf{B}_\star\widehat{\boldsymbol{u}}_\star) + q\operatorname{sign}(\widehat{\boldsymbol{u}}_\star) = \mathbf{0}$$
$$\Leftrightarrow \widehat{\boldsymbol{u}}_\star = \left(\mathbf{B}_\star^T\mathbf{B}_\star\right)^{-1}\left(\mathbf{B}_\star^T\boldsymbol{z} - q\operatorname{sign}(\widehat{\boldsymbol{u}}_\star)\right) \quad (7)$$

where the matrix inversion in the latter equation is properly defined if $\mathbf{B}_\star$ is full-rank—see the former discussion in Section IV-E.

The efficiency of HC and ICD depends on their ability to quickly identify the correct signed support. Basically, each loop of HC tends to add nonzero components to the iterates starting from the zero vector. Although the principle of ICD is essentially different, it was found that most of the changes in the iterates operated similarly, where only a few changes in the support are performed at each iteration. Consequently, algorithmic efficiency is directly related to the degree of sparsity: the sparser the solution, the faster the support is identified. In that sense, such algorithms can be linked with greedy support exploration techniques, but operate on the convex formulation of the sparse representation problem (see [23] for a detailed study of the parallelism between HC and Orthogonal Matching Pursuit). Hence, HC and ICD fundamentally differ from gradient-based strategies, which do also produce sparse solutions, but do little exploit sparsity for computational efficiency. On the other hand, support exploration algorithms would probably not be appropriate for very large size problems or less sparse solutions.

B. Homotopy Continuation in Practice

We recall the principle of HC, first proposed in [22] for the specific formulation of (5). Let $\widehat{\boldsymbol{u}}(q)$ denote the minimizer of (5) for a given value of q, and let $\overline{q}$ denote the desired value of q. HC is based on the observation that the signed support of $\widehat{\boldsymbol{u}}$ is a piecewise constant function of q. Starting at $q_0 = \left\|\mathbf{B}^T\boldsymbol{u}\right\|_\infty$, for which $\widehat{\boldsymbol{u}}(q_0) = \mathbf{0}$, HC works by decreasing q until a change appears in the signed support of $\widehat{\boldsymbol{u}}$. Indeed, for a given support, the value $q^{(m)}$ at which a change would appear on the sign of the mth component by decreasing q can be computed analytically (see [22] for details). The largest value among all $q^{(m)}$ is selected as q_1 and the signed support is updated, that is, one component switches from zero to ± 1 or from ± 1 to zero. The procedure is repeated and a non-increasing sequence of values $\{q_t\}_{t=1\ldots T}$ is built at which the support of $\widehat{\boldsymbol{u}}(q)$ changes—then, the support is updated. Algorithm stops when $\overline{q} \in [q_T, q_{T-1}]$. The signed support of $\widehat{\boldsymbol{u}}(\overline{q})$ is then identified and amplitudes are obtained by (7).

The most time-consuming part of this procedure is the computation, at each iteration t, of all $q^{(m)}$, which involve M system inversions whose size generally increase with t. Since every change in the support only operates on one component, inversions can be performed by recursively building inverse matrices. Note that HC provides all supports for q varying from q_0 to $\overline{q}$: this could help selecting *a posteriori* the regularization parameter, for example by choosing a fixed number of nonzero components. On the contrary, in our case, $\overline{q}$ is tuned according to the simple statistical rule given in Section IV-C, and fixing a certain number of components in the solution is to be avoided here. Indeed, depending on their position in the field of view, the spectra may be rich in spectral features (bright galaxies), or present only one or a few emission lines (faintest galaxies)—or even, for regions of the sky with no detectable source, contain only noise.

C. Iterative Coordinate Descent and Accelerations

1) Basic Version: ICD consists in performing componentwise minimizations of (5), which have the analytical solution

$$\arg\min_{u_m} J(\boldsymbol{u}) = \phi_q^{\mathrm{st}}\left(\boldsymbol{b}_m^T\left(\boldsymbol{z} - \sum_{p\neq m} u_p\boldsymbol{b}_p\right)\right) \quad (8)$$

where ϕ_q^{st} is the soft-thresholding function [5]:

$$\begin{cases} \phi_q^{\mathrm{st}}(u) = 0, & \text{if } |u| \leq q \\ \phi_q^{\mathrm{st}}(u) = u - \operatorname{sign}(u)q, & \text{if } |u| > q. \end{cases}$$

Basic ICD works by starting at any point in $\mathbb{R}^M$ and then repeatedly updating all coordinates successively, until some stopping rule is met. KKT conditions in (4) provides an explicit characterization of the minimizer that can be used as a strong convergence test, which reads for criterion (5):

$$\begin{cases} \text{for } \widehat{u}_m = 0: & |\boldsymbol{b}_m^T(\boldsymbol{z} - \mathbf{B}\widehat{\boldsymbol{u}})| < q \\ \text{for } \widehat{u}_m \neq 0: & \boldsymbol{b}_m^T(\boldsymbol{z} - \mathbf{B}\widehat{\boldsymbol{u}}) = q\operatorname{sign}(\widehat{u}_m). \end{cases} \quad (9)$$

In the following, we will denote by $M_{\mathrm{KKT}}(\boldsymbol{u})$ the number of coordinates in $\boldsymbol{u}$ that satisfy (9). Convergence is then declared when $M_{\mathrm{KKT}}(\boldsymbol{u}) = M$.

Standard ICD is ensured to converge toward the minimum of (5) [25]. Its efficiency for minimizing ℓ^1-penalized functionals was shown in particular in [27]–[29], and has two main reasons:

- residuals $\boldsymbol{z} - \sum_{p\neq m} u_p\boldsymbol{b}_p$ can be updated recursively and each update (8) can be performed at low cost;
- if $\widehat{\boldsymbol{u}}$ is sparse, then most updates concern zero values, which do not require any computation.

We propose the following improvements to this standard algorithm.

2) Selective Cycling [28], [46]: The most straightforward improvement consists in cycling only through the components that *need* updating. Indeed, each update (8) has a nonzero cost

due to the inner product $\boldsymbol{b}_m^T\cdot$, even if no update is performed. If the solution is sparse, however, then most zero components are quickly identified by ICD and do not need to be updated. Consequently, we consider a cycling rule (which we refer to as NZ cycling) where updates (8) are only applied to the nonzero components of the current iterate. Cycling is also periodically performed through all components, which ensures convergence: including cycling steps on nonzero components only, all of which decrease the value of J, can be viewed as the addition of *spacer steps* [47, Ch. 7] in the standard ICD, and yields a convergent procedure.

3) Support Testing: Recall that, if the signed support of the solution is known, then corresponding amplitudes can be found analytically by (7). Consequently, any signed support can be tested as the optimal one: corresponding amplitudes are computed according to (7), then optimality condition (9) is checked. Implementing support testing in ICD aims at performing a shortcut in last iterations, which generally only work at estimating amplitudes while the correct support has already been identified.

4) Local Tricks: When implementing ICD, two efficiency-limiting factors were identified: 1) some zero components are sometimes reached slowly; and 2) many successive iterates may contain two highly correlated atoms, whereas only one is active at the optimum. We propose to test the optimality of the support obtained by setting to zero each nonzero coefficient [for 1)], or each coefficient in pairs of active atoms whose correlation exceeds some threshold [for 2)]: we call this step *atom disambiguation*. In both cases, the component is set to zero in the current iterate if the number of optimal coordinates M_{KKT} is increased.

Support testing requires the computation of both (7) and (9). Hence, in practice, it is only performed periodically, and only if the current support has not been already tested before. Similarly, local accelerations are only performed when the iterate is close to convergence. Table I details the implementation of the accelerated ICD algorithm, where:

- $\mathbf{S}$ collects all signed supports for which optimality has already been tested;
- T_{all} controls the period at which all coordinates are swept in one ICD iteration;
- T_{test} controls the period at which support testing and local exploration are performed;
- $M_{\mathrm{KKT}}^{\min}$ defines the minimum number of components satisfying KKT conditions (9) required for performing optimality testing. In the following, it is set to $M-10$;
- $\mu^{\max}$ defines the correlation value between two active atoms above which zeros are tested in the corresponding support. In the following, it is set to $1-10^{-4}$;
- `tol` is a numerical tolerance parameter, used in checking KKT equalities. In the following, it is set to 10^{-4}.

TABLE I
ACCELERATED ICD ALGORITHM FOR ℓ^1-NORM-CONSTRAINED OPTIMIZATION

Initialize $t=0$, $\boldsymbol{u}^{(0)}\in\mathbb{R}^M$ and $\mathbf{S}=\emptyset$. Then: i)

1) Selection of sweep indexes.
   - If $t\equiv 0 \pmod{T_{\mathrm{all}}}$, then set $\mathcal{M}=[1,\dots,M]$;
   - else, set $\mathcal{M}=\mathcal{M}_\star=\{m|u_m^{(t)}\neq 0\}$.
2) ICD sweep. For $m\in\mathcal{M}$, update successively $u_m^{(t)}$ with soft-thresholding:
$$u_m^{(t)} = \phi_q^{\mathrm{st}}\Big(\boldsymbol{b}_m^T\big(\boldsymbol{z} - \sum_{p\in\mathcal{M},\,p<m} u_p^{(t)}\boldsymbol{b}_p \;\dots \;\dots - \sum_{p\in\mathcal{M},\,p>m} u_p^{(t-1)}\boldsymbol{b}_p\big)\Big).$$
3) If $t\equiv 0 \pmod{T_{\mathrm{test}}}$ and $\mathrm{sign}(\boldsymbol{u}^{(t)})\notin\mathbf{S}$, then test optimality of $\mathrm{sign}(\boldsymbol{u}^{(t)})$:
   - with $\star$ indexing elements in $\mathcal{M}_\star$, compute
$$\widehat{\boldsymbol{u}}_\star^{(t)} = \left(\mathbf{B}_\star^T\mathbf{B}_\star\right)^{-1}\left(\mathbf{B}_\star^T\boldsymbol{z} - q\ \mathrm{sign}(\boldsymbol{u}^{(t)})\right)$$
   and form $\widehat{\boldsymbol{u}}^{(t)}$ by $\widehat{\boldsymbol{u}}_\star^{(t)}=\widehat{\boldsymbol{u}}_\star^{(t)}$ and $\widehat{u}_m^{(t)}=0$ for $m\notin\mathcal{M}_\star$;
   - check optimality of $\widehat{\boldsymbol{u}}^{(t)}$ by computing the number $M_{\mathrm{KKT}}(\widehat{\boldsymbol{u}}^{(t)})$ of components satisfying:
$$\begin{cases} \text{for } \widehat{u}_m^{(t)}=0: & |\boldsymbol{b}_m^T(\boldsymbol{z}-\mathbf{B}\widehat{\boldsymbol{u}}^{(t)})|<q \\ \text{for } \widehat{u}_m^{(t)}\neq 0: & \boldsymbol{b}_m^T(\boldsymbol{z}-\mathbf{B}\widehat{\boldsymbol{u}}^{(t)})=q\,\mathrm{sign}(\widehat{u}_m^{(t)}) \end{cases}$$
   where the equality condition is tested up to numerical tolerance given by `tol`;
   - if $M_{\mathrm{KKT}}(\widehat{\boldsymbol{u}}^{(t)})=M$, then stop. Otherwise, add $\mathrm{sign}(\boldsymbol{u}^{(t)})$ to $\mathbf{S}$.
4) If $t\equiv 0 \pmod{T_{\mathrm{test}}}$ and $M_{\mathrm{KKT}}\geq M_{\mathrm{KKT}}^{\min}$, then try local tricks: $\forall m\in\mathcal{M}_\star$ (alternately $\forall m,p\in\mathcal{M}_\star$ such that $|\boldsymbol{b}_m^T\boldsymbol{b}_p|>\mu_{\max}$),
   - set the m^{th} (p^{th}) component to 0: form $\boldsymbol{u}^{\mathrm{test}}$ by $\boldsymbol{u}^{\mathrm{test}}=\boldsymbol{u}^{(t)}$ except $u_m^{\mathrm{test}}=0$ (except $u_p^{\mathrm{test}}=0$);
   - perform optimality test 3) with $\boldsymbol{u}^{\mathrm{test}}$ instead of $\boldsymbol{u}^{(t)}$ and add $\mathrm{sign}(\boldsymbol{u}^{\mathrm{test}})$ to $\mathbf{S}$;
   - if $M_{\mathrm{KKT}}(\boldsymbol{u}^{\mathrm{test}}) > M_{\mathrm{KKT}}(\boldsymbol{u}^{(t)})$, then keep $u_m^{(t)}=0$ ($u_p^{(t)}=0$).
5) If $M_{\mathrm{KKT}}(\widehat{\boldsymbol{u}}^{(t)})=M$, then stop. Otherwise, set $t=t+1$ and go back to i).

VI. Simulation Results

A. Restoration of MUSE-Like Spectra

Estimation results are given for the data in Fig. 1. We recall that such data are the result of complex simulations. In particular, they are not built in accordance with model (2), which is used for reconstruction. Criterion (5) is minimized for $q=4$ (see Section IV-C) and amplitudes are re-estimated according to (6). For spectrum $\boldsymbol{s}$ and corresponding noisy or reconstructed $\widetilde{\boldsymbol{s}}$, we define the signal-to-noise ratio (SNR) and the spectral angle (SA), respectively, by

$$\mathrm{SNR}_{\mathrm{dB}} = 10\log_{10}\frac{||\boldsymbol{s}||^2}{||\boldsymbol{s}-\widetilde{\boldsymbol{s}}||^2}$$
$$\text{and}\quad \mathrm{SA}_{\mathrm{deg}} = \mathrm{acos}\frac{\widetilde{\boldsymbol{s}}^T\boldsymbol{s}}{||\widetilde{\boldsymbol{s}}||\,||\boldsymbol{s}||}$$

where the latter is a common measurement of spectral similarity in hyperspectral imaging [48]. Restoration yields[3] SNR $=$ 12.1 dB and SA $=3.5\,\mathrm{deg}$, whereas for noisy data we have respectively SNR $=3.1$ dB and SA $=26.7\,\mathrm{deg}$. Note that before amplitude re-estimation with (6), we have SNR $=9.2$ dB and

[3]Current MUSE simulations only provide noiseless but already *convolved* spectra: $\boldsymbol{s}$ is thus unknown. The SNRs that would be obtained with respect to $\boldsymbol{s}$ instead of its convolved version, are thus expected to be slightly higher than the SNRs presented here.

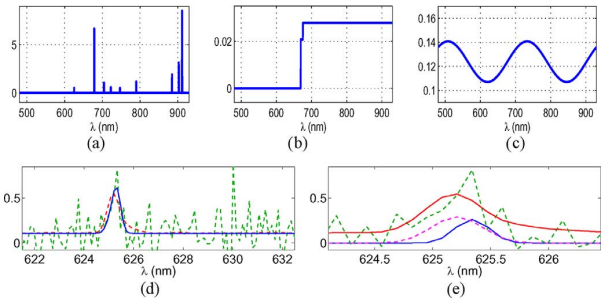


Fig. 2. Decomposition results on the MUSE-like spectrum of Fig. 1. (a)–(c) Estimated spectral components: (a) line spectrum; (b) step spectrum; and (c) continuous spectrum. (d) Zoom on the spectral line at 625 nm, with noise-free (red), noisy (green), and restored (blue) data. (e) Noise-free (red) and noisy (green) data, together with the two selected atoms (magenta dashed and blue solid curves). The noise-free data serve here as ground truth, but are already convolved by the LSF. Amplitudes are light fluxes, in $\mathrm{erg} \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-2}$ ($\times 10^{-20}$).

SA $= 4.8\,\mathrm{deg}$. Right panels in Fig. 1 show the restored spectrum, and associated spectral components are plotted in Fig. 2. The estimated line spectrum correctly locates the main lines. In particular, a faint emission line is detected at 625 nm in a very noisy environment, as shown on the zooms in Fig. 2(d) and (e), where the asymmetric line profile is estimated by two spline atoms with different widths and slightly shifted centers. A break at 673 nm is detected by two close step atoms, and the continuous spectrum is estimated with two components (the mean value and one low frequency oscillation). The estimated continuous spectrum fits the noiseless spectrum quite well, except at the highest wavelengths, where the oscillation produced by the sine atom generates a more important difference with the reference spectrum. However, given the low SNR (see the noisy data in green), this is still a rather satisfactory result. The noise reduction is clearly visible in Fig. 1(b), which also shows undetected faint lines around 700 nm and side effects caused by line profile approximation errors. Such errors, however, remain relatively small compared with the signal amplitude. In the case shown here, the number of synthesis coefficients is 23. This is a favorable case in the sense that for most spectra, the SNR will be lower than shown here—as will also be the number of detectable spectral features.

We note that all selected columns in $\mathbf{B}$ are linearly independent, so that inversion of $\mathbf{B}_\star^T \mathbf{B}_\star$ in (6) is possible. In this example, the Exact Recovery Coefficient (ERC, [42]) equals -2.81; hence, it cannot be used to claim uniqueness of the solution[4]. However, all algorithms that have been tested (see next Section VI-B) converged to the same solution.

B. Optimization Efficiency

In this section, the behavior and performance of several algorithms are compared. Minimization of (5) is performed on the example of Section VI-A by previously described HC and ICD methods and by recent gradient-based techniques: GPSR, SpaRSA, and FISTA. Many algorithms recently appeared in the literature; hence, this comparison is far from exhaustive. However, the former gradient methods were shown to outperform several other algorithms in their corresponding papers, and HC was shown to be much more efficient than standard quadratic programming solvers on several problems of similar (or smaller) sizes than the problem considered here [26], [27].

1) Implementation Details: All algorithms are implemented in Matlab and run on an AMD Opteron 2356 Dual Quad-Core processor under Linux with Central Processing Units (CPUs) clocked at 2.30 GHz. Matrix-vector products are computed only on nonzero components and Matlab's sparse data structure for matrices[5] was used if the computational cost was reduced. Parameter q was set to 4. HC provides in addition the solution paths for higher values of q, but this gain is not considered here.

GPSR and SpaRSA were implemented from the codes available at M. Figueiredo's web page,[6] and only the most favorable results are presented among four tested options (with/without "warm start" and with/without enforcing monotonicity). For FISTA, the gradient step size depends on the spectral norm of matrix $\mathbf{B}$ [15], [18], whose computation is extremely cost consuming in our case. In the following, the maximum step size that yields convergence was determined empirically for each simulation. For HC, both recursive and direct matrix inversions were implemented and only the most favorable results are presented. In general, since the solution is very sparse, direct computations were more efficient.

All algorithms are initialized at $\mathbf{0}$, and are stopped when KKT conditions (4) are satisfied with a numerical tolerance of 10^{-4} for equalities. In all tests that were performed, all algorithms yielded the same support and similar amplitudes up to numerical tolerance.

2) CPU Costs for Astrophysical Data: Fig. 3 plots the number of components with correct sign in the current iterate versus CPU time for all tested algorithms, jointly with the corresponding criterion. For HC, at each change in the support, amplitudes were computed with (7) and criterion (5) was evaluated. The ICD version described in Table I was implemented with $T_{\mathrm{all}} = 250$, $T_{\mathrm{test}} = 500$.

ICD with full cycling is not efficient, because most time is lost by unnecessary cycling along zero components. Gradient-based strategies also perform very poorly. One can see in particular that they are very slow at finding a support close to the true one. On the contrary, HC and ICD with selective cycling are much more efficient. Their performance is shown for two realizations of noise affecting the noise-free data in Fig. 1. In the first one (center row), all versions of ICD with NZ cycling (hereafter, ICD-NZ, that is, with step 1) in Table I) outperform HC. One can see in particular that mostly the half of the computational time of ICD-NZ is dedicated to the estimation of amplitudes—this can be evaluated by comparing solid and dotted vertical blue lines in Fig. 3, center left, respectively indicating convergence and correct support identification (the three dotted lines for the three ICD versions are almost superimposed). Performing support testing (step 3) in Table I) then reduces the cost almost by half. In this example, local tricks [step 4)] do not bring any significant improvement. In the second simulation (Fig. 3

[4]A positive ERC value is actually a sufficient condition for uniqueness.

[5]Indeed, for MUSE-like data, $\mathbf{H}$ is sparse and most atoms in the dictionary are support-limited; hence, $\mathbf{B}$ also has a sparse structure.

[6]http://www.lx.it.pt/~mtf

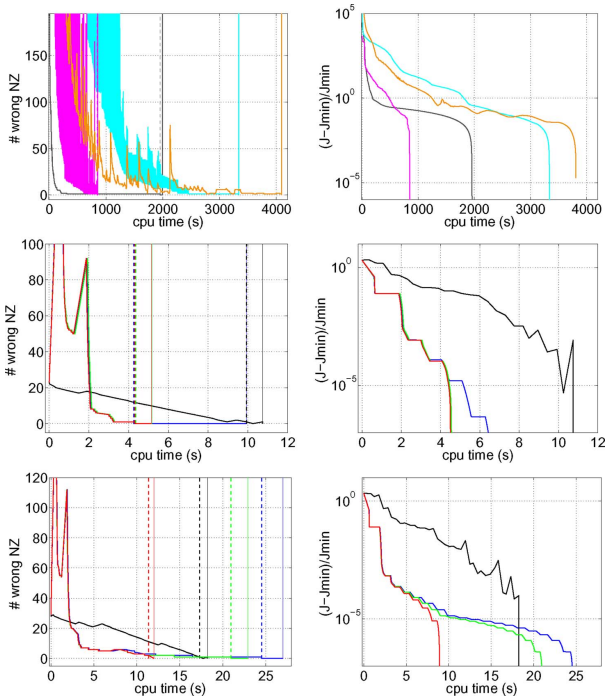


Fig. 3. Number of correctly identified components (left) and criterion value (right) versus CPU time. Top: FISTA (orange), GPSR (cyan), basic ICD (gray), and SpaRSA (magenta). Center and bottom: HC (black), ICD with NZ cycling (blue), ICD with NZ cycling and support testing (green), ICD with NZ cycling, support testing and local tricks (red). Top and center panels correspond to the same data set. Bottom panels correspond to a critical data set for ICD with slow atom disambiguation (see text). Vertical full (resp., dotted) lines on left panels indicate CPU time for convergence (resp., for correct signed support identification).

TABLE II
OPTIMIZATION COSTS FOR DIFFERENT ALGORITHMS ON ASTROPHYSICAL SPECTRA AND CS-LIKE SCENARIOS (IN SECONDS). VALUES IN PARENTHESES INDICATE TIMES FOR CORRECT SUPPORT IDENTIFICATION

	MUSE-like data NZ ≃ 23		CS scenario CS-30		CS-300	
ICD						
full cycling	1992	(1955)	3.23	(1.71)	5.58	(3.54)
NZ cycling	25.0	(11.6)	1.20	(1.15)	2.53	(2.40)
NZ cycling + support testing	11.7	(10.5)	1.35	(1.30)	2.53	(2.40)
NZ cycling + support testing + local tricks	7.2	(6.6)	1.31	(1.26)	2.51	(2.39)
HC	15.5	(15.0)	17.98	(17.52)	459	(457)
FISTA	4090	(4090)	3.69	(3.69)	7.31	(7.31)
GPSR	3340	(3337)	1.78	(1.19)	3.92	(3.08)
SpaRSA	850.9	(849.4)	1.78	(1.15)	2.98	(2.20)

bottom), critical configurations are yielded by ICD-NZ iterations caused by slow disambiguation between correlated atoms (see Section V-C4), so that 28 s are needed for ICD-NZ convergence, whereas HC takes 18 s. Adding a support testing step reduces the cost by 4 s, and local tricks yield an additional gain of 11 s.

Table II (left column) shows the computational costs for all algorithms, averaged on 15 realizations of the noise process. Costs are given for both convergence and correct support identification. Confirming the example in Fig. 3, gradient methods and ICD with full cycling yield excessive costs. Cycling only nonzero components reduces the cost of ICD, but the average cost still exceeds the one of HC. Note in particular that the correct support is obtained, in average, at less that half the total duration of the algorithm. Performing support testing then divides the cost by more than half, and adding local exploration steps still saves approximately 40% of the CPU time. The final version of our ICD algorithm then reduces by most by half the cost of HC.

3) CPU Costs for an Artificial Compressed Sensing Example: Such poor performance of gradient methods are partly due to the absence of fast operators, but this is not the only reason. Algorithms were also run on an artificial example with the same operator size than before ($\mathbf{B}$ is $3463 \times 26\,015$), but whose coefficients were randomly drawn from a Gaussian distribution with unit variance, in a "compressed-sensing" (CS) philosophy—see [17] for a similar example. Contrary to the astrophysical spectra case, the dictionary is almost *incoherent*. Data were generated by applying $\mathbf{B}$ to a sparse vector with randomly chosen locations and amplitudes of nonzero components, drawn from uniform and unit variance Gaussian distributions, respectively. Then, 5-dB white Gaussian noise was added, to reach a similar SNR to that of former MUSE-like data. Two different sparsity levels with 30 and 300 nonzero components were generated, that we denote CS-30 and CS-300. Here again, all algorithms always converged to the same solution. ICD versions were implemented with $T_{\text{all}} = 100$ and $T_{\text{test}} = 20$ for CS-30, and $T_{\text{all}} = 10$, $T_{\text{test}} = 5$ for CS-300.

CPU times averaged on 15 random realizations are given in Table II, center and right columns. All algorithms except HC run much more quickly on CS problems, especially ICD with full cycling and gradient methods, among which SpaRSA yields the lowest CPU times. We explain such differences between the two problems by the very different structures of the two dictionaries. Whereas atoms in the astrophysical dictionary are highly correlated, randomly drawn atoms are almost orthogonal. Consequently, descent directions in the CS case are much better separated, and much deeper descent steps are performed by both gradient methods and ICD. Indeed, the spectral norm of the dictionary (the square root of the maximum eigenvalue of $\mathbf{B}^{\text{t}}\mathbf{B}$, which equals 1 for orthonormal matrices) equals 2411 for the normalized astrophysical dictionary and 14 in the CS case. Recall that the maximum step size ensuring convergence of gradient methods like FISTA is inversely proportional to the spectral norm of the dictionary [15], [18].

By construction, the cost of HC is globally proportional to the number of nonzero components in the solution, and does not strongly depend on the dictionary. In both CS examples, HC is not competitive with ICD and gradient strategies, and its cost becomes prohibitive for CS-300. We note that the cost of all algorithms increases when switching from CS-30 to CS-300 data. This is a logical result for ICD because a larger support is searched (in particular, NZ cycling is performed on more components) and also for gradient-based algorithms, because products $\mathbf{B}\boldsymbol{u}$ operate in higher dimension.

ICD with NZ cycling is still the most efficient algorithm among all tested methods, but additional support testing and local exploration steps do not improve efficiency. This is an expected result because such tests are designed for coherent dictionaries, and their contribution in this case does not compensate for their additional cost. We also note that, on both CS examples, SpaRSA is the most efficient strategy for support identification. This suggests that gradient-based methods could also benefit from support testing steps.

VII. Conclusion and Further Work

Restoration of astrophysical spectra was addressed as a sparse approximation problem. A data formation model was constructed, which accounts for observational constraints. A specific dictionary was designed in accordance with astrophysical spectroscopy, where each atom corresponds to an elementary spectral feature. This allowed us to combine denoising and deconvolution with the detection of physically relevant features, such as emission or absorption lines and discontinuities. Sparse estimation was considered through the minimization of an ℓ^1-norm based criterion, where specificities were shown to require normalization of the equivalent dictionary. Results on an artificial spectrum extracted from MUSE simulations were presented, where detections of lines and of a discontinuity were achieved; however, restoration quality of the continuous spectrum was limited by the too constraining oscillating atoms present in the dictionary. Optimization was studied, and an algorithm based on the ICD scheme was proposed, with accelerations that exploit the sparsity of the solution, substantially reducing the computational cost. The procedure was shown to outperform other recent algorithms for ℓ^1-penalized optimization with our designed dictionary, and also on simulations with random matrices. Consequently, ICD-based methods can be considered as a powerful alternative to gradient-based optimization techniques, especially in cases where no fast transform algorithm can be implemented.

In terms of data modeling, refinements in the design of the dictionary should be investigated. In particular, the parametrization of the continuous spectrum is not fully satisfactory. In continuity with our sparsity-based approach, using a dictionary with low-frequency splines is currently under study. Other parametric models such as polynomials, or non-parametric models based on a Markovian smoothness-promoting penalization such as in [49] are other possibilities. Both require, however, the tuning of additional parameters compared with the "full-sparse" approach. Smoother breaks could also be added in the dictionary, e.g., with several values of a discretized slope parameter.

A crucial extension of this work for MUSE data concerns the three-dimensional restoration of hyperspectral cubes. Formulating a three-dimensional problem, accounting for the instrument spatial point spread function, would perform joint spatial and spectral deconvolution and would certainly improve the restoration quality. Specific problems include the characterization of spatially extended sources (where several pixels have proportional spectra) and spectral *unmixing* [48]. Indeed, in the case of overlapping objects, the spectrum in a given pixel is a mixture of the corresponding spectra. By processing all pixels independently, the method presented in this paper does not allow to separate such spectral components. The main issue for addressing such problems obviously concerns the related computational complexity. The implicit dimension reduction operated by a sparse decomposition in the spectral domain is consequently a key point that should be exploited in order to efficiently tackle three-dimensional problems.

From the algorithmic point of view, automatic rules for fixing the parameters of the accelerated ICD procedure should be studied, based for example on the size of the problem and the expected degree of sparsity of solutions. We have shown that the structure of the dictionary deeply impacts the behavior of algorithms, especially for gradient methods. Precise characterizations of such a dependence deserve more attention. Including acceleration steps as performed with ICD could also be studied for other optimization methods. More generally, the merging of different algorithmic structures with complementary properties should be investigated for high-dimensional and complex sparse optimization problems.

References

[1] R. Bacon *et al.*, "Probing unexplored territories with MUSE: A second generation instrument for the VLT," in *Proc. SPIE*, Jul. 2006, vol. 6269, Ground-based and Airborne Instrumentation for Astronomy.

[2] *Bayesian Approach to Inverse Problems*, J. Idier, Ed.. New York: STE and Wiley.

[3] J. Tennyson, *Astronomical Spectroscopy*. London, U.K.: Imperial College Press, 2005.

[4] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*. New York: Academic, 2008.

[5] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, vol. 81, no. 3, pp. 425–455, 1994.

[6] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors," *IEEE Trans. Inf. Theory*, vol. 45, no. 3, pp. 909–919, Apr. 1999.

[7] A. Antoniadis and J. Fan, "Regularization of wavelet approximations," *J. Amer. Statist. Assoc.*, vol. 96, pp. 939–967, 2001.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[9] S. Sardy, A. G. Bruce, and P. Tseng, "Block coordinate relaxation methods for nonparametric wavelet denoising," *J. Comput. Graph. Statist.*, vol. 9, pp. 361–379, 2000.

[10] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov. 2001.

[11] J.-L. Starck, M. Elad, and D. L. Donoho, "Redundant multiscale transforms and their application for morphological component analysis," *Adv. Electron. El. Phys.*, vol. 132, pp. 287–348, 2004.

[12] J.-J. Fuchs, "On sparse representations in arbitrary redundant bases," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1341–1344, Jun. 2004.

[13] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[14] M. Figueiredo and R. Nowak, "An EM algorithm for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 12, no. 8, pp. 906–916, Aug. 2003.

[15] I. Daubechies, M. Defrise, and C. De Mol, "An iterative thresholding algorithm for linear inverse problems with a sparsity constraint," *Commun. Pure Appl. Math.*, vol. 57, no. 11, pp. 1413–1457, 2004.

[16] M. A. T. Figueiredo, J. M. Bioucas-Dias, and R. D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. Image Process.*, vol. 16, no. 12, pp. 2980–2991, Dec. 2007.

[17] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, Jul. 2009.

[18] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imaging Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[19] S. Alliney and S. A. Ruzinsky, "An algorithm for the minimization of mixed ℓ^1 and ℓ^2 norms with application to Bayesian estimation," *IEEE Trans. Signal Process.*, vol. 42, no. 3, pp. 618–627, Mar. 1994.

[20] M. R. Osborne, B. Presnell, and B. A. Turlach, "A new approach to variable selection in least squares problems," *IMA J. Numer. Anal.*, vol. 20, no. 3, pp. 389–403, Jul. 2000.

[21] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.
[22] D. M. Malioutov, M. Çetin, and A. S. Willsky, "Homotopy continuation for sparse signal representation," in *Proc. IEEE ICASSP*, 2005, vol. 5, pp. 733–736.
[23] D. L. Donoho and Y. Tsaig, "Fast solution of ℓ_1-norm minimization problems when the solution may be sparse," *IEEE Trans. Inf. Theory*, vol. 54, no. 11, pp. 4789–4812, Nov. 2008.
[24] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *J. Comput. Graph. Stat.*, vol. 7, no. 3, pp. 397–416, Sep. 1998.
[25] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *J. Optim. Theory Appl.*, vol. 109, pp. 475–494, 2001.
[26] M. Friedlander and M. Saunders, "Discussion: The Dantzig selector: Statistical estimation when p is much larger than n, by E. Candès and T. Tao," *Ann. Statist.*, vol. 35, no. 6, pp. 2385–2391, 2007.
[27] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Ann. Appl. Stat.*, vol. 1, no. 2, pp. 302–332, 2007.
[28] S. Bourguignon, H. Carfantan, and J. Idier, "A sparsity-based method for the estimation of spectral lines from irregularly sampled data," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 575–585, Dec. 2007.
[29] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *Ann. Appl. Statist.*, vol. 2, no. 1, pp. 224–244, 2008.
[30] A. Jarno, R. Bacon, P. Ferruit, and A. Pécontal-Rousset, "Numerical simulation of the VLT/MUSE instrument," in *Proc. SPIE*, 2008, vol. 7017, Modeling, Systems Engineering, and Project Management for Astronomy III, pp. 701710–701710–8.
[31] S. Bourguignon, D. Mary, and E. Slezak, "Sparsity-based denoising of hyperspectral astrophysical data with colored noise: Application to the MUSE instrument," in *Proc. IEEE WHISPERS*, Jun. 2010, pp. 1–4, DOI 10.1109/WHISPERS.2010.5594902.
[32] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.*, vol. 37, no. 23, pp. 3311–3332, 1997.
[33] M. Aharon, E. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
[34] M. Yaghoobi, L. Daudet, and M. E. Davies, "Parametric dictionary design for sparse coding," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 3311–3332, Dec. 2009.
[35] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
[36] J. A. Högbom, "Aperture synthesis with a non-regular distribution of interferometer baselines," *Astron. Astrophys. Suppl.*, vol. 15, pp. 417–426, Jun. 1974.
[37] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
[38] G. M. Davis, S. Mallat, and Z. Zhang, "Adaptive time-frequency decompositions," *SPIE J. Opt. Eng.*, vol. 33, no. 7, pp. 2183–2189, Jul. 1994.
[39] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *Int. J. Control*, vol. 50, no. 5, pp. 1873–1896, 1989.
[40] J.-J. Fuchs, "On the application of the global matched filter to DOA estimation with uniform circular arrays," *IEEE Trans. Signal Process*, vol. 49, no. 4, pp. 702–709, Apr. 2001.
[41] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Trans. Image Process.*, vol. 2, no. 3, pp. 296–310, Mar. 1993.
[42] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
[43] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec.. 2005.
[44] I. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A recursive weighted norm minimization algorithm," *IEEE Trans. Signal Process.*, vol. 45, no. 3, pp. 600–616, Mar. 1997.
[45] E. J. Candès, Y. Eldar, D. Needell, and P. Randall, "Compressed sensing with coherent and redundant dictionaries," *Appl. Comput. Harmon. Analysis*, vol. 31, no. 1, pp. 59–73, Jul. 2011.
[46] J. H. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *J. Statist. Software*, vol. 33, no. 1, pp. 1–22, Feb. 2010.
[47] D. G. Luenberger, *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1989.
[48] C.-I. Chang, *Hyperspectral Imaging: Techniques for Spectral Detection and Classification*. New York: Kluwer/Plenum, 2003.
[49] P. Ciuciu, J. Idier, and J.-F. Giovannelli, "Regularized estimation of mixed spectra using a circular Gibbs-Markov model," *IEEE Trans. Signal Process.*, vol. 49, no. 10, pp. 2201–2213, Oct. 2001.

Sébastien Bourguignon was born in Dijon, France, in 1977. He received the diploma degree in electrical engineering from École Supérieure d'Électricité, Gif-sur-Yvette, France, the engineer degree from ETSIT, Universidad Politécnica de Madrid, Spain, in 2001, and the Ph.D. degree in signal processing from the University of Toulouse, France, in 2005.

From 2002 to 2007, he was with the Astrophysics Laboratory of Toulouse-Tarbes, France. From 2007 to 2008, he was with IFREMER, the French research institute for exploitation of the sea. He is currently a Postdoctoral Fellow at the Cassiopée Laboratory, Côte d'Azur Observatory, Nice, France. His research interests include statistical inference and estimation, sparse approximation, optimization and MCMC algorithms, and applications to multi-dimensional observational data.

David Mary received the Ph.D. degree in signal processing from the École Nationale Supérieure des Télécommunications, Paris, France, in 2003.

In 2004, he joined the Aryabhatta Research Institute, Observational Sciences, Nainital, India, and the Astronomisches Rechen Institut, Heidelberg, Germany, in 2006. Since 2007, he has been an Assistant Professor in the Laboratoire Fizeau, University of Nice Sophia Antipolis, Nice, France. His research interests include statistical estimation and detection, approximation theory, and their applications.

Éric Slezak was born in France in 1961. He received the M.S. degree in signal processing from the University of Nice Sophia Antipolis, Nice, France, in 1984 and the Ph.D. degree in physics from the University of Nice Sophia Antipolis in 1988. His thesis research was carried out at the Nice Astronomical Observatory and dealt with the development of algorithms for object detection and classification in wide-field optical images.

He was among the first to introduce multiscale techniques in observational cosmology, developping wavelet-based approaches to analyze images, describe complex objects, compute density probability functions, and perform spectro-imagery from low S/N X-ray data. Doing so, he contributed in quantifying the subclustering, segregation properties and dynamical status of several clusters of galaxies and in detecting large-scale diffuse emission related to their building histories. Since 2006, he has held a Full Astronomer position at the Côte d'Azur Observatory, Nice, France. His research interests include astronomical image segmentation and classification from multi-wavelength data analysis. Since 2009, he has led a four-year IT research project aiming to develop the signal processing methods required to analyze the massive hyperspectral datasets provided by the forthcoming integral field spectrographs in astronomy.