Contributions en imagerie computationnelle et problèmes inverses.

François Orieux

Habilitation à diriger des recherches

université PARIS-SACLAY

Contributions en imagerie computationnelle et problèmes inverses

Habilitation à diriger des recherches de l'Université Paris-Saclay

présentée et soutenue à Orsay, le 20 mars 2024, par

François ORIEUX

Composition du jury

Laure BLANC-FÉRRAUD Directrice de Recherche CNRS, I3S	Rapportrice
Loïc DENIS Professeur des Universités, Université de Saint-Etienne, LHC	Rapporteur
Laurent JACQUES Professeur, Université Catholique de Louvain	Rapporteur
Pierre-Olivier AMBLARD Directeur de Recherche CNRS, GIPSA-Lab	Examinateur
Olivier BERNÉ Directeur de Recherche, IRAP	Examinateur
Michel KIEFFER Professeur des Universités, Université Paris-Saclay, L2S	Examinateur

Habilitation à Diriger des Recherches

Contributions en imagerie computationnelle et problèmes inverses

Université Paris-Saclay, Faculté des Sciences d'Orsay

François Orieux

Soutenue publiquement le
20 mars 2024

Composition du jury

Laure Blanc-Féraud Directrice de Recherche CNRS, I3S	Rapportrice
Loïc Denis Professeur des Universités, Université de Saint-Etienne, LHC	Rapporteur
Laurent Jacques Professeur, Université Catholique de Louvain	Rapporteur
Pierre-Olivier Amblard Directeur de Recherche CNRS, GIPSA-Lab	Examinateur
Olivier Berné Directeur de Recherche, IRAP	Examinateur
Michel Kieffer Professeur des Universités, Université Paris-Saclay, L2S	Examinateur

francois.orieux@universite-paris-saclay.fr
Laboratoire des Signaux et Systèmes
Université Paris-Saclay – CNRS – CentraleSupélec
3 rue Joliot-Curie, 91190 Gif-sur-Yvette, France

Door meten, tot weten*

Heike Kamerlingh Onnes (1853 – 1926)

*De la mesure, vers la connaissance.

Table des matières

Première partie

Parcours professionnel

L'homme–Lion de Stadel, une des plus anciennes sculptures connues, environ 40 000 ans. Ivoire de mammouth. Dagmar Hollmann / Wikimedia Commons. *License : CC BY-SA 4.0.*

Synthèse 1.

1.1. Situation depuis 2014

Je suis actuellement *Maître de Conférence des Universités* en section 61 à l'Université Paris-Saclay.

Enseignement

J'enseigne à la Faculté des Sciences d'Orsay dans le Département de Physique. J'interviens en Licence, Master 1 et Master 2 dans les filières E3A et de physique avec pour l'essentiel des cours en traitement du signal, automatique ainsi que problèmes inverses et programmation.

Recherche

J'effectue ma recherche au Laboratoire des Signaux et Systèmes (L2S – Univ. Paris-Saclay, CNRS et CentraleSupélec), dans le « Groupe Problèmes Inverses » (GPI) du pôle « Signaux et Statistique ». Les sujets que je traite portent sur les problèmes inverses, les approches bayésiennes, l'estimation des hyperparamètres ou encore la quantification des incertitudes. J'applique ces travaux méthodologiques en traitement d'images, l'imagerie hyperspectrale et multispectrale, en microscopie biologique ou en astronomie.

L2S, 3 rue Joliot-Curie, 91190, Gif-sur-Yvette
01 69 85 17 47
orieux@l2s.centralesupelec.fr
pro.orieux.fr

1.2. Parcours professionnel

Depuis 2023. Co-resonsable du Master 2 *Automatique, Traitement du Signal et des Images* (ATSI) avec Hugues Mounnier.

2020 – 2022. Demi-délégation CNRS au LS2N à Nantes dans l'équipe SIMS .

2011 – 2014. *Ingénieur de Recherche CNRS*, titulaire. Membre de l'équipe scientifique pour l'analyse des données du projet Planck sur le fond diffus cosmologique (séparation de sources, auto calibration . . .).

2009 – 2011. *Post-doctorat* dans l'unité d'Analyse d'Image Quantitative de l'Institut Pasteur (18 mois). Postdoctorat au Laboratoire de Physique et d'Étude des Matériaux de l'ESPCI. Recherche sur la microscopie à haute résolution temporelle et spatiale de type SIM.

2006 – 2009. *Moniteur* à l'école Polytechnique dans les modex Image du département TREX. Encadrement de projets en traitement d'image et vision par ordinateur.

Bénéficiaire de la PEDR depuis 2020.

1.3. Formation initiale

2006 – 2009. *Doctorat* [1] en physique, spécialité *traitement du signal*, de l'Université Paris-Sud, réalisé au L2S dans le Groupe Problèmes Inverses (GPI) et soutenue le 16 novembre 2009.

Titre : Inversion bayésienne myope et non supervisée pour l'imagerie sur-résolue : Application à l'instrument SPIRE de l'observatoire spatial Herschel.

2005 – 2006. Ingénieur de l'ESEO (Angers) et M2R STI de l'Université de Rennes 1.

1: Directeur de thèse : Jean-François Giovannelli ; Co-encadrement : Thomas Rodet et Alain Abergel

1.4. Synthèse bibliométrique

IdHal : `francois-orieux`
ORCID : `0000-0001-5638-3416`

La liste complète des publications est disponible avec le lien `cv.archives-ouvertes.fr/francois-orieux`.

Résumé bibliométrique depuis le recrutement en 2014

— 11 articles de revues internationales avec comité de lecture (22 au total).
— 11 articles de conférences internationales avec comité de lecture (16 au total).
— 10 articles de conférences nationales avec comité de lecture au total.
— Communications nationales (GdR ISIS, JIONC, GRETSI, . . .)
— 1 logiciel déposé à l'APP. Contribution à la bibliothèque Python `scikit-image` [2].

2: ainsi que divers codes disponibles sur `github.com/forieux`

Quatre derniers articles

1. R. Abirizk, F. Orieux et A. Abergel, « Super-Resolution Hyperspectral Reconstruction with Majorization-Minimization Algorithm and Low-Rank Approximation, » IEEE Trans. on Comp. Imaging, p. 260–272, 2022.
2. M. Seznec, N. Gac, F. Orieux et A. S. Naik, « Real-time optical flow processing on embedded GPU : An hardware-aware algorithm to implementation strategy, » J. of Real-Time Image Proc., t. 19, n° 2, p. 317–329, 2022.
3. M. Seznec, N. Gac, F. Orieux et A. Sashala Naik, « Computing large 2D convolutions on GPU efficiently with the im2tensor algorithm, » J. Real-Time Image Proc., t. 19, n° 6, p. 1035–1047, 2022.
4. M. A. Hadj-Youcef, F. Orieux, A. Fraysse, and A. Abergel, « Fast Joint Multiband Reconstruction from Wideband Images Based on Low Rank Approximation, » IEEE Trans. Comput. Imaging, p. 922–933, 2020.

Activités scientifiques 2.

2.1. Projets

2.1.1. Projets financés

2022–2026 : Participant PEPR Origin (porteur volet IA : L. Denis, Lab. Hubert Cu- rien). 3,2 Me au total, 1 thèse accompagnée en co-encadrement pour le L2S et l'IMS.
2021–2025 : Participant ANR DarkEra (porteur N. Gac, L2S). 500 k€.
2018–2019 : Participant PEPS SKALLAS. 20 k€ – 12 mois.
2017–2018 : **Porteur** APP émergent Paris-Saclay Hyperfusion pour la fusion de données hyper et multi spectrale et le traitement de données massives. 15 k€ – 18 mois.
2017–2018 : **Porteur** APP CNES pour le financement de projet de thèse sur la fusion de données pour le JWST. Demi-financement thèse soit 50 k€ – 36 mois.
2017–2018 : Participant Défi mastodons CNRS Hyperstars. 35 k€ – 24 mois.
2017–2018 : Participant Défi Imag'In CNRS RESSOURCES. 31 k€ – 24 mois.
2014 : Participant Défi instrumentation CNRS BIBISIM . 21 k€ – 12 mois.

2.1.2. Contrats et collaborations industrielles

2022 : Thalès (Thèse CIFRE d'A. Archet) : 45 k€ sur trois ans, co-responsable avec N. Gac.
2018 : Atos-Bull, 30 k€ sur trois ans pour l'accompagnement de thèse de N. Monnier, coordonné avec N. Gac.
2018 : Thalès (Thèse CIFRE de M. Seznec) : 45 k€ sur trois ans, coordonné avec N. Gac.
2018 : ERAMET (contrat industriel), 20 k€, coordinateur avec Ali Djafari pour la segmentation d'images hyperspectrales de minéraux.

2.2. Encadrement

2.2.1. Encadrement de thèse

Thèses soutenues

— Nicolas Monnier [1]

Soutenance : 12/06/2023.

1: ExaSKA : Parallélisation sur serveur de calcul intensif pour le radiotélescope exascale SKA.

Encadrement : 30 % (Directeur N. Gac 40 % et co-encadrement C. Tasse).
Financement : Bourse Région IdF (DIM AVAC).
Publications : 2 conférences int. et 1 nat. en 1er auteur.

2: Spectro-imagerie haute résolution par inversion de données pour le JWST.

— Ralph Abi Rizk [2]

Soutenance : 08/11/2021
Encadrement : directeur à 70 % (dérogation), co-encadrement A. Abergel 30 %.
Financement : bourse CNES 50 % et de l'ED STIC Paris-Saclay 50 %.
Publications : 1 journal int., 1 conférence int. et 1 nat. en 1er auteur.

3: De l'algorithme à l'implémentation, élaboration d'un flot d'optimisations pour le calcul haute performance temps réel sur systèmes embarqués parallèles hétérogènes.

— Mickaël Seznec [3]

Soutenance : 25/10/2021.
Encadrement : 40 % (directeur N. Gac 50 %, co-encadrement Alvin Sashala Naik à Thalès 10 %).
Financement : thèse CIFRE avec Thalès TRT.
Publications : 2 journaux int., 2 conférences int. en 1er auteur.

4: Spatio spectral reconstruction from low resolution multispectral data : application to the Mid-Infrared instrument of the James Webb Space Telescope.

— Amine Hadj-Youcef [4]

Soutenance : 27/09/2018.
Encadrement : 60 % (directrice Aurélia Fraysse, co-encadrement Alain Abergel).
Financement : bourse de thèse de l'ED STITS.
Publications : 1 journal int. en 1er auteur, 1 journal int. en 6e, 2 conf. int. en 1er auteur, 1 conf. nat. en 1er auteur.
Situation : Data Scientist chez TAG Heuer en 2020.

5: Candidats (proto-)amas de galaxies à grand redshift vus par le CFHT

— Benjamin Clarenc [5]

Soutenance : 11/09/2018.
Encadrement : 50 % la 1re année (directeur Hervé Dole).
Financement : bourse CNES 50 % et de l'ED AA IdF.
Publications : 1 journal int. (5e auteur).
Situation : Ingénieur développement logiciel chez ALTEN en 2020.

Direction et co-encadrement de thèses en cours

6: Traitement conjoint des données d'imagerie et de spectroscopie de l'instrument MIRI du JWST.

— Dan Pineau [6] (depuis 2022)

Encadrement : 70% (co-directeur A. Abergel 30%).
Financement : bourse de thèse CNES 50 % et LabCom IN-CLASS 50 %.
Publications : 1 conf. nat (GRETSI 2023).

7: Flot d'optimisation semi-automatisé de réseaux de neurones sur plateforme embarquée.

— Agathe Archet [7] (depuis 2022)

Encadrement : 33% (directeur N. Gac 34%, co-encadrement N. Ventroux 33%).
Financement : Thèse CIFRE Thalès.
Publications : 1 conf. int. (DSD 2023), 3 conf. nat. (GRETSI 2021 et 2022 et GSOC 2023),

2.2.2. Encadrement de stages

2018 : Stage M2 de Ralph Abi Rizk à 100 % avec un financement à ½ AAP stage du L2S et ½ projet émergent HyperFusion,
2015 : Stage M2 de Raphael Chinchilla à 100 % avec un financement à ½ AAP stage du L2S et ½ sur bourse jeune chercheur.
2022 : Stage M1 de Baptiste Leroux à 100 %.
2019 : Stage M1 de Fabio Eid Morooka à 50 % (co-encadrement N. Gac).
2014 : Stage M1 de Dylan Fagot à 50 % (co-encadrement A. Fraysse).

2.2.3. Autres encadrements

2018 : Mickaël Seznec, ingénieur d'étude, co-encadrement avec N. Gac.
2018–2020 : Jérémy Besson, parcours recherche étudiant de CentraleSupélec, co-encadrement avec N. Gac à 50 %.
2018 – 2019 : Alexandre Journeaux, parcours recherche étudiant de CentraleSupélec, encadrement à 100 %.

2.3. Autres activités

Mandats électifs

— Membre élu national au CNU section 61 depuis septembre 2020.
— Membre élu local dans la Commission Consultative de l'Univertisé Paris-Saclay (CCUPS) section 60-61-62 depuis janvier 2022 (consultation sur les carrières, promotions, congés et délégations, primes . . .), depuis 2022.

Animations scientifiques

— Co-responsable des Séminaires de Traitement du Signal de Paris-Saclay « S^3 » de 2019 à 2023 (37 séminaires organisés avec invitation des orateurs, annonces, création du site web [8], twitter, youtube, speakerdeck, . . .).
— Création, animations et organisation du « Journal Club [9] » depuis 2020 (33 rencontres depuis).

Participation à des comités de recrutement

— Membre extérieur du comité de sélection du poste MCU 776, en 2021, en section 61 pour l'Université de Côte d'Azur.
— Membre du comité de sélection du poste MCU 1521, en 2015, en section 61 pour l'UFR de médecine de l'Université Paris-Sud.

Responsabilités administratives

— Chargé de mission sur l'interdisciplinarité au L2S de 2018 à 2023.
— Chargé de mission « Calculs scientifiques » depuis 2023.

8: s3-seminar.github.io

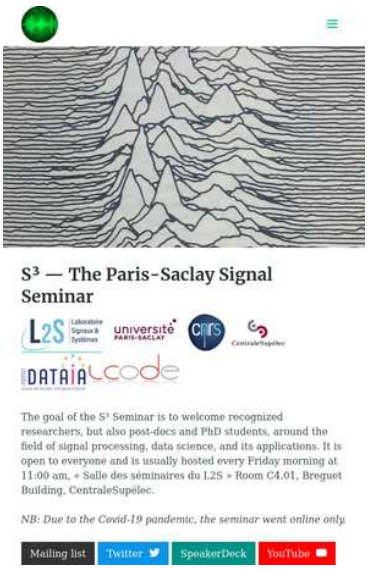


Figure 2.1. – Nouveau site web du séminaire S^3.

9: Séminaire biliographique ou groupe de lecture. Il s'agit d'un lieu de discussion scientifique autour d'un article présenté par quelqu'un, le plus souvent un doctorant.

- Membre invité du conseil de laboratoire de 2018 à 2020.
- Membre de la commission d'organisation du prix de thèse Signal-Image-Vision du Gretsi [10].
- Correspondant pour le L2S du GdR ISIS.
- Gestion du parc informatique du Groupe Problèmes Inverses (installation, mise à jour, commandes, *etc.*).
- Membre du conseil scientifique du Laboratoire Commun INCLASS (LabCom) IAS – Acri.

10: membre du club EEA depuis 2019

Travail d'expertise

- Rapporteur pour les revues : Optics Letters, IEEE trans. on Image Processing, IEEE trans. on Computational Imaging, Elsevier Digital Signal Processing, IEEE Signal Processing Letters, Biomedical Optics Express.
- Rapporteur pour le colloque Gretsi, EUSIPCO et IEEE ICIP.
- Membre du comité de suivi individuel à mi-thèse de Tran Khanh Hung, Goluck Konuko.
- Membre du comité d'évaluation à mi-thèse de Mehdi Amrouche.

Représentation et rayonnement

- Chercheur associé à l'Institut d'Astrophysique Spatiale depuis 2014.
- Orateur invité dans 9 *workshops* ou séminaires depuis 2014.
- Invitation au jury de thèse d'Antoine Marchal, soutenue le 25 sept. 2019 à l'Université Paris-Saclay (titre : Étude de la structure multiphase du milieu interstellaire neutre turbulent).
- *Médiation scientifique :* Présentation des défis et outils du traitement du signal pour l'astronomie à l'école d'été SPACEBOS sur les métiers du spatial (en 2018 et 2019). Présentation devant un public divers avec des ingénieurs, techniciens, juristes. . .

Activités pédagogiques 3.

J'effectue mon service d'enseignement dans le Département de Physique de la Faculté des Sciences d'Orsay. J'interviens notamment dans des filières d'enseignement de physique pour y dispenser les notions de traitement du signal indispensables à tout physicien expérimental, mais également dans la filière E3A [1] ou dans l'école d'ingénieur Polytech Paris-Saclay.

1: Électronique, Énergie Électrique, Automatique

3.1. Activités d'enseignement

Je suis responsable d'une majorité des UE. J'ai également rédigé presque tous les polycopiés (plusieurs disponibles en ligne) de cours, TD et TP, naturellement en m'inspirant de ceux existants ou d'autres ouvrages. Je ne présente pas les charges pédagogiques de jury, de suivi de stages ou d'apprentis par exemple.

Globalement, à part les deux premières années où l'Université Paris-Sud offre une décharge de 46h. eq. TD, j'ai effectué légèrement plus que 192h eq. TD d'enseignements par an *.

Traitement du signal (TS) – M1 Physique et Application et M1 de Mécanique des Fluides – 40h – entre 15 et 60 étudiants.

Suite à une séparation entre du M1 Physique et Application et Mécanique, cette UE s'est retrouvée dupliquée. Elle a pour objectif de parcourir l'ensemble des bases en traitement du signal pour des étudiants de physique qui seront confrontés au traitement de données réelles (notion de bruit, d'estimation, convolution, Fourier, échantillonnage, filtrage). En M1 pour des physiciens, l'objectif est d'avoir un enseignement résolument pratique avec un volume de TP important sans négliger la théorie.

Signaux et Systèmes Linéaires (SSL) – L3 Électronique, Énergie Électrique et Automatique – 33h – 20 étudiants.

Cette UE présente l'ensemble des notions de systèmes linéaires et de l'analyse harmonique pour les systèmes analogiques et numériques avec la théorie de l'échantillonnage pour faire le pont entre les deux domaines. Les notions abordées sont plutôt théoriques avec des accents mis sur les liens entre les transformations.

Outils pour le traitement du signal (OMTS) – Polytech Paris-Sud 1re année ingénieur – 22h – 30 étudiants.

Cette UE reprend essentiellement le contenu de l'UE précédente, mais pour un public ingénieur avec un contenu pratique sous forme de TP plus conséquent..

*. Le département de physique d'Orsay ne rémunère pas les heures supplémentaires.

Programmation et données numériques (PDN) – M1 Physique et Application – 26h – 60 étudiants.

L'UE programmation est majoritairement pratique et a pour vocation d'enseigner les outils numériques, avec le langage python, à des physiciens. Le niveau étant très hétérogène, ils étudient les notions de base de l'algorithmie (boucles, tests conditionnels . . .), mais également des outils plus avancés comme l'optimisation de moindres carrés non linéaires à l'aide de bibliothèques. Nous nous adaptons au rythme de chacun.

Traitement du signal (TS2) – Double L3 Math-Physique – 31h – 30 étudiants.

J'ai monté cette nouvelle UE en 2018-2019 qui a pour public des étudiants en double licence sélective Math-Physique. Les étudiants sont d'un bon niveau mathématique et ils ont déjà eu une UE traitement du signal dans son versant mathématique. Pour ma part je leur présente le versant physique ou ingénieur dans une pédagogie de type *inversée* où ils doivent exploiter les notions vues, avec mes documents mis à disposition, pour mener à bien des TPs et des projets.

Projets – 2e année IUT de Cachan – 20h.

Pendant l'année universitaire 2019 – 2020 j'ai fait un échange de service avec N. Gac de l'IUT de Cachan pour intervenir pendant 20h sur de l'encadrement de projet. J'ai proposé ainsi un nouveau projet sur du traitement de la parole avec un vocoder pour produire des effets de voix robotique [2].

2: Ce ne sont pas les vocodeurs tels que ceux employées pour coder la voix pour de la télétransmission.

Problèmes Inverses (PI) – M2 Astronomie, Astrophysique et Ingénieurie Spatiale – 40h – 15 étudiants.

J'ai mis en place cette UE, entre 2014 et 2019, qui a pour objet de présenter les bases des problèmes inverses à des étudiants qui se destinent majoritairement à une thèse en astrophysique. L'origine des étudiants est diverse (magistère de Physique de Paris-Sud, ENS Cachan, Paris 6, . . .). Je présente les bases des problèmes inverses mal posés avec les notions de fréquences, d'échantillonnage, de modèle instrument, de bruit et d'estimation.

L'UE est sous la forme de cours-projet sur une semaine. Ils vont jusqu'à la mise en oeuvre d'algorithmes d'optimisation à base de gradient sur des critères convexes différentiables. Cette UE a pris fin en 2019 suite à un changement en profondeur de la maquette après cinq ans.

Automatique (A) – M2 Outils et Système de l'Astronomie et l'Espace – 25h – 25 étudiants.

L'objectif de cette UE est de présenter les bases de l'automatique et de l'asservissement (après une UE sur l'estimation et le traitement du signal) à des profils ingénieurs pour le spatial. L'idée est de balayer les notions de systèmes linéaires, boucles ouvertes

et fermées, stabilité, régulateur et échantillonnage. Le nombre d'heures étant réduit, il s'agit essentiellement d'une introduction et de leur donner les points d'entrées qu'ils retrouveront dans le polycopié.

Traitement du signal et de l'image en biologie et en médecine (TSIBM)– M2 Automatique et Traitement du Signal et des Images – 6h.

Je suis intervenu avec Charles Soussen et Nicolas Gac dans cette UE du Master 2 ATSI entre 2016 et 2020. Nous leur présentons les outils méthodologiques et technologiques avancés pour le traitement du signal en biologie.

Pour ma part je leur présente les méthodes de reconstruction d'image en microscopie biologique [3]. Le spectre méthodologique va des problèmes inverses avec les approches bayésiennes aux méthodes d'optimisation avec contrainte par multiplieur de Lagrange.

3: Ce contenu pédagogique a été dispensé également comme vacataire plusieurs fois dans le Master 2 ISIS de Bordeaux.

3.2. Supports

3.2.1. Supports papier

Pour ces enseignements j'ai produit la majorité des supports de cours soit

- quatre polycopiés de cours, de TD et TP pour les UE de traitement du signal, signaux et systèmes linéaires, outils pour le traitement du signal et automatique.
- sept polycopiés de TP supplémentaires pour l' UE Problème inverse, Traitement du signal en double licence et pour le M2 ATSI.

3.2.2. Logiciel de présentation

Depuis 2019 je développe une application de démonstration interactive pour les étudiants, en cours et en distanciel. Il est disponible ici github.com/forieux/teachapp. Ce logiciel offre une interface interactive automatique à l'utilisateur et au programmeur en charge des des figures, graphes, sons *etc*. La figure 3.1 montre un exemple avec les exponentielles complexes. Les boutons permettent de changer les paramètres de l'exponentielle et le résultat de la projection sur les axes. Le programmeur à juste à coder la figure.

L'intérêt pédagogique est pour moi évident et le retour positif a été immédiat. Les étudiants se re-concentrent, voient une visualisation propre et pratique des équations. Je l'utilise aussi pour illustrer des résultats de TD par exemple. L'interactivité permet de mieux saisir certains phénomènes. Enfin, cela introduit une sorte de pause studieuse qui soulage tout le monde.

J'ai fait de nombreux essais et tests avec d'autres outils qui ne m'ont pas donné satisfaction étant donné mon cahier des charges.

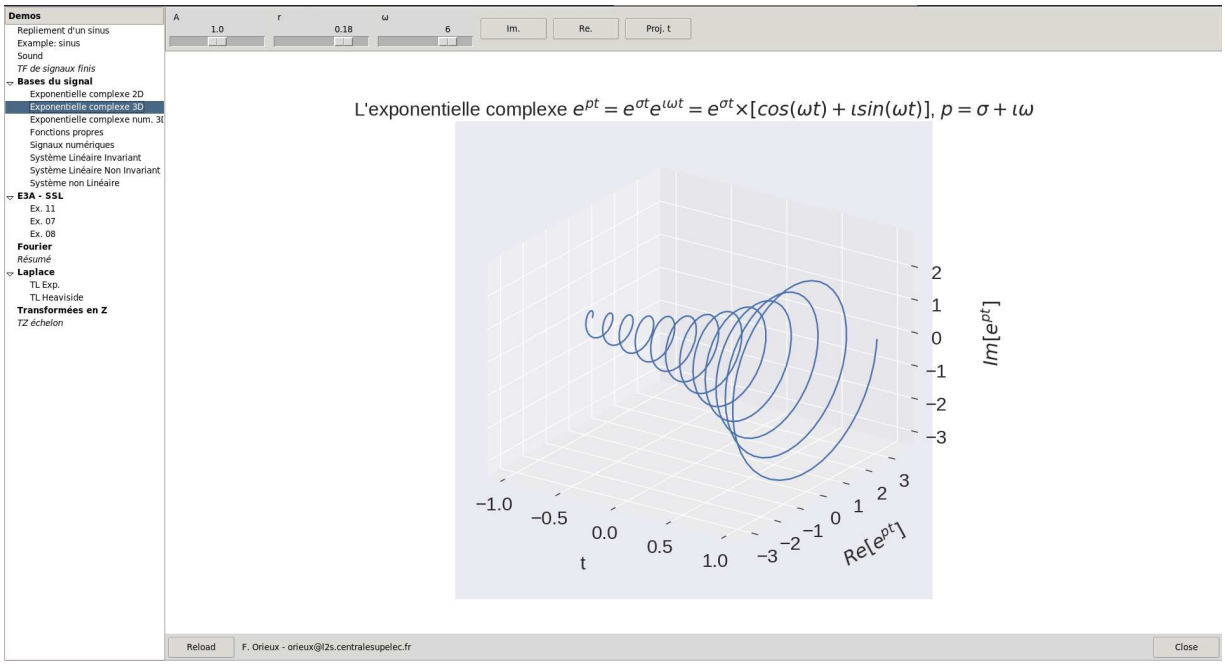


Figure 3.1. – Exemple d'illustration interactive. La liste sur la droite présente toutes les animations trouvées. La barre de contrôle en haut est générée automatiquement et permet de modifier la figure.

Je voulais un outil

- *simple* qui ne fait qu'une seule chose,
- *adapté* à la présentation *en cours* (vidéo projecteur, plein écran, sans distraction),
- *flexible* et *puissant*, c'est-à-dire qui ne présente que peu de limites dans ses capacités,
- qui n'impose pas trop de nouvel apprentissage pour produire une animation.

J'ai essayé, parmi d'autres, le Jupyter Notebook ainsi que GeoGebra. Le Jupyter Notebook est mal adapté à une présentation en classe. L'interactivité de GeoGebra est exemplaire, mais il n'est pas assez puissant et flexible pour enseigner les sciences de l'ingénieur.

J'ai donc commencé à implémenter mon propre logiciel. Il est centré sur

- Python, Matplotlib et autres bibliothèques : tout ce que fait Matplotlib est possible et la limite est uniquement Python. Je produis ainsi des graphes dynamiques, du son, des images et des vidéos.
- Pas d'apprentissage de bibliothèque en plus pour utiliser le logiciel, uniquement Python et Matplotlib.
- Génération automatique des contrôles (boutons, sliders, . . .) pour minimiser la charge.
- Interface minimale adaptée à la présentation en classe.

Chaque illustration nécessite de produire un code, mais Python et Matplotlib sont assez haut niveau. De plus les illustrations sont réutilisables.

Le logiciel est encore très jeune. J'ai de nombreuses perspectives qui nécessiteront d'autres forces de travail que la mienne. Je pense par exemple à la possibilité d'utiliser Matlab, la production d'un exécutable autonome pour plusieurs plates-formes ou la possibilité de faire des visualisations sur le web pour les étudiants en révision.

4. Liste de publications

La liste complète des publications est disponible avec le lien `cv.archives-ouvertes.fr/francois-orieux`.

IdHal : `francois-orieux`
ORCID : `0000-0001-5638-3416`

Résumé bibliographique depuis 2014

— 8 articles de revues internationales avec comité de lecture (18 au total [1], 3 associées au travail de thèse).
— 9 articles de conférences internationales avec comité de lecture (15 au total).
— 8 articles de conférences nationales avec comité de lecture au total.
— Communications nationales (GdR ISIS, JIONC, GRETSI, ...)
— 1 logiciel déposé à l'APP. Contribution à la bibliothèque Python `scikit-image` [2].

1: j'ai exclu l'article sur `scikit-image` car bien que les *contributeurs* soient dans les auteurs, c'est une contribution de code (avec révision malgré tout).

2: ainsi que divers codes disponibles sur `github.com/forieux`

Cinq publications les plus significatives

Voici une sélection, avec une part d'arbitraire, de publications les plus significatives. Le nom des doctorants est souligné. Mes critères ont été l'impact et l'originalité de la contribution ainsi que la représentation de la diversité de mes activités.

1. M. A. Hadj-Youcef, F. Orieux, A. Fraysse, and A. Abergel, « Fast Joint Multiband Reconstruction from Wideband Images Based on Low Rank Approximation, » IEEE Trans. Comput. Imaging, pp. 922–933, May 2020, doi : 10.1109/TCI.2020.2998170.

   Cette contribution est importante, car il s'agit de la publication de journal de mon premier doctorant, M. A. Hadj-Youcef. Nous y montrons que les modèles de mélange utilisés habituellement en hyperspectral peuvent être utilisés dans l'imagerie multispectrale pour le traitement conjoint des bandes pour la restauration d'images. Nous établissons également des preuves d'inversion de matrice aboutissant à un algorithme extrêmement rapide.

2. A. Marchal, M.-A. Miville-Deschênes, F. Orieux, N. Gac, C. Soussen et al. « ROHSA : Regularized Optimization for Hyper-Spectral Analysis », Astronomy and Astrophysics, EDP Sciences, 2019.

   Cet article est le résultat d'un travail interdisciplinaire entre des astrophysiciens et des traiteurs de signaux. Il s'agit d'un

problème de séparation de sources sur des données hyperspectrales, sources qui représentent des nuages de gaz chaud et froid. Le mélange provient de la projection sur la ligne de visée et la mesure indirecte. La méthode proposée est la première à réussir le démélange et à atteindre ces performances. Les astrophysiciens considèrent ce problème vieux de cinquante ans comme résolu.

3. F. Orieux, E. Sepulveda, V. Loriette, B. Dubertret et J.-C. Olivo-Marin, « Bayesian Estimation for Optimized Structured Illumination Microscopy, » IEEE Transactions on Image Processing, t. 21, n o 2, p. 601-614, fév. 2012.

   Ce travail porte sur le traitement de données de microscopie biologique en co-conception. Le microscope provoque un repliement spectral par modulation d'amplitude et notre travail est le premier à avoir montré une limite basse en nombre d'images nécessaire pour la reconstruction avec un algorithme permettant cette reconstruction de manière non supervisée.

4. F. Orieux, O. Féron et J.-F. Giovannelli, « Sampling High-Dimensional Gaussian Distributions for General Linear Inverse Problems, » IEEE Signal Processing Letters, t. 19, n o 5, p. 251-254, 2012.

   Dans cet article nous proposons un nouvel algorithme Perturbations-Optimisation (PO) pour faire l'échantillonnage de très grands champs gaussiens corrélés, dans les cas où la factorisation de Cholesky par exemple serait impossible, et si la précision respecte une certaine structure. L'originalité repose sur l'exploitation des algorithmes d'optimisation, utilisés sur des critères perturbés, estimant ainsi non pas la moyenne (ou le MAP), mais un échantillon.

5. F. Orieux, J.-F. Giovannelli et T. Rodet, « Bayesian Estimation of Regularization and Instrument Parameters for Wiener-Hunt Deconvolution, » Journal of the Optical Society of America, A, Optics, Image Science, and Vision, t. 27, n o 7, p. 1593-1607, 2010.

   Cet article présente un algorithme [3] de déconvolution où les paramètres d'observation et les hyperparamètres sont estimés. Il repose sur un échantillonneur de Gibbs et la diagonalisation dans la base de Fourier des opérateurs. L'algorithme est très rapide (moins d'une seconde sur les ordinateurs récents pour une image 1024×1024).

3: L'algorithme est disponible dans la bibliothèque scikit-image.

Articles de revues internationales avec comité de lecture

[1] R. Abirizk, F. Orieux et A. Abergel, « Super-Resolution Hyperspectral Reconstruction with Majorization-Minimization Algorithm and Low-Rank Approximation, » *IEEE Transactions on Computational Imaging*, p. 260-272, 2022.

[6] M. Seznec, N. Gac, F. Orieux et A. S. Naik, « Real-time optical flow processing on embedded GPU : an hardware-aware algorithm to implementation strategy, » *J Real-Time Image Proc*, t. 19, n° 2, p. 317-329, avr. 2022.

[7] M. Seznec, N. Gac, F. Orieux et A. Sashala Naik, « Computing large 2D convolutions on GPU efficiently with the im2tensor algorithm, » *J Real-Time Image Proc*, t. 19, n° 6, p. 1035-1047, 1er déc. 2022.

[10] M. A. Hadj-Youcef, F. Orieux, A. Fraysse et A. Abergel, « Fast Joint Multiband Reconstruction from Wideband Images Based on Low Rank Approximation, » *IEEE Trans. Comput. Imaging*, p. 922-933, 28 mai 2020.

[15] Y. Lai-Tim, L. Mugnier, F. Orieux, R. Baena-Gallé, M. Paques et S. Meimon, « Jointly Super-Resolved and Optically Sectioned Bayesian Reconstruction Method for Structured Illumination Microscopy, » *Optics Express*, t. 27, n° 23, p. 33 251-33 267, oct. 2019.

[17] A. Marchal, M.-A. Miville-Deschênes, F. Orieux et al., « ROHSA : Regularized Optimization for Hyper-Spectral Analysis, » *Astronomy & Astrophysics*, t. 626, A101, 2019.

[28] F. Orieux, V. Loriette, J.-C. Olivo-Marin, E. Sepulveda et A. Fragola, « Fast Myopic 2D-SIM Super Resolution Microscopy with Joint Modulation Pattern Estimation, » *Inverse Problems*, t. 33, n° 12, p. 125 005, 2017.

[29] A. Boucaud, M. Bocchio, A. Abergel, F. Orieux, H. Dole et M.-A. Hadj-Youcef, « Convolution Kernels for Multi-Wavelength Imaging, » *Astronomy & Astrophysics*, t. 596, A63, 2016.

[30] A. Boucaud, H. Dole, A. Abergel, H. Ayasso et F. Orieux, « PSF Homogenization for Multi-Band Photometry from Space on Extended Objects, » *EAS Publications Series*, t. 78–79, D. Mary, R. Flamary, C. Theys et C. Aime, éd., p. 275-285, 2016.

[31] O. Féron, F. Orieux et J.-F. Giovannelli, « Gradient Scan Gibbs Sampler : An Efficient Algorithm for High-Dimensional Gaussian Distributions, » *IEEE Journal of Selected Topics in Signal Processing*, t. 10, n° 2, p. 343-352, mars 2016.

[33] P. Vermeulen, F. Orieux, J.-C. Olivo-Marin et al., « Out of Focus Background Subtraction for Fast Structured Illumination Super Resolution Microscopy of Optically Thick Samples, » *Journal of microscopy*, t. 259, n° 3, p. 257-268, 2015.

[36] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias et al., « Scikit-image : image processing in Python, » *PeerJ*, t. 2, e453, 19 juin 2014.

[37] P. Collaboration, « Planck 2013 Results. I. Overview of Products and Scientific Results, » *Astronomy & Astrophysics*, t. 571, 2013.

[38] P. Collaboration, « Planck 2013 Results. VI. High Frequency Instrument Data Processing, » *Astronomy & Astrophysics*, t. 571, J. Tauber, éd., A6, oct. 2013.

[39] P. Collaboration, « Planck 2013 Results. XV. CMB Power Spectra and Likelihood, » *Astronomy & Astrophysics*, t. 571, J. Tauber, éd., A15, oct. 2013.

[41] F. Orieux, J.-F. Giovannelli, T. Rodet et A. Abergel, « Estimating Hyperparameters and Instrument Parameters in Regularized Inversion. Illustration for Herschel/Spire Map Making., » *Astronomy & Astrophysics*, t. 549, n° A83, jan. 2013.

[43] F. Orieux, O. Féron et J.-F. Giovannelli, « Sampling High-Dimensional Gaussian Distributions for General Linear Inverse Problems, » *IEEE Signal Processing Letters*, t. 19, n° 5, p. 251-254, 2012.

[44] F. Orieux, J.-F. Giovannelli, T. Rodet, A. Abergel, H. Ayasso et M. Husson, « Super-Resolution in Map-Making Based on a Physical Instrument Model and Regularized Inversion. Application to SPIRE/Herschel, » *Astronomy & Astrophysics*, t. 539, n° A38, p. 1-28, mars 2012.

[45] F. Orieux, E. Sepulveda, V. Loriette, B. Dubertret et J.-C. Olivo-Marin, « Bayesian Estimation for Optimized Structured Illumination Microscopy, » *IEEE Transactions on Image Processing*, t. 21, n° 2, p. 601-614, fév. 2012.

[50] F. Orieux, J.-F. Giovannelli et T. Rodet, « Bayesian Estimation of Regularization and Instrument Parameters for Wiener-Hunt Deconvolution, » *Journal of the Optical Society of America*, A, Optics, Image Science, and Vision, t. 27, n° 7, p. 1593-1607, 2010.

[53] J.-A. Rodon, A. Zavagno, J.-P. Baluteau, L. Anderson, E. Polehampton et et al., « Physical Properties of the Sh2-104 Hii Region as Seen by Herschel, » *Astronomy & Astrophysics*, t. 518, p. L80, 2010.

[58] T. Rodet, F. Orieux, J.-F. Giovannelli et A. Abergel, « Data Inversion for Over-Resolved Spectral Imaging in Astronomy, » *IEEE Jour. of Selected Topics in Signal Proc.*, t. 2, n° 5, p. 802-811, oct. 2008.

Congrès internationaux avec actes et comité de lecture

[2] N. Monnier, D. Guibert, C. Tasse et al., « Multi-Core Multi-Node Parallelization of the Radio Interferometric Imaging Pipeline DDFacet, » in *IEEE Workshop on Signal Processing Systems (SiPS)*, Rennes, France, nov. 2022.

[9] R. Abirizk, F. Orieux et A. Abergel, « Non-Stationary Hyperspectral Forward Model and High-Resolution, » in *Proc. of 27th IEEE Int. Conf. on Image Processing*, Abu-Dhabi, United Arab Emirates, oct. 2020, p. 5.

[12] M. Seznec, N. Gac, F. Orieux et A. S. Naik, « A new convolutions algorithm to leverage tensor cores, » Poster (San José, USA), mai 2020.

[13] M. Seznec, N. Gac, F. Orieux et A. S. Naik, « An Efficiency-Driven Approach For Real-Time Optical Flow Processing On Parallel Hardware, » in *2020 IEEE International Conference on Image Processing (ICIP)*, oct. 2020, p. 3055-3059.

[18] M. A. Hadj-Youcef, F. Orieux, A. Fraysse et A. Abergel, « Spatio-Spectral Multichannel Reconstruction from Few Low-Resolution Multispectral Data, » in *26th European Signal Processing Conference (EUSIPCO 2018)*, Rome, Italy, sept. 2018.

[22] M. A. Hadj-Youcef, F. Orieux, A. Fraysse et A. Abergel, « Restoration from Multispectral Blurred Data with Non-Stationary Instrument Response, » in *Proc. of EUSIPCO*, 2017.

[24] W. Meiniel, P. Spinicelli, E. D. Angelini et al., « Reducing Data Acquisition for Fast Structured Illumination Microscopy Using Compressed Sensing, » in *Proc. of IEEE Int. Symp. on Biomedical Imaging (ISBI 2017)*, 2017.

[27] F. Orieux et R. Chinchilla, « Semi-Unsupervised Bayesian Convex Image Restoration with Location Mixture of Gaussian, » in *25th European Signal Processing Conference (EUSIPCO)*, 2017, p. 758-762.

[32] F. Orieux, O. Féron et J.-F. Giovannelli, « Gradient Scan Gibbs Sampler : An Efficient High-Dimensional Sampler Application in Inverse Problems, » in *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE Int. Conf. On*, 2015.

[35] P. Vermeulen, F. Orieux, J. C. Olivo-Marin et al., « In Vivo Dual-Color High Resolution Sim of Optically Thick Samples with Background Subtraction, » in *FOM*, 2014.

[40] J. C. Olivo-Marin, Y. Le Montagner, F. Orieux, V. Loriette, A. Fragola et E. Angelini, « Mathematical Microscopy, » in *Information Optics*, 2013.

[46] P. Vermeulen, F. Orieux, E. Sepulveda, J. C. Olivo-Marin, A. Fragola et V. Loriette, « Dynamic SIM for High-Speed Imaging and Optical Sectioning in Living Samples, » in *FOM*, 2012.

[51] F. Orieux, J.-F. Giovannelli et T. Rodet, « Deconvolution with Gaussian Blur Parameter and Hyperparameters Estimation, » in *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*, 2010, p. 1350-1353.

[52] F. Orieux, T. Rodet et J.-F. Giovannelli, « Instrument Parameter Estimation in Bayesian Convex Deconvolution, » in *Image Processing (ICIP), 2010 17th IEEE Int. Conf. On*, IEEE, sept. 2010, p. 1161-1164.

[55] F. Orieux, T. Rodet et J.-F. Giovannelli, « Super-Resolution with Continuous Scan Shift, » in *Image Processing (ICIP), 2009 16th IEEE Int. Conf. On*, Cairo, Egypt, nov. 2009, p. 1169-1172.

[56] T. Rodet, F. Orieux, J.-F. Giovannelli et A. Abergel, « Data Inversion for Hyperspectral Objects in Astronomy, » in *WHISPERS '09 - 1st Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, 2009.

Congrès et colloques nationaux avec comité de lecture

[4] N. Monnier, F. F. Orieux, N. Gac, D. Guibert, C. Tasse et E. Raffin, « Interpolation Rapide Sky to Sky Pour l'Imagerie Radio Interférométrique, » in *GRETSI 28ème Colloque Francophone de Traitement Du Signal et Des Images*, Nancy, France, sept. 2022.

[5] M. Seznec, A. Archet, N. Gac, F. Orieux et A. S. Naik, « MobileFlow : Modèle et Mise En Œuvre Pour Une Inférence de Flot Optique Efficace, » in *28eme Colloque GRETSI Traitement Du Signal & Des Images*, Nancy, France, sept. 2022.

[14] R. Abirizk, F. Orieux et A. Abergel, « Reconstruction hyperspectrale à haute résolution à partir de mesures de spectrometrie, » in *Actes du 27e GRETSI*, Lille, août 2019, p. 4.

[16] Y. Lai-Tim, L. Mugnier, F. Orieux, R. Baena-Gallé, M. Paques et S. Meimon, « Toward a Jointly Super-Resolved and Optically Sectioned Reconstruction for Structured Illumination Retinal Imaging, » in *Actes du 27e GRETSI*, Lille, France, août 2019.

[20] R. Baena-Gallé, L. M. Mugnier et F. Orieux, « Optical Sectioning with Structured Illumination Microscopy for Retinal Imaging : Inverse Problem Approach, » in *Actes Du 26e GRETSI*, Juan-les-pins, France, sept. 2017, p. 4.

[21] M. A. Hadj-Youcef, F. Orieux, A. Fraysse et A. Abergel, « Restauration d'objets Astrophysiques à Partir de Données Multispectrales Floues et d'une Réponse Instrument Non-Stationnaire, » in *Actes Du 26e GRETSI*, 2017.

[23] W. Meiniel, P. Spinicelli, E. Angelini et al., « Accélération de La Méthode de Microscopie à Illumination Structurée à l'aide de l'acquisition Comprimée, » in *GRETSI 2017*, Juan-les-Pins, France, 2017.

[26] F. Orieux et R. Chinchilla, « Restauration d'image Par Une Approche Bayésienne Semi Non Supervisée et Le Mélange de Gaussienne, » in *Actes Du 26e GRETSI*, Juan-les-Pins, France, sept. 2017.

[47] O. Féron, F. Orieux et J.-F. Giovannelli, « Échantillonnage de Champs Gaussiens de Grande Dimension, » in *Journées de Statistique (JdS10)*, 2010.

[59] F. Orieux, T. Rodet, J.-F. Giovannelli et A. Abergel, « Inversion de Données Pour l'imagerie Spectrale Sur-Résolues En Astronomie, » in *Actes Du 21e GRETSI*, sept. 2007, p. 717-720.

Logiciels *open source*

[8] F. Orieux et R. Abirizk, *Q-MM : The Quadratic Majorize-Minimize Python Toolbox*, 4 fév. 2021.

[11] F. Orieux, *TeachApp : An Application for Interactive Demonstration in Classroom*, 2020.

[25] F. Orieux, *BSIM : Bayesian Structured Illumination Microscopie*, 2017.

[42] F. Orieux, *Python Package : Image Deconvolution in Skimage*, 2012.

[48] F. Orieux, *Matlab Package : Conjugate Gradient for Large Inverse Problems*, 2010.

[49] F. Orieux, *Matlab Package : Unsupervised Image Deconvolution*, 2010.

Deuxième partie

Travaux de recherche et perspectives

« L'Âge mûr » – Camille Claudel.

5. Introduction

Cette partie est consacrée à la présentation de mes travaux de recherche. Comme il s'agit de travaux déjà publiés, j'ai pris le parti de faire une présentation succincte et non exhaustive. J'expose tout d'abord des éléments de contexte et de problématiques puis décris succinctement les travaux réalisés avec quelques résultats. Je ne décris pas avec autant de détails l'état de l'art, les résultats et les comparaisons ; pour cela je renvoie aux articles de journaux et documents déjà publiés et joins en partie III ainsi qu'au *preprint* disponible sur HAL ou ma page web professionnelle.

Cette partie est constituée de trois chapitres. Le chapitre 6 correspond à mes travaux et mes encadrements de thèses principaux sur la reconstruction et la restauration d'image multi et hyperspectrale. Le chapitre 7 concerne le problème de reconstruction d'image en microscopie biologique. Enfin le chapitre 8 présente des travaux plus méthodologiques, sans application particulière, sur les approches bayésiennes et les algorithmes MCMC pour les problèmes inverses, notamment en image.

Enfin j'ai mené plusieurs travaux et encadrements que je ne décris pas ici pour ne pas rallonger le manuscrit et sans minimiser leur importance. Tout d'abord, j'ai une collaboration importante avec Nicolas Gac sur l'adéquation algorithme–architecture, dans laquelle j'apporte l'expertise sur les algorithmes. La thèse CIFRE que Mickaël Seznec a soutenu en octobre 2021, portait sur l'accélération sur GPU d'algorithmes de flot optique et de convolution deux images. Ces deux aspects ont chacun amené à une publication de journal [6], [7]. Cette collaboration avec Thalès se poursuit actuellement avec la thèse d'Agathe Archet sur les méthodes de *Neural Architecture Search* pour cible embarquée et la segmentation sémantique d'images aéroportées. Ensuite, Nicolas Monnier a récemment soutenu sa thèse en juin 2023, qui portait sur l'accélération d'algorithme de synthèse de Fourier pour la radioastronomie. Son travail a abouti à deux conférences internationales [4], [5]. Pour terminer, le projet Hyperstars avec Marc-Antoine Miville-Deschênes et Antoine Marchal portait sur un problème de séparation de source sur des mesures hyperspectrales [21].

Reconstruction d'images multispectrales et hyperspectrales

6.

Cette partie présente mes travaux sur la reconstruction et la restauration d'images multi et hyperspectrales. Elle est associée aux travaux d'encadrement des thèses d'Amine Hadj-Youcef (2015–2018), de Ralph Abirizk (2018–2021) et plus récemment de Dan Pineau (depuis 2022). Elle s'est effectuée en collaboration étroite avec Alain Abergel, Professeur à l'Université Paris-Saclay, à l'Institut d'Astrophysique Spatiale (IAS – Université Paris-Saclay, CNRS), et Aurélia Fraysse, Maîtresse de Conférence à l'Université Paris-Saclay, au L2S.

6.1. Introduction

Les techniques d'imagerie hyperspectrale et multispectrale sont de plus en plus répandues. Un contexte d'application majeure concerne l'observation de la terre et la télédétection [17], [33], [40], [52], mais on les retrouve également en analyse de matériaux ou de minéraux sur site [23], en imagerie médicale [45], en analyse d'œuvre d'art[36], et bien sûr en astronomie avec des observations au sol [21] et spatiales [13], [49], [55]. On peut notamment mentionner le programme Copernicus de l'ESA dont l'objet est la mise en orbite d'une vingtaine de satellites (famille Sentinelle) pour l'observation de la terre en multi et hyperspectrale.

Pour faire simple, une image hyperspectrale, illustrée figure 6.1, ou multispectrale, se représente le plus souvent comme un objet 3D, ou 2D+λ, disposant de deux dimensions spatiales et une dimension de longueur d'onde. Elle est souvent notée $\boldsymbol{x} \in \mathbb{R}^{I \cdot J \cdot L}$ avec $I \cdot J$ pixels et L longueurs d'ondes, mais la notation matricielle $\boldsymbol{X} \in \mathbb{R}^{IJ \times L}$, où chaque spectre est rangé en colonne, est également couramment employée pour certaines applications.

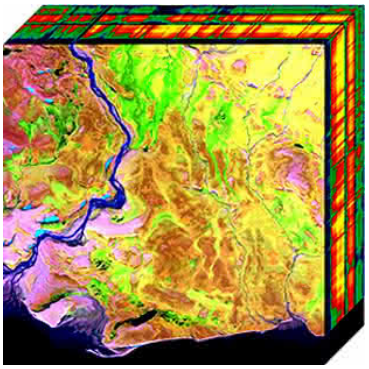


Figure 6.1. – Exemple d'image hyperspectrale (wikimedia).

Il est communément admis, dans la communauté du traitement du signal, que ce qui distingue les deux types d'imagerie ce sont les caractéristiques d'échantillonnage spatial et spectral. L'imagerie hyperspectrale est considérée comme ayant un échantillonnage en longueur d'onde plus élevé, de l'ordre de $L = 1000$ à $10\,000$ points sur l'intervalle considéré, alors que l'on dispose plutôt entre 10 et 100 points pour l'imagerie multispectrale. De même, l'échantillonnage spatial est meilleur (grossièrement d'un rapport 2 à 4) pour l'imagerie multispectrale.

6.1.1. Physique d'acquisition et instrumentation

Étant donné qu'il n'existe pas de détecteurs 3D, les images multispectrales et hyperspectrales ne sont pas acquises directement. Le rayonnement électromagnétique est nécessairement transformé

pour pouvoir être projeté sur le détecteur. Ces transformations peuvent être plus ou moins importantes ou visibles en fonction de la technologie ou du domaine d'application.

Le principe de l'observation multispectrale ou large bande, bien qu'il y ait des exceptions, est habituellement le plus simple. Après le système optique formant une image sur le plan focal, l'instrument contient des filtres en longueur d'onde, et le rayonnement est intégré sur la longueur d'onde par un détecteur 2D, typiquement une matrice CCD.

L'imagerie hyperspectrale est plus complexe, car l'objet d'intérêt est tout d'abord le spectre [1]. La présentation ci-dessous se limite aux techniques les plus courantes. Un composant essentiel sera l'élément dispersif, tel un prisme, qui donnera pour une source monochromatique une position spatiale différente, ou un angle de sortie différent, en fonction de la longueur d'onde. Les premiers instruments mesuraient les spectres point par point avec une reconstruction par balayage.

1: La différence de technologie explique la différence d'échantillonnage. Les images multispectrales ayant une largeur de bande spectrale plus importante peuvent diminuer la surface collectrice à SNR équivalent.

La technologie dite *pushbroom*, illustrée figure 6.2, a permis un bond important. Le principe est de placer une fente sur le plan focal d'un système d'imagerie pour supprimer une dimension spatiale, puis d'utiliser un système dispersif qui projettera la longueur d'onde dans la deuxième dimension spatiale du détecteur, maintenant inutilisée.

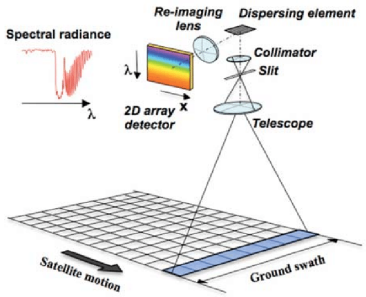


Figure 6.2. – Principe de la technique *pushbroom* (Del Bello, *et al.* (2017).

Dans le cas de capteurs aéroportés, cette dernière technologie est bien adaptée, mais le temps reste nécessaire pour acquérir les dimensions spatiales manquantes. Les techniques dites à champ intégral (*Integral Field Unit*, IFU), permettent d'observer un champ spatial 2D et non plus juste une ligne. Plusieurs technologies existent comme la matrice de microlentille ou l'utilisation de plusieurs « fentes ».

En conclusion, les images hyperspectrales sont déjà le résultat de traitement de données ayant pour objectif de reconstruire un cube 2D+λ puisque les données ne sont pas sous cette forme. Les images multispectrales quant à elles sont généralement utilisées telles quelles.

6.1.2. Problèmes méthodologiques

En termes de traitement de données on observe un ensemble de problèmes. Il y a tout d'abord des problèmes de traitement des données brutes pour reconstruire un cube échantillonné uniformément sur les trois dimensions $\boldsymbol{x}$. Ensuite, se pose des problèmes de déconvolution [55] par exemple. Enfin, il faut être capable d'utiliser le ou les cubes reconstruits $\boldsymbol{x}$ pour extraire des informations de haut niveau avec des problèmes de classification [17], de dé-mélange [46] ou encore de la caractérisation [23].

Dans le cas de l'imagerie multispectrale, le plus souvent le problème abordé concerne la segmentation en observation de la

terre et la photométrie en observation de l'espace. Par ailleurs, l'imagerie multispectrale a un bon échantillonnage spatial et peut également être « fusionnée » avec une image hyperspectrale par des méthodes dites de *pansharpening* [2], pour aboutir à une image HS avec un meilleur échantillonnage.

2: *pan* pour « tout », désignant une image en niveau de gris collectant toute la lumière visible.

Mon travail se place dans le cadre des *problèmes inverses*, à l'interface entre la mesure (ou l'instrumentation) et l'estimation. Je m'intéresse en particulier à la prise en compte de la physique d'acquisition pour la restauration ou la reconstruction des images hyper ou multispectrales. Je me suis notamment intéressé à deux aspects plus spécifiques. Tout d'abord aux effets de flou spatiaux *et* spectraux non stationnaires qui interviennent inévitablement lorsque la longueur d'onde se trouve dans l'infrarouge et au-delà, et deuxièmement à la combinaison entre ces modèles et les représentation en sous-espace pour régulariser le problème.

6.1.3. Limites et cas du JWST

Le cas d'étude des travaux est le James Webb Space Telescope* (JWST), observatoire spatial et successeur du Hubble Space Telescope (HST). Le JWST est un projet spatial hors norme, le plus cher de l'histoire avec un budget d'au moins 10 milliards de dollars, sans compter les coûts de fonctionnement de la mission. Une des particularités du JWST est la taille de son miroir, de 6,5 m, et dépliant, car trop grand pour la coiffe d'Ariane 5, voir figure 6.3. Plus précisément, avec l'IAS, nous travaillons sur le *Mid Infrared Instrument*† (MIRI) qui comprend un imageur large bande produisant des images multispectrales, et plusieurs IFU produisant des images hyperspectrales entre 5 et 25 µm.

Figure 6.3. – Le miroir plié du JWST.

Flou variable

Plusieurs problèmes se posent dans ce cas. Contrairement au cas usuel considéré en imagerie multispectrale et hyperspectrale, c'est-à-dire en télédétection et observations de la terre, le fait d'observer dans l'infrarouge fait que les phénomènes de diffraction ne sont plus négligeables étant donné les pas d'échantillonnage rencontrés. Les systèmes produisent alors des résultats flous avec une réponse qui dépend de la longueur d'onde λ. Dans le cas d'une formation d'image, la réponse de l'optique [3] au rayonnement $f(x, y)$ peut être vue comme une convolution 2D

3: dans le cas des hypothèses petits angles et de la diffraction de Fraunhofer.

$$g(x, y, \lambda) = \iint f(x', y', \lambda) h(x - x', y - y', \lambda) \, \mathrm{d}x' \, \mathrm{d}y' \qquad (6.1)$$

où g est l'image sur le plan focal et h la réponse, ou PSF pour *Point Spread Function*. Dans le cas de la formation d'un spectre $s(\lambda)$

*. www.jwst.fr

†. https://sci.esa.int/web/jwst/-/46826-miri-the-mid-infrared-instrument-on-jwst

à partir de la dispersion d'une entrée $e(\lambda)$, le problème est plus important, car le système est alors non stationnaire avec une sortie

$$s(\lambda) = \int e(\lambda')h(\lambda, \lambda')\,\mathrm{d}\lambda'. \tag{6.2}$$

La variation de la PSF au premier ordre est un facteur de dilatation avec

$$h(x, y, \lambda) = h_0\left(x\frac{\lambda_0}{\lambda}, y\frac{\lambda_0}{\lambda}\right) \tag{6.3}$$

avec h_0 la PSF à la longueur d'onde λ_0 choisie arbitrairement. Dans ce cas on voit que pour le JWST, la PSF à 25 µm est cinq fois plus large qu'à 5 µm, ce qui est considérable pour une analyse conjointe des images. La figure 6.4 montre cet effet dans le cas des PSF du JWST. Cela introduit une différence de résolution [4] dans les mesures, qu'il est nécessaire de prendre en compte pour des analyses quantitatives.

4: J'utilise « résolution » pour désigner la capacité à distinguer deux points proches, dans des images au contenu spectrale limité en hautes fréquences (absence de Dirac).

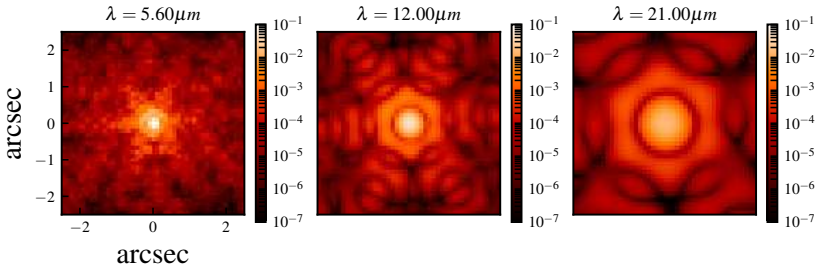


Figure 6.4. – Exemple de PSF du JWST à trois longueurs d'onde en échelle logarithmique (Hadj-Youcef 2018). Les PSF sont calculées à l'aide de WebbPSF. Leur structure spatiale dépend de la forme du miroir.

Fusion et mesures hétérogènes

Le deuxième problème concerne la reconstruction de l'image hyperspectrale $\boldsymbol{x}$ à partir d'une collection de mesures indirectes $\boldsymbol{y}_i$ avec plusieurs instruments. En effet, les effets instrumentaux sont nombreux, divers, et dans le cas du JWST, il n'existe pas un unique IFU, mais plusieurs regroupés sous le nom de MIRI.

Dans ce cas, un problème de fusion de mesures hétérogènes se pose dans le sens où

- les réponses spatiales ou spectrales ne sont pas identiques,
- les champs de vues ne sont pas identiques,
- l'échantillonnage spatial et spectral est différent selon les instruments.

Les parties suivantes présentent les travaux menés sur ces questions. Tout d'abord avec la thèse d'Amine Hadj-Youcef, puis celle de Ralph Abirizk. Je ne présente que succinctement ces travaux et renvoie aux publications de revues ou thèses pour les détails.

6.2. Reconstruction HS à partir d'images larges bandes

Comme mentionné plus haut, les images multispectrales en infrarouge et au-delà sont affectés par des effets de flous spatiaux dus à la diffraction. Ce flou dépend de la longueur d'onde, mais les mesures résultent d'une intégration sur une grande plage de longueur d'onde. La figure 6.5 montre le cas de l'imageur MIRI.

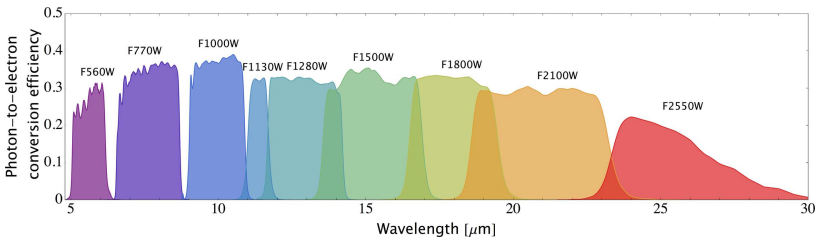


Figure 6.5. – Filtre en longueur d'onde de l'imageur MIRI (NASA). L'imageur dispose de neuf filtres large bande de 5 à 25 µm, avec des recouvrements ou non.

Dans la littérature, plusieurs approches ont été développées. Tout d'abord des traitements bande par bande en considérant une PSF différente pour chaque bande [57], [63], ce qui amène à un problème de déconvolution d'image [66]. D'autres travaux peuvent considérer des corrélations entre les images avec le travail séminal [90], par le biais de régularisation. D'autres approches, plus empiriques et dans un objectif de comparaison entre les images, consistent à dégrader les images à haute résolution [34], [35], problème appelé « homogénéisation ». Plusieurs problèmes se posent avec ces approches.

1. Tout d'abord il n'y a pas de PSF stationnaire pour les images large bande puisqu'elle dépend du spectre de chaque source ou pixel. De même, il n'existe pas de modèle unique de PSF variable. Dit autrement, l'effet de flou spatial dépend de la source.
2. Un traitement bande par bande aboutit à des résultats pour chaque bande qui peuvent toujours avoir des pouvoir de résolution au sens de Rayleigh différentes ; le bruit, les PSFS, etc. étant *a priori* différents. Si l'objectif est de les comparer entre elles pour de la photométrie par exemple, alors cela ne sert à rien.
3. Le problème posé par l'homogénéisation est évident puisqu'il va à l'inverse de l'objectif recherché par les développements instrumentaux, notamment augmenter la taille du miroir pour améliorer la résolution.

L'approche développée est originale, car elle consiste à remonter plus amont dans le problème et poser directement la question comme étant de reconstruire le cube 2D+λoriginal à partir de mesures, plutôt que de restaurer des images à partir d'images.

6.2.1. Modèle direct

Sans rentrer dans les détails, à partir d'une image hyperspectrale $\boldsymbol{x}[i, j, l]$ on obtient un ensemble d'images 2D $\boldsymbol{y}_p$ comme

$$\boldsymbol{y}_p[i, j] = \sum_l \boldsymbol{\omega}_p[l] \sum_{i',j'} \boldsymbol{x}[i', j', l]\boldsymbol{h}[i - i', j - j', l] \tag{6.4}$$

avec (i, j) les indices spatiaux et l l'indice spectral. Ainsi chaque image $\boldsymbol{y}_p$ est obtenue à partir d'un même cube $\boldsymbol{x}$ 2D+λ, avec une convolution par des PSF $\boldsymbol{h}$ qui dépendent de la longueur d'onde. Le résultat est ensuite intégré en longueur d'onde après une pondération par le filtre $\boldsymbol{\omega}_p$, avec un exemple figure 6.6.

Dans un premier temps nous avons proposé [26] un nouveau modèle avec une modélisation plus fine des spectres. Ce modèle repose sur une décomposition continue linéaire par morceaux du spectre, comme illustré figure 6.7, où les nœuds $\boldsymbol{x}_k$ ne sont pas disposés régulièrement et ne sont pas associés à un filtre en particulier.

On obtient alors une sortie modèle comme

$$\boldsymbol{y}_p = \sum_k \boldsymbol{H}_{p,k}\boldsymbol{x}_k \tag{6.5}$$

où $\boldsymbol{H}_{p,k}$ ne sont pas les PSF monochromatiques, mais des PSF intégrées correspondant aux noeuds k et l'image p. Ce modèle est linéaire, peut s'écrire $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$, et prend en considération les points 1 et 2 plus haut.

Cependant le problème de reconstruction, reposant par exemple sur un moindre carré

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2, \tag{6.6}$$

est sévèrement sous déterminé, car on cherche K nœuds à partir de $P \ll L$ images large bande. Par ailleurs, la présence de convolutions spatiales induit des problèmes de mauvais conditionnement et donc d'instabilité. Le problème inverse est mal-posé [89].

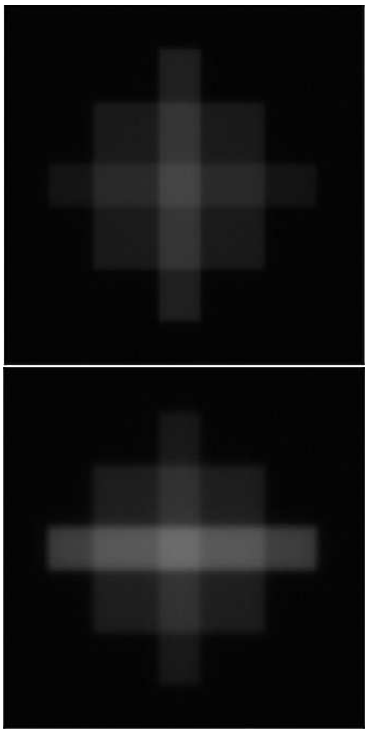


Figure 6.6. – Exemple synthétique de la simulation de deux images large bande $\boldsymbol{y}_p$ aux longueurs d'onde centrées en 10 et 20 µm. L'auguementation du flou est visible ainsi que la proportion variable des composantes formant les figures géométriques.

[26]: Hadj-Youcef et al. (2017), « Restoration from Multispectral Blurred Data with Non-Stationary Instrument Response »

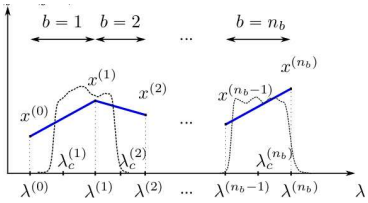


Figure 6.7. – Décomposition linéaire par morceau dans la dimension spectrale [26].

6.2.2. Régularisation et approximation par sous-espace

Nous avons proposé deux approches. La première consiste à ajouter une régularisation de douceur pour garantir l'existence et la stabilité.

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2 + \mu\|\boldsymbol{D}_{i,j}\boldsymbol{x}\|^2 + \mu\|\boldsymbol{D}_l\boldsymbol{x}\|^2. \tag{6.7}$$

Cette solution, ainsi que le modèle direct, a fait l'objet d'une publication de conférence [26]. Les résultats, obtenus par minimisation du critère par gradients conjugués, sont cependant un peu décevants, les données n'étant pas assez informatives.

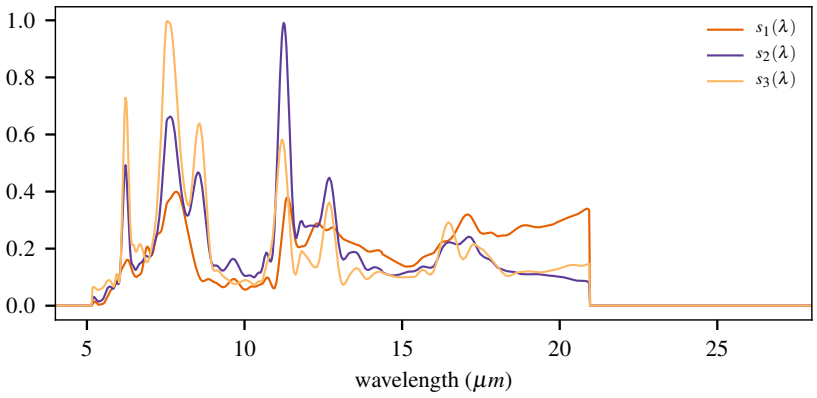


Figure 6.8. – Exemple de motifs spectraux utilisés [72]. Ces motifs sont issus d'une NMF sur des données de Spitzer/IRS, ancien IFU mesurant dans le même domaine de longueur d'onde que le JWST.

La deuxième approche, qui a fait l'objet d'un article de journal [15], est plus performante et repose sur une approximation par sous-espace

$$\boldsymbol{x}[i,j,l] = \sum_k \boldsymbol{a}_k[i,j]\boldsymbol{s}_k[l] \tag{6.8}$$

classique en imagerie hyperspectrale. Ce modèle est le plus souvent employé pour de la séparation de source ou de la segmentation. Nous l'utilisons plutôt comme moyen de régularisation en considérant les motifs spectraux $\boldsymbol{s}_k$ connus [72] et les coefficients de mélange $\boldsymbol{a}_k$ inconnus.

Dans ce cas le modèle direct devient

$$\boldsymbol{y}_p = \begin{bmatrix}\boldsymbol{H}_{p,1} & \dots & \boldsymbol{H}_{p,K}\end{bmatrix} \begin{bmatrix}\boldsymbol{a}_1 \\ \vdots \\ \boldsymbol{a}_K\end{bmatrix} = \sum_k \boldsymbol{a}_k * \boldsymbol{h}_{k,p} \tag{6.9}$$

où chaque bloc $\boldsymbol{H}_{p,k}$ correspond à la matrice de passage du coefficient k dans la bande p et correspond à une convolution spatiale, noté $*$, avec une PSF équivalente $\boldsymbol{h}_{k,p}$. L'article [15] détaille l'élaboration de ce modèle.

[15]: Hadj-Youcef et al. (2020), « Fast Joint Multiband Reconstruction from Wideband Images Based on Low Rank Approximation »

Ce modèle est puissant, car il répond très efficacement aux points 1 et 2 plus haut.

1. Chaque spectre vit dans un espace de dimensions réduit, plus faible que le nombre d'images large bande $K \leq P$. On n'observe plus de sous détermination.
2. Par ailleurs, le modèle est exact dans le sens où il correspond exactement à une convolution spatiale monochromatique pour chaque longueur d'onde. Les PSF effectives $\boldsymbol{h}_{k,p}$ décrivent l'impact spatial de chaque contenu spectral k dans les bandes p.
3. Enfin, le modèle introduit naturellement une corrélation entre les images large bande qui vient du modèle. Dit autrement, si les données de la bande $p = 1$ sont expliquées par une amplitude plus importante du motif $k = 2$, cette explication doit rester compatible avec les autres bandes (si le motif est non nul dans cette bande).

Le problème à résoudre dans ce cas s'écrit

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|^2 + \sum_k \mu_k R(\boldsymbol{a}_k) \tag{6.10}$$

avec $R(\boldsymbol{a}) = \sum_c \phi(\boldsymbol{d}_c^t \boldsymbol{a})$, où $\boldsymbol{d}_c$ produit une combinaison linéaire de pixels (comme la différence entre pixels voisins) et ϕ une fonction dérivable convexe, par exemple la fonction de Huber [81]. Cette régularisation est motivée pour rendre le problème de restauration spatiale bien posé tout en préservant les contours, avec un algorithme rapide décrit dans la partie suivante.

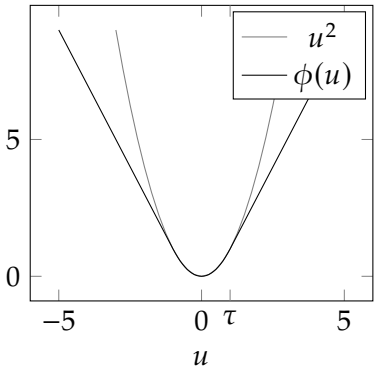


Figure 6.9. – La fonction de Huber ϕ.

6.2.3. Algorithme de reconstruction SQ

Le critère équation (6.10) est dérivable et suit une structure Semi-Quadratique (SQ) dans le cas du choix de la fonction de Huber. Les travaux de Geman et Reynolds [88] (GR) suivi de Geman et Yang [85] (GY) ont effectivement proposé de construire un critère semi-quadratique augmenté K tel que

$$\min_{\boldsymbol{b}\in\boldsymbol{B}} K(\cdot, \boldsymbol{b}) = J(\cdot) \tag{6.11}$$

avec $\boldsymbol{b}$ des variables auxiliaires. Dans ce cas, il est possible de manipuler le critère par rapport à $\boldsymbol{x}$ et $\boldsymbol{b}$ respectivement, dans un schéma de minimisation alterné [81], [83] ou échantillonnage de Gibbs [85], [88]. Dans le cas de l'optimisation, les deux formes GR et GY, appliquées au critère (6.10), donnent

$$K_{GR}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|^2 + \sum_k \mu_k \sum_c \left(b_c \left(\boldsymbol{d}_c^t \boldsymbol{a}_k \right)^2 + \psi(b_c) \right) \tag{6.12}$$

et

$$K_{GY}(\boldsymbol{a}, \boldsymbol{b}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|^2 + \sum_k \mu_k \sum_c \left(\frac{1}{2} \left(\boldsymbol{d}_c^t \boldsymbol{a}_k - b_c \right)^2 + \xi(b_c) \right). \tag{6.13}$$

La minimisation alternée conduit à la résolution d'un problème quadratique pour $\boldsymbol{a}$ et indépendant pour $\boldsymbol{b}$. Ces critères ont été beaucoup étudiés [56], [70], [74], [75], [81], [83]. Par exemple Nikolova et coll. [74] montrent le lien avec les algorithmes reposant sur la linéarisation du gradient (*Iterative Reweighted Least-Square*, IRLS). Labat et coll. [70] ont également montré que les algorithmes SQ sont des algorithmes de gradients préconditionnés, avec préconditionneur variable, et à pas fixe de valeur 1. En particulier, il n'est pas nécessaire de résoudre exactement le problème quadratique pour garantir la convergence [73]. Plus récemment, [56], [70] ont montré qu'il est plus intéressant de voir les algorithmes SQ comme des algorithmes de Majorisation-Minimisation Quadratique et d'utiliser cette approche pour une recherche de pas dans un gradient conjugué non linéaire.

Dans le cas de la fusion d'image large bande, nous avons proposé dans [15] d'utiliser plutôt l'approche GY, car le système quadratique peut être résolu exactement en utilisant des convolutions circulantes (bien que $\boldsymbol{H}$ ne le soit pas). En effet, en repartant de l'équation 6.13 on a

[15]: Hadj-Youcef et al. (2020), « Fast Joint Multiband Reconstruction from Wideband Images Based on Low Rank Approximation »

$$K_{GY}(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|^2 + \|\boldsymbol{D}\boldsymbol{a} - \boldsymbol{b}\|^2 \tag{6.14}$$

avec des matrices bloc-circulantes. Dans ce cas le minimiseur

$$\hat{\boldsymbol{a}} \propto \left(\boldsymbol{H}^t\boldsymbol{H} + \boldsymbol{D}^t\boldsymbol{D}\right)^{-1}\left(\boldsymbol{H}^t\boldsymbol{y} + \boldsymbol{D}^t\boldsymbol{b}\right) \tag{6.15}$$

nécessite l'inversion de la matrice normale $\boldsymbol{Q} = \boldsymbol{H}^t\boldsymbol{H} + \boldsymbol{D}^t\boldsymbol{D}$ qui peut être menée avec une approximation circulante des convolutions présente dans $\boldsymbol{H}$ et $\boldsymbol{D}$, voir [15] appendice A pour la démonstration. L'approche GY permet de rendre cette matrice $\boldsymbol{Q}$ indépendante de $\boldsymbol{b}$ et l'inversion peut alors être précalculée une seule fois et réutilisée à chaque mise à jour des variables $\boldsymbol{b}$ dans l'algorithme SQ.

L'application de $\boldsymbol{Q}^{-1}$ est peu coûteuse, par conséquent l'algorithme est très rapide et ne prend que quelques secondes pour l'estimation de deux à trois images de coefficients et de taille 256×256 sur un ordinateur portable standard. La qualité des résultats est excellente, à la fois sur la restauration spatiale et spectrale. La figure 6.10 montre un résultat tiré de l'article.

Enfin, un des objectifs de ce travail est également d'avoir, dans la mesure du possible, une résolution spatiale homogène à toutes les longueurs d'onde et de faire bénéficier aux grandes longueurs d'onde la bonne résolution des longueurs d'onde plus courtes. La figure 6.11 illustre cet effet sur la fonction de transfert optique (*Optical Transfert Function*) calculée comme

$$\mathrm{otf}_\nu^{\mathrm{equiv}} = \frac{\overset{\circ}{\hat{\boldsymbol{x}}}_\nu}{\overset{\circ}{\boldsymbol{x}}_\nu} \tag{6.16}$$

où ν désigne la fréquence spatiale, $\overset{\circ}{\boldsymbol{x}}$ la transformée de Fourier de la vérité terrain et $\overset{\circ}{\hat{\boldsymbol{x}}}_\nu$ la transformée de Fourier de l'estimation.

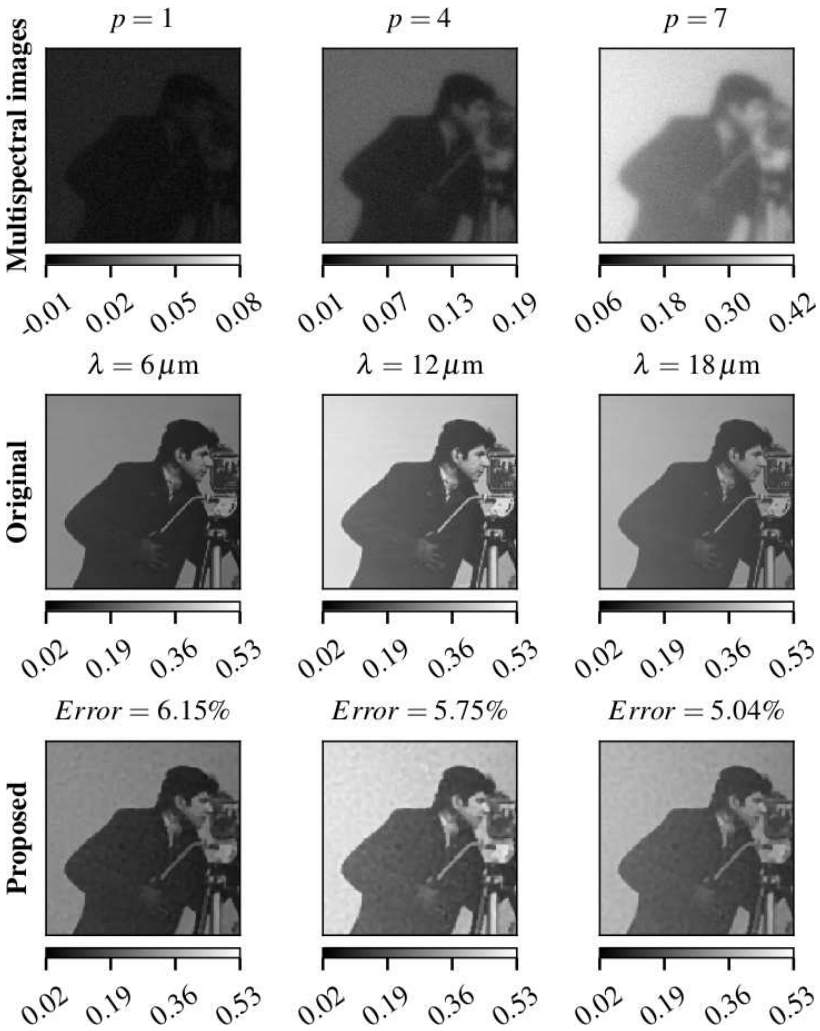


Figure 6.10. – Résultat sur le cameraman avec un gradient spatial tiré de [15]. La ligne du milieu correspond à la vérité terrain à trois longueurs d'onde. La ligne du haut correspond à trois images mesurées dans les bandes contenant ces longueurs d'onde. On voit que les dynamiques ne sont pas retrouvées avec des contrastes changeants. Par ailleurs le flou augmente avec la longueur d'onde et du bruit est visible. La ligne du bas correspond à la proposition. On observe une déconvolution très nette, homogène sur la longueur d'onde avec une bonne dynamique. La figure de l'article est plus complète avec des comparaisons avantageuses pour la méthode proposée.

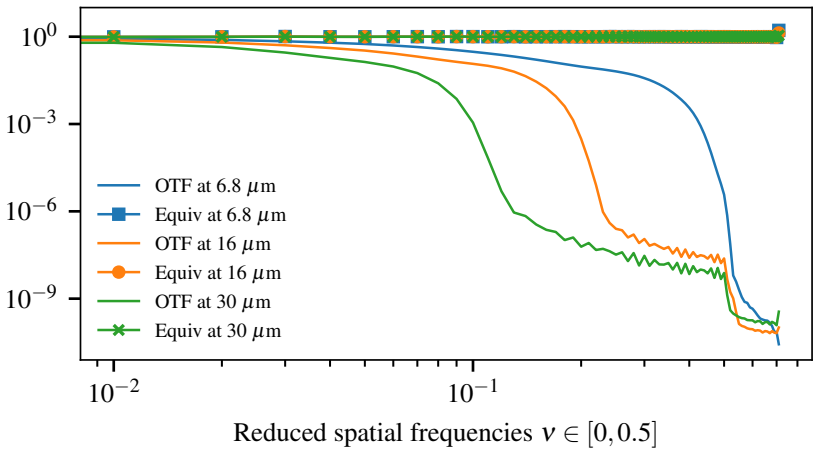


Figure 6.11. – Fonction de transfert optique équivalente, en moyenne radiale ([15]). La figure montre trois OTF originales à trois longueurs d'onde. On observe bien l'effet passe-bas qui augmente avec la longueur d'onde. La figure montre les OTF équivalentes au mêmes longueurs d'onde. Dans ce cas favorable on voit que non seulement la déconvolution est efficace, mais quasiment identique.

6.3. Reconstruction et fusion d'images HS

Dans cette partie je présente les travaux de Ralph Abirizk. Certains éléments, notamment sur le modèle direct ont fait l'objet d'une publication à ICIP 2020 [11]. D'autres, plus récents, font l'objet d'un article de journal [2].

Une des limites des travaux ci-dessus est la connaissance des motifs spectraux s_k. Il est possible d'utiliser des mesures d'instruments précédents, mais l'exercice est vite limité. Par exemple, l'intervalle d'observation spectrale n'est pas nécessairement identique [5], et des mesures ne sont pas nécessairement disponibles pour toutes les observations.

L'idée est donc d'utiliser des mesures d'IFU pour combler ce manque d'information. Malheureusement dans le cas du JWST/-MIRI, il n'y a pas un, mais plusieurs IFU, qui souffrent par ailleurs de plusieurs limitations.

1. Tout d'abord ces instruments souffrent également du problème de flou spatial, dépendant de la longueur d'onde.
2. La diffraction introduit également un problème de flou spectral mentionné équation (6.2). En outre la réponse dépend de la longueur d'onde ‡ mais également de la position spatiale de la source comme le montre la figure 6.12, ce qui entraîne une dégénérescence.
3. Dans le cas du JWST/MIRI il y a également des problèmes de repliement de spectre dus à un échantillonnage spatial insuffisant. L'instrument utilise alors plusieurs pointages avec décalage ce qui conduit à un problème de super résolution.
4. Enfin les pas d'échantillonnages spatiaux et spectraux sont différents entre les IFU.

En conclusion, il convient de fusionner les données provenant de ces différents IFU qui ne couvrent pas tous exactement les mêmes champs de vue spatiaux et spectraux §.

[11]: Abirizk et al. (2020), « Non-Stationary Hyperspectral Forward Model and High-Resolution »

[2]: Abirizk et al. (2022), « Super-Resolution Hyperspectral Reconstruction with Majorization-Minimization Algorithm and Low-Rank Approximation »

5: on voit figure 6.8 que les motifs issus de Spitzer ne vont pas jusqu'à 25 µm.

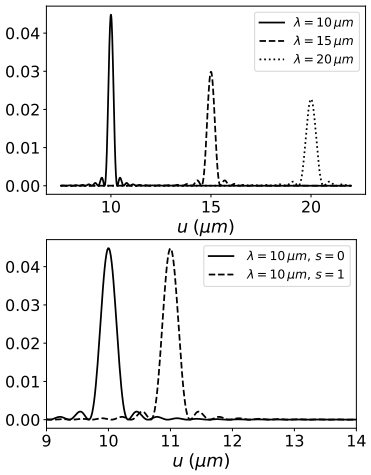


Figure 6.12. – Réponse spectrale d'un réseau de diffraction ([11]) sur un capteur de dimension spatiale u. La figure du haut montre le flou non stationnaire avec trois sources monochromatiques. La figure du bas montre la réponse pour une même source monochromatique, mais avec une position spatiale différente.

6.3.1. Modèle direct

Le modèle instrument est significativement plus complexe. Un IFU partage le même miroir que l'imageur. Par contre des fentes sont présentes sur le plan focal, suivi d'un élément dispersif pour finir par le détecteur. Dans [11] et dans la thèse actuelle de R. Abirizk nous avons abouti au modèle

$$\boldsymbol{y}_{c,p}[i',l'] = \sum_{i,j}\sum_{l\in\mathcal{L}_c}\left(\boldsymbol{x}[i,j,l] \underset{i,j}{*} \boldsymbol{h}[i,j,l]\right) \boldsymbol{w}_c[i-i_p, j-j_p, l] \times \boldsymbol{h}_{c,j,p}[l',l]. \quad (6.17)$$

‡. heureusement

§. jwst-docs.stsci.edu/mid-infrared-instrument/miri-observing-modes/miri-medium-resolution-spectroscopy/

où i' est l'axe spatial après convolution par $\boldsymbol{h}$, $\boldsymbol{y}$ sont les mesures 2D pour un canal c et un pointage p, $\boldsymbol{x}$ l'image HS inconnue et $\boldsymbol{h}$ la PSF du miroir, $\boldsymbol{w}_c$ le fenêtrage spatiospectral du canal [6] c, $\boldsymbol{h}_{c,j,p}$ la réponse du réseau de diffraction qui dépend du canal ainsi que de la position de la source dans la fente, déterminée par j et le pointage p, et enfin le pointage déterminé par (i_p, j_p). La somme sur l correspond au flou spectral. La somme sur i, j correspond à l'intégration capteur pour modéliser la super-résolution. Pour cela, toutes les dimensions (fenêtres, PSF, ...) sont multiples d'un pas d'échantillonnage fin choisi pour $\boldsymbol{x}$.

6: canal est un terme générique qui regroupe des paramètres instrumentaux spécifiques, comme une fente donnée.

Enfin, tout comme pour l'imageur, un modèle de mélange ou de réduction de dimension comme l'équation (6.8) page 33 a été considéré. Ce modèle présente à nouveau de nombreux avantages.

1. Le nombre d'inconnues est réduit ce qui favorise le rapport signal sur bruit.
2. La corrélation des données est mieux prise en compte puisque des mesures sur les courtes longueurs d'onde influent également sur les grandes longueurs d'onde. On peut par exemple reconstruire en dehors des champs de vue de certains IFU, ce qui n'est pas possible sans la fusion de données.
3. La résolution spectrale est fixée par les motifs spectraux ce qui permet la restauration des raies spectrales par exemple.
4. Si les motifs sont considérés connus, alors il y a moins d'hyperparamètres à régler.

6.3.2. Reconstruction par algorithmes MM

Le modèle direct équation (6.17) est complexe, mais reste linéaire. Avec le modèle de mélange linéaire, il s'agit d'un véritable problème de fusion de données.

La reconstruction des cartes d'abondance $\boldsymbol{a}$ peut s'écrire à nouveau comme la minimisation d'un critère mixte

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \|\boldsymbol{y} - \boldsymbol{H}_{HS}\boldsymbol{a}\|^2 + \sum_k \mu_k R(\boldsymbol{a}_k) \tag{6.18}$$

mais cette fois-ci le modèle d'observation $\boldsymbol{H}_{HS}$ est bien plus complexe et n'a pas de structure particulière, la formulation GY ne présentant pas d'intérêt particulier. Pour cela nous nous sommes orientés vers les évolutions récentes de ces algorithmes basés sur les principes de Majorisation-Minimisation Quadratique [56], [70], développés notamment au LS2N à Nantes.

Ces algorithmes permettent d'optimiser des critères sous la forme

$$J(\boldsymbol{a}) = \sum_q \mu_q \Psi_q(\boldsymbol{V}_q \boldsymbol{a}_q - \boldsymbol{\omega}_q) \tag{6.19}$$

avec $\boldsymbol{V}_q$ un opérateur linéaire, $\boldsymbol{\omega}_q$ un terme de données fixé, μ_q des hyperparamètres et $\Psi_q(\boldsymbol{u}) = \sum_i \phi_q(u_i)$. Par ailleurs, la fonction

scalaire ϕ doit être

1. $\mathscr{C}^1$, paire, coercive,
2. $\phi(\sqrt{\cdot})$ est concave sur $\mathbb{R}^+$,
3. et $0 < \dot{\phi}(u)/u < +\infty, \forall u \in \mathbb{R}$.

Il s'agit des mêmes conditions pour avoir une formulation GR. Ces structures permettent de définir des algorithmes efficaces reposant sur des majorantes tangentes quadratiques Q [75], [92]

$$Q\left(\boldsymbol{a}, \boldsymbol{a}^k\right) = J\left(\boldsymbol{a}^k\right) + \nabla J\left(\boldsymbol{a}^k\right)^t \left(\boldsymbol{a} - \boldsymbol{a}^k\right) + \frac{1}{2}\left(\boldsymbol{a} - \boldsymbol{a}^k\right)\boldsymbol{A}^{(k)}\left(\boldsymbol{a} - \boldsymbol{a}^k\right) \quad (6.20)$$

où

$$\boldsymbol{A}^{(k)} = \sum_q \mu_q \boldsymbol{V}_q^T \operatorname{diag}\left(\boldsymbol{b}_q^k\right)\boldsymbol{V} \quad (6.21)$$

et

$$\boldsymbol{b}_q^k = \frac{\dot{\phi}(\boldsymbol{V}_q\boldsymbol{x}^k - \boldsymbol{\omega}_q)}{\boldsymbol{V}_q\boldsymbol{x}^k - \boldsymbol{\omega}_q}.$$

Comme le critère d'origine est différentiable (et convexe si ϕ est convexe), l'optimisation peut être faite avec un gradient conjugué non linéaire [78], ou plus efficacement avec des méthodes à sous-espace. Dans ce dernier cas, l'algorithme itératif est

$$\boldsymbol{a}^{(k+1)} = \boldsymbol{a}^{(k)} + \alpha^{(k)}\boldsymbol{g}^{(k)} + \sum_{s=1}^{S} \beta^{(k,s)}\boldsymbol{d}^{(k-s)} \quad (6.22)$$

où $\boldsymbol{g}^{(k)} = \nabla J\left(\boldsymbol{a}^{(k)}\right)$ est le gradient au point courant, $\boldsymbol{d}^{(k-s)}$ les directions de recherche précédentes, $\alpha^{(k)}$ et $\beta^{(k,s)}$ des scalaires.

Nous avons utilisé l'algorithme 3MG (pour *Majorize-Minimize Memory Gradient*) de Chouzenoux et coll. [56] qui exploite la structure du critère (6.14) pour calculer le pas α et le paramètre de conjugaison $\beta^{(s)}$ par Majorisation-Minimisation Quadratique, avec démonstration de la convergence. En effet, l'équation (6.22) peut être écrite comme

$$\boldsymbol{a}^{(k+1)} = \boldsymbol{a}^{(k)} + \boldsymbol{D}^{(k)}\boldsymbol{s}^{(k)} \quad (6.23)$$

où $\boldsymbol{D}^{(k)}$ est le sous-espace de dimension S, contenant le gradient courant et les directions précédentes, et $\boldsymbol{s}^{(k)}$ un vecteur de pas de dimension S. Contrairement à une recherche de pas traditionnelle, ces derniers peuvent être obtenus avec une formule explicite.

$$\boldsymbol{s}^{(k)} = {\boldsymbol{B}^{(k)}}^{-1}\nabla J^{(k)}\left(\boldsymbol{a}^{(k)}\right) \quad (6.24)$$

avec $\boldsymbol{B}^{(k)} = {\boldsymbol{D}^{(k)}}^t\boldsymbol{A}^{(k)}\boldsymbol{D}^{(k)}$ une matrice $S \times S$. L'équation (6.24) correspond à la première itération de minimisation d'une majorante quadratique de dimension S tangente aux points courant $\boldsymbol{a}^{(k)}$.

[2]: Abirizk et al. (2022), « Super-Resolution Hyperspectral Reconstruction with Majorization-Minimization Algorithm and Low-Rank Approximation »

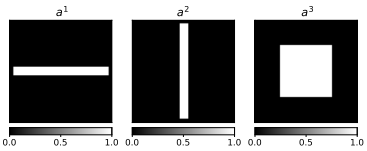


Figure 6.13. – Coefficients de mélange pour la croix.

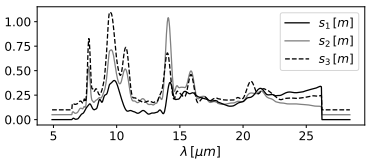


Figure 6.14. – Motifs spectraux pour la croix, tiré de mesures de Spitzer/IRS [72].

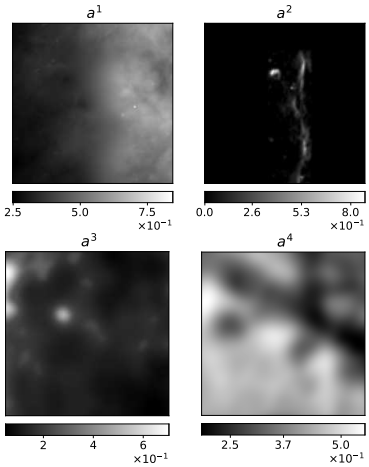


Figure 6.15. – Coefficients de mélange pour la barre d'Orion [14]. Ils correspondent à la répartition spatiale de différents composés chimiques.

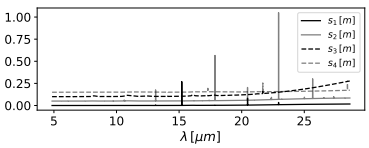


Figure 6.16. – Motifs spectraux pour la barre d'Orion [14]. Les motifs sont décalés verticalement pour une meilleure lecture. Ils correspondent à différents composés chimiques.

6.3.3. Résultats

La méthode a été appliquée à l'instrument MIRI du JWST. Les données ont été simulées avec le modèle direct, une validation plus aboutie avec par exemple le simulateur officiel devant faire l'objet de travaux futurs. Il s'agit encore de résultats de l'article de journal [2].

Les tests ont été faits sur deux exemples de données. Le premier est semi-synthétique avec des motifs spectraux, figure 6.14, issus de mesures de Spitzer/IRS. Les cartes de coefficients de mélange, figure 6.13, représentent une croix facilitant la lecture des résultats. Le deuxième exemple, figures 6.15 et 6.16, est plus réaliste et provient de simulations [14], obtenues à partir de mesures réelles, de la barre d'Orion.

Les figures 6.17 et 6.18 montrent quelques résultats préliminaires obtenus sur l'objectif final, c'est-à-dire la reconstruction du cube hyperspectral et non la reconstruction des cartes de coefficients. Il n'existe pas d'algorithme existant avec le même objectif, c'est-à-dire la fusion de données hétérogènes pour l'imagerie HS. Nous comparons donc à une méthode standard qui consiste à appliquer l'algorithme de base *Shift & Add* [80] de super-resolution[7], qui permet de générer un sous-cube HS pour chaque canal, suivi d'une déconvolution. avec une *a priori* TV spatiale, sur chaque image monochromatique. L'objectif, en quelque sorte, est de prendre en compte la convolution spatiale, mais sans modèle instrument complet ni modèle de sous-espace. Il faut noter que l'on ne peut pas réellement comparer aux mesures brutes qui ne vivent pas dans le même espace.

Les figures montrent clairement l'intérêt de la prise en compte du modèle instrument. Tout d'abord, une seule image HS est reconstruite avec l'ensemble des mesures. De plus le modèle de sous-espace permet de prendre en compte les corrélations des mesures de différents instruments pour améliorer la résolution spatiale, de manière homogène, sur tout l'intervalle de longueur d'onde.

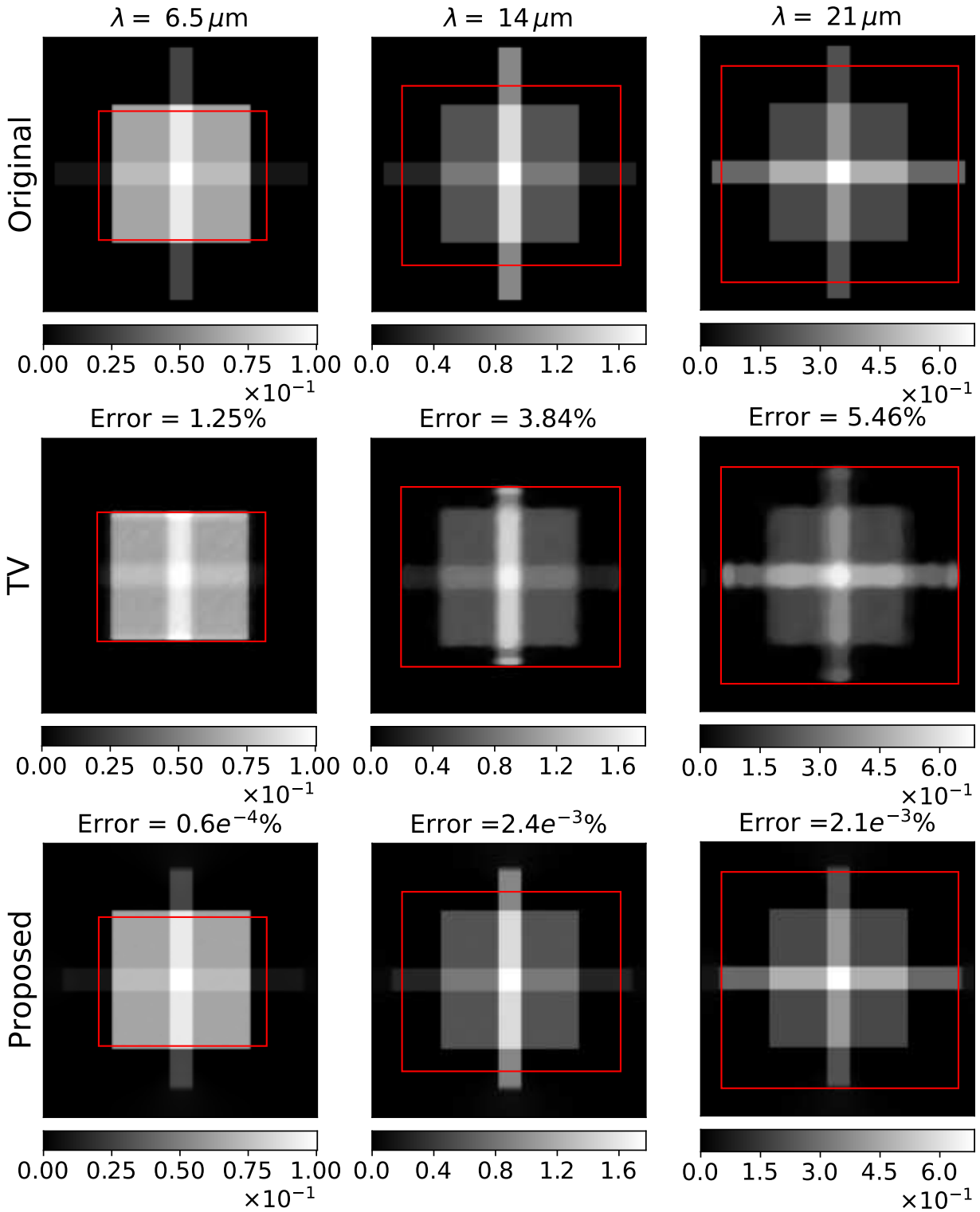


Figure 6.17. – Reconstruction de la croix. La première ligne montre l'objet vrai à trois longueurs d'onde observées par trois canaux (instrument) différents. On observe une proportion différente de chaque élément motif. Le carré rouge montre le champ de vue pour chaque canal. La deuxième ligne montre le résultat de la méthode standard basée uniquement sur la super-résolution – déconvolution. On observe également une dégradation de la résolution spatiale avec les longueurs d'onde croissantes. La troisième ligne est la méthode proposée. Un unique objet est estimé avec un champ de vue total, notamment dans la première colonne, correspondant au plus grand champ de vue. Par ailleurs la reconstruction est bonne y compris dans ces zones alors qu'il n'y a pas de mesures directes. Enfin la résolution spatiale est beaucoup plus homogène à toutes les longueurs d'onde puisque celle-ci est déterminée par la résolution des coefficients de mélange.

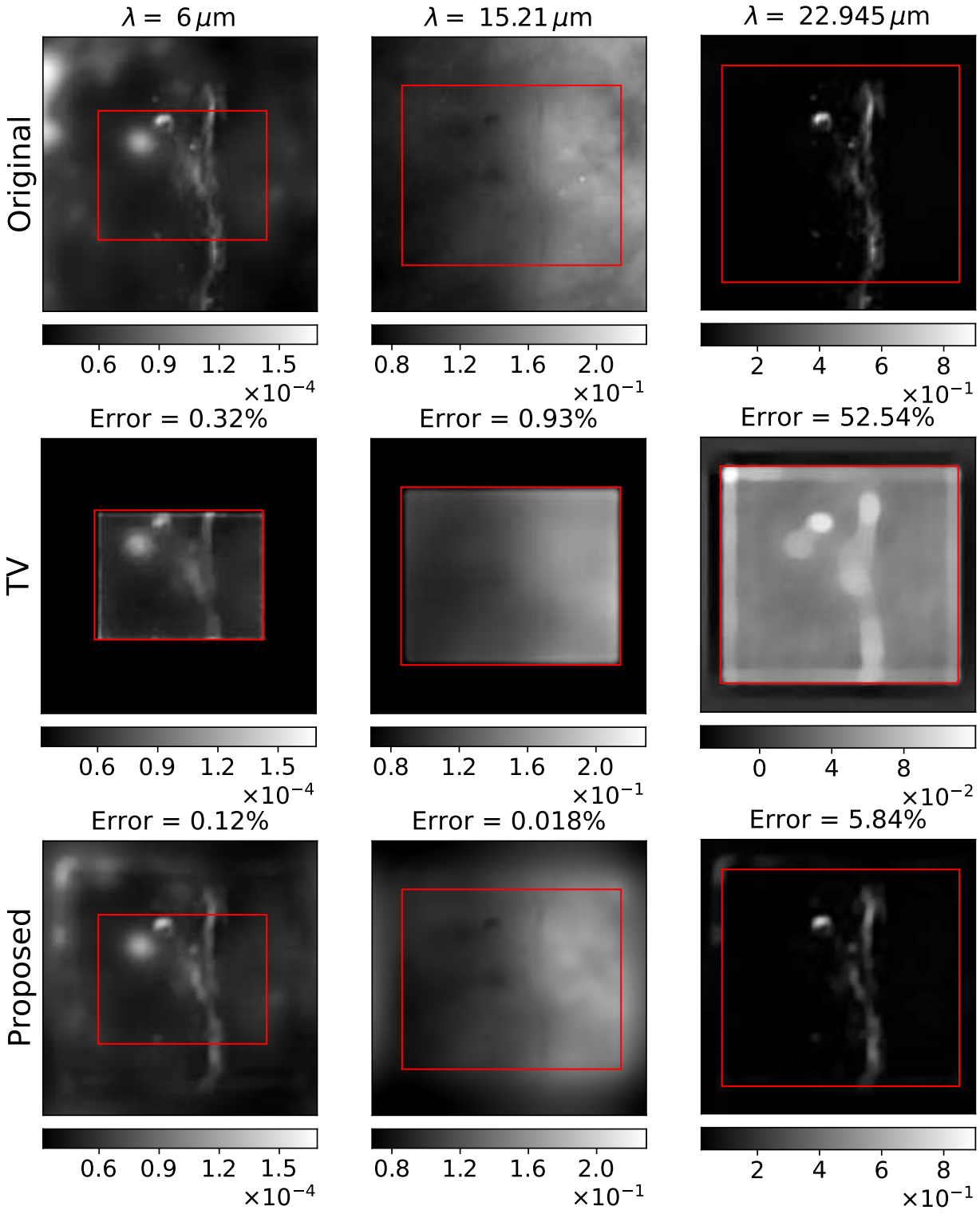


Figure 6.18. – Reconstruction de la barre d'Orion à trois longueurs d'onde. La première ligne correspond à la vérité terrain, la deuxième aux données « déconvolées » et la troisième à la méthode proposée. Le carré rouge correspond au champ de vue des instruments aux longueurs d'ondes considérées. Globalement les commentaires de la figure 6.17 s'appliquent. On peut ajouter que la prise en compte d'un modèle de sous-espace ou de mélange améliore significativement les résultats à 22 µm qui souffre d'une contamination importante due au flou spectral introduit par les instruments. Par contre les sources ponctuelles visibles à 15 µm ne sont pas correctement restaurées.

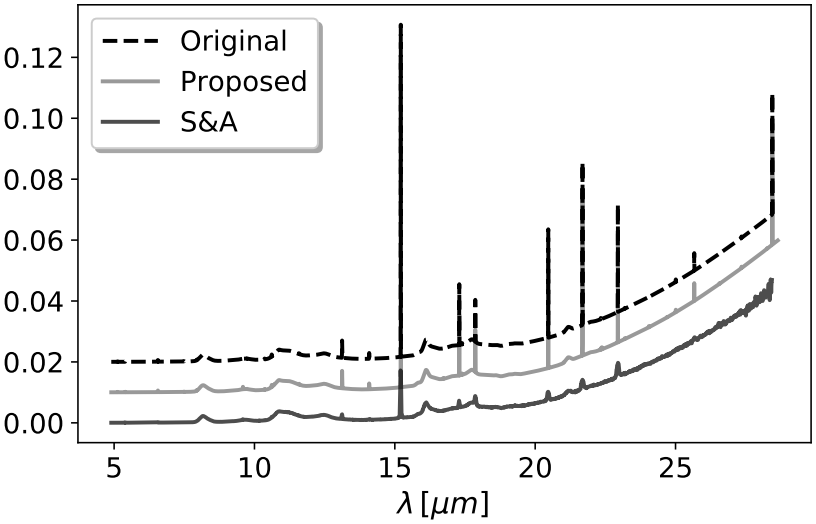


Figure 6.19. – Reconstruction du spectre pour un pixel de la barre d'Orion. Les spectres sont décalés verticalement pour faciliter la lecture. On observe principalement l'effet de flou présent dans la méthode *Shift & Add* qui ne prend pas en compte les effets instrumentaux. La méthode proposée utilisant des motifs spectraux est bien évidemment aussi résolue que souhaité et sans bruit, par construction.

6.4. Conclusion

J'ai présenté dans ce chapitre une partie des travaux que j'ai menés sur l'imagerie hyperspectrale et multispectrale depuis 2015. Je me suis concentré sur les travaux pour lesquels j'ai mené un encadrement important, avec notamment la direction de la thèse de Ralph Abirizk.

Le chapitre suivant présente mes travaux sur un autre problème d'imagerie computationnelle en microscopie optique pour la biologie.

7: appelé habituellement *coaddition* en astronomie

Super-Résolution d'image en microscopie biologique

7.

Je présente ici les travaux menés en collaboration avec Jean-Christophe Olivo-Marin de l'Institut Pasteur dans l'équipe Analyse d'Images Biologiques, Alexandra Fragola et Vincent Loriette au LPEM de l'ESPCI, et Laurent Mugnier chercheur à l'ONERA. Ce sont des travaux sur des méthodes de reconstruction d'images en Microscopie par Illumination Structurée.

7.1. Introduction

7.1.1. La microscopie plein champ

La microscopie, optique notamment, fait l'objet de nombreux développements instrumentaux innovants. Une différence notable en microscopie, par rapport aux travaux du chapitre précédent, est que l'on maîtrise l'illumination de la scène, ce qui ouvre de nombreuses possibilités.

La microscopie *widefield*, ou plein champ, est l'instrumentation la plus simple. En éclairement par transmission, elle consiste à éclairer derrière l'échantillon, la lumière transmise à travers étant collectée par le microscope et envoyée dans les oculaires ou sur une caméra pour enregistrement. La figure 7.1 montre un exemple.

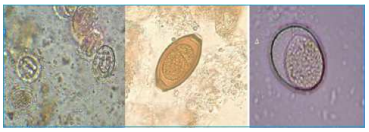


Figure 7.1. – Exemple d'image de microscopie plein champ par transmission (biofaculte.blogspot.com).

La microscopie par fluorescence repose sur l'utilisation de fluorophores, molécules fluorescentes comme la GFP (*Green Fluorescent Protein*) attachée à des molécules d'intérêts. La fluorescence est alors exploitée avec un laser éclairant à la longueur d'onde d'excitation et des filtres couleurs pour ne conserver que la longueur d'onde émise. En utilisant plusieurs fluorophores et filtres, comme sur la figure 7.2, il devient possible d'imager plusieurs molécules et processus biologiques en concordance.

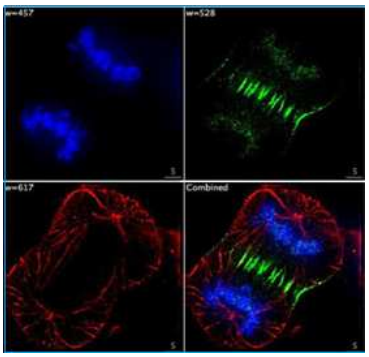


Figure 7.2. – Exemple d'image de microscopie plein champ par fluorescence (biofaculte.blogspot.com).

Il existe de très nombreuses modalités et variations techniques, mais, dans le contexte de mon travail, un microscope optique est un système qui peut être modélisé en intensité comme

$$g(x,y,z) = h(x,y,z) \underset{x,y,z}{*} [I(x,y,z) \times f(x,y,z)] \qquad (7.1)$$

où $(x,y,z) \in \mathbb{R}^3$ sont les coordonnées spatiales 3D, $I : \mathbb{R}^3 \to \mathbb{R}^+$ l'illumination, $f \in \mathbb{R}^+$ l'objet observé, $g \in \mathbb{R}^+$ l'image formée et $h \in \mathbb{R}^+$ la réponse impulsionnelle ou PSF (*Point Spread Function*) de l'instrument[1]. Avec ce modèle, l'objet d'intérêt est 3D et le microscope, vu comme une lentille, est un système linéaire invariant avec une réponse impulsionnelle 3D également. Dans les faits la mesure est faite sur un plan focal z_n pilotable et aboutit à la mesure

1: On néglige ici les effets non stationnaires comme les aberrations

d'une image

$$g_{z_n}(x,y) = \int h(x,y,z) \underset{x,y}{*} [i(x,y,z) \times f(x,y,z)]\,\mathrm{d}z \qquad (7.2)$$

correspondant à l'intégration sur z d'un cube qui a été convolué suivant x et y seulement.

Les résultats dépendant fortement de la PSF h qui présente plusieurs caractéristiques. Pour illustrer, la figure 7.3 présente une PSF générée avec *PSF Generator* [2]. Tout d'abord on observe bien la nature 3D de la PSF. L'image de gauche correspond à la PSF théorique classique dite d'Airy et correspond à l'observation par l'équation (7.2) d'une source ponctuelle, soit un Dirac, placé à l'origine $x = y = z = 0$. L'image du milieu correspond à l'observation par le même modèle d'une source fortement défocalisée avec $z \neq 0$. Ces deux PSF sont dites « latérales » pour signifier qu'elles sont pour un z fixé. L'image de droite est la PSF axiale c'est-à-dire une coupe avec y fixé.

2: bigwww.epfl.ch/algorithms/psfgenerator/

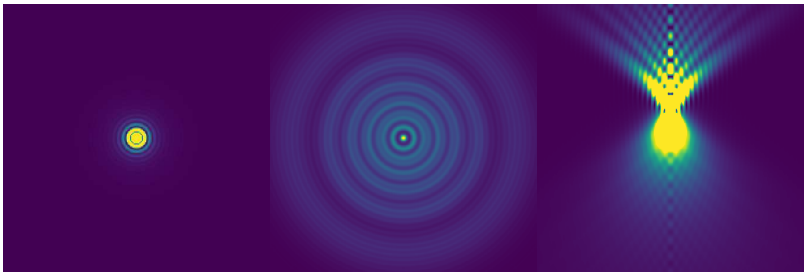


Figure 7.3. – Exemple de PSF (générée par *PSF Generator*). L'image de droite correspond à la PSF latérale focalisée, la figure d'Airy, avec une source ponctuelle dans le plan focal. L'image du milieu à la PSF latérale défocalisée, avec une source ponctuelle hors focus. La figure de droite est la PSF axiale avec l'axe z sur l'axe vertical. Les images de gauche et droite sont saturées pour améliorer la lecture.

On remarque la nature oscillatoire de la PSF, mais le plus notable c'est l'élargissement de la PSF avec la defocalisation. En outre, par conservation de la quantité de lumière, toutes les PSF latérales sont d'énergie équivalente. On peut tirer plusieurs conclusions.

1. La PSF la plus petite est la figure d'Airy correspondant au plan focalisé. Cela dit, l'étalement de la PSF n'est pas négligeable et la résolution spatiale est limitée à environ 200 nm en microscopie optique [3].
2. Avec le modèle d'observation équation (7.2), tous les plans sont imagés et sommés avec un flou d'autant plus important que les plans sont défocalisés. Cela forme un fond très basses fréquences appelé fond « hors-focus », visible par exemple figure 7.4.
3. Avec juste une image, on a alors un problème de séparation de source et entre l'image au plan focal et l'image de fond, et de déconvolution de l'image au plan focal. Le problème est d'autant plus important que l'échantillon est épais.

3: D'où l'usage de la microscopie électronique par exemple qui utilise des électrons plutôt que des photons visibles ayant une longueur d'onde bien plus grande.

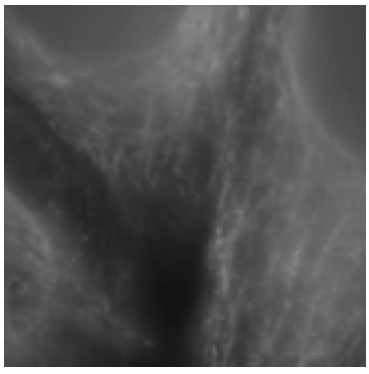


Figure 7.4. – Imagerie *widefield* de tubuline.

Ce problème de lumière hors focus et de flou peut également être expliqué par la réponse en fréquences, ou OTF (*Optical Transfert*

Function), illustré figure 7.5. La figure montre le support de l'OTF qui forme une sorte de tore, avec une fréquence de coupure finie. En outre, l'OTF suivant l'axe k_z, avec $k_x = k_y = 0$, correspond à un Dirac. Autrement dit seul le continu passe et les fréquences spatiales suivant l'axe z sont essentiellement perdues. C'est le *missing cone*.

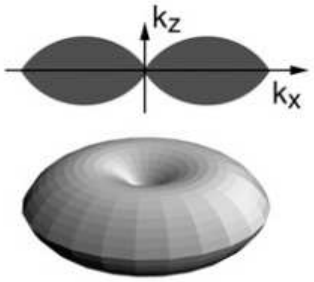


Figure 7.5. – Support de l'OTF 3D.

7.1.2. Techniques de super-résolution

Une des premières techniques d'amélioration est la microscopie confocale (apparue en 1980 pour les premières versions commerciales) qui consiste à placer un trou (*pinhole*) sur un plan focal du chemin optique, permettant de ne sélectionner que les rayons provenant d'un point précis de l'objet d'intérêt. On supprime ainsi la lumière hors focus et la résolution est légèrement améliorée. L'image entière est acquise par balayage. Cette technique, initialement lente, a gagné en rapidité avec l'utilisation des *spinning disk* possédant plusieurs *pinhole* et l'amélioration des détecteurs [4] Le rythme des innovations s'est ensuite accentué depuis les années 2000 où plusieurs techniques dites de *super-résolution* sont apparues, avec notamment le STED et le PALM/STORM, deux techniques ayant amené à l'attribution du prix Nobel de Chimie à S. W. Hell, E. Betzig et W. Moerner, et la SIM.

4: qui sont des photos multiplicateurs.

Le principe de la microscopie STED, pour *Stimulated Emission Depletion* et illustrée figure 7.6, est, après illumination et excitation des fluorophores par le faisceau, d'envoyer une impulsion de déplétion avec la forme d'un anneau pour bloquer l'émission des fluorophores, aboutissant à un faisceau plus petit. Une image est ensuite acquise par balayage.

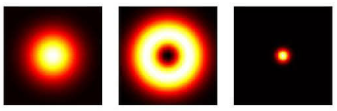


Figure 7.6. – Principe de la stimulation STED (wikipedia).

La microscopie PALM/STORM utilise des fluorophores dits photo-activables. Un premier laser active aléatoirement un sous-ensemble des fluorophores comme fluorescent, avec une faible densité. En supposant que les fluorophores sont tous disjoint au sens de Rayleigh, on les localise précisément par ajustement de la PSF. Des images successives sont acquises, avec à chaque fois une nouvelle répartition aléatoire des molécules fluorescentes.

Les deux techniques précédentes permettent d'atteindre des niveaux de résolution jusqu'à 5 nm, soit l'échelle moléculaire. Elles sont par contre lentes et assez photo-toxiques pour les échantillons biologiques.

7.2. Algorithmes de reconstruction en Microscopie SIM

La microscopie SIM, pour *Structured Illumination Microscopy*, est une alternative plus rapide et moins photo-toxique avec une amélioration de la résolution latérale et axiale, moindre cependant

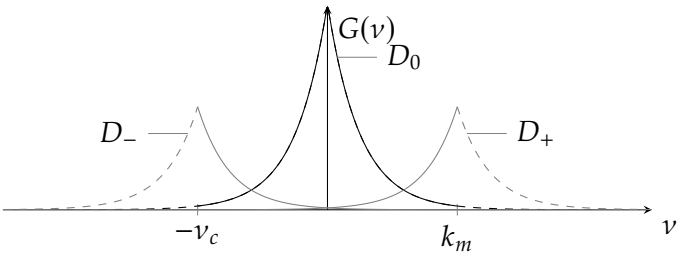


Figure 7.7. – Illustration du principe de la SIM avec k_m la fréquence de modulation et ν_c la fréquence de coupure de l'OTF.

qu'avec le STED ou le PALM. Le principe fondamental de cette microscopie est de faire une modulation d'amplitude de l'échantillon avant son observation par l'instrument et le flou spatial. On provoque ainsi un repliement volontaire des fréquences spatiales qui étaient précédemment filtrées et sont mesurées ainsi de manière indirecte. On peut ainsi atteindre le double de la résolution, ou plus exactement reconstruire des fréquences jusqu'à deux fois la fréquence de coupure [5]

5: Il est possible d'aller plus loin avec des fréquences de modulation supérieures, mais c'est très difficile à cause des limites instrumentales.

La modulation d'amplitude est quasiment exclusivement des grilles sinusoïdales

$$i(x,y) = 1 + \alpha\cos\big(2\pi k_x x + 2\pi k_y y\big) \tag{7.3}$$

avec $0 \leq \alpha \leq 1$, de fréquences (k_x, k_y). Plusieurs techniques existent pour générer ces franges, mais les deux plus courantes sont la grille physique ou l'interférence de deux faisceaux générés par un SLM (*Spatial Light Modulator*).

Sur le principe, en repartant de l'équation (7.1) en 2D seulement, et en passant dans l'espace de Fourier on obtient

$$G(\nu_x,\nu_y) = H(\nu_x,\nu_y) \times \Big[I(\nu_x,\nu_y) \underset{x,y}{*} F(\nu_x,\nu_y)\Big] \tag{7.4}$$

ce qui avec

$$I(\nu_x,\nu_y) = \delta(\nu_x,\nu_y) + \frac{\alpha}{2}\delta(\nu_x - k_x, \nu_y - k_y) + \frac{\alpha}{2}\delta(\nu_x + k_x, \nu_y + k_y) \tag{7.5}$$

donne

$$G(\nu_x,\nu_y) = D_0(\nu_x,\nu_y) + D_-(\nu_x,\nu_y) + D_+(\nu_x,\nu_y) \tag{7.6}$$

avec

$$\begin{aligned} D_0(\nu_x,\nu_y) &= H(\nu_x,\nu_y)F(\nu_x,\nu_y), \\ D_-(\nu_x,\nu_y) &= H(\nu_x,\nu_y)F(\nu_x - k_x, \nu_y - k_y), \\ D_+(\nu_x,\nu_y) &= H(\nu_x,\nu_y)F(\nu_x + k_x, \nu_y + k_y). \end{aligned} \tag{7.7}$$

Le terme D_0 correspond au spectre de l'image *widefield* standard. Les deux autres termes sont des répliques décalées dans Fourier par modulation d'amplitude comme illustré figure 7.7 qui montre comment les hautes fréquences supprimées en *widefield* se trouvent maintenant dans le support de l'OTF par *aliasing*.

Partant du principe que le support de l'OTF contient trois inconnues, D_0, D_- et D_+, l'algorithme traditionnel utilise alors trois mesures avec modulations différentes $i_n(x,y) = 1+\cos\left(2\pi k_x x + 2\pi k_y y + \phi_n\right)$ de phases $\phi_n = 0, 2\pi/3$ et $-2\pi/3$ respectivement. Dans ce cas, on résout le système avec des combinaisons linéaires d'images.

$$\begin{aligned} D_0 &= \frac{1}{3}\left(G_0 + G_1 + G_2\right), \\ D_- &= \frac{1}{3}\left(G_0 + e^{4\imath\pi/3}G_1 + e^{-4\imath\pi/3}G_2\right), \\ D_+ &= \frac{1}{3}\left(G_0 + e^{-4\imath\pi/3}G_1 + e^{4\imath\pi/3}G_2\right). \end{aligned} \tag{7.8}$$

La reconstruction classique [82] se fait ensuite par recalage dans le pan de Fourier de D_- et D_+ puis un filtre de Wiener, régularisé donc, est appliqué pour prendre en compte l'OTF. À partir de là, plusieurs problèmes se posent.

— Tout d'abord il n'est pas nécessaire d'avoir trois mesures images, soit neuf en 2D, car D_- et D_+ contiennent la même information par symétrie hermitienne. Autrement dit, quatre mesures uniquement sont nécessaires pour reconstruire les hautes fréquences, mais dans ce cas le système d'équations (7.8) n'est plus utilisable. Ce point est important pour améliorer la rapidité du système.
— Par ailleurs, la connaissance des paramètres de la modulation doit être assez précise lorsque k_m s'approche de ν_c au risque d'introduire des artefacts de reconstruction.
— Enfin d'autres questions se posent sur l'estimation des paramètres de régularisation, le modèle de bruit ou la prise en compte de plan hors focus, le modèle équation (7.4) étant 2D.

7.2.1. Reconstruction SIM à 4 images

Dans le cas 2D, le pavage des termes D_- et D_+ donne la figure 7.8 obtenue avec trois orientations, *i.e.*, trois couples de fréquence de modulation (k_x, k_y). Avec trois orientations, la méthode classique nécessite un minimum de neuf images, trois par orientation. En plus du raisonnement ci-dessus, on peut constater que dans ce cas, de la redondance est introduite, car la surface du plan de Fourier reconstruit est au maximum quatre fois plus grande (le rayon étant doublé). Autrement dit, l'information nécessaire est déjà présente avec une image *widefield* et trois images modulées, une par orientation.

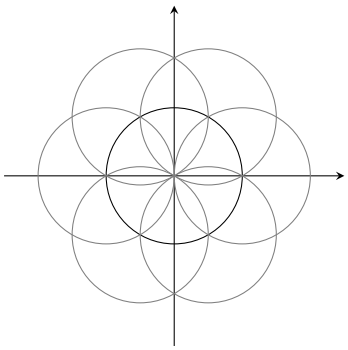


Figure 7.8. – Pavage de l'extension fréquentielle en SIM harmonique. Le cercle noir correspond au support de l'OTF.

Ce jeu de données ne permet pas d'utiliser le système équation (7.8). Dans [54] nous avons proposé, avec Jean-Christophe Olivo-Marin de l'institut Pasteur, Alexandra Fragola et Vincent Loriette du LPEM de l'ESPCI une méthode alternative reposant sur une approche inverse avec un modèle explicite des données

[54]: Orieux et al. (2012), « Bayesian Estimation for Optimized Structured Illumination Microscopy »

et un estimateur Bayésien. Cette méthode est la première à avoir montré la faisabilité d'une reconstruction à quatre images, avec un algorithme et des résultats.

L'approche repose en particulier sur le premier modèle direct qui a été proposé pour la microscopie SIM. Il s'écrit

$$\boldsymbol{g}_i = \boldsymbol{S}\boldsymbol{C}\boldsymbol{M}_i\boldsymbol{x} + \boldsymbol{n}_i \tag{7.9}$$

où $\boldsymbol{x} \in \mathbb{R}^N$ est l'image inconnue, $\boldsymbol{g}_i \in \mathbb{R}^M$ est une image avec une grille (modulation) particulière, $\boldsymbol{C}$ le flou introduit par le microscope modélisé comme une convolution (mise en œuvre par transformée de Fourier ou non), $\boldsymbol{S} \in \mathbb{R}^{N\times M}$ un facteur de sous-échantillonnage nécessaire pour reconstruire une image avec plus de pixels que les mesures ($N \geq M$), et $\boldsymbol{M}_i \in \mathbb{R}^{N\times N}$ une matrice de modulation d'amplitude diagonale. Le terme $\boldsymbol{n}$ correspond à du bruit. En concaténant l'ensemble des données, on obtient $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$ soit un modèle direct linéaire, mais non stationnaire.

Cette approche présente plusieurs avantages. Tout d'abord elle utilise moins d'approximations et celles introduites sont mieux maîtrisées puisqu'elles peuvent être comparées à ce que l'on sait de l'instrument. Par ailleurs, cette modélisation permet une plus grande flexibilité par exemple sur la forme de la modulation qui ne doit plus être strictement une grille sinusoïdale.

Ce modèle a été utilisé [54] dans une approche bayésienne où l'on infère sur la loi *a posteriori*

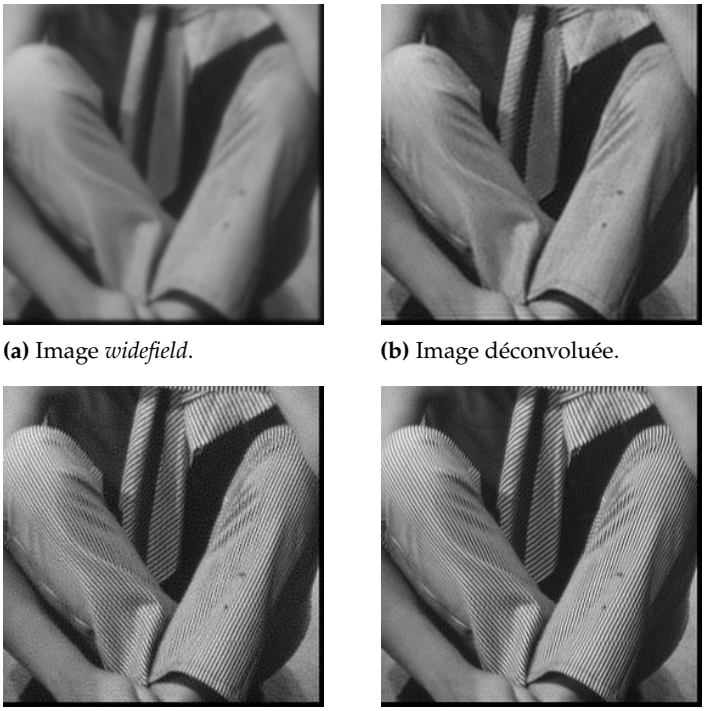
$$\begin{aligned} \mathrm{p}(\boldsymbol{x}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{n}} \mid \boldsymbol{g}) &\propto \mathrm{p}(\boldsymbol{g} \mid \boldsymbol{x}, \gamma_{\boldsymbol{n}})\, \mathrm{p}(\boldsymbol{x} \mid \gamma_{\boldsymbol{x}})\, \mathrm{p}(\gamma_{\boldsymbol{n}})\, \mathrm{p}(\gamma_{\boldsymbol{x}}) && (7.10)\\ &\propto {\gamma_{\boldsymbol{x}}}^{\frac{N-1}{2}-1} {\gamma_{\boldsymbol{n}}}^{\frac{M}{2}-1} \\ &\quad \exp\left(-\frac{\gamma_{\boldsymbol{n}}}{2}\|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{x}\|^2\right) \exp\left(-\frac{\gamma_{\boldsymbol{x}}}{2}\|\boldsymbol{D}\boldsymbol{x}\|^2\right) && (7.11) \end{aligned}$$

en considérant une vraisemblance gaussienne, un *a priori* gaussien, et des lois non informatives pour les paramètres de précision $\gamma_{\boldsymbol{n}}$ et $\gamma_{\boldsymbol{x}}$. Dans cette loi, les hyperparamètres sont considérés inconnus et doivent être estimés. On considère également un *a priori* de douceur sur l'image avec $\boldsymbol{D}$ un opérateur de différences pour régulariser le mauvais conditionnement venant notamment de la convolution spatiale. Nous avons ensuite proposé un échantillonneur de Gibbs pour approcher l'estimateur de la moyenne *a posteriori*

$$\hat{\boldsymbol{x}}_{\text{EAP}} = \mathbb{E}[\boldsymbol{x}] \approx \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{x}_k. \tag{7.12}$$

Pour cela il faut être capable de simuler des échantillons de loi gaussiennes de grande dimension et corrélées

$$\boldsymbol{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{7.13}$$



(a) Image *widefield*. **(b)** Image déconvoluée.

(c) Reconstruction à 4 images. **(d)** Reconstruction à 9 images.

Figure 7.9. – Reconstruction de « Barbara ». On observe tout d'abord qu'il n'y a pas de différences perceptibles entre la reconstruction à 4 images et celle à 9 images. Deuxièmement on voit bien l'apport de l'imagerie SIM par rapport à l'imagerie *widefield* et la déconvolution. Cette dernière notamment ne permet pas de reconstruire des fréquences au-delà de la fréquence de coupure, comme les rayures sur le pantalon. Au contraire la SIM est capable de les reconstruire puisque ces dernières sont mesurées indirectement par repliement de spectre.

avec $\boldsymbol{\Sigma}_k = \gamma_{\boldsymbol{n}_k}\boldsymbol{H}^t\boldsymbol{H} + \gamma_{\boldsymbol{x}_k}\boldsymbol{D}^t\boldsymbol{D}$ et $\boldsymbol{\mu}_k = \boldsymbol{\Sigma}_k^{-1}\boldsymbol{H}^t\boldsymbol{g}$. La non-stationnarité et la grande dimension du problème empêchent l'utilisation des algorithmes classiques reposant sur la factorisation de Fourier ou de Cholesky par exemple. Pour cela nous avons proposé l'algorithme PO décrit dans le chapitre 8.

Les résultats sont au rendez-vous et les images produites avec uniquement quatre images au lieu de neuf présentent la même qualité (au rapport signal sur bruit près) et les mêmes structures sont reconstruites.

Ces travaux [54] sont les premiers à proposer une méthode pour la reconstruction SIM à partir d'un nombre réduit d'images. Cet article est cité essentiellement pour cette raison. Il n'aborde pas cependant d'autres problèmes que sont l'estimation des paramètres de la modulation et la lumière parasite provenant des plans hors focus.

[54]: Orieux et al. (2012), « Bayesian Estimation for Optimized Structured Illumination Microscopy »

7.2.2. Estimation des paramètres de la modulation

Pour une bonne reconstruction, que ce soit avec la méthode classique, ou la méthode proposée, il faut connaître précisément les paramètres de la modulation $(k_{x,n}, k_{y,n}, \phi_n)$ ainsi que la profondeur de modulation α_n pour chaque modulation i_n. Dans le cas contraire des artefacts apparaissent [79]. Il s'agit d'un problème myope, car il est impossible de calibrer correctement les paramètres si la fréquence de modulation est proche de la fréquence de coupure [6]. Par ailleurs, ces paramètres doivent être estimés pour

6: Les SIM commerciaux ont le plus souvent des fréquences de coupure plus faibles.

chaque image, y compris si certaines valeurs se répètent (dans le cas à neuf images ou de vidéos par exemple). Il s'agit donc d'un problème *myope* ou d'*auto-calibration* où les paramètres du modèle $\boldsymbol{H}$ doivent être estimés en même temps que l'inconnue $\boldsymbol{x}$.

La plupart des articles traitant de microscopie SIM survolent ou ne mentionnent pas l'estimation de ces paramètres. Deux travaux notables existent. Tout d'abord Shroff et coll. [67] propose d'estimer les phases par *cross-correlation* entre le terme D_- et D_0. Malheureusement les fréquences sont obtenues uniquement en repérant la fréquence avec un maximum, donc sur la grille de la TFD, et la méthode nécessite neuf images pour effectuer la séparation des motifs D_*. Par ailleurs, Schaefer et coll. [79] propose une approche se basant sur deux volets. Tout d'abord ils fournissent une étude des artefacts introduits par une mauvaise estimation des paramètres et de leur forme, pour ensuite proposer une méthode d'optimisation visant à minimiser ces artefacts dans la reconstruction. Cette approche est intéressante, mais est limitée aux structures d'artefacts étudiés (produit de deux sinus par exemple).

[30]: Orieux et al. (2017), « Fast Myopic 2D-SIM Super Resolution Microscopy with Joint Modulation Pattern Estimation »

Dans [30] nous avons proposé une méthode alternative reposant sur la minimisation du critère joint

$$\hat{\boldsymbol{x}}, \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{x},\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{H}(\boldsymbol{\theta})\boldsymbol{x}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{x}\|^2 \tag{7.14}$$

où $\boldsymbol{\theta} = (k_{x,1}, k_{y,1}, \phi_1, \alpha_1, k_{x,2}, \dots)$, sous contraintes de support pour $\boldsymbol{\theta}$. Ces contraintes sont essentiellement le pas d'échantillonnage de la TFD pour les fréquences de modulations $\boldsymbol{k}$, centré sur la fréquence maximum, $[0, 2\pi]$ pour la phase ϕ et $[0, 1]$ pour le contraste α. On peut noter que le problème est, pour $\boldsymbol{x}$ fixe, indépendant par image $\boldsymbol{g}_i$. Ce problème s'avère particulièrement complexe par la corrélation entre les variables $\boldsymbol{k}$ et ϕ premièrement, et la présence de minima locaux y compris dans les contraintes.

La méthode que nous avons proposée utilise une deuxième innovation offerte par l'approche « problème inverse » et repose sur la flexibilité du modèle direct $\boldsymbol{H}(\boldsymbol{\theta}) = \boldsymbol{S}\boldsymbol{C}\boldsymbol{M}(\boldsymbol{\theta})$. En effet on peut utiliser n'importe quelle modulation dans cette approche pour peu qu'elle injecte les mêmes informations. Il s'avère que le microscope du LPEM de l'ESPCI, comme la plupart des microscopes reposent sur des SLM pour produire les franges d'interférences. Naturellement ces SLM produisent dans l'échantillon des franges comme

$$i(x, y) = 1 + \alpha\cos\big(2\pi k_x x + 2\pi k_y y + \phi\big) + \beta\cos\left(\pi k_x x + \pi k_y y + \frac{\phi}{2}\right) \tag{7.15}$$

7: Les franges étant en intensité, le résultat dans Fourier est l'autocorrélation de trois Dirac en $\boldsymbol{k}$, $-\boldsymbol{k}$ et 0, donnant l'équation (7.15) dans l'espace image.

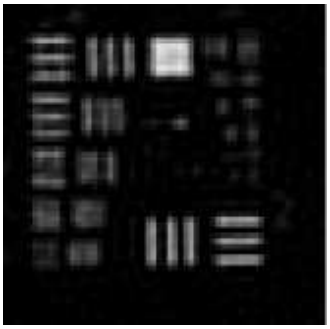
c'est-à-dire avec une modulation supplémentaire de fréquence moitié[7]. C'est un dispositif supplémentaire qui permet de sup-

primer ce deuxième jeu de franges, dispositif introduit, car cette deuxième modulation oblige l'acquisition de cinq images au lieu de trois avec l'algorithme classique, soit quinze en tout.

À l'inverse, l'introduction d'une modulation supplémentaire n'empêche pas l'utilisation de seulement quatre images, mais la présence des mêmes paramètres permet d'améliorer significativement le rapport signal sur bruit, tout en simplifiant le montage optique. L'amélioration du rapport signal sur bruit permet de diminuer la présence de minima locaux, qui restent présents malheureusement, et d'améliorer la robustesse de l'estimation.

L'algorithme en lui-même reste relativement basique dans [30] avec une minimisation alternée entre $\boldsymbol{x}$ et $\boldsymbol{\theta}$, un gradient conjugué pour $\boldsymbol{x}$, une recherche par quadrillage pour se prémunir des minima locaux pour (k_x, k_y, ϕ) et une solution explicite pour α car le problème est linéaire. L'algorithme possède cependant encore une sensibilité à l'initialisation.

(a) Image reconstruite avec la méthode [67].



(b) Image reconstruite avec la méthode proposée.

(c) TF de la figure 7.10a.

(d) TF de figure 7.10b.

Figure 7.10. – Exemple de reconstruction obtenue avec la méthode proposée pour l'estimation des paramètres. La figure 7.10a utilise la méthode de Schroff [67]. On constate qu'elle n'est pas adaptée pour estimer les paramètres pour la super-résolution où la fréquence de modulation est proche de la fréquence de coupure. Concernant la méthode proposée, bien que visuellement on ne distingue pas d'artefacts dans 7.10b, la transformée de Fourier 7.10d montre quelque résidu de grille.

7.2.3. Reconstruction avec coupe optique

La présence de plan hors focus est gênante pour l'interprétation des images. Par ailleurs, elle produit des artefacts de reconstruction avec les méthodes évoquées plus haut, car ces dernières considèrent une modulation 2D sur tout le champ, ce qui limite la SIM à l'observation d'échantillon fin avec la microscopie TIRF par exemple [8]. Être capable d'imager uniquement un plan de l'échantillon s'appelle une coupe optique. La microscopie confocale le

8: technique où seulement 200 nm sont éclairés sous la lamelle.

fait naturellement, de même que l'imagerie par feuille de lumière. La microscopie SIM initiale est également utilisée pour faire une coupe optique [25], [50] en exploitant la présence de frange d'interférence seulement dans le plan de mise au point. Cependant, la coupe optique et la super-résolution conjointe n'ont été que peu étudiées ensemble.

Cette partie présente des travaux qui n'ont été présentés qu'au JIONC en 2014 [47] et publiés dans un logiciel déposé à l'Agence de Protection des Programmes [28]. Une autre partie s'est faite en collaboration avec Roberto Baena-Gallé [25], Yann Lai-Tim et Laurent Mugnier [19], pour la SIM en ophtalmologie. En repartant de l'équation (7.2) on peut avoir un nouveau modèle de mesures

[19]: Lai-Tim et al. (2019), « Jointly Super-Resolved and Optically Sectioned Bayesian Reconstruction Method for Structured Illumination Microscopy »

$$\boldsymbol{g}_n = \boldsymbol{S}\boldsymbol{C}\boldsymbol{M}_n\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{n}_n = \boldsymbol{H}_n\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{n}_n \tag{7.16}$$

où $\boldsymbol{x}$ représente une image 2D de coupe optique, $\boldsymbol{b}$ la lumière provenant des plans défocalisés, *i.e.*, la somme des plans spatialement convolués par les PSF respectives à chaque défocus, et $\boldsymbol{n}$ un terme de bruit. L'image widefield correspond dans ce cas à $\boldsymbol{w} = \boldsymbol{S}\boldsymbol{C}\boldsymbol{x} + \boldsymbol{b}_w$.

Il s'agit donc d'un problème de séparation de source, car l'équation (7.16) est la combinaison linéaire de deux images inconnues $\boldsymbol{x}$ et $\boldsymbol{b}$. Pour résoudre la dégénérescence, plusieurs possibilités existent.

Approche par régularisation

Dans [30], [47] nous exploitons le modèle (7.16) et le fait que la plan au focus $\boldsymbol{x}$ est le seul à être modulé en amplitude contrairement à $\boldsymbol{b}$ et que le fond $\boldsymbol{b}$ est le même dans chaque image $\boldsymbol{g}_n$. Par ailleurs nous avons ajouté une régularisation de douceur sur $\boldsymbol{b}$ pour forcer le fond à être très basse fréquence. Enfin nous avons ajouté un *a priori* de positivité sur la coupe optique $\boldsymbol{x}$

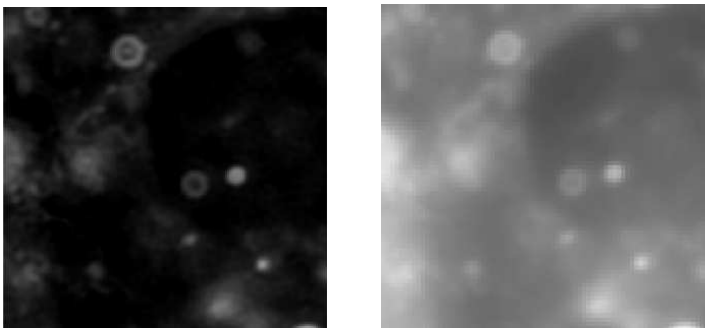
$$\begin{aligned} \hat{\boldsymbol{x}}, \hat{\boldsymbol{b}} = \underset{\boldsymbol{x},\boldsymbol{b}}{\arg\min}\ & \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{x}\|^2 + \mu\|\boldsymbol{D}\boldsymbol{b}\|^2 \\ & \text{s.c.} \quad x_p \geq 0, \forall p. \end{aligned} \tag{7.17}$$

Pour résoudre la contrainte, nous avons proposé une méthode reposant sur les multiplicateurs de Lagrange et des variables auxiliaires $\boldsymbol{s}$ qui portent la contrainte dans le critère augmenté (soit l'ADMM)

$$\begin{aligned} J(\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{s}, \boldsymbol{l}) = & \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{x} - \boldsymbol{b}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{x}\|^2 + \mu\|\boldsymbol{D}\boldsymbol{b}\|^2 \\ & + c\|\boldsymbol{x} - \boldsymbol{s}\|^2 + \boldsymbol{l}^t(\boldsymbol{x} - \boldsymbol{s}) \\ & \text{s.c.} \quad x_p - s_p = 0, s_p \geq 0, \forall p. \end{aligned} \tag{7.18}$$

avec $c \geq 0$. La minimisation alternée est relativement facile pour chaque variable, excepté pour $\boldsymbol{x}$ qui est quadratique, mais ne peut être résolue explicitement. Dans ce cas on peut utiliser un algorithme itératif de type gradient conjugué. La figure 7.11 montre un exemple de reconstruction.

Dans les faits la régularisation sur $\boldsymbol{b}$ est importante et l'erreur de modélisation sur $\boldsymbol{D}$ et μ introduit des transferts (*leakeage*) entre $\boldsymbol{b}$ et $\boldsymbol{x}$.



(a) Coupe optique **(b)** Image widefield

Figure 7.11. – Exemple de reconstruction sur mesures réelles avec une coupe optique par régularisation sur des cellules HEK. On observe bien la reconstruction en forme de *donuts* permise par la reconstruction, bien moins visible avec l'imagerie *widefield*, ainsi que l'effet de coupe optique.

Approche par soustraction

Dans [25] l'approche est différente et exploite une caractéristique de la microscopie SIM. En effet en repartant de l'équation (7.16) et (7.3) page 48 on peut écrire le modèle

$$\begin{aligned} \boldsymbol{g}_n &= \boldsymbol{S}\boldsymbol{C}\boldsymbol{M}_n\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{n}_n \\ &= \boldsymbol{S}\boldsymbol{C}(\boldsymbol{I} + \boldsymbol{G}_n)\boldsymbol{x} + \boldsymbol{b} + \boldsymbol{n}_n \\ &= \boldsymbol{S}\boldsymbol{C}\boldsymbol{G}_n\boldsymbol{x} + \boldsymbol{w} + \boldsymbol{n}_n \end{aligned}$$

avec $\boldsymbol{G}_n$ une matrice diagonale contenant uniquement les termes en cosinus de (7.3), et $\boldsymbol{w}$ l'image *widefield*. Par ce biais on peut traiter le problème de reconstruction de la coupe optique $\boldsymbol{x}$ avec un nouveau modèle de données auquel on retranche l'image *widefield* [9]

$$\begin{aligned} \hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|(\boldsymbol{g} - \boldsymbol{w}) - \boldsymbol{S}\boldsymbol{C}\boldsymbol{G}\boldsymbol{x}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{x}\|^2 \\ \text{s.c.} \quad x_p \geq 0, \forall p. \end{aligned} \tag{7.19}$$

L'avantage certain c'est qu'il n'y a plus deux sources à estimer, mais uniquement la coupe optique et cela repose sur la physique d'acquisition et non sur des modèles *a priori* de douceur. Le désavantage c'est qu'il y a une incohérence et une approximation dans le modèle. En effet, la partie basse fréquence de $\boldsymbol{x}$ se trouve être dans $\boldsymbol{w}$ ce qui est négligé en considérant $\boldsymbol{w}$ comme une nuisance soustraite. Ce travail, ainsi que le suivant ont été portés par l'ONERA en collaboration avec L. Mugnier avec qui nous avons collaboré.

[25]: Baena-Gallé et al. (2017), « Optical Sectioning with Structured Illumination Microscopy for Retinal Imaging : Inverse Problem Approach »

9: si celle-ci est accessible. Dans le cas d'une acquisition à neuf images elle correspond à la moyenne. Dans une observation à quatre images comme dans [30] c'est une des images mesurées.

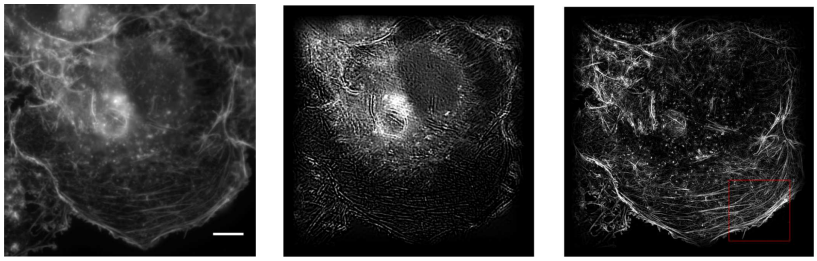
Approche par modélisation de PSF

[20]: Lai-Tim et al. (2019), « Toward a Jointly Super-Resolved and Optically Sectioned Reconstruction for Structured Illumination Retinal Imaging »

Enfin, le travail de thèse de Yann Lai-Tim [20] propose une autre approche reposant sur les propriétés des PSF 3D des microscopes. En effet, les PSF défocalisées ont des propriétés d'écrasement des hautes fréquences. En particulier, si f_c est la fréquence de coupure maximale dans le plan au focus, la fréquence $f_c/2$ est la plus atténuée dans les plans hors focus. Dans ce cas, placer une modulation d'amplitude à $f_m = f_c/2$ permet de distinguer le mieux les plans hors focus et la coupe optique. De cette manière on peut modéliser les données comme

$$\boldsymbol{g}_n = \boldsymbol{S}\boldsymbol{C}_f\boldsymbol{M}_n\boldsymbol{x}_f + \boldsymbol{S}\boldsymbol{C}_o\boldsymbol{x}_o + \boldsymbol{n}_n \tag{7.20}$$

où $\boldsymbol{x}_o$ et $\boldsymbol{C}_o$ est l'image et la convolution correspondant aux plans défocalisés, et $\boldsymbol{x}_f$ et $\boldsymbol{C}_f$ la PSF correspondant à la coupe optique. La méthode ensuite repose sur un *a priori* de douceur et de positivité. En particulier, la douceur est définie dans le plan de Fourier comme une loi de puissance dont les paramètres sont estimés.



(a) Image widefield **(b)** Image hors-focus estimée **(c)** Coupe optique reconstruite

Figure 7.12. – Résultat obtenu avec l'algorithme BOSSA-SIM

7.3. Conclusion

J'ai présenté ici les travaux que j'ai menés personnellement ou en collaboration sur un problème de reconstruction d'image en microscopie biologique à haute résolution. C'est un problème sur lequel j'ai commencé à travailler durant mon postdoc et que j'ai poursuivi une fois à l'IAP puis lorsque j'ai été recruté au L2S, avec une baisse d'activité sur ce thème ces dernières années pour me concentrer plus sur l'imagerie multi et hyperspectrale.

Dans la prochaine partie je présente des travaux réalisés plus en amont sur des algorithmes MCMC pour la résolution de problèmes inverses.

Algorithmes MCMC pour les problèmes inverses

8.

Je présente ici les travaux menés en collaboration avec Jean-François Giovannelli, Olivier Féron, Raphael Chinchilla et plus récemment avec Saïd Moussaouï et Jérôme Idier. Il s'agit de travaux méthodologiques sur l'inférence bayésienne pour le traitement du signal. Ces travaux ont occupé moins de place que souhaité, raison pour laquelle j'ai effectué une demi-délégation CNRS au LS2N de Nantes.

8.1. Introduction

La résolution de problèmes inverses, jusqu'à l'émergence récente des méthodes par apprentissage machine, repose sur deux approches principales. La première, approche dite variationnelle, repose sur la définition de la solution comme le minimiseur d'un critère $J(\boldsymbol{x})$

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} J(\boldsymbol{x}).$$

Ce critère modélise l'ensemble des informations à disposition pour résoudre le problème. Dans le cas des problèmes inverses, le plus souvent le critère prend la forme

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} Q(\boldsymbol{x};\boldsymbol{y}) + R(\boldsymbol{x})$$

avec un terme Q d'attache aux données $\boldsymbol{y}$, souvent un moindre carré, et un terme de régularisation R modélisant les informations supplémentaires ou des caractéristiques désirées de la solution telles que la régularité, des contraintes ou la parcimonie. Bien sûr, en fonction de la structure du critère, différents algorithmes existent pour tenter d'obtenir le minimiseur [1].

La deuxième repose sur les méthodes bayésiennes [84] où l'information disponible est modélisée par une loi *a posteriori*

$$\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}) = \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})\,\mathrm{p}(\boldsymbol{x})/\mathrm{p}(\boldsymbol{y}) \tag{8.1}$$

avec une vraisemblance $\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})$ correspondant au terme d'attache aux données et un *a priori* se relatant à la régularisation. Une fois la loi définie, il faut choisir un estimateur, le plus courant étant le *maximum a posteriori*, ou MAP, correspondant au minimiseur du critère précédent

$$\hat{\boldsymbol{x}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{x}} \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}) = \arg\min_{\boldsymbol{x}} -\log\big(\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y})\big) = \arg\min_{\boldsymbol{x}} J(\boldsymbol{x}).$$

avec $Q(\boldsymbol{x};\boldsymbol{y}) \propto -\log\big(\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y})\big)$ et $R(\boldsymbol{x}) \propto -\log\big(\mathrm{p}(\boldsymbol{x})\big)$. Pour calculer le MAP on utilise alors un algorithme d'optimisation adapté au

1: en supposant qu'il existe et est unique sinon le problème est *mal posé*.

critère.

L'autre estimateur classique est l'*espérance a posteriori*, ou EAP,

$$\hat{\boldsymbol{x}}_{\text{EAP}} = \int_{\boldsymbol{x}} \mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y})\, \mathrm{d}\boldsymbol{x} \tag{8.2}$$

plus souvent considéré dans une approche bayésienne stricte, car il correspond au minimiseur de l'erreur quadratique moyenne (EQM) [2]

$$\hat{\boldsymbol{x}}_{\text{EAP}} = \arg\min_{\hat{\boldsymbol{x}}} \mathbb{E}\left[\|\hat{\boldsymbol{x}} - \boldsymbol{x}^*\|^2\right] \tag{8.3}$$

sous la loi *a posteriori* avec $\boldsymbol{x}^*$ la vraie valeur.

2: ou coût quadratique (MMSE pour *Minimum Mean Square Error*), le MAP correspondant au coût *tout ou rien*.

La difficulté principale, en particulier pour les problèmes inverses, avec l'EAP équation (8.2) est la nécessité de calculer une intégrale sur un espace de grande taille d'une part, avec une loi *a posteriori* habituellement complexe d'autre part. Dans ce cas le choix repose souvent sur l'utilisation d'algorithmes MCMC qui ont la réputation d'être peu rapides.

De mon point de vue, autant il semble indéniable que ces algorithmes sont plus lents que les algorithmes d'optimisation pour fournir une approximation correcte de l'estimateur, autant leur lenteur est exagérée [3]. Une raison possible pourrait être l'investissement inférieur des efforts de recherche de la communauté. À titre d'exemple, la requête « *inverse problems optimization* » renvoie dix fois plus de résultats sur `Google Scholar` que la requête « *inverse problems MCMC* ».

3: ou l'efficacité relative des algorithmes d'optimisation sur estimée.

Pourtant les algorithmes MCMC présentent plusieurs intérêts. Tout d'abord, en plus de minimiser l'EQM, ils permettent de calculer tout autre moment de la loi *a posteriori* comme la déviation standard pour quantifier les incertitudes par exemple, ou des corrélations entre paramètres. De plus, sans aborder les contraintes techniques, ils permettent d'estimer dans un cadre cohérent les paramètres de nuisance comme les hyper paramètres.

Dans la suite je présente les travaux que j'ai menés pour utiliser les algorithmes MCMC sur des problèmes inverses de grandes dimensions.

8.2. MCMC pour l'inversion non supervisée

L'inversion non supervisée s'intéresse à la reconstruction de l'inconnue dans le cas où certains paramètres $\boldsymbol{\theta}$ du modèle sont considérés inconnus en plus de l'objet $\boldsymbol{x}$. Ils sont appelés paramètres de nuisance parfois, par rapport aux paramètres d'intérêt. Dans un cadre bayésien, cela conduit à une loi

$$\mathrm{p}(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\, \mathrm{p}(\boldsymbol{x})\, \mathrm{p}(\boldsymbol{\theta}). \tag{8.4}$$

Ensuite tout dépend du modèle bien évidemment et donc du cadre applicatif. Pour les problèmes inverses mal posés, certaines structures sont récurrentes. Je présente dans la suite celles qui reposent sur des lois conditionnelles $\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\theta})$ Gausiennes.

8.2.1. Contexte

Dans le cadre des problèmes inverses en dimension finie, le plus courant est de considérer une vraisemblance Gaussienne [4]

$$\mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x}) \propto |\boldsymbol{\Sigma}_n|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^t \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\right)$$

avec $\boldsymbol{H}$ le modèle d'observation et $\boldsymbol{\Sigma}_n$ modélisant la corrélation du bruit affectant les mesures $\boldsymbol{y}$[5]. Cette vraisemblance conduit, dans le cas des problèmes inverses mal posés, à des estimateurs non définis ou instables. La solution consiste le plus souvent à ajouter des *a priori* par exemple gaussiens

$$\mathrm{p}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{m}-\boldsymbol{D}\boldsymbol{x})^t \boldsymbol{\Sigma}_x^{-1}(\boldsymbol{m}-\boldsymbol{D}\boldsymbol{x})\right)$$

où $\boldsymbol{D}$ est un opérateur d'analyse, $\boldsymbol{m}$ une valeur de rappel, nulle le plus souvent, et $\boldsymbol{\Sigma}_x$ une corrélation a priori.

Il faut remarquer tout d'abord que la densité a priori pour $\boldsymbol{x}$ est dans ce cas

$$\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{Q}_x^{-1}\boldsymbol{D}^t\boldsymbol{\Sigma}_x^{-1}\boldsymbol{m}, \boldsymbol{Q}_x^{-1}\right)$$

avec $\boldsymbol{Q}_x = \boldsymbol{D}^t\boldsymbol{\Sigma}_x^{-1}\boldsymbol{D}$ et que cette structure peut imposer des contraintes, comme son existence tout d'abord ($\boldsymbol{Q}_x$ ne devant pas être singulière), mais également sur sa manipulation a posteriori. Par conséquent sa fonction de partition n'est pas proportionnelle à $|\boldsymbol{\Sigma}_x|^{-\frac{1}{2}}$ mais à $|\boldsymbol{Q}_x|^{-\frac{1}{2}}$ [6].

La loi *a posteriori*

$$\begin{aligned}\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{y}) &\propto \mathrm{p}(\boldsymbol{y} \mid \boldsymbol{x})\,\mathrm{p}(\boldsymbol{x}) \\ &\propto \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})^t\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{x})\right)\exp\left(-\frac{1}{2}(\boldsymbol{m}-\boldsymbol{D}\boldsymbol{x})^t\boldsymbol{\Sigma}_x^{-1}(\boldsymbol{m}-\boldsymbol{D}\boldsymbol{x})\right)\end{aligned}$$

est également une loi gaussienne de moyenne

$$\bar{\boldsymbol{m}} = \boldsymbol{Q}^{-1}\left(\boldsymbol{H}^t\boldsymbol{\Sigma}_n^{-1}\boldsymbol{y} + \boldsymbol{D}^t\boldsymbol{\Sigma}_x^{-1}\boldsymbol{m}\right) \tag{8.5}$$

et d'inverse covariance

$$\boldsymbol{Q} = \boldsymbol{H}^t\boldsymbol{\Sigma}_n^{-1}\boldsymbol{H} + \boldsymbol{D}^t\boldsymbol{\Sigma}_x^{-1}\boldsymbol{D}. \tag{8.6}$$

Au vu du rôle équivalent entre les deux termes, on considére

4: sous-entendu que le modèle direct est linéaire.

5: Dans ce cas, les paramètres de nuisances $\boldsymbol{\theta}$ que l'on souhaite estimer peuvent être des paramètres de $\boldsymbol{H}$ ou $\boldsymbol{\Sigma}_n$.

6: Ces points peuvent paraître anecdotiques mais sont souvent négligés ou « mis sous le tapis » avec parfois des conséquences sur les conclusions prises.

maintenant simplement des lois gaussiennes de moyenne

$$\bar{\boldsymbol{m}} = \boldsymbol{Q}^{-1} \sum_k \boldsymbol{H}_k^t \boldsymbol{\Sigma}_k^{-1} \boldsymbol{m}_k \tag{8.7}$$

et de précision

$$\boldsymbol{Q} = \sum_k \boldsymbol{H}_k^t \boldsymbol{\Sigma}_k^{-1} \boldsymbol{H}_k \tag{8.8}$$

fréquemment rencontrées dans la résolution de problèmes inverses.

La simulation de ces lois pour produire des échantillons peut s'avérer complexe et très coûteuse. Deux problèmes principaux tout en étant liés. Tout d'abord lorsque le problème est de grande taille et que ni les opérateurs $\boldsymbol{H}_k$ ni les covariances $\boldsymbol{\Sigma}_k$ ne présentent de structures particulières, le calcul de la moyenne directement n'est pas faisable à cause de l'inversion $\boldsymbol{Q}$. De même, pour produire un échantillon il est nécessaire de calculer une excursion par rapport à la moyenne comme $\boldsymbol{x}^{(i)} = \bar{\boldsymbol{m}} + \boldsymbol{\eta}^{(i)}$, avec $\boldsymbol{\eta}$ suivant une loi normale centrée de précision $\boldsymbol{Q}$. Cette fois la difficulté repose plus sur la factorisation de cette matrice ou de son inverse.

8.2.2. Algorithmes existants

Depuis longtemps la simulation de loi normale est faisable en utilisant la propriété qu'une transformation linéaire de loi normale est une loi normale.

— Si l'on dispose de la factorisation de Cholesky de la covariance $\boldsymbol{Q}^{-1} = \boldsymbol{L}\boldsymbol{L}^t$ alors on obtient aisément un échantillon avec $\boldsymbol{\eta} = \boldsymbol{L}\boldsymbol{\epsilon}$, avec $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Seulement cette factorisation est faisable uniquement pour des dimensions réduites.

— Si l'on dispose de la factorisation de Cholesky de la précision $\boldsymbol{Q} = \boldsymbol{L}\boldsymbol{L}^t$ alors on obtient un échantillon avec la résolution d'un système linéaire $\boldsymbol{L}\boldsymbol{\eta} = \boldsymbol{\epsilon}$. Les contraintes sont les mêmes que pour le cas précédent[7].

7: la résolution du système étant plutôt aisée, $\boldsymbol{L}$ étant triangulaire.

Les deux cas précédents sont classiques. On peut également mentionner deux autres possibilités.

— Des cas particuliers où $\boldsymbol{Q}$ serait diagonalisable aisément dans une base $\boldsymbol{Q} = \boldsymbol{F}^*\boldsymbol{\Lambda}\boldsymbol{F}$, comme celle de Fourier par exemple pour les problèmes de déconvolution. Dans ce cas un échantillon s'obtient avec $\boldsymbol{\eta} = \boldsymbol{F}^\dagger\boldsymbol{\Lambda}^{-1/2}\boldsymbol{F}\boldsymbol{\epsilon}$. Ce cas est restreint à la disponibilité d'une telle base.

— L'échantillonnage de Gibbs en simulant pixel par pixel [91], chaque pixel suivant une loi gaussienne scalaire conditionnellement aux autres. Ce cas de figure est très lent et nécessite de nombreuses itérations pour atteindre la convergence.

Enfin on peut mentionner des algorithmes plus récents, dits hybrides, qui exploitent des informations de courbure comme en optimisation pour orienter un échantillonnage de Gibbs.

Langevin Monte-Carlo

L'algorithme **LMC** pour Langevin Monte-Carlo [58], [86] propose de simuler des échantillons d'une loi π par résolution d'une équation différentielle stochastique de type Langevin

$$X = \nabla \log \pi(X) + \sqrt{2}\dot{W}$$

où W est un mouvement brownien. Dans ce cas, la distribution stationnaire ρ de X est π. Dans le cas d'une résolution par discrétisation[8] on utilise l'équation récurrente[9]

$$X_{k+1} := X_k + \tau \nabla \log \pi(X_k) + \sqrt{2\tau}\xi_k,$$

avec ξ_k suivant une loi centrée réduite et τ un pas. Dans ce cas on parle d'**ULA** pour Unajusted Langevin Algorithm [18] car X n'a plus π comme loi stationnaire. Si on ajoute une étape Metropolis-Hastings on parle de **MALA** pour Metropolis Adjusted ce qui permet d'avoir la bonne convergence. Notons également que **P-MALA** pour Prox-MALA a été proposé [38] comme adaptation pour gérer les distributions π non continûment différentiables.

8: le cas pratique quasi exclusif par conséquent.

9: dite de Euler–Maruyama.

Hamiltonian Monte-Carlo

HMC pour *Hamiltonian Monte Carlo* [41], [59], associe x à la position d'une particule possédant une quantité de mouvement p dont le déplacement est dicté par un Hamiltonien $H(x,p) = -\log \pi(x) + p^t \boldsymbol{M}^{-1} p/2$ avec $\boldsymbol{M}$ une matrice de masse. La résolution se fait également par discrétisation avec L pas de taille Δ_t appelés *leap frog*, terminés par une étape de Metropolis-Hastings. Le réglage de Δ_t peut se faire comme avec une marche aléatoire adaptative.

L'algorithme **NUTS** pour *No U-Turn Sampler* [44] propose de régler L automatiquement. Avec les outils récents de différentiation automatique permettant de calculer directement le gradient du potentiel à partir du code informatique, NUTS est devenu l'algorithme standard pour les problèmes de grande dimension présents dans les *toolbox* telles que STAN ou PyMC ou autre PPL *.

Discussion

Ces deux derniers algorithmes sont populaires, mais peuvent en pratique s'avérer décevants, en particulier pour les problèmes inverses souvent mal conditionnés. MALA est l'équivalent d'un algorithme du premier ordre et peut nécessiter de nombreuses itérations pour converger, bien qu'il soit possible d'utiliser une version preconditionnée. Par ailleurs, ce sont deux algorithmes qui convergent à pas fixe, et dans les problèmes mal conditionnés le pas est souvent petit, imposant également de nombreuses itérations. Dans les cas non différentiables, c'est la constante de Lipschitz qui imposera un pas petit[10].

10: Par exemple dans le travail [38] sur l'exemple de déconvolution sur une image 128×128, l'auteur mentionne une période de chauffe d'un million d'échantillons et de 20 000 échantillons pour calculer l'EAP, sans préciser de temps de calcul.

*. Probabilistic Programming Language

8.3. Perturbation, Optimisation

[53]: Orieux et al. (2012), « Sampling High-Dimensional Gaussian Distributions for General Linear Inverse Problems »

Pour la simulation de loi normale, nous avons proposé dans [53] un algorithme qui exploite la structure classique rencontrée dans la résolution de problèmes inverses Eq. (8.8). Dans ce cas la loi que l'on souhaite simuler est gaussienne $\mathcal{N}(\bar{\boldsymbol{m}}, \boldsymbol{Q}^{-1})$. L'utilisation de Cholesky est impossible par la taille de $\boldsymbol{Q}$ et la nécessité de l'inverser.

Comme mentionné dans [39], une alternative est la simulation de $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q})$ suivie de la résolution de $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{\eta}$ soit $\boldsymbol{x} = \boldsymbol{Q}^{-1}\boldsymbol{\eta}$. Dans ce cas on a $\mathbb{V}[\boldsymbol{x}] = \mathbb{E}\left[\boldsymbol{Q}^{-1}\boldsymbol{\eta}\boldsymbol{\eta}^t\boldsymbol{Q}^{-1}\right] = \boldsymbol{Q}^{-1}\boldsymbol{Q}\boldsymbol{Q}^{-1} = \boldsymbol{Q}^{-1}$. Pour résumer, la méthode doit être capable

- de simuler avec une covariance $\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q}$ (pour obtenir une covariance $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$),
- résoudre le système linéaire avec la matrice normale $\boldsymbol{Q}$ et
- calculer la moyenne $\bar{\boldsymbol{m}}$.

Ces trois points sont résolus par les algorithmes **PO** pour Perturbation–Optimisation. Notamment le troisième point nécessite souvent l'inversion de la matrice normale.

L'algorithme consiste à générer des échantillons suivant les lois *a priori* $\mathrm{p}(\boldsymbol{\epsilon}_k) = \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{\Sigma}_k^{-1}\right)$, ce qui revient à perturber les moyennes $\tilde{\boldsymbol{m}}_k = \boldsymbol{m} + \boldsymbol{\epsilon}_k$. Ensuite on construit un critère perturbé (étape P)

$$J(\boldsymbol{x}) = \sum_k \|\tilde{\boldsymbol{m}}_k - \boldsymbol{H}_k\boldsymbol{x}\|^2_{\boldsymbol{\Sigma}_k^{-1}} \tag{8.9}$$

qui a pour minimum

$$\bar{\boldsymbol{m}} = \boldsymbol{Q}^{-1}\sum_k \boldsymbol{H}_k\boldsymbol{\Sigma}_k^{-1}\tilde{\boldsymbol{m}}_k. \tag{8.10}$$

Dans le cas où $\boldsymbol{Q}$ ne peut pas être inversé facilement, ce minimum peut être calculé par un algorithme d'optimisation comme un gradient conjugué linéaire [87] (étape O). La preuve pour montrer que (8.10) est plutôt facile : $\bar{\boldsymbol{m}}$ est un vecteur gaussien, car combinaison linéaire de vecteurs $\boldsymbol{\epsilon}_k$ gaussiens, dont le calcul de la moyenne et de la covariance se fait à l'aide de calculs d'algèbre linéaire [53].

[53]: Orieux et al. (2012), « Sampling High-Dimensional Gaussian Distributions for General Linear Inverse Problems »

Variante RJPO

Dans sa version initiale, que l'on pourrait qualifier d'EPO pour *Exact PO*, on suppose que l'optimisation est menée à son terme et que le résultat est bien $\bar{\boldsymbol{m}}$. Seulement PO est le plus souvent utilisé dans les cas où la dimension est grande et où justement les itérations d'un algorithme d'optimisation sont habituellement tronquées. Dans ce cas on parle de TPO pour *Truncated PO* et le résultat est connu pour être biaisé. Je présente partie 8.4 une proposition alternative pour tenter de garder de bonnes propriétés dans ce cas.

Cependant l'alternative la plus aboutie pour le moment est

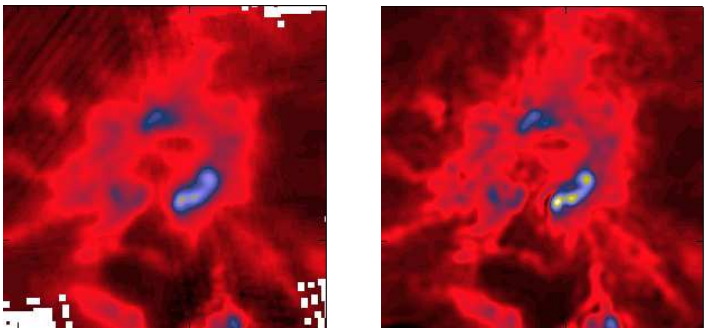


Figure 8.1. – Reconstruction à partir de données réelles de l'instrument SPIRE d'Herschel. À gauche les données brutes et à droite la reconstruction par inversion et estimation des hyperparamètres.

la variante RJPO [39] qui propose une variante reposant sur les sauts réversibles à dimension constante pour calculer le critère d'acceptation d'un Metropolis-Hastings permettant de garantir que l'échantillon produit est bien sous la bonne loi. Les auteurs proposent également une variante adaptative qui permet de régler automatiquement le niveau de résolution du système linéaire.

Discussion

Cet algorithme se compare directement aux algorithmes présentés précédemment, notamment NUTS. Les deux exploitent des principes d'optimisation et d'utilisation du gradient pour obtenir des échantillons, avec des perturbations stochastiques. La comparaison s'arrête là [16] et les divergences sont multiples. PO est un algorithme exact qui produit des échantillons indépendants quand NUTS est un algorithme MCMC classique. Par ailleurs PO est pour la simulation de lois gaussiennes et utilise cette structure alors que NUTS est plus générale.

En termes d'efficacité PO est plus efficace, un test rapide utilisant PyMC [11], en petite dimension pour utiliser EPO, montre que ce dernier est environ deux fois plus rapide. NUTS doit produire plus d'échantillons corrélés et donc faire plus d'appels au calcul du gradient. Par ailleurs ce test s'est fait sur un problème jouet bien conditionné. Les différences s'accentuent sur des problèmes de grande taille mal-posés et RJPO qui nécessite uniquement une résolution partielle du système.

11: utilisant lui même PyTensor pour compiler le code calculant le gradient

8.3.1. Applications

L'algorithme PO a été appliqué avec succès sur plusieurs problèmes d'inversion et de reconstruction non supervisée. Il a été appliqué sur des problèmes de super-résolution et déconvolution [53], en microscopie [30] (voir le chap. 7), en segmentation avec estimation de paramètres de texture [48] en astronomie par exemple [65], [66], illustré figure 8.1

Un des apports de ces algorithmes d'échantillonnage efficaces est la possibilité de les inclure dans un échantillonneur de Gibbs qui estime les hyperparamètres, mais également les autres moments de la loi, comme la variance de l'objet d'intérêt pour l'estimation

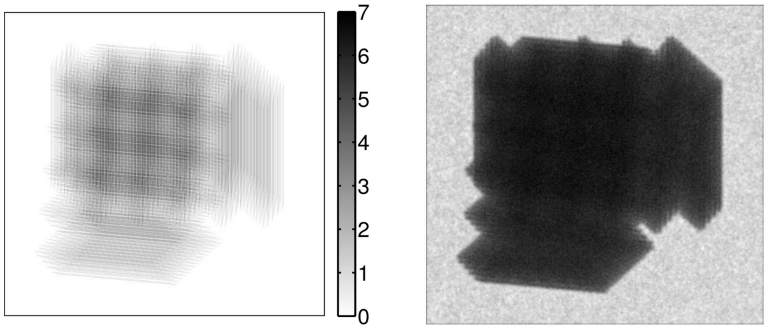


Figure 8.2. – Carte d'estimation de l'incertitude de reconstruction en super-résolution non supervisée. L'image de gauche est la carte de redondance, où plus un pixel est sombre, plus il est observé. L'image de droite est la carte d'incertitude, c'est-à-dire la diagonale de la matrice de covariance *a posteriori*. On peut observer que plus un pixel est observé, moins il y a d'incertitude.

d'une carte d'incertitude. La figure 8.2 est une illustration de cette estimation sur un cas de super-résolution.

8.4. Gradient Scan

Une des limitations de l'algorithme PO est la nécessité d'obtenir le minimum du critère (8.9). Malheureusement pour des problèmes de grande dimension, l'optimisation est habituellement tronquée après N itérations et l'itérée ne suit pas la bonne loi dans ce cas. La solution de base est de fixer un nombre important d'itérations. Dans ce cas l'algorithme « marche » mais cela reste insatisfaisant.

Plusieurs approches ont été développées. Tout d'abord si on utilise un algorithme de gradient conjugué, décrit 1, on peut écrire l'itérée courante

$$\boldsymbol{x}^{(n)} = \boldsymbol{x}^{(0)} + \sum_{i=0}^{n} \alpha^{(i)} \boldsymbol{d}^{(i)}$$

avec les formules du gradient conjugué pour le pas α et la direction $\boldsymbol{d}$. Malheureusement bien que les formules soient explicites il n'y a pas encore, pour le moment, de loi identifiée pour $\boldsymbol{x}^{(n)}$.

Algorithme 1 : Gradient conjugué linéaire pour la résolution de $\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b}$.

Set $\boldsymbol{x}^{(0)}$ and $n \leftarrow 0$;
$\boldsymbol{r} \leftarrow \boldsymbol{b} - \boldsymbol{Q}\boldsymbol{x}^{(0)}$;
$\boldsymbol{d}^{(0)} \leftarrow \boldsymbol{r}$;
$\delta^{(0)} \leftarrow \boldsymbol{r}^t\boldsymbol{r} = \|\boldsymbol{r}\|^2$;
repeat
  $\boldsymbol{q} \leftarrow \boldsymbol{Q}\boldsymbol{d}$;
  $\alpha \leftarrow \delta^{(n)}/\boldsymbol{d}^t\boldsymbol{q}$;
  $\boldsymbol{x}^{(n+1)} \leftarrow \boldsymbol{x}^{(n)} + \alpha\boldsymbol{d}$;
  $\boldsymbol{r} \leftarrow \boldsymbol{r} - \alpha\boldsymbol{q}$;
  $\delta^{(n+1)} \leftarrow \boldsymbol{r}^t\boldsymbol{r}$;
  $\beta \leftarrow \delta^{(n+1)}/\delta^{(n)}$;
  $\boldsymbol{d} \leftarrow \boldsymbol{r} + \beta\boldsymbol{d}$;
  $n \leftarrow n + 1$;
until *Some criterion is meet*;

Nous avons proposé dans [42] et [37] l'algorithme *Gradient Scan Gibbs Sampler* reposant sur les idées du *Gradient Direction Sampler* (CDS) de Colin Fox [68] et du *random scan* [77].

L'idée principale repose sur le fait que l'algorithme du gradient conjugué produit une factorisation de la matrice $\boldsymbol{Q}$ au cours des itérations. En effet, un ensemble de directions conjuguées [12] $\{\boldsymbol{d}_1, \dots, \boldsymbol{d}_N\}$, forme une base de $\mathbb{R}^N$ avec

12: ou $\boldsymbol{Q}$ orthogonales, tel que $\boldsymbol{d}_n^t \boldsymbol{Q} \boldsymbol{d}_m = 0$ pour n et $m = 1, \dots, N$, avec $n \neq m$

$$\boldsymbol{x} = \sum_{n=1}^{N} \alpha_n \boldsymbol{d}_n \quad \text{avec} \quad \alpha_n = \frac{\boldsymbol{d}_n^t \boldsymbol{Q} \boldsymbol{x}}{\boldsymbol{d}_n^t \boldsymbol{Q} \boldsymbol{d}_n}.$$

Par conséquent, si $\boldsymbol{x} \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{Q}^{-1}\right)$ est un vecteur gaussien de moyenne $\boldsymbol{m}$ et de précision $\boldsymbol{Q}$ alors les « pas » α_n sont, conditionnellement aux $\boldsymbol{d}_n$, également gaussiens de loi

$$\alpha_n \sim \mathcal{N}\left(\frac{\boldsymbol{d}_n^t \boldsymbol{Q} \boldsymbol{m}}{\boldsymbol{d}_n^t \boldsymbol{Q} \boldsymbol{d}_n}, \frac{1}{\boldsymbol{d}_n^t \boldsymbol{Q} \boldsymbol{d}_n}\right).$$

Et réciproquement si on simule des pas α_n suivant la loi ci-dessus, alors $\boldsymbol{x} = \sum_{n=1}^{N} \alpha_n \boldsymbol{d}_n$ est un vecteur gaussien de moyenne $\boldsymbol{m}$ et covariance $\boldsymbol{Q}^{-1}$.

[42]: Orieux et al. (2015), « Gradient Scan Gibbs Sampler : An Efficient High-Dimensional Sampler Application in Inverse Problems »

Dans [42] et [37] nous proposons deux contributions reposant sur cette factorisation. Tout d'abord une robustesse supérieure à [68] dans le cas où l'itérée courante est proche d'une direction conjuguée, car dans ce cas l'algorithme a tendance à rester sur cette direction. On obtient ce résultat à l'aide d'une perturbation sur la direction initiale.

La deuxième contribution est l'inclusion de l'algorithme, avec des itérations tronquées, c'est-à-dire avec N inférieur à la dimension de $\boldsymbol{x}$, lorsqu'il est inclus dans un échantillonneur de Gibbs pour, par exemple, échantillonner les hyperparamètres. Je renvoie à [42] pour les détails.

Cet algorithme présente plusieurs avantages. Tout d'abord il est garanti convergent malgré la troncature des itérations, car dans les directions explorées c'est un échantillonneur de Gibbs qui simule exactement suivant les lois conditionnelles *a posteriori* pour α_n (scalaires gaussiennes). Ensuite les preuves ne dépendent pas du nombre de directions explorées ce qui permet potentiellement d'accélérer l'algorithme global. Il n'y a cependant pas d'étude sur la qualité ou la vitesse d'exploration de l'espace des paramètres et dans les faits, la qualité est sensiblement inférieure à RJPO. Cependant RJPO est un algorithme qui produit des échantillons *indépendants* au sein de l'algorithme MCMC, ce qui est *a priori* plus efficace en termes d'exploration de l'espace et plus coûteux en calculs numériques.

8.5. Déconvolution convexe

Les algorithmes présentés ici s'attachent à simuler une loi gaussienne. Celle-ci correspond à un terme d'attache aux données et une régularisation quadratique dans le cadre variationnel. Une voie de recherche est dès lors de simuler efficacement des lois non gaussiennes. Cependant cette tâche est significativement plus complexe en particulier si l'on souhaite simuler les hyper paramètres.

Problème posé

Si l'on conserve l'attache aux données gaussienne, la loi *a priori* peut s'écrire

$$\mathrm{p}(\boldsymbol{x} \mid \boldsymbol{\theta}) = K_{\boldsymbol{\theta}}^{-1} \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x})) \tag{8.11}$$

avec $E_{\boldsymbol{\theta}}$ l'énergie de Gibbs. De nombreuses énergies E_θ sont possibles, mais celles qui nous intéresses sont en relation avec les pénalités non quadratiques comme les pénalités $\ell_1(\boldsymbol{x}) = |\boldsymbol{x}|$ ou $\ell_{2,1}$ convexes, soit quadratiques à l'origine, puis à tendance linéaire comme le potentiel de Huber [13].

13: Le potentiel de Huber, $\ell_{2,1}$ par construction

$$\varphi(u) = \begin{cases} \frac{1}{2}u^2, & \text{si } u \leq \delta, \\ \delta|u| - \frac{\delta^2}{2}, & \text{sinon.} \end{cases}$$

Dans ce cas plusieurs difficultés se posent. Tout d'abord il y a une difficulté si E induit des corrélations dans $\boldsymbol{x}$ avec par exemple

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mu \sum_{c \in \mathscr{C}} \phi\left(\boldsymbol{d}_c^t \boldsymbol{x}\right).$$

Dans ce cas l'ensemble des vecteurs calculant les combinaisons linéaires de $\boldsymbol{x}$ produisent des corrélations *a priori* contrairement à une pénalité indépendante, par exemple

$$E_{\boldsymbol{\theta}}(\boldsymbol{x}) = \mu \sum_i |x_i|$$

dans le cas du *Bayesian Lasso* [71], permettant de favoriser la parcimonie. La difficulté est que ces corrélations déterminent la fonction de partition

$$K_{\boldsymbol{\theta}} = \int \exp(-E_{\boldsymbol{\theta}}(\boldsymbol{x}))\, \mathrm{d}x$$

rendant rapidement son calcul impossible contrairement au cas séparable. De plus, les corrélations *a priori* sont présente *a posteriori* ce qui peut compliquer la manipulation de cette dernière, pour l'échantillonnage notamment.

L'estimation des hyperparamètres $\boldsymbol{\theta}$ présente également une difficulté importante, car la fonction de partition $K_{\boldsymbol{\theta}}$, à nouveau, dépend de ces paramètres. Pour une inférence non supervisée, où les paramètres $\boldsymbol{\theta}$ sont également inconnus, il faut connaître l'expression de K en fonction de ces derniers, ou alors utiliser des algorithmes d'intégration numérique qui ne nécessitent pas sa connaissance avec un coût induit plus important.

Enfin la dernière difficulté concerne l'exploration *a posteriori*

d'une loi ayant ce type d'*a priori*. Les algorithmes disponibles pour produire des échantillons de loi de grande dimension avec des corrélations sont pour LMC [58], [86] et HMC [41], [59] présentés partie 8.2.2 pour des lois log concave différentiable. Dans le cas non différentiable on peut mentionner l'algorithme P-MALA qui est un algorithme de Langevin utilisant l'approximation de Moreau pour utiliser des algorithmes proximaux [38]. Plus récemment Vono *et. al.* [8] ont proposé un algorithme inspiré du *variable spliting* de l'ADMM, le *Split Gibbs Sampler*, mais cet algorithme échantillonne une loi approchante, une approximation, et non la loi cible.

Proposition

Les algorithmes LMC sont des algorithmes du premier ordre utilisant le gradient pour produire un échantillon. Outre les problèmes soulevés plus haut, ils sont peu étudiés lorsque l'on souhaite également estimer les hyper paramètres, c'est à dire que conditionnellement à $\boldsymbol{\theta}$ la loi change. Dans [29], nous avons proposé d'exploiter plutôt une structure conditionnellement gaussienne.

[29]: Orieux et al. (2017), « Semi-Unsupervised Bayesian Convex Image Restoration with Location Mixture of Gaussian »

Dans ce cas de figure la loi *a priori* est défini comme

$$\begin{aligned} \mathrm{p}(\boldsymbol{x} \mid \gamma_x, \gamma_b) &= \int_b \mathrm{p}(\boldsymbol{x} \mid \gamma_x, \boldsymbol{b})\, \mathrm{p}(\boldsymbol{b} \mid \gamma_b)\, \mathrm{d}\boldsymbol{b} \\ &\propto \int_b \exp\left(-\frac{\gamma_x}{2}\|\boldsymbol{D}\boldsymbol{x} - \boldsymbol{b}\|^2\right) \prod_i \mathrm{p}(\boldsymbol{b}_i \mid \gamma_b)\, \mathrm{d}\boldsymbol{b} \end{aligned}$$

avec $\boldsymbol{b}$ une variable latente, auxiliaire, ou cachée suivant les points de vue ; et séparable. L'*a priori* sur $\boldsymbol{x}$ est définie comme la marginale sur $\boldsymbol{b}$ de la loi jointe. Il y a deux avantages : tout d'abord les algorithmes MCMC permettent de faire naturellement des marginalisations, donc le modèle échantillonné est bien $\mathrm{p}(\boldsymbol{x})$, et la loi *a posteriori* conditionnelle pour $\boldsymbol{x}$ est gaussienne si la vraisemblance est gaussienne. Le deuxième avantage est que l'on dispose d'algorithmes efficaces pour la simulation de loi conditionnellement gaussienne.

La loi ci-dessus dépend de la loi de mélange. Dans [29] nous avons utilisé la même loi que celle proposée par [69] c'est à dire $\mathrm{p}(\boldsymbol{b}_i)$ comme une loi de Laplace et $\mathrm{p}(\boldsymbol{x})$ une loi log-concave [14] LogErf. Avec un modèle de données $\boldsymbol{H}$ circulant et un opérateur de régularisation $\boldsymbol{D}$ également, on peut produire des algorithmes très rapides, de l'ordre de quelques secondes pour une image 256×256. La figure 8.3 présente un résultat obtenu.

14: La convolution de deux lois log-concave étant log-concave.

Par ailleurs cet algorithme permet également, comme algorithme MCMC, d'estimer les hyperparamètres, excepté la variance du bruit pour le moment, ainsi que les moments d'ordre deux pour estimer l'incertitude. La figure 8.4 présente le résultat correspondant à la figure 8.3.

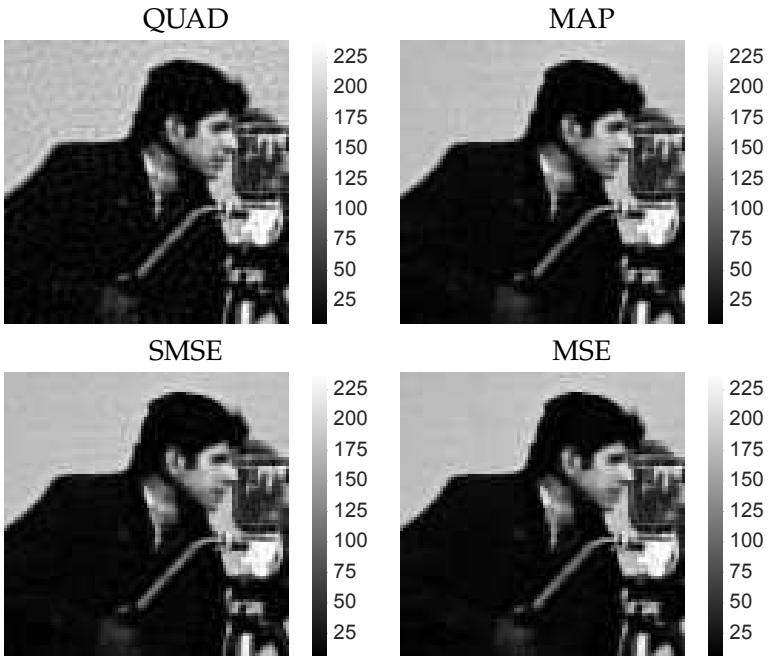


Figure 8.3. – La figure présente un résultat de déconvolution non supervisée (avec estimation des hyperparamètres) sur le cameraman utilisant le même modèle, mais avec des estimateurs différents. L'image QUAD est la restauration avec un *a priori* quadratique. L'image MAP est le résultat de l'optimisation du critère convexe défini comme l'antilogarithme de la loi jointe, à hyperparamètres fixés au meilleur au sens de l'erreur quadratique de reconstruction (connaissant le vrai). C'est un résultat classique de restauration d'image avec préservation de bord, le potentiel n'étant pas Huber, mais LogErf, tout à fait similaire. L'image SMSE correspond au MSE à hyperparamètres fixés également, obtenue avec un algorithme MCMC. On constate que les résultats entre le MAP et le MSE sont similaires. Enfin l'image MSE est le résultat de l'algorithme proposé [29] *avec* estimation des hyperparamètres, excepté la variance du bruit. Le résultat est de qualité équivalente à ceux réglés connaissant le vrai, tout en étant non supervisé.

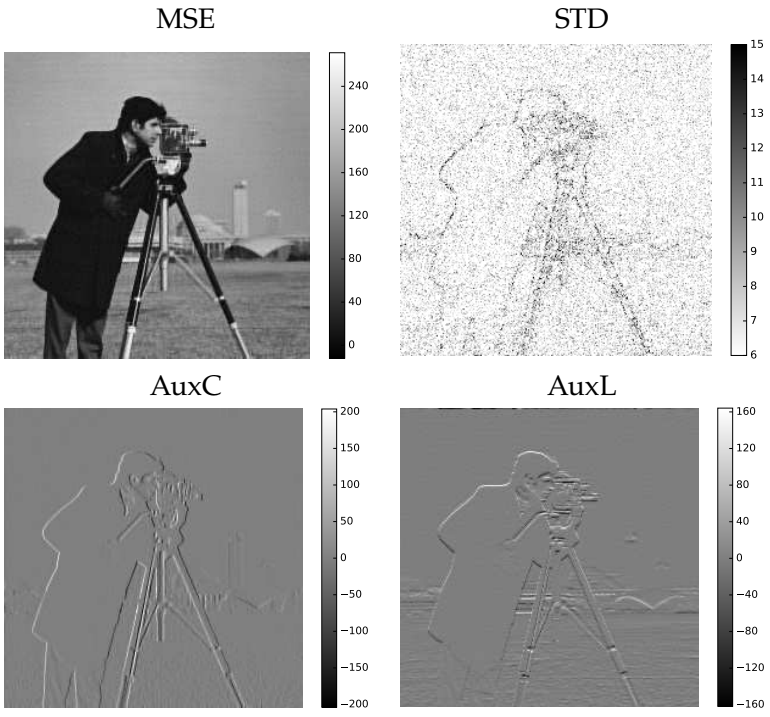


Figure 8.4. – La figure MSE est le résultat d'une déconvolution non supervisée. La figure STD présente l'estimation de la diagonale de la covariance *a posteriori*, soit la variance de chaque pixel. On constate que les bords sont les plus sujets à l'incertitude, ce qui semble logique puisque c'est pour eux que l'information est le plus dégradée par la convolution produisant les mesures. Les figures AuxC et AuxL sont les estimations des variables latentes $\boldsymbol{b}$ correspondant à l'*a priori* sur les gradients en colonne et en ligne, respectivement. Ces résultats sur une image 256×256 ont été obtenus en moins d'une minute avec une période de chauffe de 100 itérations.

9. Perspectives

9.1. Perspectives méthodologiques

Les trois prochaines parties présentent des perspectives méthodologiques. Les deux premières sont dans la continuité de mes précédents travaux. La première 9.1.1, sur l'estimation des hyperparamètres, s'envisage à plus court terme, car j'ai déjà mené quelques travaux préliminaires exposés partie 8.5. La deuxième 9.1.2 sur l'accélération d'algorithme MCMC est plutôt à moyen terme et repose sur les travaux initiés en délégation, mais qui ont besoin encore de temps. Enfin la troisième 9.1.3 sur l'apprentissage machine est plus risquée et à plus long terme. C'est un sujet beaucoup plus exploratoire avec une littérature foisonnante qui nécessite beaucoup plus d'investissement. Les perspectives applicatives sont dans la continuité de mes travaux actuels.

9.1.1. Estimation d'hyperparamètres avec des modèles non gaussiens

Comme présenté dans le chapitre 8, les algorithmes MCMC sont envisageables pour l'estimation des hyper paramètres ou des moments d'ordre supérieur comme la variance. Cependant les problèmes en grande dimension, comme les reconstructions d'images hyperspectrales par exemple, posent de sévères contraintes de temps de calcul sur les algorithmes. En parallèle, les modèles instrument ou de régularisation gagnent en complexité pour augmenter leur capacité de représentation. Dans le chapitre 8 j'ai présenté essentiellement des algorithmes pour la simulation d'échantillons gaussiens qui possèdent plusieurs limitations.

Elle est continue donc inadaptée pour des variables discrètes. Les variables discrètes se rencontrent pour des problèmes de segmentation par exemple [61] avec des variables de classe ou d'étiquette. Elles sont également rencontrées dans des problèmes de parcimonie avec les modèles Bernoulli-Gaussien [3], [60].

Le cas d'usage le plus souvent rencontré est la préservation des bords, c'est-à-dire la présence de fort gradient dans le signal. La loi gaussienne correspond à la pénalité ℓ_2 connue pour ne pas modéliser correctement cette information et introduire des artefacts de rebond au niveau de ces transitions. Dans ce cas on utilise alors plutôt une régularisation R avec préservation des

bords à l'aide d'une pénalité de type $\ell_{2,1}$

$$R(\boldsymbol{x}) = \sum_{c \in \mathcal{C}} \phi_c(\boldsymbol{d}_c^t \boldsymbol{x})$$

où ϕ_c est par exemple la pénalité de Huber, ou alors la pénalité ℓ_1 donnant la *variation totale* [76]

$$R(\boldsymbol{x}) = |\nabla \boldsymbol{x}|.$$

par exemple isotrope

$$|\nabla \boldsymbol{x}| = \sum_{i,j} \sqrt{\left|x_{i+1,j} - x_{i,j}\right|^2 + \left|x_{i,j+1} - x_{i,j}\right|^2}$$

. Dans les deux cas, cela conduit à une densité de probabilité qui n'est pas gaussienne (tout en restant log concave).

Pour finir, la loi gaussienne est également sans contraintes de support ce qui dans certaines situations est limitant. Dans le cas de l'imagerie computationnelle, on rencontre souvent des problèmes de positivité ($x > 0$) ou de non-négativité ($x \geq 0$) des pixels de l'image. Les contraintes peuvent également concerner le support ou tout simplement la valeur de paramètres de modèle comme une largeur de réponse impulsionnelle.

Les différentes limitations mentionnées ont naturellement fait l'objet de travaux. On peut mentionner [69] pour la simulation de champ non stationnaire non gaussien avec un mélange de gaussienne sur la moyenne, ainsi que [38] pour l'algorithme P-MALA qui propose la simulation de champ non gaussien présentant une pénalité non différentiable. Concernant les travaux avec des variables d'étiquette, énormément de travaux ont été menés sur les champs de Potts [61]. Enfin les algorithmes de type Langevin ou hamiltoniens comme NUTS sont construits pour simuler des champs de grande dimension ayant un potentiel différentiable. Cependant les algorithmes P-MALA ou NUTS peuvent s'avérer peu efficaces dans le cas non supervisé.

Pour ma part je souhaite mener des développements pour exploiter autant que possible l'algorithme PO ou assimilé, présenté partie 8.3, pour simuler des champs Gaussiens. C'est un algorithme efficace pour les problèmes en grandes dimensions et qui a fait ses preuves dans les cas non supervisés où l'on estime les hyper paramètres et paramètres des modèles.

Plusieurs pistes sont possibles. Tout d'abord le modèle de [29] qui est un mélange de gaussienne sur la moyenne

$$\mathrm{p}(x \mid \gamma_x, \gamma_b) = \int_b g(x \mid b, \gamma_x)\, \mathrm{p}(b \mid \gamma_b)\, \mathrm{d}b$$

où $g(x|b,\gamma) = \mathcal{N}(b, 1/\gamma_x^{1/2})$, et qui présente quelques difficultés.

Tout d'abord l'algorithme n'est pas capable d'estimer correctement le niveau de bruit, mais uniquement les deux paramètres de l'*a priori*[1]. Deuxièmement, l'algorithme d'échantillonnage des moyennes latentes $\boldsymbol{b}$ par inversion de fonction de répartition présente facilement des erreurs numériques. Proposer un autre algorithme plus robuste et qui serait quasiment aussi efficace est un défi. Une première approche pourrait être d'exploiter NUTS ou les algorithmes de rejet. Par ailleurs la loi de mélange $\mathrm{p}(b)$ est un degré de liberté qu'il pourrait être intéressant d'exploiter pour simplifier l'échantillonnage ou sophistiquer le modèle.

1: soit le niveau pour $\boldsymbol{x}$ et celui des moyennes latentes $\boldsymbol{b}$

Une autre approche serait le mélange de Gaussiennes par la variance

$$\mathrm{p}(x \mid \gamma_v) = \int_b g(x \mid 0, v)\, \mathrm{p}(v \mid \gamma_v)\, \mathrm{d}v.$$

Ce modèle s'approche plus du type Geman & Reynolds [75], [88] et est moins exploité. Dans le cas indépendant en petite dimension, il a été exploité pour la version bayésienne de LASSO puisque la loi de Laplace s'écrit comme un mélange par la variance [76]. Cela dit en grandes dimensions avec corrélation le problème reste ouvert et comme dans [88] amène une covariance dont la structure est brisée par les variances $\{v_i\}$. Par contre de nombreuses densités connues (Laplace, Gamma, Inverse Gamma, ...) s'écrivent comme mélange par la variance et la loi marginale ne dépend que d'un hyper paramètre.

Concernant les algorithmes MCMC avec contraintes pour les problèmes inverses en grande dimension, la littérature est moins prolixe. Tout d'abord il faut noter qu'il y a une difficulté sur la modélisation. Tout d'abord pour le cas scalaire on peut distinguer la gaussienne positive, définie sur $x \geq 0$ mais avec une probabilité nulle en 0, et le cas « non-négatif » ou « rectifié », où il y a une masse de probabilité non nulle en zéro[2]. Le deuxième cas s'apparente à un modèle Bernoulli gaussien. En multidimensionnel, le problème est plus complexe et les modèles facilement incorrectement définiæ. On peut mentionner les travaux récents [3] ou encore les travaux de Bardsley et Fox [51] qui pourraient être des points de départ de l'état de l'art.

2: un pixel « positif ou nul », correspondant au cas des contraintes actives ou non en optimisation.

Certaines de ces perspectives font l'objet d'une nouvelle collaboration avec J.-F. Giovannelli dans le cadre du projet PEPR Origin pour lequel un financement de thèse a permis le recrutement de Pierre Minier en thèse à partir de septembre 2023.

9.1.2. Accélération d'algorithmes

L'algorithme PO nécessite de résoudre un problème d'optimisation pour produire un échantillon gaussien. Je propose dans la partie précédente de l'exploiter pour produire des échantillons non gaussiens. Cela dit, le fait de résoudre un problème d'optimisation est coûteux, quoique bien moindre qu'une approche par

factorisation de Cholesky par exemple.

Plusieurs voies d'accélérations sont possibles. Tout d'abord PO produit des échantillons indépendants, ce qui, en dehors d'un échantillonneur de Gibbs qui simulerait également d'autres variables, fait de PO un algorithme qui n'est pas une chaîne de Markov. C'est un atout, mais l'algorithme pourrait être retravaillé pour accepter une dépendance à l'échantillon précédent qui permettrait de gagner en rapidité. C'est le cas dans la forme présentée [53] où l'initialisation est l'échantillon précédent et les itérations sont tronquées avec cependant une perte des garanties de convergence. Au contraire RJPO [39] est garanti malgré la troncature, mais produit des échantillons indépendants et est plus lent.

Une deuxième voie d'accélération est l'utilisation de préconditionneurs. Le problème à résoudre est un système linéaire

$$\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{b} \tag{9.1}$$

avec $\boldsymbol{Q} = \sum_k \boldsymbol{M}_k^t \boldsymbol{\Sigma}_k^{-1} \boldsymbol{M}_k$ et $\boldsymbol{b} = \sum_k \boldsymbol{M}_k^t \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{y} + \boldsymbol{\epsilon}_k)$ (voir partie 8.3). La matrice normale $\boldsymbol{Q}$ présente certaines structures, qui ont été utilisées pour établir PO, mais qui pourraient également servir à construire un préconditionneur $\boldsymbol{P}^{-1}$ défini positif. Dans ce cas on ne cherche pas à résoudre, (9.1) mais

$$\boldsymbol{P}^{-1}\boldsymbol{Q}\boldsymbol{x} = \boldsymbol{P}^{-1}\boldsymbol{b} \tag{9.2}$$

3: comme un préconditionneur circulant pour un problème de déconvolution non circulante.

dans l'idée où $\boldsymbol{P}^{-1} \approx \boldsymbol{Q}^{-1}$ tout en étant facilement calculable et rapide à appliquer. La plupart des approches proposées reposent soit sur la construction de préconditionneur ad hoc en fonction du problème [3], soit sur des approches numériques comme des préconditionneurs faits de factorisation QR tronquée, soit de préconditionneur simple, mais estimé comme des préconditionneurs diagonaux [43]. Je pense que pour les problèmes inverses de grandes dimensions, il y a des alternatives qui peuvent être à mi-chemin entre les approches purement numériques et les approches ad hoc où les préconditionneurs exploitent la structure, Tœplitz par exemple [32], tout en répondant à un niveau d'optimalité en minimisant une norme d'erreur, $\|\boldsymbol{P}^{-1}\boldsymbol{Q} - \boldsymbol{I}\|_F$ étant la plus classique.

Il s'agit actuellement d'un projet de recherche mené en collaboration avec J. Idier et S. Moussaoüi du LS2N de Nantes. J'ai mené un projet de délégation sur la période 2021–2023. L'épidémie de Covid-19 a perturbé significativement le déroulé du projet, mais nous avons mené des premiers développements avec succès en utilisant [64] comme point de départ.

9.1.3. Apprentissage Machine pour les problèmes inverses

> « Je découvris qu'un réseau neural est un instrument que l'on utilise uniquement pour résoudre des problèmes trop complexes pour être compris en eux-mêmes. S'il est assez souple, un tel réseau peut être configuré à l'aide de boucles de rétroaction pour imiter n'importe quel système ou presque, pour produire les motifs de sortie à partir des mêmes données en entrée. Mais cela n'éclaire en rien la nature du dispositif que l'on stimule ainsi.
>
> " La compréhension, affirma la conférencière, est un concept très surfait. Personne ne *comprend* vraiment comment un œuf fertilisé se transforme en un être humain. Quelle attitude devrions-nous adopter ? Cesser d'avoir des enfants jusqu'à ce que l'ontogenèse soit complètement décrite par une série d'équations différentielles ? " »
>
> La nouvelle *En apprenant à être moi*, Greg Egan, 2013.

L'apprentissage machine, et les techniques *data based* reposant majoritairement sur les réseaux de neurones, est devenu incontournable. Les succès en classification, régression ou génération sont importants. Dans le cadre de l'imagerie computationnelle, des problèmes inverses au sens large, plusieurs techniques sont actuellement explorées.

Tout d'abord les approches de « bout en bout » où la base de données d'apprentissage contient les mesures et l'image à reconstruire [27] ont naturellement été les premières proposées. Elles supposent l'abandon de tout modèle et nécessitent de très gros réseaux et un volume important de données. Une autre approche qui s'inspire des approches par *variables spliting* et l'algorithme ADMM est le *Plug & Play* [22] ou le *Regularisation by Denoising* [31]. Avec ces approches, la solution au problème de minimisation des variables auxiliaires dans l'algorithme itératif ADMM s'assimile à un débruitage. On remplace alors ce débruitage par un « débruiteur arbitraire » comme un réseau de neurones. Dans ce cas de figure on perd rapidement toute propriété de convergence.

Une troisième approche est le déroulage de boucle ou *unrolling*. Dans ce cas on assimile les itérations d'un algorithme itératif à une couche d'un réseau de neurones. Plutôt que d'avoir une itération fixée par les modèles on va plutôt apprendre les itérations. L'idée initiale à été proposée par Gregor & LeCun en 2010 [62] pour FISTA, renommé LISTA pour *Learned*. Depuis de nombreuses variantes sont apparues tentant de conserver le modèle direct, l'apprentissage d'une couche dupliquée ou de N couches différentes, ou encore d'apprendre d'autres algorithmes comme un algorithme

Bayésien Variationnel [1]. Enfin on peut également mentionner les approches où c'est l'*a priori* qui est appris dans une approche variationnelle en analyse [10], ou synthèse [24].

Cela dit, l'apprentissage pour les problèmes inverses, ou lorsque des approches utilisant des modèles fiables existent, pose quelques difficultés. Tout d'abord les modèles sont non linéaires et possèdent un nombre très important de paramètres. Par conséquent il est établi que les problèmes d'optimisation utilisant ces modèles sont non convexes et sensibles à l'initialisation. Les pas, ou *learning rate*, sont ajustés pour accélérer la minimisation, sans nécessairement avoir de garantie. En pratique, le critère lui-même ne devient finalement plus une référence sur laquelle se reposer pour définir un estimateur suivi d'un algorithme convergeant, mais une sorte de guide dont on exploite la capacité à globalement orienter les poids d'un modèle vers un comportement souhaité. La métrique de performance est alors l'unique référence. Autrement dit il est souvent difficile de dire quoi que ce soit du résultat autre que « mon intuition initiale quant à la performance du modèle se vérifie dans mes métriques lors de mes tests. » Sans vouloir enlever l'importance et la valeur des tests et des métriques, cela me semble prendre le contre-pied de la tendance historique en traitement de l'information. Sauf si « la compréhension est un concept très surfait » je pense qu'un juste milieu est nécessaire [4].

4: Je pense notamment à un séminaire où l'orateur présente les approches *data based* supérieures à celles basées sur un modèle puisqu'il suffit alors de générer des données.

Par ailleurs, l'apprentissage dans sa forme actuelle nécessite l'accès à des moyens de calcul importants, ce qui n'est pas très inclusif, peu économe en énergie, ni très écologique malheureusement. La taille de la base de données est également un point critique. Ce dernier point est régulièrement problématique lorsque les bases ne correspondent pas au problème traité. Les techniques de *few-shot learning* ou *transfert learning* restent applicables si le *domain gap* n'est pas très important.

Pour conclure, sans tomber dans un effet de mode ou dans une position uniquement critique, je pense que du travail reste à faire pour positionner correctement l'apport de l'apprentissage dans les problématiques de reconstruction, d'estimation, et d'inversion qui ne sont pas par nature des problèmes d'apprentissage. L'accès à des bases de données, la sophistication des modèles, permit notamment par la maturation de l'autodifférentiation de fonctions définies par du code [5], ne sont pas à laisser de côté pour ceux qui veulent encore comprendre. Pour ma part deux pistes m'intéressent. Tout d'abord les approches basées sur le *prior learning* où l'on remplace un *a priori* ad hoc de style TV par un *a priori* appris. La formulation basée sur les *implicit layers* [9] me semble très intéressante. La question du *domain gap* demeure cependant s'il n'y a pas de base de données importante comme en astrophysique. Je pense que les approches utilisant les réseaux génératifs comme les GAN ou les VAE introduisent plus d'*a priori* que nécessaire,

5: technologie à l'impact peut être sous évaluée

permettant une sur performance mais sur des données de test très marquées. Enfin, je m'intéresse à l'accélération des algorithmes et solutions existantes par l'assistance d'outils exploitant les solutions venant de l'apprentissage. On peut par exemple penser à des pré conditionneurs.

9.2. Perspectives applicatives

9.2.1. Fusion d'images multi et hyperspectrale

Le chapitre 6 présente des travaux sur la reconstruction d'images multispectrales et hyperspectrales qui présentent des problèmes de flou spatiaux ou spectraux dus à l'effet de diffraction de la lumière. Pour le moment les travaux ont porté sur des reconstructions indépendantes. Cependant dans certaines situations des observations de la même scène sont faites avec les deux types d'instruments, ou d'images. Dans ce cas de figure, l'image multispectrale présente le plus souvent une bonne résolution spatiale avec un grand champ, quand l'image hyperspectrale sera plutôt de champ réduit avec une plus faible résolution spatiale, mais une bonne résolution spectrale [6]. C'est notamment le cas du JWST.

6: Ces différences sont dues aux différences de technologie des instruments.

Une perspective est donc naturellement de faire de la fusion de données, proposée par Claire Guilloteau [12] en 2020 pour le JWST et les instruments NirCam et NirSpec (proche infrarouge). La méthode proposée repose sur un modèle de mélange et la prise en compte d'un flou spatial variant avec la longueur d'onde. Le problème est résolu en minimisant un moindre carré régularisé. Depuis 2022, Dan Pineau est en thèse en collaboration avec l'IAS pour travailler sur le problème de la fusion de données. Il travaille notamment sur la résolution exacte du problème posé dans [12] ainsi qu'une meilleure régularisation préservant les bords.

À plus long terme, je pense que le réglage des hyperparamètres constitue une véritable difficulté pour ces problèmes de reconstructions et de fusion. En effet, dans ce cas si l'on souhaite utiliser un modèle à sous-espace pour l'objet, ou un modèle de mélange, il faut un à deux hyperparametres par carte d'abondance, en fonction de la régularisation choisie. Le réglage à la main devient très compliqué. Les développements méthodologiques présentés dans les parties 8 et 9.1 seraient de bons points de départ.

9.2.2. Synthèse de Fourier en imagerie radio

Durant la thèse de Nicolas Monnier, soutenue en juin 2023, nous avons travaillé sur un problème de synthèse de Fourier pour le Square Kilometer Array SKA. Dans ce cas le modèle direct s'écrit

$$\boldsymbol{y} = \boldsymbol{SFx} \tag{9.3}$$

où $\boldsymbol{x}$ est l'image du ciel, $\boldsymbol{F}$ la transformée de Fourier discrète et $\boldsymbol{S}$ un opérateur d'appariement entre les coefficients de Fourier $\overset{\circ}{\boldsymbol{x}}$ et les mesures, ou visibilités, $\boldsymbol{y}$. Si $\boldsymbol{S}$ est un simple opérateur de sous-échantillonnage alors on peut réduire le problème à une « simple » déconvolution sur $\tilde{\boldsymbol{x}}$

$$\tilde{\boldsymbol{x}} = \boldsymbol{F}^H \boldsymbol{S}^H \boldsymbol{y} = \boldsymbol{F}^H \boldsymbol{S}^H \boldsymbol{S} \boldsymbol{F} \boldsymbol{x} \tag{9.4}$$

puisque dans ce cas $\boldsymbol{F}^H \boldsymbol{S}^H \boldsymbol{S} \boldsymbol{F} = \boldsymbol{C}$ est une convolution circulante.

Un modèle plus juste n'est pas celui eq. (9.3) mais plutôt $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$ où $\boldsymbol{H}$ est une transformée de Fourier irrégulière ou non uniforme puisque les coordonnées des visibilités $\boldsymbol{y}$ ne sont pas sur une grille. Par ailleurs d'autres effets de géométrie ou de calibration rentrent en jeu. De nombreux travaux portent sur ces problématiques, mais une approche standard consiste à utiliser le modèle eq. (9.3) mais où $\boldsymbol{S}$ n'est pas un sous-échantillonnage, mais un opérateur d'interpolation. Dans ce cas $\boldsymbol{C}$ n'est plus une convolution et le problème est divisé en deux étapes. Une étape « exacte », la boucle majeure (*major loop*) où l'on calcule le gradient d'un moindre carré $\|\boldsymbol{y} - \boldsymbol{S}\boldsymbol{F}\boldsymbol{x}\|^2$ soit

$$\delta = 2\boldsymbol{F}^H \boldsymbol{S}^H (\boldsymbol{S}\boldsymbol{F}\boldsymbol{x} - \boldsymbol{y})$$

et entre-temps une boucle mineure (*minor loop*), qui approche le modèle et procède à une déconvolution (le plus souvent avec CLEAN).

Je propose de poursuivre les travaux de N. Monnier qui a notamment travaillé sur l'accélération du calcul du gradient en exploitant la forme développée $\boldsymbol{F}^H \boldsymbol{S}^H \boldsymbol{y} - \boldsymbol{F}^H \boldsymbol{S}^H \boldsymbol{S} \boldsymbol{F} \boldsymbol{x}$. Je pense qu'il peut être avantageux de revenir sur la séparation boucle mineure, boucle majeure, qui ne possède pas de garantie de convergence vers une solution stable, tout en maintenant un coût raisonnable dans les algorithmes itératifs, en exploitant par exemple les travaux sur les préconditionneurs.

Je souhaite poursuivre ces travaux avec Nicolas Gac, professeur à Paris-Saclay au SATIE, avec qui j'ai une collaboration de longue date ainsi que l'observatoire de Nancay qui procède a des mesures radio sous de nombreuses formes et développe de nouveaux prototypes.

9.2.3. Microscopie optique

La microscopie optique pour la biologie est un domaine très actif. Avec l'Institut Pasteur, l'ESPCI, ou l'ONERA, j'ai proposé plusieurs méthodes originales pour le traitement d'images pour le SIM, présentées chapitre 7 page 45. Ces traitements se sont essentiellement concentrés sur le traitement de mesures 2D. Je vois plusieurs perspectives possibles notamment le passage à la reconstruction 3D à partir d'un ensemble de mesures 2D tout en maintenant un niveau faible d'images pour reconstruire le cube

(quinze avec l'algorithme usuel). L'amélioration de la régularisation voire l'utilisation de l'IA font également partie des perspectives possibles.

Bibliographie

[1] Y. Huang, E. Chouzenoux et J.-C. Pesquet, « Unrolled Variational Bayesian Algorithm for Image Blind Deconvolution, » *IEEE Transactions on Image Processing*, t. 32, p. 430-445, 2023 (cf. p. 74).

[2] R. Abirizk, F. Orieux et A. Abergel, « Super-Resolution Hyperspectral Reconstruction with Majorization-Minimization Algorithm and Low-Rank Approximation, » *IEEE Transactions on Computational Imaging*, p. 260-272, 2022 (cf. p. 37, 40).

[3] M. Amrouche, H. Carfantan et J. Idier, « Efficient Sampling of Bernoulli-Gaussian-Mixtures for Sparse Signal Restoration, » *IEEE Transactions on Signal Processing*, t. 70, p. 5578-5591, 2022 (cf. p. 69, 71).

[4] N. Monnier, D. Guibert, C. Tasse et al., « Multi-Core Multi-Node Parallelization of the Radio Interferometric Imaging Pipeline DDFacet, » in *IEEE Workshop on Signal Processing Systems (SiPS)*, Rennes, France, nov. 2022 (cf. p. 25).

[5] N. Monnier, F. Orieux, N. Gac, C. Tasse, E. Raffin et D. Guibert, « Fast Sky to Sky Interpolation for Radio Interferometric Imaging, » in *2022 IEEE International Conference on Image Processing (ICIP)*, oct. 2022, p. 1571-1575 (cf. p. 25).

[6] M. Seznec, N. Gac, F. Orieux et A. S. Naik, « Real-time optical flow processing on embedded GPU : an hardware-aware algorithm to implementation strategy, » *J Real-Time Image Proc*, t. 19, n° 2, p. 317-329, avr. 2022 (cf. p. 25).

[7] M. Seznec, N. Gac, F. Orieux et A. Sashala Naik, « Computing large 2D convolutions on GPU efficiently with the im2tensor algorithm, » *J Real-Time Image Proc*, t. 19, n° 6, p. 1035-1047, 1er déc. 2022 (cf. p. 25).

[8] M. Vono, D. Paulin et A. Doucet, « Efficient MCMC Sampling with Dimension-Free Convergence Rate Using ADMM-type Splitting, » *Journal of Machine Learning Research*, t. 23, n° 25, p. 1-69, 2022 (cf. p. 67).

[9] D. Gilton, G. Ongie et R. Willett, « Deep Equilibrium Architectures for Inverse Problems in Imaging, » *IEEE Transactions on Computational Imaging*, t. 7, p. 1123-1133, 2021 (cf. p. 74).

[10] T. Oberlin et M. Verm, « Regularization via Deep Generative Models : An Analysis Point of View, » in *2021 IEEE International Conference on Image Processing (ICIP)*, sept. 2021, p. 404-408 (cf. p. 74).

[11] R. Abirizk, F. Orieux et A. Abergel, « Non-Stationary Hyperspectral Forward Model and High-Resolution, » in *Proc. of 27th IEEE Int. Conf. on Image Processing*, Abu-Dhabi, United Arab Emirates, oct. 2020, p. 5 (cf. p. 37).

[12] C. Guilloteau, T. Oberlin, O. Berné et N. Dobigeon, « Fusion of Hyperspectral and Multispectral Infrared Astronomical Images, » in *2020 IEEE 11th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, juin 2020, p. 1-5 (cf. p. 75).

[13] C. Guilloteau, T. Oberlin, O. Berné et N. Dobigeon, « Hyperspectral and Multispectral Image Fusion Under Spectrally Varying Spatial Blurs – Application to High Dimensional Infrared Astronomical Imaging, » *IEEE Transactions on Computational Imaging*, t. 6, p. 1362-1374, 2020 (cf. p. 27).

[14] C. Guilloteau, T. Oberlin, O. Berné, É. Habart et N. Dobigeon, « Simulated JWST Data Sets for Multispectral and Hyperspectral Image Fusion, » *AJ*, t. 160, n° 1, p. 28, juin 2020 (cf. p. 40).

[15] M. A. Hadj-Youcef, F. Orieux, A. Fraysse et A. Abergel, « Fast Joint Multiband Reconstruction from Wideband Images Based on Low Rank Approximation, » *IEEE Trans. Comput. Imaging*, p. 922-933, 28 mai 2020 (cf. p. 33, 35, 36).

[16] M. Vono, N. Dobigeon et P. Chainais, « High-Dimensional Gaussian Sampling : A Review and a Unifying Approach Based on a Stochastic Proximal Point Algorithm, » 4 oct. 2020 (cf. p. 63).

[17] N. Audebert, B. Le Saux et S. Lefevre, « Deep Learning for Classification of Hyperspectral Data : A Comparative Review, » *IEEE Geoscience and Remote Sensing Magazine*, t. 7, n° 2, p. 159-173, juin 2019 (cf. p. 27, 28).

[18] A. Durmus, S. Majewski et B. Miasojedow, « Analysis of Langevin Monte Carlo via Convex Optimization, » *The Journal of Machine Learning Research*, t. 20, n° 1, p. 2666-2711, 2019 (cf. p. 61).

[19] Y. Lai-Tim, L. Mugnier, F. Orieux, R. Baena-Gallé, M. Paques et S. Meimon, « Jointly Super-Resolved and Optically Sectioned Bayesian Reconstruction Method for Structured Illumination Microscopy, » *Optics Express*, t. 27, n° 23, p. 33 251-33 267, oct. 2019 (cf. p. 54).

[20] Y. Lai-Tim, L. Mugnier, F. Orieux, R. Baena-Gallé, M. Paques et S. Meimon, « Toward a Jointly Super-Resolved and Optically Sectioned Reconstruction for Structured Illumination Retinal Imaging, » in *Actes du 27e GRETSI*, Lille, France, août 2019 (cf. p. 56).

[21] A. Marchal, M.-A. Miville-Deschênes, F. Orieux et al., « ROHSA : Regularized Optimization for Hyper-Spectral Analysis, » *Astronomy & Astrophysics*, t. 626, A101, 2019 (cf. p. 25, 27).

[22] E. T. Reehorst et P. Schniter, « Regularization by Denoising : Clarifications and New Interpretations, » *IEEE Transactions on Computational Imaging*, t. 5, n° 1, p. 52-67, mars 2019 (cf. p. 73).

[23] T. Bui, B. Orberger, S. B. Blancher et al., « Building a Hyperspectral Library and Its Incorporation into Sparse Unmixing for Mineral Identification, » in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, juill. 2018, p. 4261-4264 (cf. p. 27, 28).

[24] V. Lempitsky, A. Vedaldi et D. Ulyanov, « Deep Image Prior, » in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT : IEEE, juin 2018, p. 9446-9454 (cf. p. 74).

[25] R. Baena-Gallé, L. M. Mugnier et F. Orieux, « Optical Sectioning with Structured Illumination Microscopy for Retinal Imaging : Inverse Problem Approach, » in *Actes Du 26e GRETSI*, Juan-les-pins, France, sept. 2017, p. 4 (cf. p. 54, 55).

[26] M. A. Hadj-Youcef, F. Orieux, A. Fraysse et A. Abergel, « Restoration from Multispectral Blurred Data with Non-Stationary Instrument Response, » in *Proc. of EUSIPCO*, 2017 (cf. p. 32).

[27] K. H. Jin, M. T. McCann, E. Froustey et M. Unser, « Deep Convolutional Neural Network for Inverse Problems in Imaging, » *IEEE Transactions on Image Processing*, t. 26, n° 9, p. 4509-4522, sept. 2017 (cf. p. 73).

[28] F. Orieux, *BSIM : Bayesian Structured Illumination Microscopie*, 2017 (cf. p. 54).

[29] F. Orieux et R. Chinchilla, « Semi-Unsupervised Bayesian Convex Image Restoration with Location Mixture of Gaussian, » in *25th European Signal Processing Conference (EUSIPCO)*, 2017, p. 758-762 (cf. p. 67, 68, 70).

[30] F. Orieux, V. Loriette, J.-C. Olivo-Marin, E. Sepulveda et A. Fragola, « Fast Myopic 2D-SIM Super Resolution Microscopy with Joint Modulation Pattern Estimation, » *Inverse Problems*, t. 33, n° 12, p. 125 005, 2017 (cf. p. 52-55, 63).

[31] Y. Romano, M. Elad et P. Milanfar, « The Little Engine That Could : Regularization by Denoising (RED), » *SIAM J. Imaging Sci.*, t. 10, n° 4, p. 1804-1844, jan. 2017 (cf. p. 73).

[32] C. Wang, H. Li et D. Zhao, « Preconditioning Toeplitz-plus-diagonal linear systems using the Sherman–Morrison–Woodbury formula, » *Journal of Computational and Applied Mathematics*, t. 309, p. 312-319, jan. 2017 (cf. p. 72).

[33] C. Yi, Y.-Q. Zhao, J. Yang, J. C.-W. Chan et S. G. Kong, « Joint Hyperspectral Superresolution and Unmixing With Interactive Feedback, » *IEEE Trans. Geosci. Remote Sensing*, t. 55, n° 7, p. 3823-3834, juill. 2017 (cf. p. 27).

[34] A. Boucaud, M. Bocchio, A. Abergel, F. Orieux, H. Dole et M.-A. Hadj-Youcef, « Convolution Kernels for Multi-Wavelength Imaging, » *Astronomy & Astrophysics*, t. 596, A63, 2016 (cf. p. 31).

[35] A. Boucaud, H. Dole, A. Abergel, H. Ayasso et F. Orieux, « PSF Homogenization for Multi-Band Photometry from Space on Extended Objects, » *EAS Publications Series*, t. 78–79, D. Mary, R. Flamary, C. Theys et C. Aime, éd., p. 275-285, 2016 (cf. p. 31).

[36] L. de Viguerie, M. Alfeld et P. Walter, « La lumière pour une imagerie chimique des peintures, » *Reflets phys.*, n° 47-48, p. 106-111, mars 2016 (cf. p. 27).

[37] O. Féron, F. Orieux et J.-F. Giovannelli, « Gradient Scan Gibbs Sampler : An Efficient Algorithm for High-Dimensional Gaussian Distributions, » *IEEE Journal of Selected Topics in Signal Processing*, t. 10, n° 2, p. 343-352, mars 2016 (cf. p. 65).

[38] M. Pereyra, « Proximal Markov chain Monte Carlo algorithms, » *Stat Comput*, t. 26, n° 4, p. 745-760, juill. 2016 (cf. p. 61, 67, 70).

[39] C. Gilavert, S. Moussaoui et J. Idier, « Efficient Gaussian Sampling for Solving Large-Scale Inverse Problems Using MCMC, » *IEEE Transactions on Image Processing*, t. 63, n° 1, p. 11, 2015 (cf. p. 62, 63, 72).

[40] L. Loncan, L. B. de Almeida, J. M. Bioucas-Dias et al., « Hyperspectral Pansharpening : A Review, » *IEEE Geoscience and Remote Sensing Magazine*, t. 3, n° 3, p. 27-46, sept. 2015 (cf. p. 27).

[41] R. A. Norton et C. Fox. « Efficiency and computability of MCMC with Langevin, Hamiltonian, and other matrix-splitting proposals. » (13 jan. 2015), adresse : http://arxiv.org/abs/1501.03150 (visité le 10/05/2023), preprint (cf. p. 61, 67).

[42] F. Orieux, O. Féron et J.-F. Giovannelli, « Gradient Scan Gibbs Sampler : An Efficient High-Dimensional Sampler Application in Inverse Problems, » in *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE Int. Conf. On*, 2015 (cf. p. 65).

[43] E. Chouzenoux, J.-C. Pesquet et A. Repetti, « Variable Metric Forward–Backward Algorithm for Minimizing the Sum of a Differentiable Function and a Convex Function, » *J Optim Theory Appl*, t. 162, n° 1, p. 107-132, juill. 2014 (cf. p. 72).

[44] M. D. Hoffman et A. Gelman, « The No-U-Turn Sampler : Adaptively Setting Path Lengths in Hamiltonian Monte Carlo, » *Journal of Machine Learning Research*, t. 15, n° 47, p. 1593-1623, 2014 (cf. p. 61).

[45] G. Lu et B. Fei, « Medical Hyperspectral Imaging : A Review, » *JBO*, t. 19, n° 1, p. 010 901, jan. 2014 (cf. p. 27).

[46] W. Ma, J. M. Bioucas-Dias, T. Chan et al., « A Signal Processing Perspective on Hyperspectral Unmixing : Insights from Remote Sensing, » *IEEE Signal Processing Magazine*, t. 31, n° 1, p. 67-81, jan. 2014 (cf. p. 28).

[47] F. Orieux, A. Fragola, V. Loriette et J.-C. Olivo-Marin, « Approche Bayésienne Pour La Microscopie Par Illumination Structurée, » 19 mars 2014 (cf. p. 54).

[48] C. Vacar, J.-F. Giovannelli et Y. Berthoumieu, « Bayesian Texture and Instrument Parameter Estimation From Blurred and Noisy Images Using MCMC, » *IEEE Signal Processing Letters*, t. 21, n° 6, p. 707-711, juin 2014 (cf. p. 63).

[49] F. Soulez, E. Thiébaut et L. Denis, « Restoration of Hyperspectral Astronomical Data with Spectrally Varying Blur, » *EAS Publications Series*, t. 59, D. Mary, C. Theys et C. Aime, éd., p. 403-416, 2013 (cf. p. 27).

[50] B. Thomas, M. Momany et P. Kner, « Optical sectioning structured illumination microscopy with enhanced sensitivity, » *J. Opt.*, t. 15, n° 9, p. 094 004, sept. 2013 (cf. p. 54).

[51] J. M. Bardsley et C. Fox, « An MCMC method for uncertainty quantification in nonnegativity constrained inverse problems, » *Inverse Problems in Science and Engineering*, t. 20, n° 4, p. 477-498, juin 2012 (cf. p. 71).

[52] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon et al., « Hyperspectral Unmixing Overview : Geometrical, Statistical, and Sparse Regression-Based Approaches, » *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, t. 5, n° 2, p. 354-379, avr. 2012 (cf. p. 27).

[53] F. Orieux, O. Féron et J.-F. Giovannelli, « Sampling High-Dimensional Gaussian Distributions for General Linear Inverse Problems, » *IEEE Signal Processing Letters*, t. 19, n° 5, p. 251-254, 2012 (cf. p. 62, 63, 72).

[54] F. Orieux, E. Sepulveda, V. Loriette, B. Dubertret et J.-C. Olivo-Marin, « Bayesian Estimation for Optimized Structured Illumination Microscopy, » *IEEE Transactions on Image Processing*, t. 21, n° 2, p. 601-614, fév. 2012 (cf. p. 49-51).

[55] S. Bongard, F. Soulez, E. Thiebaut et E. Pécontal, « 3-D deconvolution of hyper-spectral astronomical data, » *Monthly Notices of the Royal Astronomical Society*, t. 418, n° 1, p. 258-270, 21 nov. 2011 (cf. p. 27, 28).

[56] E. Chouzenoux, J. Idier et S. Moussaoui, « A Majorize–Minimize Strategy for Subspace Optimization Applied to Image Restoration, » *IEEE Trans. on Image Process.*, t. 20, n° 6, p. 1517-1528, juin 2011 (cf. p. 34, 38, 39).

[57] L. Denis, E. Thiébaut et F. Soulez, « Fast Model of Space-Variant Blurring and Its Application to Deconvolution in Astronomy, » in *2011 18th IEEE International Conference on Image Processing*, IEEE, 2011, p. 2817-2820 (cf. p. 31).

[58] M. Girolami et B. Calderhead, « Riemann Manifold Langevin and Hamiltonian Monte Carlo Methods, » *Journal of the Royal Statistical Society Series B : Statistical Methodology*, t. 73, n° 2, p. 123-214, 1er mars 2011 (cf. p. 61, 67).

[59] R. M. Neal, *MCMC Using Hamiltonian Dynamics*. 10 mai 2011 (cf. p. 61, 67).

[60] C. Soussen, J. Idier, D. Brie et J. Duan, « From Bernoulli–Gaussian Deconvolution to Sparse Signal Restoration, » *IEEE Transactions on Signal Processing*, t. 59, n° 10, p. 4572-4584, 2011 (cf. p. 69).

[61] H. Ayasso et A. Mohammad-Djafari, « Joint NDT Image Restoration and Segmentation Using Gauss–Markov–Potts Prior Models and Variational Bayesian Computation, » *IEEE Transactions on Image Processing*, t. 19, n° 9, p. 2265-2277, 2010 (cf. p. 69, 70).

[62] K. Gregor et Y. LeCun, « Learning Fast Approximations of Sparse Coding, » in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, sér. ICML'10, Madison, WI, USA : Omnipress, 21 juin 2010, p. 399-406 (cf. p. 73).

[63] P. Guillard, T. Rodet, S. Ronayette et al., « Optical Performance of the JWST/MIRI Flight Model : Characterization of the Point Spread Function at High Resolution, » in *Space Telescopes and Instrumentation 2010 : Optical, Infrared, and Millimeter Wave*, International Society for Optics and Photonics, t. 7731, 2010, 77310J (cf. p. 31).

[64] M. K. Ng et J. Pan, « Approximate Inverse Circulant-plus-Diagonal Preconditioners for Toeplitz-plus-Diagonal Matrices, » *SIAM J. Sci. Comput.*, t. 32, n° 3, p. 1442-1464, jan. 2010 (cf. p. 72).

[65] F. Orieux, T. Rodet et J.-F. Giovannelli, « Instrument Parameter Estimation in Bayesian Convex Deconvolution, » in *Image Processing (ICIP), 2010 17th IEEE Int. Conf. On*, IEEE, sept. 2010, p. 1161-1164 (cf. p. 63).

[66] F. Orieux, « Inversion bayésienne myope et non-supervisée pour l'imagerie sur-résolue. Application à l'instrument SPIRE de l'observatoire spatial Herschel., » thèse de doct., Université Paris-Sud 11, nov. 2009 (cf. p. 31, 63).

[67] S. A. Shroff, J. R. Fienup et D. R. Williams, « Phase-Shift Estimation in Sinusoidally Illuminated Images for Lateral Superresolution, » *Journal of the Optical Society of America A*, t. 26, n° 2, p. 413-424, 2009 (cf. p. 52, 53).

[68] C. Fox, « A Conjugate Direction Sampler for Normal Distributions, with a Few Computed Examples, » University of Otago, 2008-1, sept. 2008 (cf. p. 65).

[69] J.-F. Giovannelli, « Unsupervised Bayesian Convex Deconvolution Based on a Field With an Explicit Partition Function, » *IEEE Trans. on Image Process.*, t. 17, n° 1, p. 16-26, jan. 2008 (cf. p. 67, 70).

[70] C. Labat et J. Idier, « Convergence of Conjugate Gradient Methods with a Closed-Form Stepsize Formula, » *J Optim Theory Appl*, p. 18, 2008 (cf. p. 34, 38).

[71] T. Park et G. Casella, « The Bayesian Lasso, » *Journal of the American Statistical Association*, t. 103, n° 482, p. 681-686, 2008 (cf. p. 66).

[72] O. Berné, C. Joblin, Y. Deville et al., « Analysis of the emission of very small dust particles from Spitzer spectro-imagery data using blind signal separation methods, » *A&A*, t. 469, n° 2, p. 575-586, 2 1er juill. 2007 (cf. p. 33, 40).

[73] C. Labat et J. Idier, « Convergence of Truncated Half-Quadratic and Newton Algorithms with Application to Image Restoration, » 2007, p. 15 (cf. p. 34).

[74] M. Nikolova et R. H. Chan, « The Equivalence of Half-Quadratic Minimization and the Gradient Linearization Iteration, » *IEEE Trans. on Image Process.*, t. 16, n° 6, p. 1623-1627, juin 2007 (cf. p. 34).

[75] M. Allain, J. Idier et Y. Goussard, « On Global and Local Convergence of Half-Quadratic Algorithms, » *IEEE Trans. on Image Process.*, t. 15, n° 5, p. 1130-1142, mai 2006 (cf. p. 34, 39, 71).

[76] T. Chan, S. Esedoglu, F. Park et A. Yip, « Total Variation Image Restoration : Overview and Recent Developments, » *Handbook of mathematical models in computer vision*, p. 17-31, 2006 (cf. p. 70, 71).

[77] R. A. Levine et G. Casella, « Optimizing random scan Gibbs samplers, » *Journal of Multivariate Analysis*, t. 97, n° 10, p. 2071-2100, 1er nov. 2006 (cf. p. 65).

[78] J. Nocedal et S. J. Wright, *Numerical Optimization* (Springer Series in Operations Research), 2nd ed. New York : Springer, 2006, 664 p. (cf. p. 39).

[79] L. H. Schaefer, D. Schuster et J. Schaffer, « Structured Illumination Microscopy : Artefact Analysis and Reduction Utilizing a Parameter Optimization Approach, » *J Microsc*, t. 216, n° 2, p. 165-174, nov. 2004 (cf. p. 51, 52).

[80] S. C. Park, M. K. Park et M. G. Kang, « Super-Resolution Image Reconstruction : A Technical Overview, » *IEEE Signal Processing Magazine*, t. 20, n° 3, p. 21-36, mai 2003 (cf. p. 40).

[81] J. Idier, « Convex Half-Quadratic Criteria and Interacting Auxiliary Variables for Image Restoration, » *IEEE Trans. on Image Process.*, t. 10, n° 7, p. 1001-1009, juill. 2001 (cf. p. 34).

[82] M. G. L. Gustafsson, « Surpassing the Lateral Resolution Limit by a Factor of Two Using Structured Illumination Microscopy, » *Journal of Microscopy*, p. 6, 2000 (cf. p. 49).

[83] P. Charbonnier, L. Blanc-Feraud, G. Aubert et M. Barlaud, « Deterministic Edge-Preserving Regularization in Computed Imaging, » *IEEE Transactions on Image Processing*, t. 6, n° 2, p. 298-311, fév. 1997 (cf. p. 34).

[84] E. T. Jaynes, *Probability Theory - The Logic of Science*. 1996 (cf. p. 57).

[85] D. Geman et C. Yang, « Nonlinear image recovery with half-quadratic regularization, » *IEEE Trans. on Image Process.*, t. 4, n° 7, p. 932-946, juill. 1995 (cf. p. 34).

[86] J. Besag, « Comments on "Representations of Knowledge in Complex Systems" by U. Grenander and MI Miller, » *J. Roy. Statist. Soc. Ser. B*, t. 56, n° 591-592, p. 4, 1994 (cf. p. 61, 67).

[87] J. R. Shewchuk, « An Introduction to the Conjugate Gradient Method Without the Agonizing Pain, » School of Computer Science Carnegie Mellon University, Pittsburgh, 4 août 1994 (cf. p. 62).

[88] D. Geman et G. Reynolds, « Constrained restoration and the recovery of discontinuities, » *IEEE Trans. Pattern Anal. Machine Intell.*, t. 14, n° 3, p. 367-383, mars 1992 (cf. p. 34, 71).

[89] G. Demoment, « Image reconstruction and restoration : overview of common estimation structures and problems, » *IEEE Trans. Acoust., Speech, Signal Processing*, t. 37, n° 12, p. 2024-2036, déc. 1989 (cf. p. 32).

[90] N. Galatsanos et R. Chin, « Digital Restoration of Multichannel Images, » *IEEE Transactions on Acoustics, Speech, and Signal Processing*, t. 37, n° 3, p. 415-421, mars 1989 (cf. p. 31).

[91] S. Geman et D. Geman, « Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, » *IEEE Trans. Pattern Anal. Mach. Intell.*, t. PAMI-6, n° 6, p. 721-741, nov. 1984 (cf. p. 60).

[92] H. Voß et U. Eckhardt, « Linear convergence of generalized Weiszfeld's method, » *Computing*, t. 25, n° 3, p. 243-251, 1er sept. 1980 (cf. p. 39).

Troisième partie

Sélection d'articles

Retour de chasse. Les deux chiens traînent les phoques morts sur la glace - Diamond Jenness ©courtesy galerie Lumière des roses.

Super-Resolution Hyperspectral Reconstruction with Majorization-Minimization Algorithm and Low-Rank Approximation

Ralph Abi-Rizk, François Orieux and Alain Abergel

***Abstract*—Hyperspectral imaging (HSI) has become an invaluable imaging tool for many applications in astrophysics or Earth observation. Unfortunately, direct observation of hyperspectral images is impossible since the actual measurements are 2D and suffer from strong spatial and spectral degradations, especially in the infrared. We present in this work an original method for high-resolution hyperspectral image reconstruction from heterogeneous 2D measurements degraded by integral field spectroscopy (IFS) instrument. A fundamental part of this work is developing a forward model that accounts for the limitations of the IFS instrument, such as wavelength-dependent spatial and spectral blur, subsampling, and inhomogeneous sampling steps. The reconstruction method inverts the forward model using a deterministic regularization framework for edge-preserving. It fuses information from different observations and spectral bands for resolution enhancements. We rely on the Majorize-Minimize memory gradient (3MG) optimization algorithm to solve the inverse problem while considering a low-rank approximation for the unknown to handle the high-dimensionality of the problem.**

***Index Terms*—Inverse Problems – Super Resolution – Hyperspectral Imaging – Deconvolution – Spectral unmixing**

I. Introduction

HYPERSPECTRAL imaging (HSI) simultaneously collects high-resolution spectra at different spatial locations. It is wildly used for remote sensing applications in numerous domains such as in astrophysics [1], fluorescence microscopy [2], military [3], medical diagnosis [4], and others. HSI products are 3-dimensional (3D) images (i, j, l) where (i, j) are the two spatial dimensions and l the spectral dimension. Unfortunately, direct observations of hyperspectral (HS) images are not straightforward because 3D detectors do not exist. Instead, HS instruments, primarily relying on dispersive spectrometers, are designed to acquire measurements projected onto 2D detectors. In particular, HS instruments based on Integral Field Spectroscopy (IFS) [5] simultaneously observe the field of view (FOV) of the 3D input image through several thin slits in parallel. The dispersed wavelength from each slit is projected onto 2D detectors, spanning a spatial dimension along one axis and a spectral dimension along the other axis. Consequently, a reconstruction stage is required to estimate the 3D input image from the collected 2D measurements. Although having a high spectral resolution, the 2D measurements suffer from spatial and spectral limitations during the acquisition process, such as blurring, sampling, and noise. First, because of the diffraction [6], the optical response known as the Point Spread function (PSF) introduces a wavelength-dependent spatial blurring. Second, the response of the dispersing system introduces a spectral blurring, which is also wavelength-dependent [7]. Finally, the spatial sampling on the detector is often insufficient at all wavelengths. To enhance the spatial resolution lost at the detector, a "dithering" method is considered [8], [9], consisting of observing the same scene multiple times by slightly shifting the measuring instrument. The resulted multi-frame measurements lead to a Super Resolution (SR) problem [10].

Several multi-frame SR algorithms have been addressed to reconstruct a discrete 3D input image from a set of measurements degraded by the HS instrument. The state-of-art approach for multi-frame SR 3D reconstruction is based on the shift and addition (S&A) method [8], [11]. It merges the overall sampled and aliased measurements to provide a single reconstructed 3D image with an enhanced spatial resolution. Even though the S&A provides fast and non-iterative algorithms, it does not consider spatial and spectral blurring. Hence, it can be followed by a deblurring step, such as a Total Variation (TV) regularization [12], [13]. This technique is efficient for monochromatic image reconstruction. However, for HS image reconstruction, the deblurring step treats the spatial and spectral dimensions separately without considering the correlations between spectral bands.

Multi-frame SR reconstruction algorithms for HS images have also been treated as an inverse problem allowing a joint process of the spatial and spectral information from all the measurements. Such approaches rely on an explicit forward model that considers the limitations of the HS instrument and some additional priors about the 3D input image [14], [15]. Other tensor-based methods for HS reconstruction have been proposed in [16] for SR, deblurring and denoising. Most of these approaches assume a low-rank structure, where the input is represented with a small number of spectral components.

[14] developed a forward model that simulates optically blurred, sampled, and aliased HS multi-frame measurements. They proposed a SR reconstruction algorithm based on the projection onto convex sets (POCS) method [17] that relies on the forward model to restore the observed HS image,

R. Abi-Rizk and F. Orieux are at the Laboratoire des Signaux et Systèmes, Université Paris-Saclay, CNRS, CentraleSupélec, 3 rue Joliot-Curie, 91190 Gif-Sur-Yvette, France, e-mail: *name@l2s.centralesupelec.fr*

A. Abergel is at the Institut d'Astrophysique Spatiale, Université Paris-Saclay, CNRS, 91405 Orsay, France

This project is co-funded by CNES. This work was performed using HPC resources from the "Mésocentre" computing center of CentraleSupélec and École Normale Supérieure Paris-Saclay supported by CNRS and Région Île-de-France (http://mesocentre.centralesupelec.fr/)

approximated in the low-dimensional subspace. [15] handled the multi-frame SR reconstruction in the principal component analysis (PCA) domain. They used the first few principal components [18] to estimate motion and to reconstruct the 3D input image via the maximum a posteriori (MAP) [19]. However, the spatial blurring considered in these works is stationary, and the spatial and spectral fields of view are homogeneous.

Another solution for the spatial resolution enhancement is to perform a fusion of spatially sampled HS measurements with an auxiliary image of the same scene with high spatial resolution, if available, such as panchromatic (PAN) [20], or multispectral (MS) [1]. In particular, the HS-MS fusion has been excessively addressed in the inverse problem framework [1], [21], [22]. It relies on minimizing an objective function associated with two data fitting terms for HS and MS, respectively, and some priors about the 3D input image. [1] proposed an HS-MS fusion method while accounting for wavelength-dependent spatial blur. They provide fast algorithms in the Fourier domain while assuming a low-rank structure of the astronomical input image. However, the works proposed in [1], [14], [15] consider 3D measurements with uniform spatial and spectral sampling steps and without accounting for a wavelength-dependent spectral PSF.

We present in this work a complex forward model based on an IFS instrument dedicated to astronomical observations in the infrared spectral range. The input is a 3D spatio-spectral image with a high resolution, approximated in a low-rank subspace, and simulates a set of multi-frame measurements projected onto different 2D detectors of different characteristics. The proposed model allows different input observations from different IFS instruments and pointing. Moreover, it considers wavelength-dependent spatial and spectral blurring, as well as heterogeneous spatial and spectral sampling on the detectors. We then propose a reconstruction method that uses the forward model and relies on the regularized least square approaches with convex edge-preserving regularization [23]. To solve the inversion problem, we choose the iterative Majorize-Minimize Memory Gradient (3MG) optimization algorithm [24] tested on two synthetic 3D input images with different spatial and spectral distributions. The results show significant improvement of the spatial and spectral resolutions compared to the shift-and-add (S&A) algorithm [8], [11] followed by a spatial Total-Variation (TV) regularization for each wavelength [12], and the classic l_2 regularization [25].

The paper is organized as follows. The section II discusses the proposed methodology, first for the instrument model developed for the IFS instrument (Section II-A), and second for the forward model based on the linear mixing model (LMM) [26] (Section II-B). The SR multi-frame reconstruction algorithm is presented in Section III. In Section IV, we present the reconstruction results with the proposed algorithm and provide a comparison with other reconstruction algorithms. Finally, we provide a conclusion in Section V.

II. Forward Model for Heterogenous Hyperspectral Data Fusion

A. Observation Model

This section presents a new observation model of IFS instruments used for spectral data fusion. It considers a series of components that modify and degrade the observed 3D input image (HS image), resulting in a set of blurred, truncated, and aliased 2D multi-frame measurements.

The original discretized input image is denoted $\boldsymbol{x}[i,j,l]$, with two spatial dimensions $(i,j) \in [1,\dots,I] \times [1,\dots,J]$ denoting the pixel index, and one spectral dimension $l \in [1,\dots,L]$. It is supposed uniformly sampled with spatial steps (T_i, T_j), and spectral step T_l.

1) Spatial Filtering: Because of the diffraction phenomenon [6], the observed 3D input is spatially blurred by the response of the optical system, also known as the point spread function (PSF). The PSF, denoted $\boldsymbol{h}$, is spectrally non-invariant, with an increasing blur as the wavelength increases. We suppose that the monochromatic PSF is known from simulations [27], calibration, or previous data processing steps. The PSF is also assumed to be spatially stationary at all wavelengths. Thus, the spatial filtering is carried out by a 2D spatial convolution between the 3D input image and a discrete wavelength-dependent PSF, sampled with the same sampling step of the input, (T_i, T_j), writing

$$\boldsymbol{x}_{\text{opt}} = \boldsymbol{x}[i,j,l] \underset{i,j}{*} \boldsymbol{h}[i,j,l]. \tag{1}$$

We will see in the next section that the model includes spatially truncated observation since the field of view of the IFS instrument is smaller than that of the 3D input image. Consequently, the spatial convolution is calculated using the discrete spatial Fourier transform for fast computation [28] without introducing periodic patterns to the blurred image.

2) Spatio-Spectral Field of View: We consider here a spectral data fusion problem where $\boldsymbol{x}_{\text{opt}}$ is observed with various spatial and spectral fields of view grouped in distinct spectral channels $c \in [1,\dots,C]$. Each channel can possess a different FOV, spectral range, and sampling step size. In addition, the IFS instrument observes the FOV of each channel simultaneously through several slits in parallel. The number and size of the slits depend on the channel c. The spectral selection into channels and the spatial selection into slits result from a multiplication between $\boldsymbol{x}_{\text{opt}}$ and the channel windows $\boldsymbol{w}_c$ that writes

$$\boldsymbol{w}_c[i,j,l] \neq 0, \quad \forall (i,j,l) \in (\mathcal{I}_c, \mathcal{J}_c, \mathcal{L}_c), \tag{2}$$

with $(\mathcal{I}_c, \mathcal{J}_c, \mathcal{L}_c)$ a rectangular subset of $[1,\dots,I] \times [1,\dots,J] \times [1,\dots,L]$, and 0 otherwise. This hypothesis implies that all window sizes $(\Delta_{i,c}, \Delta_{j,c})$ are multiples of the step sizes of the 3D input image $\boldsymbol{x}$: $(\Delta_{i,c}, \Delta_{j,c}) = (n_c T_i, m_c T_j)$, with $(n_c, m_c) \in \mathbb{N}^2$. This is an approximation, minored if the step sizes are small.

In addition, the instrument enables different pointing (spatial shifts or dithering), indexed by p, that can be shared between channels. We also consider that the pointing positions $(\Delta_{i,p}, \Delta_{j,p})$ are multiples of the sampling steps of $\boldsymbol{x}$: $(\Delta_{i,p}, \Delta_{j,p}) = (i_p T_i, j_p T_j)$, with $(i_p, j_p) \in \mathbb{N}^2$.

Finally, the spatio-spectral field of view for a particular pointing writes

$$\boldsymbol{x}_{c,p}[i,j,l] = \boldsymbol{x}_{\text{opt}}[i,j,l] \times \boldsymbol{w}_c[i-i_p, j-j_p, l]. \quad (3)$$

3) Spectral Blurring: The spatio-spectral cubes $\boldsymbol{x}_{c,p}[i,j,l]$ are projected onto 2D detectors, through diffraction gratings for instance. For a monochromatic punctual source at the wavelength $\lambda = lT_l$ from the channel c and the pointing p, we consider the usual spectral response for grating spectrometers writing [29]

$$h_{c,j,p}(l',l) \propto \text{sinc}^2\left(\pi W\left(\frac{l'T_{l'} - q_{j,p}}{lT_l} - 1\right)\right). \quad (4)$$

Here $\lambda' = l'T_{l'}$ is the spatial position on the detector (in wavelength units) with a sampling step $T_{l'}$, and $q_{j,p} \in [-\Delta_{j,c}/2, \Delta_{j,c}/2]$ is the *relative* spatial position of the input source determined by the spatial position j and the pointing p. This spectral response is independent of the other spatial position indexed by i. Moreover, as pointed by Eq. (5) below, the proposed model is not limited to the spectral PSF detailed in Eq. (4). Other models or responses inferred from calibrations can also be considered.

Several particularities of the spectral response, or more generally of the dispersion system, are inferred from Eq. (4). First, as illustrated in Fig. 1, the spatial position on the detector $\lambda' = l'T_{l'}$ of the spectral response depends on the wavelength input $\lambda = lT_l$. Second, the spectral response is *not stationary* and becomes broader with the increase of the wavelength. Finally, the relative spatial position $q_{j,p}$ alters the spatial position of the spectral response on the detector by shifting its maximum position.

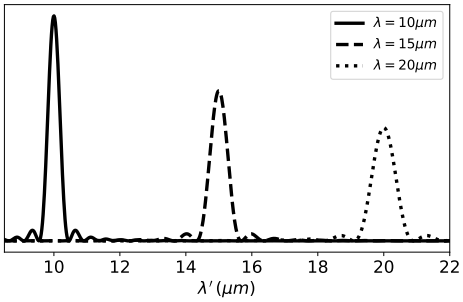


Fig. 1. Grating outputs (from Eq. (4)) for three monochromatic point sources at different wavelengths ($\lambda = lT_l$). The position on the detector ($\lambda' = l'T_{l'}$) depends on λ, while the width increases linearly with the wavelength.

Eq. (4) also depends on the parameter W. It defines the width of the spectral response, which controls the spectral resolution. Assuming that the instrument is calibrated with a known spectral resolution $R = \lambda/\Delta\lambda$, with $\Delta\lambda$ the full width at half maximum (FWHM), basic calculus leads to $W \approx 2.8R/\pi$. Finally, the grating is a non-stationary linear system.

After discretization of the response, all individual sources $\boldsymbol{x}_{c,p}[i,j,l]$ contribute on the detector resulting in an output $\boldsymbol{g}_{c,p}$ of the dispersing system that writes

$$\boldsymbol{g}_{c,p}[i,j,l'] = \sum_{l \in \mathcal{L}_c} \boldsymbol{x}_{c,p}[i,j,l] \boldsymbol{h}_{c,j,p}[l',l], \quad (5)$$

with $l' \in \mathcal{L}'_c$ the wavelength index sampled on the detector and $\boldsymbol{h}_{c,j,p}[l',l]$ obtained from Eq. 4. The model in Eq. (5) is not a convolution since the spectral response is not stationary. Consequently, homogeneous spectral sampling is not required and can vary across the channels c. The spectral resolution of the output is fixed by the spectral response and depends on the sampled wavelength l'.

4) Detector Integration: The 3D spatially and spectrally blurred cube $\boldsymbol{g}_{c,p}[i,j,l']$ depends on the spatial index j whereas detectors are 2D. Since the system is linear, all the contributions of sources within the spatial window $\mathcal{J}_{c,p}$, determined by the channel c and pointing p, are summed on the detector. Consequently, without dithering and super-resolution, the width of the window $\boldsymbol{w}_c$ along the j-axis of the measurements determines the spatial resolution along this axis.

Moreover, the spatial sampling step T_i along the i-axis of the object $\boldsymbol{g}$ is different and smaller than the sampling step of the detector. Therefore, for all pointing p, we consider that the spatial sampling step $T_{i'}$ of the detectors for each channel c is a multiple $d_c \in \mathbb{N}$ of T_i with $T_{i'} = d_c T_i$, as in the classical image Super-Resolution [10]. Consequently, the 2D measurement $\boldsymbol{y}_{c,p}$ for channel c and pointing p writes

$$\boldsymbol{y}_{c,p}[i',l'] = \sum_{i=i'd_c}^{(i'+1)d_c-1} \sum_{j \in \mathcal{J}_{c,p}} \boldsymbol{g}_{c,p}[i,j,l'] \quad (6)$$

where i' and l' are the spatial and spectral indexes on the 2D detector, respectively. In practice, the summation along the i-axis is computed as a convolution between $\boldsymbol{g}$ and a square impulse response of size d_c, followed by subsampling every d_c elements.

To conclude, we have developed a non-stationary but linear forward model involving relatively complex components that writes

$$\boldsymbol{y}_{c,p}[i',l'] = \sum_{i,j} \sum_{l \in \mathcal{L}_c} \left(\boldsymbol{x}[i,j,l] \underset{i,j}{*} \boldsymbol{h}[i,j,l]\right) \boldsymbol{w}_c[i-i_p, j-j_p, l] \times \boldsymbol{h}_{c,j,p}[l',l], \quad (7)$$

and accounts for several effects:

- 2D spatial convolutions with spectrally varying PSF described in Eq. (1) to model the optics,
- spatio-spectral windowing defined in Eq. (3) that models different spatial pointing, and different spatio-spectral selections,
- spectral blurring with a non-stationary response described in Eq. (5),
- and spatial and spectral sampling with specific steps for each detector which are larger than the sampling steps of the 3D input, described in Eq. (6).

The next section presents the combination of this observation model with a subspace representation of the 3D input image.

B. LMM forward model

1) Linear Mixing Model: Without additional information, the reconstruction of $\boldsymbol{x}$ corresponds to the estimation of

each voxel $\boldsymbol{x}[i,j,l]$ from the set of measurements $\{\boldsymbol{y}_{c,p}\}$ for all channels c and all pointing p. However, the spectral information contained in the 3D input images can be complex, with spectral rays, non-monochromatic spectral features, and continuum. Moreover, this spectral information is generally highly correlated between spatial pixels over the whole measured spectral range. Therefore, dimension reduction methods such as Principal Component Analysis (PCA) [18] or Non-negative Matrix Factorization (NMF) [30] can be very efficient on 3D images with a high spectral resolution, such as HS images.

Here we propose to write the unknown 3D image $\boldsymbol{x}$ using a Linear Mixing Model writing

$$\boldsymbol{x}[i,j,l] = \sum_{m=1}^{M} \boldsymbol{a}_m[i,j] \times \boldsymbol{s}_m[l] \tag{8}$$

where the spectral distribution at each spatial position (i,j) is a linear combination of M spectral components $\boldsymbol{s}_m$, known *a priori* or learned from the measurements, and unknown proportions $\boldsymbol{a}_m$. For our purposes, this is a subspace approximation as the number of spectral components M is much lower than the number of spectral bands L, as observed with dimension reduction methods [1]. For earth observation with segmentation problems, the spectral components $\boldsymbol{s}_m$ are pure spectra called end-members, and the $\boldsymbol{a}_m$ coefficients are called abundances [26]. In that case, additional constraints are usually imposed on $\boldsymbol{a}_m$ such as the non-negativity and sum-to-one constraints. This is not our case since we are only interested in the subspace approximation to reconstruct the original unknown 3D image $\boldsymbol{x}$, and not in the physical meaning of the spectral components $\boldsymbol{s}_m$.

The linear mixing model preserves the spatial and spectral distributions of the 3D input image and has many advantages:

- The subspace approximation significantly reduces the number of unknowns that we want to estimate.
- As a consequence it is expected to increase the Signal-to-Noise (SNR) ratio on the reconstructed object.
- The reconstruction problem is limited to the estimation of the mixing coefficients $\boldsymbol{a}_m$, which requires only a spatial regularization to enhance the spatial resolution of the estimated 3D image.
- As the reconstruction of $\boldsymbol{x}$ is a linear combination of the estimated $\boldsymbol{a}_m$ and the known $\boldsymbol{s}_m$, the final spectral resolution of the reconstructed object is the spectral resolution of the spectral components $\mathbf{s}_m$.
- The spectral information is fully given regardless of the channel characteristics, while the spatio-spectral reconstruction of $\boldsymbol{x}$ directly from the measurements, without considering the mixing model, cannot exploit the complete spectral information at all spatial positions since the observation model (see Sec. II-A2) considers different FOV depending on the channel.
- The estimated unknowns $\boldsymbol{a}_m$ depends solely on the spatial information. Hence, all spectral related terms (that depends on l and l') can be precomputed.

[1]If $M \gg L$, the reconstruction problem with an overcomplete dictionary is considered, leading to variable selection methods, often done with sparsity, outside the scope of this work.

2) Final Forward Model: By combining the linear mixing model in Eq. (8) with the observation model in Eq. (7) we obtain

$$\boldsymbol{y}_{c,p}[i',l'] = \sum_{i,j} \sum_{l\in\mathcal{L}_c} \left(\left[\sum_{m=1}^{M} \boldsymbol{a}_m[i,j]\boldsymbol{s}_m[l] \right] \underset{i,j}{*} \boldsymbol{h}[i,j,l] \right) \boldsymbol{w}_c[i-i_p, j-j_p, l] \times \boldsymbol{h}_{c,j,p}[l',l]. \tag{9}$$

The above equation can be directly used to compute the forward output. However, since we want to estimate the mixing coefficients $\boldsymbol{a}_m$ and not the full 3D input image $\boldsymbol{x}$, the known spectral components $\boldsymbol{s}_m$ can be included in the observation model. Consequently, a new spectral dependent forward model is formulated that directly links the mixing coefficients to the measurements. For that purpose, all spectral operations related to l and l' can be combined and precomputed.

First a spatial PSF cube *that depends on* m is computed for each spectral component with $\boldsymbol{h}_m[i,j,l] = \boldsymbol{s}_m[l]\boldsymbol{h}[i,j,l]$. The model then writes

$$\boldsymbol{y}_{c,p}[i',l'] = \sum_{i,j} \sum_{l\in\mathcal{L}_c} \left(\sum_{m=1}^{M} \boldsymbol{a}_m[i,j] \underset{i,j}{*} \boldsymbol{h}_m[i,j,l] \right) \boldsymbol{w}_c[i-i_p, j-j_p, l] \times \boldsymbol{h}_{c,j,p}[l',l]. \tag{10}$$

Second, the spectral blurring of $\boldsymbol{s}_m$ introduced by the spectral response $\boldsymbol{h}_{c,j,p}$ in Eq. (4) can be precomputed with

$$\boldsymbol{h}_{m,c,j,p}[i,j,l'] = \sum_{l\in\mathcal{L}_c} \boldsymbol{s}_m[l]\boldsymbol{h}[i,j,l]\boldsymbol{w}_c[l]\boldsymbol{h}_{c,j,p}[l',l] \tag{11}$$

where $\boldsymbol{h}_{m,c,j,p}[i,j,l']$ is a spatio-spectral PSF cube that depends on the spectral template number m, the spectral window $\mathcal{L}_c$, and relative position j within the spatial window $\mathcal{J}_{c,p}$ (as described by Eq. (4)). Finally, the forward model writes

$$\boldsymbol{y}_{c,p}[i',l'] = \sum_{i} \sum_{j\in\mathcal{J}_{c,p}} \boldsymbol{w}_c[i-i_p, j-j_p] \left(\sum_{m=1}^{M} \boldsymbol{a}_m[i,j] \underset{i,j}{*} \boldsymbol{h}_{m,c,j,p}[i,j,l'] \right). \tag{12}$$

Compared to Eq. (9), the final forward model is relatively simplified with the following steps:

- First, the 2D mixing coefficients $\boldsymbol{a}_m$ are convoluted by a collection of 2D+λ PSF $\boldsymbol{h}_{m,c,j,p}$ that depends on the spectral component number m, the channel c, and the relative spatial position within the channel (j,p).
- After summation on m, the cube is spatially windowed for each pointing p.
- Then, the high-resolution window is spatially detector integrated (subsampled), resulting in 2D measurements $\boldsymbol{y}_{c,p}$ with a low spatial resolution.

3) Matrix Formulation: The model in Eq. (12) is linear and represents the overall multi-frame 2D measurements $\boldsymbol{y}_{c,p}$ in terms of the unknown mixing coefficients $\boldsymbol{a}$

$$\boldsymbol{y}_{c,p} = \boldsymbol{H}_{c,p}\boldsymbol{a} = \boldsymbol{\Sigma}_{i,j}\boldsymbol{W}_c\boldsymbol{\Sigma}_m\boldsymbol{C}_{m,c,j,p}\boldsymbol{a} \tag{13}$$

where $\boldsymbol{C}$ is a convolution operator, $\boldsymbol{\Sigma}_m$ a summation on the spectral template number m, $\boldsymbol{W}_c$ a windowing or raw selection and $\boldsymbol{\Sigma}_{i,j}$ a sum on i, j to model detector integration.

Consequently, the adjoint operator writes

$$\boldsymbol{e}_{c,p} = \boldsymbol{H}_{c,p}^T \boldsymbol{y}_{c,p} = \boldsymbol{C}_{m,c,j,p}^T \boldsymbol{\Sigma}_m^T \boldsymbol{W}_c^T \boldsymbol{\Sigma}_{i,j}^T \boldsymbol{y}_{c,p} \tag{14}$$

where $\boldsymbol{\Sigma}^T$ is a duplication operator, $\boldsymbol{W}_c^T$ a zero filling operator and $\boldsymbol{C}_{m,c,j,p}^T$ a convolution with flipped response.

The overall measurements writes

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{a} = [\boldsymbol{H}_{0,0}^T, \ldots, \boldsymbol{H}_{C,P}^T]^T \boldsymbol{a} \tag{15}$$

and the full-adjoint operator writes

$$\boldsymbol{e} = \sum_{c,p} \boldsymbol{H}_{c,p}^T \boldsymbol{y}_{c,p} = \boldsymbol{H}^T \boldsymbol{y} \tag{16}$$

which is the sum of all retro-propagated measurements.

The next section describes the proposed reconstruction formalized as an inverse problem approach with efficient Quadratic Majorize-Minimize algorithm [24], [31].

III. Inverse problem

Our new forward model combines multiple observations with a full complex linear model $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{a}$. The operator $\boldsymbol{H}$ takes into account (1) spectral-dependent spatial blurring, (2) multiple channel observations with different fields of view, (3) spectral blurring, and (4) heterogeneous spatial and spectral samplings. Therefore, the mixing coefficient $\boldsymbol{a}$ reconstruction is an ill-posed inverse problem that includes data fusion, deconvolution, and multi-frame super-resolution steps.

A. Proposed reconstruction

We propose a new multi-frame SR algorithm that relies on the complete forward model in Eq. (12), with a reconstruction solution defined as the minimizer of an objective function combining a data fidelity term and a regularization term $R(a)$ expressed as

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a}} \left(\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|^2_{\boldsymbol{\Sigma}^{-1}} + R(\boldsymbol{a}) \right) \tag{17}$$

where $\boldsymbol{\Sigma}$ is the noise covariance matrix. We suppose here that the noise is Gaussian but not necessarily identically distributed or independent. However, the application of the inverse covariance matrix must be feasible. This model can approximate in some way Poisson noise with high photon count with independent non-identically distributed Gaussian noise. In that case, $\boldsymbol{\Sigma}^{-1}$ is diagonal with values that should be derived from measurements. Many regularization methods and algorithms have been proposed in the literature. For instance, l_2 regularization [25] is the seminal approach with fast algorithms but fails to preserve the high gradient values of the solution. The Total Variation (TV) regularization [12] or dictionary-based approaches with sparsity constraints [32] has been broadly used but can introduce cartoon-like effects and provide relatively slow algorithms. More recently, prior learning from data with machine learning approaches [33] has been widely explored but requires many measurements to be competitive.

In this work, the major degradation effects of the forward model are the spectral-dependent spatial and spectral blurring. However, by choosing a linear mixing model with known spectral components $\boldsymbol{s}$, our reconstruction algorithm does not require spectral regularization. On the other hand, the spatial blurring is significant, especially at long wavelengths. Therefore, we propose a convex spatial regularization to preserve strong spatial gradient in the image. The objective function, denoted $J(\boldsymbol{a})$, writes

$$J(\boldsymbol{a}) = \|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|^2_{\boldsymbol{\Sigma}^{-1}} + \mu \sum_{c \in \mathcal{C}} \phi(\boldsymbol{v}_c^T \boldsymbol{a}) \tag{18}$$

where $\boldsymbol{v}_c$ are first-order differences in the two spatial dimensions, $c \in \mathcal{C}$ a clique index that is a linear combination of pixels, μ is the spatial regularization parameter, and ϕ is a strictly differentiable convex loss like the Huber loss, allowing the use of fast optimization algorithms like [24], often faster than those based on non-differentiable loss [34]. Moreover, these algorithms are not restricted to convex loss.

B. Quadratic Majorization Minimization

The objective function $J(\boldsymbol{a})$ is an instance of the more general criterion [23], [35] with:

$$J(\boldsymbol{a}) = \sum_q \mu_q \Psi_q(\boldsymbol{\mathcal{V}}_q \boldsymbol{a}_q - \boldsymbol{\omega}_q) \tag{19}$$

where $\boldsymbol{a}$ is the unknown, $\boldsymbol{\mathcal{V}}_q$ is a linear operator, $\boldsymbol{\omega}_q$ is a data fixed vector, μ_q are scalar hyper-parameters, and $\Psi_q(\boldsymbol{u}) = \sum_c \phi_q(u_c)$.

In addition, we suppose the following assumptions for the scalar function ϕ_q [24]:

1) $\mathcal{C}^1$, even, coercive,
2) $\phi_q(\sqrt{\cdot})$ is concave on $\mathbb{R}^+$,
3) and $0 < \dot{\phi}_q(u)/u < +\infty$, $\forall u \in \mathbb{R}$.

This objective function structure is chosen to allow efficient algorithms that use majorization with quadratic surrogate functions Q which write [36]

$$Q\left(\boldsymbol{a}, \boldsymbol{a}^k\right) = J\left(\boldsymbol{a}^k\right) + \nabla J\left(\boldsymbol{a}^k\right)^T \left(\boldsymbol{a} - \boldsymbol{a}^k\right) + \frac{1}{2}\left(\boldsymbol{a} - \boldsymbol{a}^k\right) \boldsymbol{A}^{(k)} \left(\boldsymbol{a} - \boldsymbol{a}^k\right) \tag{20}$$

where

$$\boldsymbol{A}^{(k)} = \sum_q \mu_q \boldsymbol{\mathcal{V}}_q^T \operatorname{diag}(\boldsymbol{b}_q^k) \boldsymbol{\mathcal{V}}_q$$

and

$$\boldsymbol{b}_q^k = \frac{\dot{\phi}(\boldsymbol{\mathcal{V}}_q \boldsymbol{x}^k - \boldsymbol{\omega}_q)}{\boldsymbol{\mathcal{V}}_q \boldsymbol{x}^k - \boldsymbol{\omega}_q}. \tag{21}$$

Lemma 3.1: [36] Let J be the objective function defined in Eq. (19) and $\boldsymbol{a}^k \in \mathbb{R}^N$. If the assumption holds, then the quadratic surrogate function Q in (20) is a tangent majorant for J at $\boldsymbol{a}^k$, for all $\boldsymbol{a} \in \mathbb{R}^N$,

$$\begin{cases} Q(\boldsymbol{a}, \boldsymbol{a}^k) \geq J(\boldsymbol{a}), \\ Q(\boldsymbol{a}^k, \boldsymbol{a}^k) = J(\boldsymbol{a}^k). \end{cases} \tag{22}$$

The proposed criterion in Eq. (18) is an instance of Eq. (19) with $q = \{1, 2\}$, $\mu_1 = 1$, $\Psi_1(\cdot) = \|\cdot\|^2_{\boldsymbol{\Sigma}^{-1}}$, $\boldsymbol{\omega}_1 = \boldsymbol{y}$, $\boldsymbol{\mathcal{V}}_1 = \boldsymbol{H}$,

$\boldsymbol{\mathcal{V}}_2 = \boldsymbol{V} = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_\mathcal{C}]^T$. In case of quadratic loss like Ψ_1, the curvature matrix $\boldsymbol{A}$ does not change, includes $\boldsymbol{\Sigma}$, and $\boldsymbol{b}$ variables equal 1. In previous work [37], we considered the half quadratic (HQ) strategy proposed by Geman and Reynolds (GR) [35] that use quadratic surrogate function on all space $\mathbb{R}^N$. However, in our case, the computational cost remains important since our forward model is complex and works on high-dimensional input object and data. We therefore choose to adopt recent and efficient algorithms based on this Majorize-Minimize but for the step strategy applied on subspace optimization [24], [38].

C. Subspace Optimization With Majorize-Minimize Step

Since the criterion in Eq. (18) is differentiable (and convex if ϕ is convex), the optimization can be done with Non-Linear Conjugate Gradient [39], or more efficiently with subspace optimization methods that are known to be the most efficient for this kind of criterion [40]. In the latter case, the iterative algorithm writes

$$\boldsymbol{a}^{(k+1)} = \boldsymbol{a}^{(k)} + \alpha^{(k)}\boldsymbol{g}^{(k)} + \sum_{z=1}^{Z} \beta^{(k,z)}\boldsymbol{d}^{(k-z)} \tag{23}$$

where $\boldsymbol{g}^{(k)} = \nabla J(\boldsymbol{a}^{(k)})$ is the gradient of $J(\boldsymbol{a}^{(k)})$ at the iteration k, $\boldsymbol{d}^{(k-z)}$ are the previous descent directions, and $\alpha^{(k)}$ and $\beta^{(k,z)}$ are scalars.

We particularly rely on the Majorize-Minimize Memory Gradient (3MG) algorithm [24] that exploits the structure of the criterion in Eq (18) to compute both the steps α and the conjugacy parameter $\beta^{(z)}$ via the Majorize-Minimize strategy, like in Sec. III-B. Therefore, Eq. (23) can be rewritten as

$$\boldsymbol{a}^{(k+1)} = \boldsymbol{a}^{(k)} + \boldsymbol{D}^{(k)}\boldsymbol{\alpha}^{(k)} \tag{24}$$

where $\boldsymbol{D}^{(k)}$ is the subspace of dimension Z and $\boldsymbol{\alpha}^{(k)}$ a vector of steps of size Z. Contrary to the traditional line search strategy, finding the steps can be done with the Quadratic Majorize-Minimize strategy in this subspace leading to an explicit formula of $\boldsymbol{\alpha}^{(k)}$ with

$$\boldsymbol{\alpha}^{(k)} = -{\boldsymbol{U}^{(k)}}^{-1} \nabla\, J^{(k)}\Big|_{\boldsymbol{\alpha}} \left(\boldsymbol{a}^{(k)} + \boldsymbol{D}^{(k)}\boldsymbol{\alpha}^{(k)}\right) \tag{25}$$

with $\boldsymbol{U}^{(k)} = {\boldsymbol{D}^{(k)}}^t \boldsymbol{A}^{(k)} \boldsymbol{D}^{(k)}$ a $Z \times Z$ matrix. Eq. (25) corresponds to the explicit solution of a quadratic loss that majorizes the original criterion (Eq. (18)) in the subspace generated by $\boldsymbol{D}^{(k)}$. In our experiments, we choose $Z = 2$, that is a subspace of size 2 where search direction consists of the gradient and the previous search $\boldsymbol{D}^{(k-1)}\boldsymbol{\alpha}^{(k-1)}$. Note that this strategy leads to an efficient algorithm that mainly needs the computation of the gradient (therefore, one application of $\boldsymbol{H}$ and $\boldsymbol{H}^T$ per iteration, the heaviest part of the whole algorithm) and the inversion of a $Z \times Z$ matrix. Moreover, this algorithm has guaranteed convergence, and we refer to [24] for more details.

IV. Experimental Results

This section tests the proposed reconstruction algorithm on two synthetic 3D spatio-spectral images with various spatial and spectral distributions. The developed forward model is general but primarily adapted for the Medium-Resolution Spectrometer of the Mid-Infrared Instrument (MIRI/MRS) onboard the James Webb Space Telescope (JWST), measuring in the infrared spectral range, from 4.9 and 28.3 μm [41].

We compare the proposed algorithm to the state-of-art, which is the shift-and-add (*S&A*) reconstruction algorithm [42], [8] followed by a TV regularization. First, the *S&A* method shifts the overall measurements $\boldsymbol{y}_{c,p}[i', l']$ from all channels c and pointing p in order to align them (after a pre-processing step of the raw data). The results are then co-added, resulting in a reconstructed hyperspectral image with enhanced spatial resolution. This method corresponds to minimizing a least square criterion with

$$J(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{S}\boldsymbol{x}\|^2 \tag{26}$$

with $\boldsymbol{y}^t = [\boldsymbol{y}_{0,0}^T, \ldots, \boldsymbol{y}_{C,P}^T]$ and $\boldsymbol{S}^t = [\boldsymbol{S}_{0,0}^T, \ldots, \boldsymbol{S}_{C,P}^T]$. $\boldsymbol{S}_{c,p}$ is a sampling and summation matrix that models detector sampling but neglect blurring. The solution then writes

$$\boldsymbol{x}_{S\&A} = (\boldsymbol{S}^T\boldsymbol{S})^{-1} \sum_{c,p} \boldsymbol{S}_{c,p}^T \boldsymbol{y}_{c,p} \tag{27}$$

where $\boldsymbol{S}_{c,p}^T$ is an upsampling matrix, and $(\boldsymbol{S}^T\boldsymbol{S})^{-1}$ is a diagonal normalization matrix that counts the number of times a pixel is measured.

Since the *S&A* algorithm does not account for the blurring, it is usually followed by a deconvolution step. In this work, we chose a TV regularization for spatial deconvolution at each wavelength l', implemented with the primal-dual Chambolle-Pock algorithm [43] writing

$$\hat{\boldsymbol{x}}_{S\&A}^{l'} = \arg\min_{\boldsymbol{x}} \left(\|\boldsymbol{y} - \boldsymbol{H}^{l'}\boldsymbol{x}^{l'}\|_2^2 + \mu |\nabla \boldsymbol{x}^{l'}|_1 \right) \tag{28}$$

where $\boldsymbol{H}^{l'}$ is a spatial convolution operator for the wavelength l' and $\nabla \boldsymbol{x}^{l'}$ is the first-order difference of the spatial image for the same wavelength. This method does not allow data fusion since the spatial information is treated separately for every wavelength.

Finally, to highlight the importance of the edge-preserving regularization choice (see section III-A), especially for 3D input images with sharp edges, our algorithm is also compared to the classic l_2 regularization [25]

$$\hat{\boldsymbol{a}}_{\mathrm{q}} = \arg\min_{\boldsymbol{a}} \left(\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{a}\|_{\boldsymbol{\Sigma}^{-1}}^2 + \mu \|\boldsymbol{V}\boldsymbol{a}\|^2 \right) \tag{29}$$

solved via the conjugate-gradient optimization algorithm [44].

A. Setup of the experiment

We denote Obj_1 the first 3D input image, representing a synthetic object for which the results are easily interpretable. Obj_1 lives in a low-dimensional space, and expressed as a linear combination of $M = 3$ known spectral components $\mathbf{s}_m$ computed from astrophysical measurements [45], weighted by mixing coefficients $\boldsymbol{a}_m$ with sharp edges (see figure 3).

The second 3D input image, Obj_2, represents an astrophysical simulation of the photodissociation region located in the "*Orion bar*" also with known abundances and spectral components both available from [46]. It is made of $M = 4$

complex spectral distributions, containing sharp spectral lines and continuum emission, each weighted by their corresponding mixing coefficient a_m, which presents structures in a wide range of spatial scales (see figure 4).

Both input images are represented on a 3D Cartesian grid with $I \times J = 120 \times 120$ pixels with a spatial sampling step of $T_i = T_j = 0.1$ arcseconds. The spectral dimension for both input images measures the infrared spectral range from $4.85\,\mu$m to $28.5\,\mu$m. Obj_1 counts $L = 3500$ wavelengths uniformly sampled with a step $T_l = 6.7\,10^{-3}\,\mu$m, whereas, Obj_2 counts $L = 3551$ wavelengths non-uniformly sampled with a step T_l varying from $2.4\,10^{-3}$ to $1.4\,10^{-2}\,\mu$m.

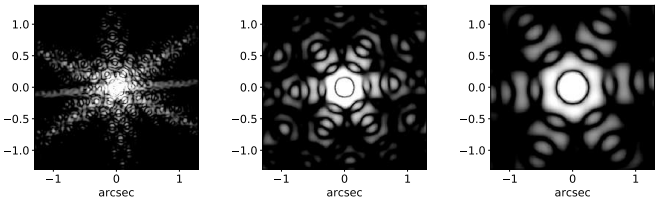
The spectral dimension is divided by the HS instrument into four distinct spectral channels, with different spectral ranges, FOV, slit width, and numbers (see table I). The optical component of the HS instrument is limited by the diffraction [6] with a PSF assumed known. The analytic form of the PSF for a monochromatic wavelength λ can be theoretically obtained from the Fourier transform of the aperture function of the telescope. The PSF width depends on the wavelength, and its FWHM is $\simeq \lambda/D$ (radians), with D referring to the diameter of the telescope aperture. However, for the JWST, there is no exact analytical description of the aperture function. Thus, the PSF is numerically computed using the WebbPSF [27] package, developed by the Space Telescope Science Institute (STScI). Fig. 2 shows three monochromatic PSF at 5, 15, and $25\,\mu$m in logarithmic scales. It highlights the importance of considering a wavelength-dependent PSF in our model, especially since the FWHM of the PSF increases by a factor of 5 between the shortest and longest wavelength.

To allow multi-frame measurements, the forward model considers multiple observations of the same 3D input image, with a dithering pattern of 8 pointing directions. For a particular pointing, the light inside each spatio-spectral selection $\boldsymbol{x}_c$ is dispersed and projected onto 2D detectors with different spectral resolution R, and different spatial and spectral step sizes depending on the channel. The spatial step size of the measurements $T_{i'}$ is fixed by the spatial sampling of the detector. Given the large size of the MRS detectors, we consider in this work that the spectral step size of the measurements is $T_{l'} = 4 \times T_{l'}^{\text{MRS}}$ and the spectral resolution is $R = R^{\text{MRS}}/4$, in order to reduce the computational cost of the problem, where $T_{l'}^{\text{MRS}}$ and R^{MRS} are the spectral sampling of the detector and the spectral resolution of the actual MRS instrument, respectively, provided in [41]. The values of $T_{i'}$, $T_{l'}$, and R, along with the dimension in pixels of the measurements (I', L') are given in Table I. Finally, the simulated measurements $\boldsymbol{y}_{c,p}$ are degraded with an additive Gaussian noise with a standard deviation σ_n fixed to have a Signal-to-Noise Ratio (SNR) equals to 30 dB for each channel

$$\text{SNR} = 10 \log_{10} \left(\|\boldsymbol{y}_{c,p}\|_2^2 / N_{cp} \sigma_n^2 \right) \tag{30}$$

where $N_{c,p}$ is the size of $\boldsymbol{y}_{c,p}$. Hence, the additive Gaussian noise is non-identically distributed over the totality of the measurements $\boldsymbol{y}$.

The algorithms are implemented in Python with the Numpy library and Q-MM [2] toolbox [31] for Quadratic Majorization-Minimization, with a single CPU at 5GHz with 32 GB of memory.



(a) $\lambda = 5\,\mu$m (b) $\lambda = 15\,\mu$m (c) $\lambda = 25\,\mu$m

Fig. 2. PSF at different wavelengths of JWST/MIRI (logarithmic scale) simulated with the WebbPSF package [27].

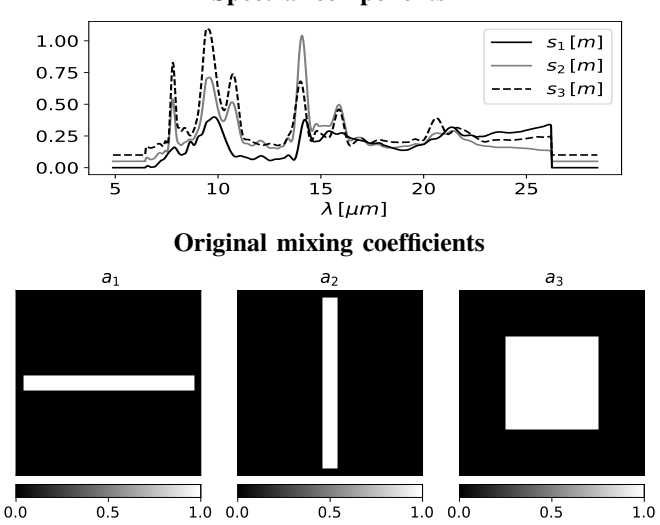


Fig. 3. Obj_1: Spectral components [top], mixing coefficients [bottom].

B. Estimation results for $\hat{\boldsymbol{a}}_m$

The estimation of the mixing coefficients $\hat{\boldsymbol{a}}_m$ depends on the overall blurred, sampled, and noisy measurements $\boldsymbol{y}$. We used the same model for the simulation and the inversion of the measurements to reconstruct the mixing coefficients for Obj_1, shown in Fig. 5 [top]. In contrast, we considered the following errors in the model to reconstruct the mixing coefficients for Obj_2 shown in Fig. 5 [bottom]: (1) the wavelength-dependent spatial PSF is spectrally shifted with an offset of $+0.25\,\mu$m (the PSF used for the reconstruction are therefore wider than the PSF used for the simulation), (2) the response of the gratings (see Eq. 4 and Fig. 1) is approximated by a Gaussian PSF, and (3) the Gaussian noise is non-identically distributed during the simulation process whereas it is identically distributed during the inversion. The red frames in Fig. 5 represent the largest observed FOV, corresponding to the *Channel 4* (see Tab. I). We are interested in reconstructing $\hat{\boldsymbol{a}}_m$ inside this FOV even if for the other channels (or wavelengths), no measurements have been made. We test our reconstruction algorithm

[2] https://github.com/forieux/qmm/

Ch.	Spectral range (μm)	FOV (pixels)	Slit width (arcsec)	Slit number	$T_{i'}$ (arcsec)	$T_{l'}$ (μm)	I' pixels	L' pixels	R
1	4.9 – 7.7	34×42	$2 \times T_j$	21	$2 \times T_i$	$4\,10^{-3}$	17	750	867
2	7.4 – 11.7	42×51	$3 \times T_j$	17	$2 \times T_i$	$6\,10^{-3}$	21	750	760
3	11.5 – 18.1	57×64	$4 \times T_j$	16	$3 \times T_i$	$9\,10^{-3}$	19	750	596
4	17.7 – 28.5	72×72	$5 \times T_j$	12	$3 \times T_i$	$1.6\,10^{-2}$	24	750	410

TABLE I
CHARACTERISTICS SPECIFIC TO THE FOUR SPECTRAL CHANNELS OF THE IFS INSTRUMENT CONSIDERED IN THIS WORK.

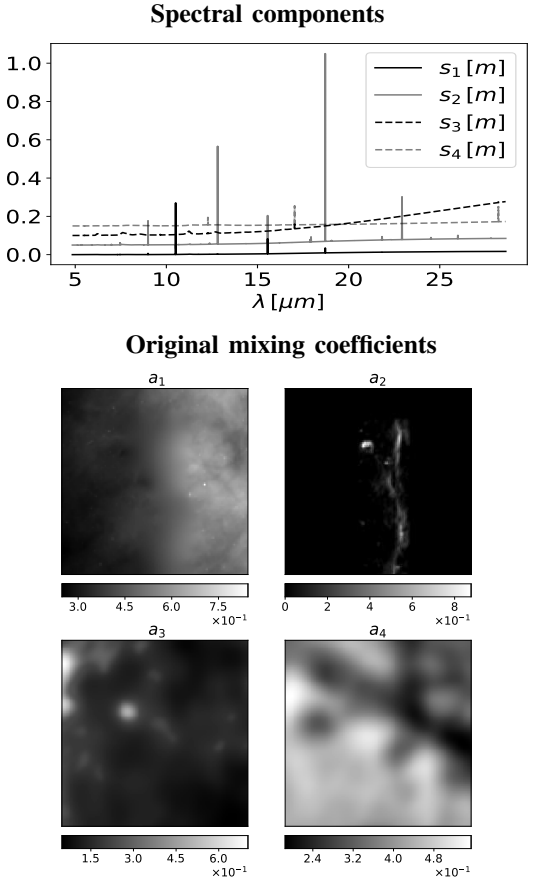


Fig. 4. Obj_2: Spectral components [top], mixing coefficients [bottom].

with a sufficient number of iterations (see Table II) to ensure convergence towards the solution $\hat{\boldsymbol{a}}_m$. The reconstruction of $\boldsymbol{a}_m$ for Obj_2 is computationally more expensive than that for Obj_1, since Obj_2 has one more spectral component.

Our proposed algorithm is based on minimizing the regularized objective function in Eq. (18) with a convex regularization function ϕ for edge-preserving. We particularly focus on the Huber potential function [23] with

$$\phi(\delta, T) = \begin{cases} \delta^2, & \text{if } |\delta| \leq T. \\ 2T|\delta| - T^2 & \text{otherwise, } T \in \mathbb{R}^+. \end{cases} \tag{31}$$

The Huber function is continuously differentiable with a quadratic form below a fixed threshold T to promote smoothness to the solution and a linear form above T to preserve the high gradient values. Consequently, two regularization parameters μ and T must be tuned to ensure the best reconstruction. Their values are reported in Table II. In practice, we have minimized the normalized least square error between the original $\boldsymbol{a}_m$ mixing coefficient and the estimated ones $\hat{\boldsymbol{a}}_m$ for both objects with

$$\text{Error}(\mu, T) = \|\boldsymbol{a}_m - \hat{\boldsymbol{a}}_m(\mu, T)\|_2 / \|\boldsymbol{a}_m\|_2. \tag{32}$$

The spatial distribution differs between $\boldsymbol{a}_m$ for Obj_2, particularly between $\boldsymbol{a}_{m=2}$ which contains sharp edges and $\boldsymbol{a}_{m\neq 2}$ which are smoother. Therefore, we have used one set of parameters (μ, T) for $m = 2$, and another one for $m \neq 2$ (see Table II).

The comparison between figures 3-4 and figure 5 shows that the reconstructed mixing coefficients $\hat{\boldsymbol{a}}_m$ for both objects are unmixed and deconvoluted while preventing noise amplification without excessive penalization of sharp edges. The normalized least square errors are as small as $0.01\,\%$ and $0.98\,\%$, for Obj_1 and Obj_2, respectively.

	Iterations	Runtime [s] per iteration	μ	T
Obj_1	455	7.8	18	0.025
Obj_2	633	10.6	0.005 ($m \neq 2$) 0.0005 ($m = 2$)	1.5 ($m \neq 2$) 0.001 ($m = 2$)

TABLE II
ITERATION NUMBERS AND HYPERPARAMETER VALUES Obj_1 AND Obj_2.

C. Hyperspectral Reconstruction

This section compares the original and the reconstructed 3D images with the proposed and state-of-the-art algorithms. The reconstructed HS images are obtained from the estimated mixing coefficients $\hat{\boldsymbol{a}}_m$ using Eq. (8). In addition to the l_2 reconstruction, the proposed results are compared to the "Shift and Add" algorithm (*S&A*) Eq. (27) followed by a TV deconvolution using Eq. (28).

We first showcase in Fig. 6 the spectral distribution at the center of the FOV of the original and the reconstructed images computed with the *S&A* and the proposed algorithms. Qualitatively the spectral distribution of the reconstructed image with the proposed algorithm matches the original spectral distribution over the whole measured range. On the other hand, the *S&A* algorithm fails to fully reconstruct the spectral distribution, particularly the spectral lines in Obj_2, which appears broader and less intense than the original ones. Such results are expected since the *S&A* algorithm does not consider the spectral blurring initially introduced by the wavelength dispersion system.

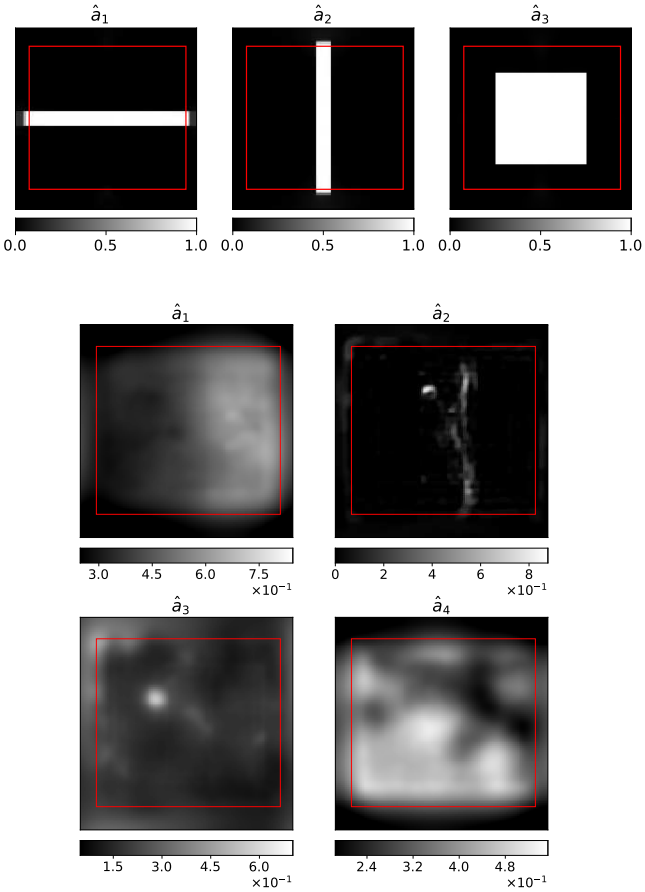


Fig. 5. Estimated mixing coefficients $\hat{\boldsymbol{a}}_m$ for Obj_1 [top], Obj_2 [bottom].

Fig. 9 illustrates the spatial distribution of the reconstructed 3D image for Obj_1 for three monochromatic images at $\lambda = 6.5$, 14 and 21 μm, respectively, belonging to channels 1, 3, and 4. The loss of spatial information caused by the detector integration is compensated in the multi-frame SR reconstruction using the *S&A* algorithm. However, since this method does not consider spectral variations of the PSF, the reconstructed images are blurred, especially at long wavelengths. We proceed by applying a TV deconvolution for each monochromatic image. This added step allows better preservation of sharp edges and smaller errors but fails to restore spatial details at a small scale since the regularization is applied separately for each monochromatic image and does not account for the correlations between spectral bands. On the other hand, the proposed algorithm shows a good performance with smaller error values for all monochromatic images. The improvement of spatial resolution is striking, and the spatial dynamic range appears fully reconstructed. Moreover, the edges in the image are well preserved, whereas the l_2 approach introduces smoothing and ringing artifacts.

Fig. 10 shows the spatial distribution of the reconstructed 3D image for Obj_2, for three monochromatic images at $\lambda = 6.5$ μm which corresponds to continuum emission, and at $\lambda = 17$ and 18.7 μm which corresponds to two different spectral lines. As mentioned earlier, the reconstructed spectral lines with the *S&A* algorithm are spectrally broadened

	No model errors	Model errors
PSNR	65	40
SSIM	0.99	0.99
SAM	4×10^{-4}	4×10^{-3}

TABLE III
THE PSNR, SSIM AND SAM FOR THE RECONSTRUCTED Obj_1 USING THE PROPOSED ALGORITHM, WITH AND WITHOUT ERRORS IN THE MODEL

because the spectral response is not considered. Hence, for a fair comparison between the algorithms, we have spectrally integrated the reconstructed HS image with the *S&A* algorithm over the broad reconstructed spectral line, then we proceed with the TV deconvolution for the integrated images. In all cases, our proposed algorithm shows the best qualitative reconstructions with the lowest errors for all monochromatic images. The l_2 approach gives good results with comparable errors for the smooth images but fails to preserve the sharp edges, particularly at $\lambda = 18.7$ μm as illustrated in Fig. 7. Analogously to Obj_1, the *S&A* and TV algorithms fail to fully reconstruct small-scale spatial details.

Finally, we compare the proposed algorithm to the HS reconstruction algorithm proposed in [37]. The latter directly estimates the full 3D input from the measurements by performing a joint spatial and spectral reconstruction with a ℓ_2–ℓ_1 regularization on spatial and spectral difference (like vector-TV) and hyperparameter set to optimize the recontruction error. We validate the reconstruction on a third image Obj_3[3] that has more complex backgrounds such as small-scale spatial features and discontinuities. Fig. 11 illustrates the spatial distribution of the reconstructed 3D image for Obj_3 at $\lambda = 14\,\mu$m.

The proposed algorithm restores the spatial dynamic with an error of 2.48% while *S&A* has an error of 7.06% and [37] with vector-TV prior has an error of 3.9%. Moreover, the proposed LMM algorithm requires only the reconstructions of $\hat{\boldsymbol{a}}_m$. In contrast, the algorithm in [37] has to estimate the full HSI $\boldsymbol{x}$, yielding more errors especially at long wavelengths since the blurring is more critical. More important, the spatial hyperparameters may not be adapted at every wavelength, like for the Obj_2 case, and the introduction of additional hyperparameters makes the tunning very difficult.

D. Quality Metrics

To better evaluate the spatial and spectral performances of the reconstruction algorithms, we use three quantitative measurements:

1) the Spectral Angular Mapper (SAM) [47] measuring the spectral distortion, in radians, of the m^{th} pixel

$$\text{SAM}(m) = \arccos\left(\frac{< \boldsymbol{x}_m, \hat{\boldsymbol{x}}_m >}{\|\boldsymbol{x}_m\|_2 \|\hat{\boldsymbol{x}}_i\|_m}\right). \qquad (33)$$

where $\boldsymbol{x}_m$ and $\hat{\boldsymbol{x}}_m$ are the spectral vector of the m^{th} spatial location ($m \in [0, \ldots, I] \times [0, \ldots, J]$) of the original and reconstructed 3D images, respectively. The

[3]available here http://lesun.weebly.com/hyperspectral-data-set.html

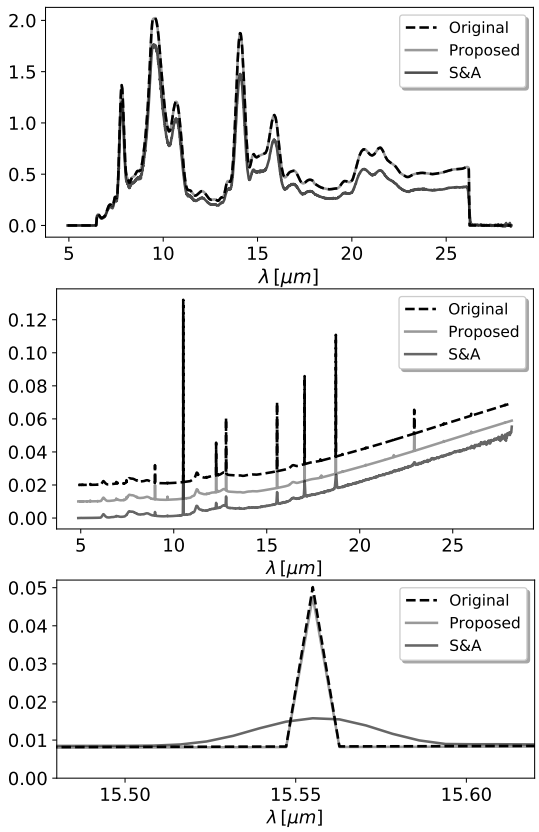


Fig. 6. Spectral distribution of the reconstructed HS image with the proposed and *S&A* algorithms, at the central spatial position (60,60) for Obj_1 [top], Obj_2 [middle: note that the three spectral distributions are shifted for clarity], zoom on a spectral line for Obj_2 [bottom].

Fig. 7. Zoom on sharp edges for Obj_2 at 18.7μm: Original [left], proposed [center], l_2 approach [right].

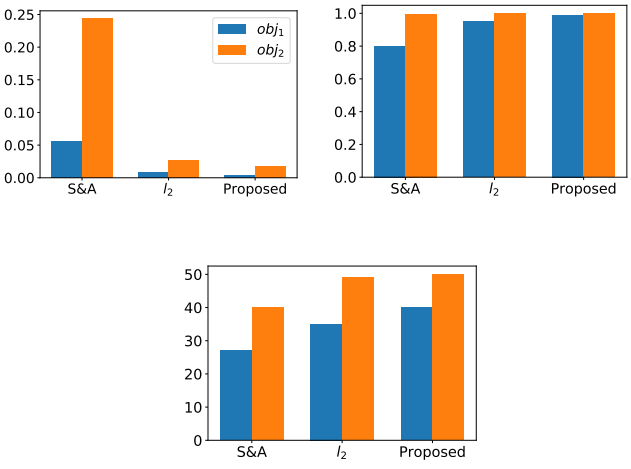


Fig. 8. Quality metrics for Obj_1 (blue) and Obj_2 (orange): Global SAM [top left], global SSIM [top right], global PSNR [bottom]

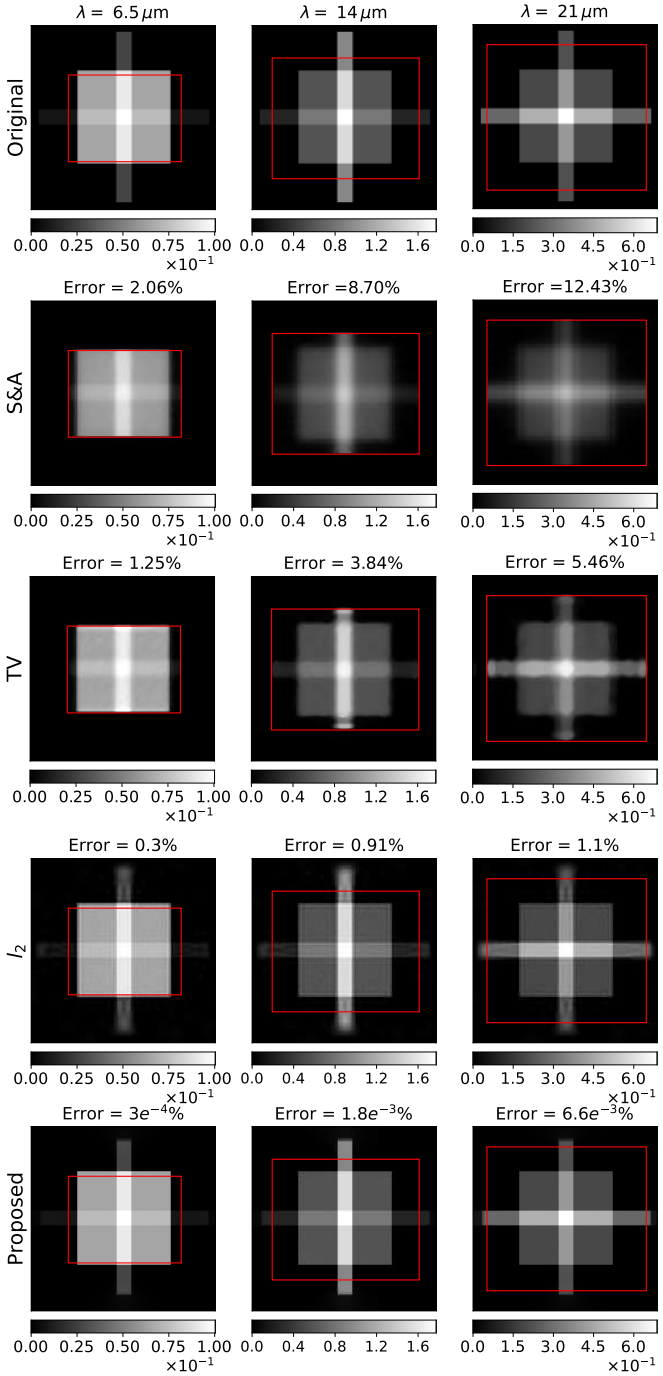


Fig. 9. Spatial reconstruction for Obj_1 : Original images at 6.5, 14, and 21 μm [1st row], *S&A* [2nd row], TV restoration [3rd row], l_2 approach [4th row], proposed [5th row].

further the SAM value is from 0, the greater the spectral distortion.

2) the peak signal-to-noise ratio (PSNR)

$$\mathrm{PSNR}(l) = 10\log_{10}\left(\frac{\max(\boldsymbol{x})^l}{\|\boldsymbol{x}^l - \hat{\boldsymbol{x}}^l\|^2}\right). \qquad (34)$$

$\mathrm{PSNR}(l)$ denotes the PSNR of the spatial image at the l^{th} spectral band of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$.

3) the Structural Similarity Index (SSIM) [48], computed

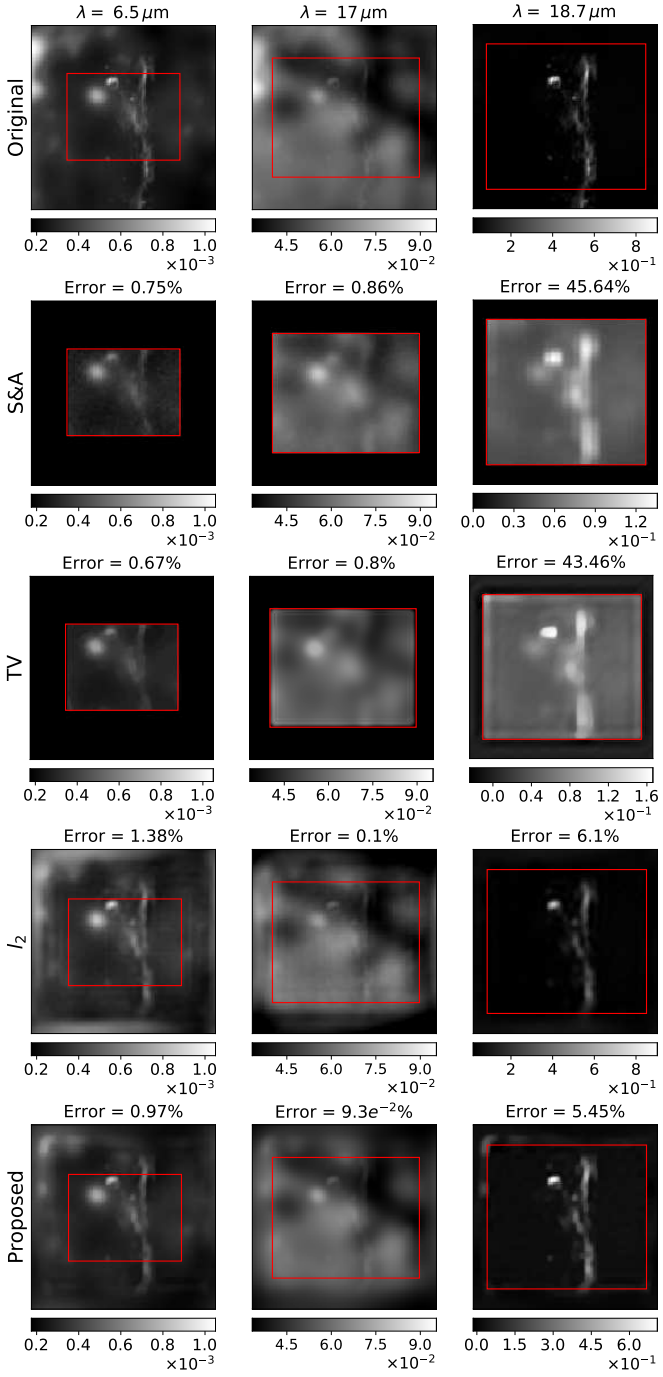


Fig. 10. Spatial reconstruction for Obj_2 : Original images at 6.5, 17, and 18.7 μm [1st row], *S&A* [2nd row], TV [3rd row], l_2 approach [4th row], proposed [5th row].

for each l^{th} spectral band, whose value varies between 0 and 1. The higher the value, the better the similarity. The global SAM is computed by averaging the whole image, while the global PSNR and SSIM are computed by averaging all spectral bands.

Table III compares the proposed reconstruction of Obj_1 with and without model errors in terms of average SAM, SSIM, and PSNR. We illustrate in Fig. 8 the average SAM, SSIM, and PSNR for Obj_1 and Obj_2 while considering the

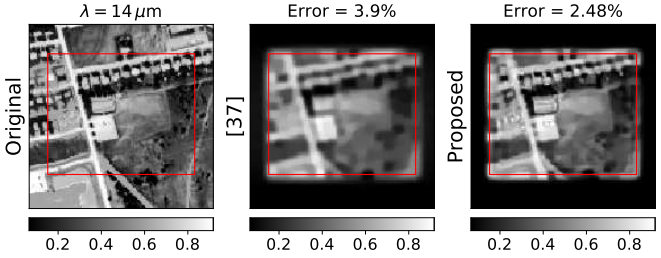


Fig. 11. Spatial reconstruction for Obj_3: Original image [left], spatial spectral reconstruction with [37] [center], proposed [right].

errors in the model during the reconstruction process. The proposed algorithm shows the best spectral reconstruction with the lowest distortion for both objects. The spectral distortion is very high for the *S&A* algorithm applied to Obj_2, due to the presence of sharp spectral lines which are hardly recovered by this algorithm. Moreover, our algorithm shows the best similarity index, especially for Obj_1. Overall, the proposed algorithm shows the best spatio-spectral reconstruction for both scenes, with the highest average PSNR values.

V. Conclusion

We present in this work a new model for hyperspectral data fusion based on low-rank approximation and a new efficient Super-Resolution algorithm for hyperspectral reconstruction based on a Quadratic Majorization-Minimization optimization algorithm.

The first contribution is to develop an explicit forward model based on IFS instruments. This model takes as input a high spatio-spectral resolution image, approximated on a low-rank subspace, and acquires a set of 2D measurements projected onto different detectors with different characteristics. The complex forward model takes into account (1) different observations of the same scene with sub-pixel shifts, (2) wavelength-dependent spatial and spectral PSFs, (3) different spectral channels and IFUs, observing the input with different spectral ranges and different numbers of slits of different sizes, and finally (4) heterogeneous spatio-spectral samplings.

The second contribution is a fusion of the multi-frame blurred and sampled 2D measurements, acquired from different spectral channels, in order to restore the single HSI observed input. The algorithm is based on the regularized least square approach with convex edge-preserving regularization, and solved via the iterative Majorize-Minimize Memory Gradient (3MG) [24] optimization algorithm, with freely provided code [4].

Our method allows joint spectral unmixing with spatial and spectral enhancements. The known spectral components serve as a spectral regularization to our approach and prevent spectral distortion, whereas the multi-frame observations and the enforced spatial regularization allow restoring the original spatial distribution without excessive penalization of high gradient values.

Our work is validated with relative errors below 1% over the whole reconstructed HS images for an SNR = 30 dB. Our

[4] http://github.com/forieux/qmm

algorithm outperformed qualitatively and quantitatively the l_2 approach, as well as the standard *S&A* and TV deconvolution algorithms.

Several perspectives can be considered. First, non-convex regularization or data-learned prior may be envisaged for better resolution of the reconstruction. The spatial and spectral resolutions of the reconstructed 3D image can also be enhanced by performing a fusion between the IFS measurements with a high spectral but low spatial resolution, considered in this work, and multispectral measurements with a high spatial but low spectral resolution, observing the same scene. The fusion problem can be solved in the inverse problem framework [1]. In addition, the spectral components of the LMM are, in many applications, not provided *a priori* and must be extracted or learned directly from the measurements along with the mixing coefficients [49]. In a wider perspective, we would like to estimate the hyperparameters jointly with the 3D input image instead of being fixed by hand. The problem can be formulated in the Bayesian framework where the solution is deduced from *a posteriori* law for the unknown hyperparameters and the 3D input image [50].

ACKNOWLEDGMENT

We thank Jérôme Idier (LS2N – CNRS) and Saïd Moussaouï (LS2N – École Centrale de Nantes) for fruitful discussions about MM optimization, Olivier Berné (IRAP – CNRS) for Obj_1 spectra, and Claire Guilloteau (INSA Rouen) for providing the Orion bar spectra and maps (Obj_2).

REFERENCES

[1] C. Guilloteau, T. Oberlin, O. Berné, and N. Dobigeon, "Hyperspectral and multispectral image fusion under spectrally varying spatial blurs – application to high dimensional infrared astronomical imaging," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1362–1374, 2020.

[2] J. G. Dwight and T. S. Tkaczyk, "Lenslet array tunable snapshot imaging spectrometer (latis) for hyperspectral fluorescence microscopy," *Biomed. Opt. Express*, vol. 8, no. 3, pp. 1950–1964, Mar 2017. [Online]. Available: http://www.osapublishing.org/boe/abstract.cfm?URI=boe-s8-s3-s1950

[3] M. Shimoni, R. Haelterman, and C. Perneel, "Hypersectral imaging for military and security applications: Combining myriad processing and sensing techniques," *IEEE Geoscience and Remote Sensing Magazine*, vol. 7, no. 2, pp. 101–117, 2019.

[4] G. Lu and B. Fei, "Medical hyperspectral imaging: a review," *Journal of Biomedical Optics*, vol. 19, no. 1, pp. 1 – 24, 2014. [Online]. Available: https://doi.org/10.1117/1.JBO.19.1.010901

[5] S. Vives, E. Prieto, Y. Salaun, and P. Godefroy, "New technological developments in integral field spectroscopy," in *Advanced Optical and Mechanical Technologies in Telescopes and Instrumentation*, E. Atad-Ettedgui and D. Lemke, Eds., vol. 7018, International Society for Optics and Photonics. SPIE, 2008, pp. 959 – 968. [Online]. Available: https://doi.org/10.1117/12.789576

[6] J. W. Goodman, *Introduction to Fourier Optics McGraw-Hill Series in Electrical and Computer Engineering*, 1996, vol. 8, no. 5. [Online]. Available: http://stacks.iop.org/1355-s5111/8/i=5/a=014?key=crossref.ad20ea108e8f625cb0486bf680f74198

[7] J.-P. Pérez, *Optique - Fondements et applications*. Paris, France:Dunod, 2004.

[8] A. S. Fruchter and R. N. Hook, "Drizzle: A method for the linear reconstruction of undersampled images," *Publications of the Astronomical Society of the Pacific*, vol. 114, no. 792, pp. 144–152, feb 2002. [Online]. Available: https://doi.org/10.1086%2F338393

[9] R. Hook and A. Fruchter, "Dithering, sampling and image reconstruction," vol. 216, p. 521, 01 2000.

[10] Sung Cheol Park, Min Kyu Park, and Moon Gi Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, May 2003.

[11] M. Elad and Y. Hel-Or, "A fast super-resolution reconstruction algorithm for pure translation motion and common space-invariant blur," *IEEE Transactions on Image Processing*, vol. 10, pp. 1187–1193, 01 2001.

[12] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992. [Online]. Available: https://www.sciencedirect.com/science/article/pii/016727899290242F

[13] H. Shen and L. Zhang, "A map-based algorithm for destriping and inpainting of remotely sensed images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 5, pp. 1492–1502, 2008.

[14] T. Akgun, Y. Altunbasak, and R. M. Mersereau, "Super-resolution reconstruction of hyperspectral images," *IEEE Transactions on Image Processing*, vol. 14, no. 11, pp. 1860–1875, 2005.

[15] H. Zhang, L. Zhang, and H. Shen, "A super-resolution reconstruction algorithm for hyperspectral images," *Signal Processing*, vol. 92, no. 9, pp. 2082 – 2096, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168412000345

[16] Y. Chang, L. Yan, X.-L. Zhao, H. Fang, Z. Zhang, and S. Zhong, "Weighted low-rank tensor recovery for hyperspectral image restoration," *IEEE transactions on cybernetics*, vol. 50, no. 11, pp. 4558–4572, 2020.

[17] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. Am. A*, vol. 6, no. 11, pp. 1715–1726, Nov 1989. [Online]. Available: http://josaa.osa.org/abstract.cfm?URI=josaa-s6-s11-s1715

[18] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, 2016.

[19] R. Schultz and R. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society*, vol. 5, pp. 996–1011, 02 1996.

[20] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," vol. 53, 07 2014.

[21] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 7, pp. 3658–3668, 2015.

[22] M. Simões, J. M. Bioucas-Dias, L. B. Almeida, and J. Chanussot, "A convex formulation for hyperspectral image superresolution via subspace-based regularization," *CoRR*, vol. abs/1411.4005, 2014. [Online]. Available: http://arxiv.org/abs/1411.4005

[23] J. Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1001–1009, July 2001.

[24] E. Chouzenoux, J. Idier, and S. Moussaoui, "A majorize–minimize strategy for subspace optimization applied to image restoration," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1517–1528, 2011.

[25] A. N. Tikhonov, A. Goncharsky, V. V. Stepanov, and A. G. Yagola, *Numerical Methods for the Solution of Ill-Posed Problems*, ser. Mathematics and Its Applications. Springer Netherlands, 1995. [Online]. Available: https://www.springer.com/gp/book/9780792335832

[26] J. Adams, M. Smith, and P. Johnson, "Spectral mixture modeling: A new analysis of rock and soil types at the viking lander 1 site," *Journal of Geophysical Research*, vol. 91, pp. 8098–8112, 1986.

[27] M. D. Perrin, R. Soummer, E. M. Elliott, M. D. Lallo, and A. Sivaramakrishnan, "Simulating point spread functions for the James Webb Space Telescope with WebbPSF," in *Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*, M. C. Clampin, G. G. Fazio, H. A. MacEwen, and J. M. O. Jr., Eds., vol. 8442, International Society for Optics and Photonics. SPIE, 2012, pp. 1193 – 1203. [Online]. Available: https://doi.org/10.1117/12.925230

[28] B. Hunt, "A matrix theory proof of the discrete convolution theorem," *IEEE Transactions on Audio and Electroacoustics*, vol. 19, no. 4, pp. 285–288, 1971.

[29] O. K. Ersoy, *Diffraction, Fourier optics and imaging*. John Wiley & Sons, 2006, vol. 30.

[30] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[31] F. Orieux and R. Abirizk, "Q-mm: The quadratic majorize-minimize python toolbox." [Online]. Available: https://github.com/forieux/qmm

[32] Y. Zhao, J. Yang, Q. Zhang, L. Song, Y. Cheng, and Q. Pan, "Hyperspectral imagery super-resolution by sparse representation and spectral regularization," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–10, 2011.

[33] T. Zhang, Y. Fu, L. Wang, and H. Huang, "Hyperspectral image reconstruction using deep external and internal learning," in *2019 IEEE/CVF*

International Conference on Computer Vision (ICCV), 2019, pp. 8558–8567.
[34] E. Chouzenoux and J.-C. Pesquet, "Convergence rate analysis of the majorize–minimize subspace algorithm," *IEEE Signal Processing Letters*, vol. 23, no. 9, pp. 1284–1288, 2016.
[35] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 3, pp. 367–383, 1992.
[36] M. Allain, J. Idier, and Y. Goussard, "On global and local convergence of half-quadratic algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 5, pp. 1130–1142, 2006.
[37] R. Abi-rizk, F. Orieux, and A. Abergel, "Non-stationary hyperspectral forward model and high-resolution," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 2975–2979.
[38] C. Labat and J. Idier, "Doi 10.1007/s10957-007-9306-x convergence of conjugate gradient methods with a closed-form stepsize formula," *Journal of Optimization Theory and Applications*, vol. 136, pp. 43–60, 01 2008.
[39] J. Nocedal, "Conjugate gradient methods and nonlinear optimization," *Linear and nonlinear conjugate gradient-related methods*, pp. 9–23, 1996.
[40] M. Zibulevsky, "Sesop-tn: Combining sequential subspace optimization with truncated newton method," Computer Science Department, Technion, Tech. Rep., 2008.
[41] M. Wells, J. W. Pel, A. Glasse, G. Wright, G. Kroes, R. Azzollini, S. Beard, B. Brandl, A. Gallie, V. C. Geers, A. M. Glauser, P. Hastings, T. Henning, R. Jager, K. Justtanont, B. Kruizinga, F. Lahuis, D. Lee, I. Martinez Delgado, and D. Wright, "The mid-infrared instrument for the james webb space telescope, vi: The medium resolution spectrometer," *Publications of the Astronomical Society of the Pacific*, vol. 127, pp. 646–664, 07 2015.
[42] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Robust shift and add approach to superresolution," in *Applications of Digital Image Processing XXVI*, A. G. Tescher, Ed., vol. 5203, International Society for Optics and Photonics. SPIE, 2003, pp. 121 – 130. [Online]. Available: https://doi.org/10.1117/12.507194
[43] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, 05 2011.
[44] J. R. Shewchuk *et al.*, "An introduction to the conjugate gradient method without the agonizing pain," 1994.
[45] O. Berné, C. Joblin, Y. Deville, J. D. Smith, M. Rapacioli, J. P. Bernard, J. Thomas, W. Reach, and A. Abergel, "Analysis of the emission of very small dust particles from Spitzer spectro-imagery data using blind signal separation methods," *Astronomy and Astrophysics - A&A*, vol. 469, pp. 575–586, July 2007, 14 pages, 11 figures, to appear in A&A. [Online]. Available: https://hal.archives-souvertes.fr/hal-s00287344
[46] C. Guilloteau, T. Oberlin, O. Berné, E. Habart, and N. Dobigeon, "Simulated jwst data sets for multispectral and hyperspectral image fusion," *The Astronomical Journal*, vol. 160, no. 1, p. 28, Jun 2020. [Online]. Available: http://dx.doi.org/10.3847/1538-s3881/ab9301
[47] Chein-I Chang, "An information-theoretic approach to spectral variability, similarity, and discrimination for hyperspectral image analysis," *IEEE Transactions on Information Theory*, vol. 46, no. 5, pp. 1927–1932, 2000.
[48] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
[49] S. Henrot, C. Soussen, M. Dossot, and D. Brie, "Does deblurring improve geometrical hyperspectral unmixing?" *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1169–1180, 2014.
[50] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4355–4368, 2009.

Ralph Abi Rizk received his Ph.D degree in image and signal processing from Laboratoire signaux et systèmes (Université Paris-Saclay, CNRS, CentraleSupélec), France in 2021. He is currently a post-doctoral researcher at Laboratoire des Sciences du Numérique de Nantes, Centrale Nantes, France. His research interests include inverse problem approaches, hyperspectral image reconstruction, and ultrasound image processing for non destructive testing (NDT).

François Orieux is Assistant Professor at the Université Paris-Saclay in the Laboratoire des Signaux et Systèmes, Groupe Problèmes Inverses (Université Paris-Saclay, CNRS, CentraleSupélec), France. He is also an associate researcher with the Institut d'Astrophysique Spatiale (Univ. Paris- Saclay, CNRS). He received his Ph.D. degree in signal processing at Université Paris-Sud, Orsay, France. His research focuses on Bayesian methodological approaches for ill-posed inverse problem resolution with examples of applications in astrophysics or biological microscopy.

Alain Abergel is professor of Physics and Astrophysics at the Paris-Saclay University. He is an astrophysicist of the interstellar medium at the Institut d'Astrophysique Spatiale (Paris-Saclay University, CNRS), and is involved in the scientific exploitation of several space missions for astrophysics (ISO, Spitzer, Herschel, Planck, JWST, ...).

Fast Joint Multiband Reconstruction from Wideband Images Based on Low-Rank Approximation

M. Amine Hadj-Youcef, François Orieux, Alain Abergel, Aurélia Fraysse

***Abstract*—Multispectral imaging systems are increasingly used in many scientific fields. However multispectral images generally present spectral and spatial limitations: the spectral information within each band is lacking because of spectral integration over the band, and the spatial resolution is limited due to the spatial convolution by spectrally variant Point Spread Functions which introduce a spatial variant blur. To address the ill-posed inverse problem of reconstruction from wideband images, we propose a new approach combining a precise instrument model for the degraded multispectral images together with a spectral approximation on a low-rank subspace of the object. The reconstruction is based on the minimization of a convex objective function composed of a data fidelity and an edge-preserving regularization term. The proposed half-quadratic algorithm alternates between the minimization of a quadratic and a separable problem, and we show that both closed-form solutions are available and tractable. Therefore, even with a non-stationary data model, the algorithm is very fast and results are obtained in a few seconds.**

Several tests are performed for multispectral data to be taken by MIRI, the mid-infrared imager of the future James Webb Space Telescope (JWST). The reconstruction results show a significant increase in spatial and spectral resolutions compared to state-of-the-art methods. Our proposed algorithm allows us to recover the spectroscopic information contained in the wideband multispectral images and to provide hyperspectral images with a homogenized spatial resolution over the entire spectral range.

***Index Terms*—Inverse Problems. Multispectral Imaging. Hyperspectral Imaging. Deconvolution. Image Reconstruction**

I. Introduction

MULTISPECTRAL imaging systems are used in many fields, e.g., astrophysics [1], remote sensing [2], medicine [3] or microscopy [4]. These imaging systems produce integrated multispectral images by observing a two spatial and one spectral dimensions, that is a 2D+λ object. Multispectral images have the benefits over hyperspectral images to have much larger field of view, better spatial resolution and higher sensitivity. However, they suffer from several undesirable spatial and spectral degradations. Firstly, spectral information is lacking because of the detector integration over wide spectral bands, resulting in an important subsampling. Secondly, the spatial resolution is limited by a 2D convolution of the 2D+λ object with the spectrally varying optical response or PSF (Point Spread Function). Generally, due to the diffraction theory [5], the longer the wavelength the wider the PSF. Moreover, the spectral content of the input object depends on the pixel position in general. As a result, the blurring of multispectral images, which are integrated over a wide spectral range, is spatially varying. Furthermore the observations are made on wide bands, increasing the effect of spatial varying PSF.

The state-of-the-art approach generally neglects spectral and spatial variations of the PSF within a band [6] as well as correlations between bands, the reconstruction becoming an independent 2D deconvolution of each image [7]. In that case, the approach remains limited to systems with narrow spectral bands, which is generally not the case for multispectral imaging systems. Also, each spectral image is processed independently, and the spatial resolution is different for each processed image.

Other works focus on the variability of the PSF, especially for image deconvolution (e.g., [8], [9]) where the shift-variant PSF is approximated with a linearly interpolated PSF. However, such a technique is generally not applicable to multispectral images since the spatial variations of the blurring are mainly due to the spatial variations of the spectral content of the object.

Another technique commonly used for astronomical images (e.g., [10], [11]) is the PSF homogenization between images obtained from different spectral bands or instruments. It consists of convolving images with appropriate kernels such that they appear as if they were measured with the same band or instrument. This approach is straightforward and simple, however, it introduces an additional blur and does not allow spectral reconstruction.

To address the problems of strong spectral subsampling and limitations of the spatial resolution due to the 2D convolution by a varying PSF, and to perform joint processing of all wide spectral bands, we propose to consider a linear spectral model together with a precise model of the multispectral instrument. Our goal is to derive an estimate of the 2D+λ object with the best spatial and spectral resolutions from the multispectral images. The proposed spectral modeling allows the description of the complexity of the input object spectra with a small number of components. This is common for hyperspectral images, in particular with a linear mixing model [12]–[14]. It was first proposed in [15] for the analysis of in-situ Mars surface images and, since then, has been used in several applications and methods [14], [16]–[19], also with spatial and spectral correlations and image enhancement methods based on total variation [20], [21]. In the context of multispectral imaging, the spectral modeling is less common except in pansharpening techniques [22]. It also is not adapted

M. Amine Hadj-Youcef, François Orieux and Aurélia Fraysse are with Laboratoire des Signaux et Systèmes (L2S), Université Paris-Saclay, CNRS, CentraleSupélec, 3 rue Joliot-Curie, 91 192 Gif-sur-Yvette, France. e-mail: hadjyoucef.amine@gmail.fr

Alain Abergel is with the Institut d'Astrophysique Spatiale, Université Paris-Saclay, CNRS, 91405 Orsay, France.

Manuscript received September, 2019; Accepted May 18, 2020.

to spectrally varying PSF. Therefore we combine spectral modeling with an instrument model that correctly describes the spectral dependence of the PSF (and the detector sampling) to build a forward model of the data.

We show that all spectral and spatial information are embedded in the forward model, without any additional approximation. We also propose a fast convex reconstruction algorithm, where the two steps of the half-quadratic iterative algorithm are closed-form expression and tractable, for the joint processing of all wideband images. Our algorithm is very fast and can restore spatial details with identical resolution in all bands and better performances than the state-of-the-art TV-based image restoration. The proposed algorithm is tested for the reconstruction of three different hyperspectral objects that cover most of the encountered cases, including a simulated astrophysical object. The results clearly show a strong improvement in spectral and spatial resolutions compared to the state-of-the-art method [23]–[25]. The imaging system considered for this application is the Mid-InfraRed Instrument (MIRI) imager [26] on board the future James Webb Space Telescope (JWST) [27], which will be from 2021 the most ambitious telescope ever launched in space.

The paper is organized as follows. Section II presents the proposed methodology, first the instrument model developed for the imaging system (Section II-A), then the object representation model based on the linear mixing model (Section II-B) and finally the forward model (Section II-C). The reconstruction algorithm based on regularization methods is presented in Section III. Reconstruction results, including a comparison with a state-of-the-art method, and a discussion are presented in Section IV for the MIRI imager of the JWST. Finally, we conclude our work and provide perspectives in Section V.

II. Data model

A. Instrument Model

The hyperspectral object of interest is defined by

$$\phi(x, y, \lambda) : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}$$

where $(x, y) \in \mathbb{R}^2$ represent the spatial dimensions and $\lambda \in \mathbb{R}_+$ represents the spectral one.

The imaging system consists of an optical system, a spectral filter, and a detector. Because of the optical diffraction theory, the optical system response is carried out by a 2D convolution [5] between the hyperspectral object ϕ and the spectral-variant PSF, h, assumed to be known, as $\iint_{\mathbb{R}^2} \phi(x', y', \lambda) h(x - x', y - y', \lambda)\, \mathrm{d}x'\mathrm{d}y'$. Hence, the final 2D images are impacted by a non-stationary spectrally integrated PSF that depends on the unknown spectral content of the object and also limits the spatial resolution of the images. Note that we assume that the monochromatic PSFs are spatially invariant. This may not be the case, but the amplitudes of the spatial variations of the monochromatic PSFs over the detector are generally lower than the amplitudes of the spectral variations of the PSF over the wide spectral bands of multispectral images.

For wideband imaging, the blurred object is spectrally filtered over P wide spectral bands $\omega_p(\lambda)$, $p = 1, \ldots, P$, where ω_p is a spectral windowing, generally given by the product of the filter transmission and the detector quantum efficiency.

Then the object is integrated within each band and sampled pixel-by-pixel on the detector matrix, thus forming a discrete spectral image. Therefore, spectroscopic information of the hyperspectral object is reduced to only P discrete values. This represents a severe degradation of the hyperspectral object since P is usually a small number.

Spatial integration at pixel (i, j) corresponds to multiplication by a square indicator function b_{d} on a two-dimensional sampling grid $\mathcal{G}_{\mathrm{d}} = \{iT_x, jT_y\}_{i;j=1}^{N_i;N_j}$ where N_i, N_j and T_x, T_y are the number of pixels and the spatial sampling steps along x and y, respectively. An additive term $n_{i,j}^p$ is added to the data to account for the detector noise. Finally, the complete equation of the imaging system model is given by

$$g_{i,j}^p = \iiint \iint \phi(x', y', \lambda)\, h(x - x', y - y', \lambda)\, \mathrm{d}x'\mathrm{d}y' \\ \omega_p(\lambda)\, b_{\mathrm{d}}(x - iT_x, y - jT_y)\, \mathrm{d}x\mathrm{d}y\mathrm{d}\lambda + n_{i,j}^p. \quad (1)$$

The model in (1) establishes a relation between the continuous hyperspectral object ϕ and the discrete images $g^p, p = 1, \ldots, P$ through the instrument response. It includes a spectral windowing and five integration for spatial convolution and spatio-spectral sampling.

B. Object Model

The object model is critical since the spectral information is seriously lacking in multispectral images. However it is important to note that the spatial structure of multispectral images depends on the spectral content of the object ϕ through the spectral dependence of the PSF h. In a previous work, [28] we proposed a spline model for spectral distribution and introduced strong smoothing prior in order to overcome the spectral subsampling, without the possibility of accurate spectral retrieval.

Instead, in order to overcome the lack of spectral information in the data, we propose here to model the hyperspectral object with a low rank approximation thanks to the linear mixing model of [15] considering a small number of components while keeping high spectral resolution. Therefore, the object is represented by a sum of M high-resolution known spectral components, $s^m(\lambda), m = 1, \ldots, M$, weighted by M mixing coefficients $f_{k,l}^m$ associated with each spatial position (k, l). Hence the object is decomposed, thanks to an indicator function b_{f}, on a two-dimensional grid $\mathcal{G}_{\mathrm{f}} = \left\{kT'_x, lT'_y\right\}_{k;l=1}^{N_k;N_l}$, where N_k, N_l, T'_x and T'_y are the number of samples and the sampling steps according to dimensions x and y, respectively. This yields

$$\phi(x, y, \lambda) = \sum_{m,k,l=1}^{M,N_k,N_l} f_{k,l}^m\, b_{\mathrm{f}}(x - kT'_x, y - lT'_y) s^m(\lambda). \quad (2)$$

Since only multispectral data are available, the spectral components $s^m(\lambda)$ are supposed known, because there is not enough data to estimate them jointly. However, they can be extracted

from previous hyperspectral measurements of similar objects with a principal component analysis (PCA) [29], non-negative matrix factorization (NMF) [30], or by elements of a given dictionary if, for instance, a material is known to be present, which is often the case. Nonetheless, we suppose that no specific meaning is attached to s_m except for their ability to represent the spectral variability that lives in a low rank space.

C. Forward Model

The forward model is obtained by substituting (2) in (1) yielding the linear model

$$g_{i,j}^{p} = \sum_{m=1}^{M} \sum_{k=1}^{N_k} \sum_{l=1}^{N_l} H_{i;j,k;l}^{p,m} \; f_{k,l}^{m} + n_{i,j}^{p} \tag{3}$$

with

$$H_{i;j,k;l}^{p,m} = \int \omega_p(\lambda) s^m(\lambda) \iint \Bigg[h(x', y', \lambda) \underset{x',y'}{*} b_{\mathrm{f}}(x' - kT_x', y' - lT_y') \Bigg] b_{\mathrm{d}}(x - iT_x, y - jT_y) \mathrm{d}x \mathrm{d}y \mathrm{d}\lambda, \tag{4}$$

where $\underset{x',y'}{*}$ indicates the 2D convolution. Combining integrals leads to

$$H_{ijkl}^{pm} = \int \omega_p(\lambda) s^m(\lambda) \bar{h}\left(kT_x' - iT_x, lT_x' - jT_x, \lambda\right) \mathrm{d}\lambda \tag{5}$$

with

$$\bar{h}\left(kT_x' - iT_x, lT_x' - jT_x, \lambda\right) = \iint h(x', y', \lambda) \underset{x',y'}{*} b_{\mathrm{f}}(x' - kT_x', y' - lT_y') \; b_{\mathrm{d}}(x - iT_x, y - jT_y) \, \mathrm{d}x \mathrm{d}y, \tag{6}$$

that is a spatial super-resolution model since this model computes the exact contribution of each pixel on each detector element. However, the computation of each H_{ijkl}^{pm} element can be computationally heavy. Nevertheless, taking $T_x' = T_x$ and $T_y' = T_y$ simplify Eq. (5) as

$$\begin{aligned} H_{ijkl}^{pm} &= \int \omega_p(\lambda) s^m(\lambda) \bar{h}\left((k-i)T_x, (l-j)T_x, \lambda\right) \mathrm{d}\lambda \\ &= H_{k-i,l-j}^{pm} \end{aligned} \tag{7}$$

and in that case, the linear model becomes a numerical spatial convolution.

By denoting $\boldsymbol{g}^p$ the vector of data in the band p, the forward model can be written as

$$\boldsymbol{g}^p = \sum_{m=1}^{M} \boldsymbol{H}^{p,m} \, \boldsymbol{f}^m + \boldsymbol{n}^p, \quad p = 1, 2, \ldots, P, \tag{8}$$

where the p-th image $\boldsymbol{g}^p \in \mathbb{R}^{N_i N_j}$ is a sum of M discrete spatial convolutions of mixing coefficients $\boldsymbol{f}^m \in \mathbb{R}^{N_k N_l}$, $m = 1, \ldots, M$, with convolution matrix $\boldsymbol{H}^{p,m} \in \mathbb{R}^{N_i N_j \times N_k N_l}$, plus an additive noise $\boldsymbol{n}^p \in \mathbb{R}^{N_i N_j}$. The convolution matrix $\boldsymbol{H}^{p,m}$ models the spatial impact of the spectral distribution m on the p-th image, describing the *spatial variation* of the response.

We propose to process data from all the bands together in order to reconstruct the total spectral information, instead of a band per band separated processing [25], [31]. This has the advantage of taking into account correlations between bands. Therefore, by concatenating all images in one vector, we obtain the following multi-observations forward model

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{n}, \tag{9}$$

where $\boldsymbol{g}^T = [\boldsymbol{g}^1, \ldots, \boldsymbol{g}^P]^T$, $\boldsymbol{f}^T = [\boldsymbol{f}^1, \ldots, \boldsymbol{f}^M]^T$, and $\boldsymbol{n}^T = [\boldsymbol{n}^1, \ldots, \boldsymbol{n}^P]^T$.

The full system observation matrix

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{H}^{1,1} & \cdots & \boldsymbol{H}^{1,M} \\ \vdots & \ddots & \vdots \\ \boldsymbol{H}^{P,1} & \cdots & \boldsymbol{H}^{P,M} \end{bmatrix}$$

is a non-square non-Tœplitz matrix with Tœplitz block components $\boldsymbol{H}^{p,m}$ representing the contribution of templates m to image p. For computational efficiency, the convolutions are done in the Fourier domain [32].

III. Reconstruction algorithm

A. Variational formulation

The problem of reconstructing f defined in Eq. (9) is ill-posed because of the convolution, leading to noise amplification. The common approach in this case is to add prior information about the solution, as in the regularized least square method [33]. Therefore, the solution $\widehat{\boldsymbol{f}}$ is obtained as a minimizer of an objective function $\mathcal{J}(\boldsymbol{f})$

$$\widehat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\operatorname{argmin}} \left\{ \mathcal{J}(\boldsymbol{f}) = \mathcal{Q}(\boldsymbol{f}, \boldsymbol{g}) + \mu \mathcal{R}(\boldsymbol{f}) \right\}, \tag{10}$$

where $\mathcal{Q}(\boldsymbol{f}, \boldsymbol{g})$ is a data fidelity term, $\mathcal{R}(\boldsymbol{f})$ a regularization term added to correct the ill-conditioning of the problem and $\mu \geq 0$ a regularization parameter to tune the trade-off between both these terms.

The noise is modeled by an identically independent Gaussian distribution for the sake of clarity, leading to $\mathcal{Q}(\boldsymbol{f}, \boldsymbol{g}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2$. The proposed algorithm can be extended to circulant spatially correlated Gaussian noise without restriction. Concerning the regularization term, many possibilities have been explored in the literature. For instance, one can consider Tikhonov regularization [34], total variation [35], wavelet-domain regularization [36], [37], or half-quadratic regularization [38], [39]. Here we are interested in the reconstruction of a spatially smooth hyperspectral object with sharp edges. The prior information is then introduced by penalizing the horizontal and vertical differences between neighboring pixels of each mixing coefficient. In that case, a multichannel regularization term is defined as

$$\mathcal{R}(\boldsymbol{f}) = \sum_{m,k,l=1}^{M,N_k,N_l} \varphi \underbrace{\left(\boldsymbol{f}_{k+1,l}^m - \boldsymbol{f}_{k,l}^m\right)}_{[\boldsymbol{D}_v \boldsymbol{f}^m]_{k,l}} + \varphi \underbrace{\left(\boldsymbol{f}_{k,l+1}^m - \boldsymbol{f}_{k,l}^m\right)}_{[\boldsymbol{D}_h \boldsymbol{f}^m]_{k,l}}, \tag{11}$$

where φ is the penalty function. $\boldsymbol{D}_h$ and $\boldsymbol{D}_v$ are first-order finite difference operators, with circularity conditions $f_{N_k+1,l}^m = f_{1,l}^m$ and $f_{k,N_l+1}^m = f_{k,1}^m$. A classical choice is the quadratic function $\varphi(x) = x^2$ that gives a differentiable

objective function and an explicit solution. However, spatial sharp edges are smoothed.

To overcome this limitation we propose to use a non-quadratic penalty function. Several methods are found in the literature such as methods based on partial differential equation [40], total variation (TV) (ℓ_1-norm of the gradient) [35], [41] half-quadratic regularization ($\ell_2\ell_1$-norm) [38], [39], [42] or majoration-minimization algorithm [43]. We use the method developed in [39] for several reasons. Firstly, the minimization of the objective function is done through alternating quadratic and separable minimization problems. Secondly, the quadratic solution is directly tractable thanks to the invertibility of the Hessian matrix, leading to a very fast algorithm. Thirdly, a variety of penalty functions can be used, such as the convex Huber function

$$\varphi(x) = \begin{cases} x^2, & \text{if } |x| < s,\ s \in \mathbb{R} \\ 2s|x| - s^2, & \text{otherwise,} \end{cases} \tag{12}$$

which is used in the rest of this work to prevents the cartoon-like effect given by TV penalization. The parameter $s > 0$ is a threshold parameter that defines the transition from a quadratic to a linear penalization. The half-quadratic regularization proposed by *Geman & Yang* in [39] consists of introducing $N_k \times N_l$ horizontal and vertical *auxiliary* variables, b, such that the penalty function φ is expressed as the minimum wrt. b of the sum of a quadratic function $(x-b)^2/2$ and an auxiliary function $\xi(b)$ (that depends on φ)

$$\varphi(x) = \inf_b \psi(b) = \inf_b \frac{1}{2}(x-b)^2 + \xi(b), \quad \forall x \in \mathbb{R}. \tag{13}$$

The construction relies on convex duality, with $\xi(b) = 2s|b|$ for Huber potential, as described in Appendix B. The practical role of these auxiliary variables b can be seen as shifting the quadratic function to a suitable position such as the cost at high gradient values is lower compared to the cost of the quadratic regularization. The auxiliary function relies on the convex duality and Legendre-Fenchel transform [38], [39], [42], [44]. Consequently, an augmented objective function $\mathcal{J}^*$ is defined such that

$$\inf_{\boldsymbol{b}_h, \boldsymbol{b}_v} \mathcal{J}^*(\boldsymbol{f}, \boldsymbol{b}_h, \boldsymbol{b}_v) = \mathcal{J}(\boldsymbol{f}). \tag{14}$$

Here $\boldsymbol{b}_h$ and $\boldsymbol{b}_v$ are vector representations of the stack of auxiliary variables along the horizontal and vertical directions. Therefore, the multichannel half-quadratic solution is obtained by minimizing the augmented objective function

$$\left(\widehat{\boldsymbol{f}}, \widehat{\boldsymbol{b}}_h, \widehat{\boldsymbol{b}}_v\right) = \underset{\boldsymbol{f}, \boldsymbol{b}_h, \boldsymbol{b}_v}{\operatorname{argmin}} \mathcal{J}^*(\boldsymbol{f}, \boldsymbol{b}_h, \boldsymbol{b}_v). \tag{15}$$

Since the criterion is globally convex [44], the computation of the joint minimizer of $\mathcal{J}^*(\boldsymbol{f}, \boldsymbol{b}_h, \boldsymbol{b}_v)$ with respect to $(\boldsymbol{f}, \boldsymbol{b}_h, \boldsymbol{b}_v)$ is achieved by iterating the following two-stage process until convergence

$$\begin{cases} \widehat{\boldsymbol{f}}^{(k)} = \underset{\boldsymbol{f}}{\operatorname{argmin}}\ \mathcal{J}^*\left(\boldsymbol{f}, \boldsymbol{b}_h^{(k-1)}, \boldsymbol{b}_v^{(k-1)}\right), & (16) \\ \widehat{\boldsymbol{b}}_h^{(k)}, \widehat{\boldsymbol{b}}_v^{(k)} = \underset{\boldsymbol{b}_h, \boldsymbol{b}_v}{\operatorname{argmin}}\ \mathcal{J}^*\left(\boldsymbol{f}^{(k)}, \boldsymbol{b}_h, \boldsymbol{b}_v\right). & (17) \end{cases}$$

B. Fast mixing coefficients $\widehat{\boldsymbol{f}}$ update

From (16) we have the quadratic criterion

$$\widehat{\boldsymbol{f}} = \underset{\boldsymbol{f}}{\operatorname{argmin}} \left\{ \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|_2^2 + \mu\left(\left\|\overline{\boldsymbol{D}}_h \boldsymbol{f} - \boldsymbol{b}_h\right\|_2^2 + \left\|\overline{\boldsymbol{D}}_v \boldsymbol{f} - \boldsymbol{b}_v\right\|_2^2 \right) \right\} \tag{18}$$

with $\overline{\boldsymbol{D}}_h = \operatorname{diag}\{\boldsymbol{D}_h, \ldots, \boldsymbol{D}_h\}$. and $\overline{\boldsymbol{D}}_v = \operatorname{diag}\{\boldsymbol{D}_v, \ldots, \boldsymbol{D}_v\}$, two diagonal-block matrices. The minimizer is explicit and is obtained by canceling the gradient. This yields

$$\widehat{\boldsymbol{f}} = \underbrace{\left(\boldsymbol{H}^T\boldsymbol{H} + \mu\left(\overline{\boldsymbol{D}}_h^T\overline{\boldsymbol{D}}_h + \overline{\boldsymbol{D}}_v^T\overline{\boldsymbol{D}}_v\right)\right)^{-1}}_{\boldsymbol{Q}} \underbrace{\left(\boldsymbol{H}^T\boldsymbol{g} + \mu\left(\overline{\boldsymbol{D}}_h^T\boldsymbol{b}_h + \overline{\boldsymbol{D}}_v^T\boldsymbol{b}_v\right)\right)}_{\boldsymbol{q}}, \tag{19}$$

where the Hessian matrix $\boldsymbol{Q} \in \mathbb{R}^{MN_kN_l \times MN_kN_l}$ is a block-circulant matrix, and $\boldsymbol{q} \in \mathbb{R}^{MN_kN_l}$ is a multichannel vector.

A common computational approach of $\widehat{\boldsymbol{f}}$ in Eq. (19) relies on solving the linear system $\boldsymbol{Q}\boldsymbol{f} = \boldsymbol{q}$ without requiring the inversion of $\boldsymbol{Q}$ thanks to iterative algorithms, e.g., conjugate gradient. On the contrary, in this paper the closed form solution is computed thanks to a fast and exact inversion of $\boldsymbol{Q}$ (see Appendix A for details).

C. Auxiliary variables $\widehat{\boldsymbol{b}}_h, \widehat{\boldsymbol{b}}_v$ update

From (17) we have

$$\widehat{\boldsymbol{b}}_h, \widehat{\boldsymbol{b}}_v = \underset{\boldsymbol{b}_h, \boldsymbol{b}_v}{\operatorname{argmin}} \sum_{m,k,l=1}^{M,N_k,N_l} \psi\left([\boldsymbol{b}_h]_{k,l}\right) + \psi\left([\boldsymbol{b}_v]_{k,l}\right) \tag{20}$$

where ψ is a convex and differentiable function defined in Appendix B. Moreover, the update equation for each auxiliary variable is explicit and separable with

$$\left[\widehat{\boldsymbol{b}}_*^m\right]_{k,l} = \underset{[\boldsymbol{b}_*^m]_{k,l}}{\operatorname{argmin}}\ \psi\left([\boldsymbol{b}_*^m]_{k,l}\right). \tag{21}$$

The computation of the minimizers in (21) is straightforward and it is detailed in Appendix B. Finally, we obtain

$$\widehat{\boldsymbol{b}}_* = \overline{\boldsymbol{D}}_*\boldsymbol{f} - \frac{1}{2}\varphi'\left(\overline{\boldsymbol{D}}_*\boldsymbol{f}\right) \tag{22}$$

where φ' is the first derivative of the Huber function given by

$$\varphi'(x) = \begin{cases} 2x, & \text{if } |x| < s, \\ 2s\,\operatorname{sign}(x), & \text{otherwise.} \end{cases}$$

The proposed Fast Joint Multiband Reconstruction (FJMR) algorithm is summarized in a pseudo-algorithm form in Algorithm 1.

IV. Experimental Results

In this section we present tests and comparisons of the proposed algorithm for the reconstruction of three hyperspectral objects having different spatial and spectral distributions. One

Algorithm 1 The FJMR algorithm

procedure FJMR($\boldsymbol{H}, \boldsymbol{D}_h, \boldsymbol{D}_v, \boldsymbol{g}, \mu$)
  Initialization: $\widehat{\boldsymbol{b}}_h = \widehat{\boldsymbol{b}}_v = \boldsymbol{0}$
  $\overline{\boldsymbol{F}} = \mathrm{diag}\{\boldsymbol{F}, \ldots, \boldsymbol{F}\}$
  $\overline{\boldsymbol{D}}_h = \mathrm{diag}\{\boldsymbol{D}_h, \ldots, \boldsymbol{D}_h\}$
  $\overline{\boldsymbol{D}}_v = \mathrm{diag}\{\boldsymbol{D}_v, \ldots, \boldsymbol{D}_v\}$
  Compute the Hessian matrix
  $\boldsymbol{Q} \leftarrow \boldsymbol{H}^T\boldsymbol{H} + \mu\left(\overline{\boldsymbol{D}}_h^T\overline{\boldsymbol{D}}_h + \overline{\boldsymbol{D}}_v^T\overline{\boldsymbol{D}}_v\right)$
  Compute the block diagonal matrix $\boldsymbol{\Upsilon}$
  $\boldsymbol{\Upsilon} \leftarrow \overline{\boldsymbol{F}}^\dagger \boldsymbol{Q}^{-1}\overline{\boldsymbol{F}}$
  while criterion is not reached **do**
    1 — Compute the solution (mixing coefficients)
    $\boldsymbol{q} \leftarrow \boldsymbol{H}^T\boldsymbol{g} + \mu\left(\overline{\boldsymbol{D}}_h^T\widehat{\boldsymbol{b}}_h + \overline{\boldsymbol{D}}_v^T\widehat{\boldsymbol{b}}_v\right)$
    $\widehat{\boldsymbol{f}} \leftarrow \boldsymbol{Q}^{-1}\boldsymbol{q} = \overline{\boldsymbol{F}}^\dagger\boldsymbol{\Upsilon}\overline{\boldsymbol{F}}\boldsymbol{q}$
    2 — Update the auxiliary variables in parallel
    $\widehat{\boldsymbol{b}}_h \leftarrow \overline{\boldsymbol{D}}_h\widehat{\boldsymbol{f}} - \frac{1}{2}\varphi'\left(\overline{\boldsymbol{D}}_h\widehat{\boldsymbol{f}}\right)$
    $\widehat{\boldsymbol{b}}_v \leftarrow \overline{\boldsymbol{D}}_v\widehat{\boldsymbol{f}} - \frac{1}{2}\varphi'\left(\overline{\boldsymbol{D}}_v\widehat{\boldsymbol{f}}\right)$
  end while
  return $\widehat{\boldsymbol{f}}$
end procedure

is the model of an astrophysical object and the other ones are synthetic objects. The multispectral imaging system we are considering is the MIRI imager [26] on board the JWST [27]. Note that we suppose the common case where the spectral components s^m are known and extracted from hyperspectral measurements or within a dictionary. We compare our results to the state-of-the-art TV deconvolution implemented with the primal dual Chambolle-Pock algorithm [25] that minimizes

$$\hat{\boldsymbol{f}}^p = \underset{\boldsymbol{f}^p}{\mathrm{argmin}} \ \|\boldsymbol{g}^p - \boldsymbol{H}^p\boldsymbol{f}^p\|_2^2 + \mu\|\nabla\boldsymbol{f}^p\|_1 \qquad (23)$$

for each band p, with $\nabla\boldsymbol{f}^p$ the spatial gradient of the image band p. Here $\boldsymbol{H}^p$ is the PSF integrated over the spectral band p assuming a flat spectrum. To complement the results we compare also our algorithm to the l_2 reconstruction defined as

$$\hat{\boldsymbol{f}}^p = \underset{\boldsymbol{f}^p}{\mathrm{argmin}} \ \|\boldsymbol{g}^p - \boldsymbol{H}^p\boldsymbol{f}^p\|_2^2 + \mu\|\nabla\boldsymbol{f}^p\|_2^2 \qquad (24)$$

that is the classical regularized least-squares method, with a conjugate-gradient as optimization algorithm.

The hyperparameter μ is hand tuned in order to minimize the ℓ_2 reconstruction error. The algorithms are coded using Python 2.7 and executed on a laptop machine with 16 GB of RAM and a processor Intel Core i7 CPU working at 2.50 GHz.

A. The MIRI Imager of the JWST

The optical system of the JWST is equipped with a 6.5 meters primary mirror composed of 18-hexagonal segments. The analytic expression of the monochromatic PSF at one wavelength can theoretically be obtained by computing the Fourier transform of the transmittance of the telescope aperture, in accordance with the diffraction theory [5]. However, it is also necessary to take into account misalignments of the 18 segments and the optical path differences. Therefore the monochromatic PSFs are computed with *WebbPSF* [45], [46], the official PSF simulator for the JWST developed by the Space Telescope Science Institute (STScI)[1]. No analytical formula of the PSF h is available, therefore all calculations to derive the convolution matrix $\boldsymbol{H}^{p,m}$ (Eq. (7)) are done numerically. Fig. 1 displays a few monochromatic PSFs computed at 6, 12, and 18 μm. We clearly see the spectral dependence of the PSF, *i.e.*, the longer the wavelength the wider the PSF as expected from the diffraction theory.

The multispectral images are integrated over $P = 9$ spectral widebands covering a spectral range from 5 to 30 μm as shown in Fig. 2. The MIRI imager provides images with a field of view (FOV) of 74×113 arcsecond^2 using a unique infrared detector with a pixel scale of 0.11 arcsecond, *i.e.*, a FOV per pixel of 0.11×0.11 arcsecond^2. Spatial variations of the PSF across the imager FOV have been measured on the flight model of the MIRI imager [6] with a width variation of the PSF across the field of view inferior to 5%. Thus the hypothesis of spatially invariant monochromatic PSFs is justified.

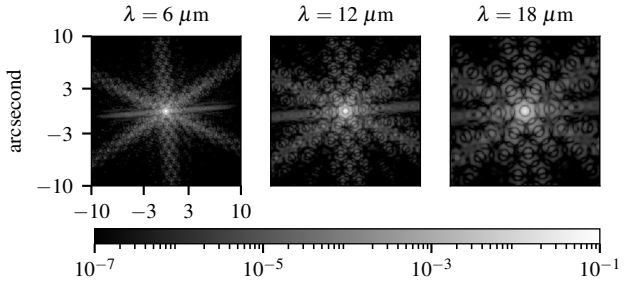


Fig. 1. Monochromatic PSF of the JWST/MIRI imager simulated at 6, 12, and 18 μm using *WebbPSF* [45] and displayed in logarithmic scale.

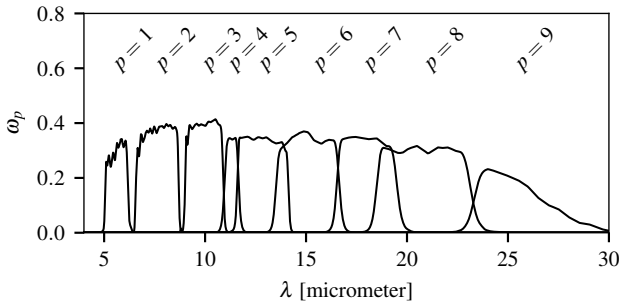


Fig. 2. Wide spectral bands ω_p of the JWST/MIRI Imager [47], from [48].

[1] http://www.stsci.edu/

B. Description of the hyperspectral objects

Our algorithm has been tested using three hyperspectral objects which cover a significant range of spatial and spectral structures expected for astrophysical observations.

The first one, denoted by Obj_1, is an astrophysical simulation of the *HorseHead nebula* [49], modeling a cloud of matter (dust and gas) illuminated by a bright star[2], with $N_k = N_l = 256$ spatial samples, and $N_\lambda = 1000$ spectral samples uniformly distributed from 1 to 30 μm. Therefore, Obj_1 is a full 2D+λ cube that is not constructed with the object model in Eq. (2), and original coefficients are not available. For the reconstruction, M spectral components s^m are extracted with a PCA [29]. For that object only $M = 3$ components (shown in Fig. 3) are sufficient to explain 99.99% of the variance of the spectra. Note that other extraction techniques could be used such as the NMF [30], or blind source separations [18], but this is not necessary in our case.

Two hyperspectral objects with more complex spatial and spectral distributions, Obj_2 and Obj_3, are synthesized. Fig. 4 and Fig. 5 display the spectral components and the original mixing coefficients used to synthesize Obj_2 and Obj_3, respectively. For both objects, the spectral components are taken from [18]. They have been computed from real data obtained by the spectrometer of the *Spitzer* Space Telescope [50] which covers the same spectral range as the MIRI imager. For the mixing coefficients of Obj_2 we take $M = 3$ rectangular patterns of different size with sharp edges, each associated to one of the three spectral components. For Obj_3 we take $M = 2$ mixing coefficients in order to create a complex high-frequency spatial structure with a smooth horizontal gradient.

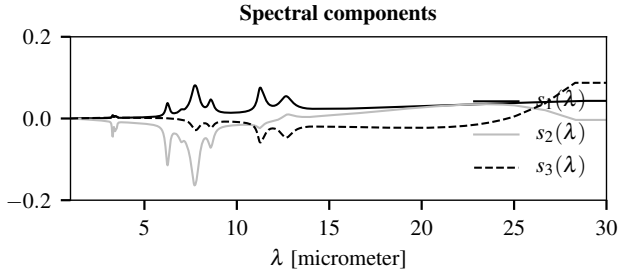


Fig. 3. Three spectral components extracted from Obj_1 with a PCA. The third component is negative due to the PCA formalism. The curves are in input sky unit.

C. Simulation of the multispectral data

The $P = 9$ images for Obj_1, Obj_2, and Obj_3 are simulated using the MIRI imager instrument model in (1), and *not the forward model with the mixing model in* (9). We degrade the images with an additive zero-mean, white, Gaussian noise of different levels of Signal-to-Noise Ratio (SNR), that is 5, 10, 20, 30, and 40 dB, defined as $\mathrm{SNR} = 10\log_{10}\left(\|\boldsymbol{g}\|_2^2 / N\sigma_n^2\right)$, where σ_n is the standard deviation of the noise, and N is the total number of pixels in $\boldsymbol{g}$.

[2]This 2D+λ object has been computed using state-of-the-art interstellar dust models and radiative transfer codes.

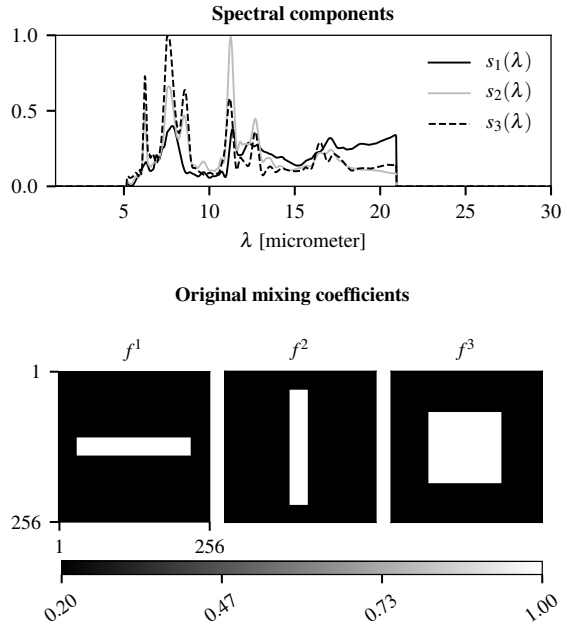


Fig. 4. Spectral components and mixing coefficients for Obj_2. The curves are in input sky unit.

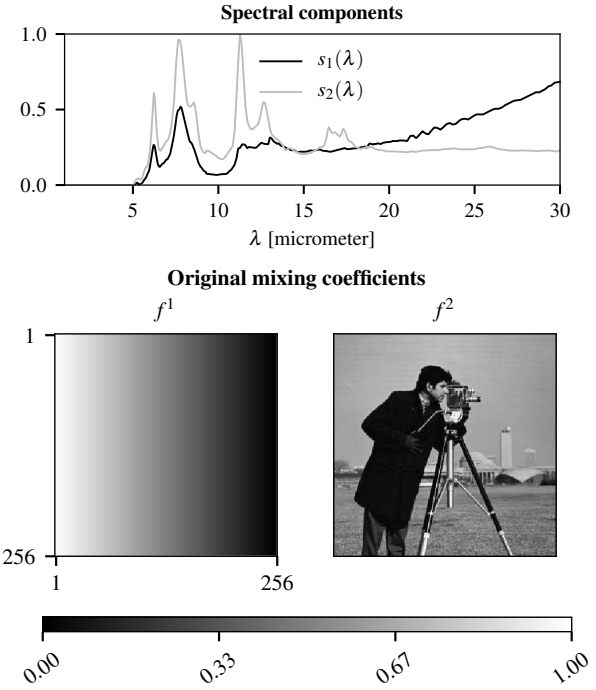


Fig. 5. Spectral components and mixing coefficients for Obj_3. The curves are in input sky unit.

The simulated multispectral images with SNR = 30 dB and $p = 1, 4, 7$ are displayed in the first row of Figs. 8, 9, and 10, for Obj_1, Obj_2 and Obj_3, respectively. As expected, the blur increases for increasing wavelengths due to the convolution by a wavelength-variable PSF. The images have different intensities due to the spectral distribution integrated within

TABLE I
RECONSTRUCTION ERRORS WITH 50 ITERATIONS.

Object	Error [%]	Runtime [s]	μ	s
$\widehat{\mathrm{Obj}}_1$	0.67	4.34	5.99×10^{-2}	1.17
$\widehat{\mathrm{Obj}}_2$	1.91	4.26	4.64×10^{-2}	0.03
$\widehat{\mathrm{Obj}}_3$	4.91	3.08	5.99×10^{-2}	0.01

the spectral bands and the width of each band. In any case these images are spatially degraded and contain poor spectral information.

D. Estimation results for $\widehat{\boldsymbol{f}}$, $\widehat{\boldsymbol{b}}_h$ and $\widehat{\boldsymbol{b}}_v$

For the reconstruction, the parameters μ and s are chosen in order to minimize the reconstruction error (in %) on sampled full 2D+λ cube (not coefficients)

$$\mathrm{Error}(\mu, s) = \|\phi_{\mathrm{orig}} - \phi_{\boldsymbol{f}}(\mu, s)\|_2 / \|\phi_{\mathrm{orig}}\|_2 \times 100, \quad (25)$$

with values reported in Table I. The spatial distributions of the estimated mixing coefficients $\widehat{\boldsymbol{f}}$ are shown in Fig. 6 for Obj_1 (top row), Obj_2 (middle row) and Obj_3 (bottom row). For Obj_1 we see that $\widehat{\boldsymbol{f}}^1$ has a higher intensity than $\widehat{\boldsymbol{f}}^2$ and $\widehat{\boldsymbol{f}}^3$, due to the domination of the first spectral component in the spectral distribution of Obj_1. For Obj_2 and Obj_3, the mixing coefficients appear properly unmixed and deconvolved.

The spatial distributions of the estimated auxiliary variables $(\widehat{\boldsymbol{b}}_h, \widehat{\boldsymbol{b}}_v)$, shown in Fig. 7 for Obj_2 mimic the contours of the mixing coefficients, as expected from the half-quadratic minimization.

E. Hyperspectral Reconstruction Results

The reconstructed objects are computed using the linear mixing model in (2) and using the estimated mixing coefficients presented above. In Table I we present the reconstruction errors for the three objects.

All reconstruction errors are below 5%. The smallest error (0.67%) is obtained for $\widehat{\mathrm{Obj}}_1$ which contains neither sharp edge nor complex spectral features. For $\widehat{\mathrm{Obj}}_2$ and $\widehat{\mathrm{Obj}}_3$, reconstruction errors are 1.91% and 4.91% respectively, because of their more complex spatial and spectral contents.

For a better illustration of our reconstruction results and a comparison to results obtained using the state-of-the-art TV deconvolution algorithm, we discuss separately the spatial distribution and the spectral distribution obtained for each object.

a) Spatial distribution: The spatial distributions are illustrated in Figs. 8, 9, and 10 by taking three monochromatic images at wavelengths $\lambda = 6, 12, 18\,\mu$m, belonging to the three spectral bands $p = 1, 4$ and 7, respectively. Figs. 9 and 10 are displayed with a spatial zoom to highlight details.

The proposed reconstructions show good performance. The dynamic range and the spatial distribution of the monochromatic images are well reconstructed with errors around 0.62% for Obj_1, 2.20% for Obj_2, and 5.75% for Obj_3. This illustrates the efficiency of the proposed algorithm for reconstruction at all wavelengths. The comparison between the first three lines of Figs. 9 and 10 illustrates the striking improvement of the spatial resolution. Our algorithm correctly recovers the sharp edges and small-scale gradients contained in the original objects. Thanks to the mixing model, the spatial resolution at all wavelengths is determined by the spatial resolution of the estimated mixing coefficients. Therefore, the reconstructed monochromatic images do not show any increasing blur with increasing wavelengths, unlike the input images and the images computed with the TV deconvolution. Moreover, TV deconvolution (which takes between 70 and 100 seconds for one image and 500 iterations) can only be done image by image and therefore cannot restore spatial details at small scales and at long wavelengths. In addition, the color bars show that the dynamic range is not properly restored with TV deconvolution. This is due to the integration over the wide spectral bands, and to the spectral dependence of the PSF within each wideband image which is neglected. The l_2 approach suffer the same problems that TV deconvolution with lesser quality results, as expected for this method.

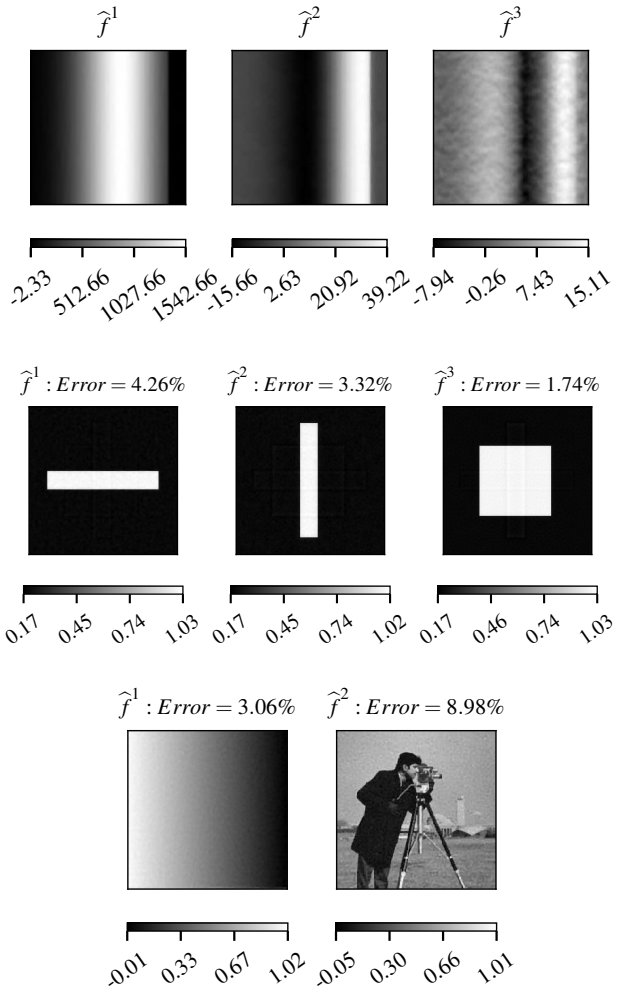


Fig. 6. Estimation results of the mixing coefficients associated to $\widehat{\mathrm{Obj}}_1$ (top), $\widehat{\mathrm{Obj}}_2$ (middle), and $\widehat{\mathrm{Obj}}_3$ (bottom). No error wrt. mixing coefficients is available for Obj_1 since the original ones do not exist.

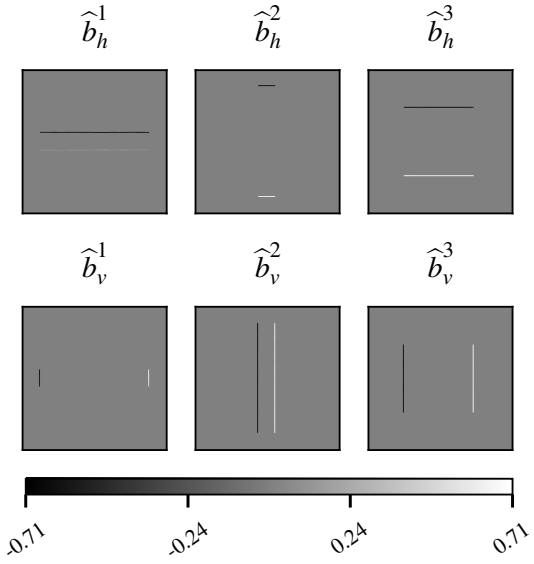


Fig. 7. Estimated auxiliary variables, horizontal and vertical, for $\widehat{\mathrm{Obj}_2}$.

Finally, the Fig. 12 shows the improvement of the spatial resolution at different wavelength values. Figures show the circular mean of the true Optical Transfer Function (OTF), that is the Fourier transform of the PSF, wrt. the "equivalent" OTF, which is the division between the true sky and the reconstructed sky in Fourier domain (for Obj_3 at 30 dB). The figure shows that in addition to the deconvolution effect and the high frequency restoration, the resolution is almost identical at all wavelength.

b) Spectral distribution: The spectral distributions for one spatial position are illustrated in Fig. 11, with a comparison between the spectrum of the original object and the reconstructed ones using the proposed and the TV algorithms. We see that the proposed algorithm produces spectra that correspond very closely to the original spectra at all wavelengths. This is due to:

- the linear mixing model with M known spectral components which allow disentangling the spectral information integrated within each wideband image,
- the spectral variant PSF to model the instrument response accurately, hence the observation matrix $\boldsymbol{H}$,
- the spectral correlations between images exploited in joint reconstruction.

In contrast, the TV deconvolution method produces a poor spectral reconstruction since the spectral information within each band is not modeled (and implicitly flat).

F. Influence of the Noise Level

The reconstruction errors wrt. the noise levels are shown in Fig. 13. As expected, the proposed algorithm is sensitive to the noise with a decrease of the reconstruction errors for an increasing SNR. The reconstruction of Obj_2 and Obj_3 is very sensitive to the noise since the noise corrupts the sharp edges in the multispectral data, and make their restoration more

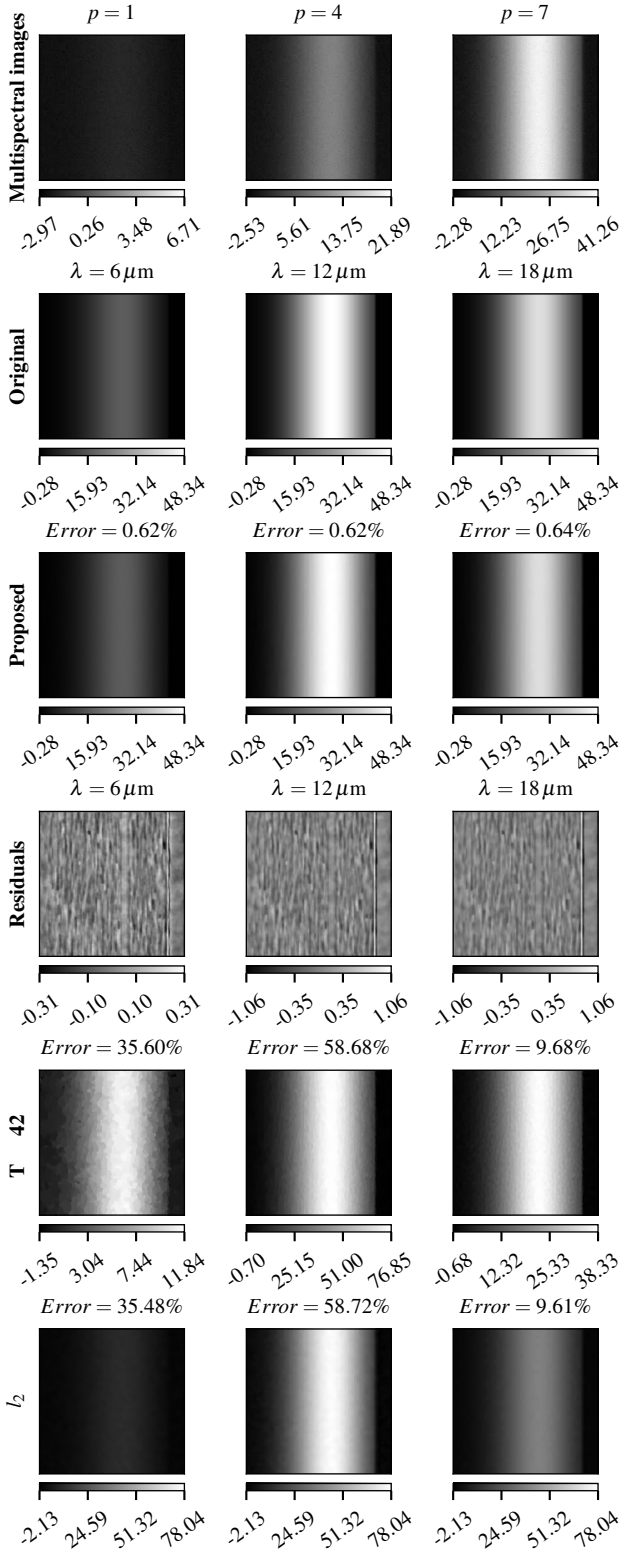


Fig. 8. Reconstruction for Obj_1 at 30 dB: [1st row] Simulated multispectral images data at bands $p = 1$, 4, and 7 with a $\mathrm{SNR} = 30$ dB. [2nd row] Original monochromatic images at 6, 12 and 18 μm, contained in bands $p = 1, 4, 7$. [3rd row] Reconstruction at 6, 12 and 18 μm. [4th row] Residuals. [5th row] Reconstruction with TV restoration.

difficult. This is less the case for Obj_1 which is dominated by a smoother spatial distribution.

V. Conclusion

We present an efficient method for restoring from multi-spectral images the spectral and spatial information which are

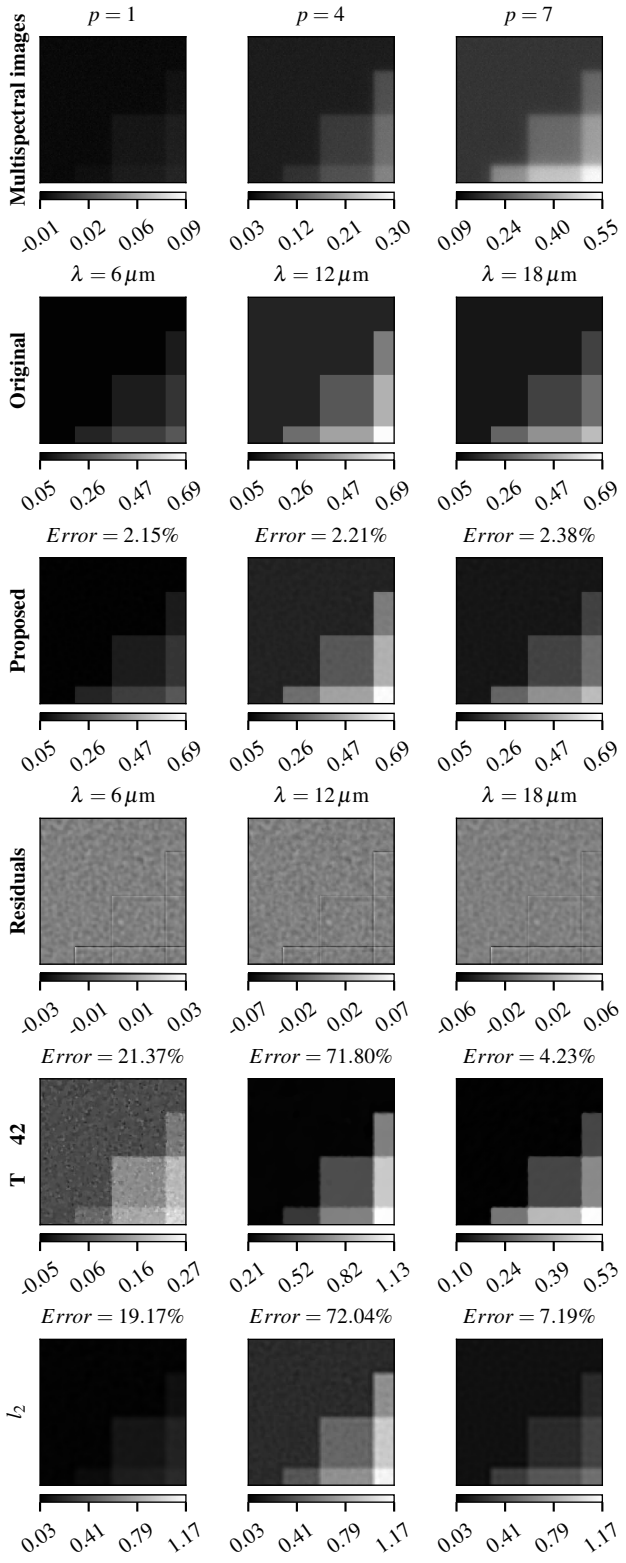


Fig. 9. Same as in Fig. 8 for Obj_2.

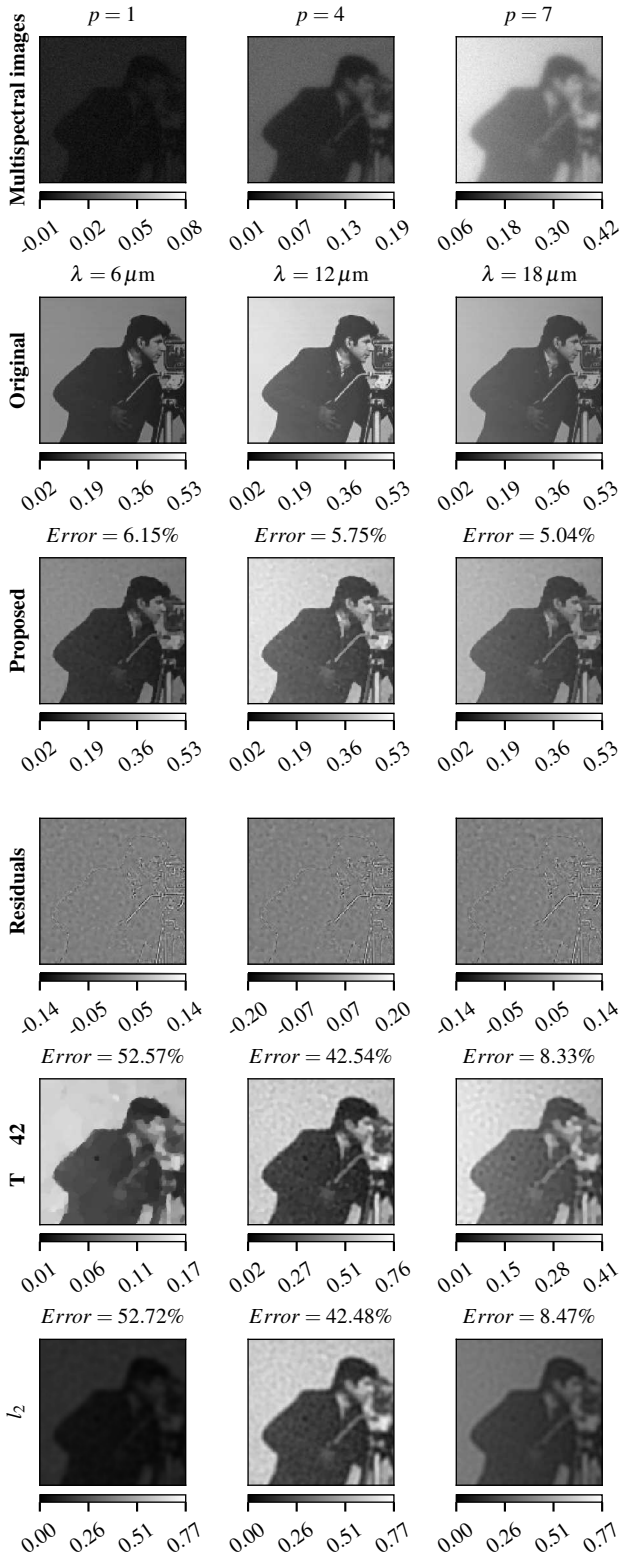


Fig. 10. Same as in Fig.8 for the Obj_3.

degraded in the acquisition process because of the spectral integration over wide bands, and the 2D convolution by the PSF (whose width linearly increases with wavelengths), which introduces a blur varying in space.

Our first contribution is a new data model that combines (1) a low rank subspace approximation of the fully resolved hyperspectral input object, and (2) a complete instrument model that takes into account the spectral variations of the PSF and the spectral response of the instrument. Then, a linear multi-observation forward model is derived, where data images appear as the sum of direct 2D convolutions of mixing coefficients allowing fast computation.

Our second contribution is a Fast Joint Multiband Reconstruction (FJMR) algorithm, an edge-preserving variational algorithm to process the full multispectral wideband images which are taken at a different spatial resolution. The proposed half-quadratic algorithm is iterative but we show that each sub-step is closed form and tractable with exact computation, especially for the quadratic step, even if the forward model is not stationary. Therefore, the algorithm is very fast with less than 4 seconds to process 9 images of size 256^2 on a standard laptop.

The performance of the reconstruction algorithm is validated for three hyperspectral objects, including an astrophysical simulated object, having different spatial and spectral distributions that cover most of the encountered cases, in the context of the MIRI infrared imager of the JWST. In all experiments, the relative errors are below 5 % for SNR $= 30$ dB. In addition, the reconstruction results using the proposed algorithm significantly outperform the TV deconvolution. The sharp edges and small-scale gradients which are contained in the original objects but blurred in the multispectral images are correctly recovered. Thanks to our model where spatial resolution is defined by mixing coefficient only, the spatial resolution is homogenized at all wavelengths. Moreover, our algorithm allows us to recover the spectroscopic information contained within each band but lost in the data because of the spectral integration over the bands.

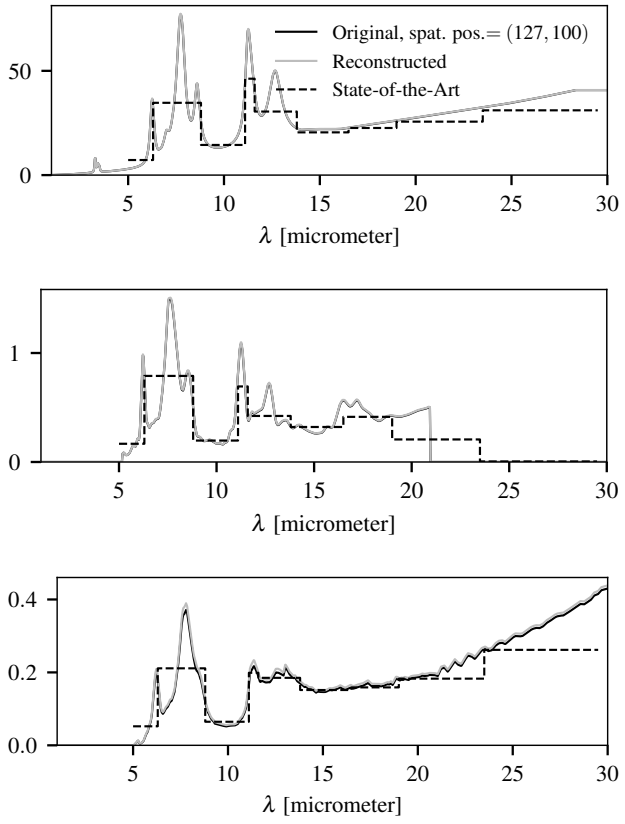


Fig. 11. Spectral distribution of the reconstruction results for one spatial position (127,100) for Obj_1 (top), Obj_2 (middle) and Obj_3 (bottom), with TV deconvolution as state-of-the-art method. The curves are in input sky unit.

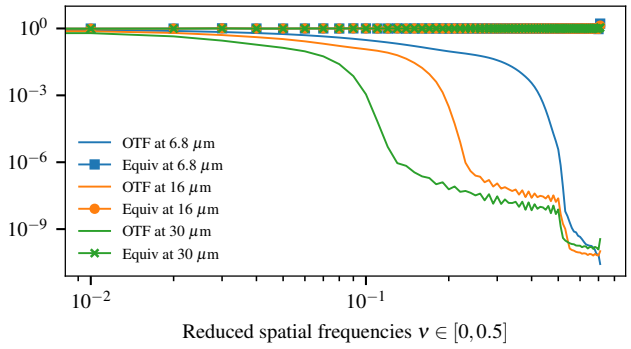


Fig. 12. Circular mean of the OTFs and the "equivalent" OTFs at three wavelengths, for Obj_3 at 30 dB. The "equivalent" OTF is the ratio between the true sky and the estimated sky at specified wavelengths. The three "equivalent" OTFs are superposed, indicating that the resolution is almost identical at these wavelengths.

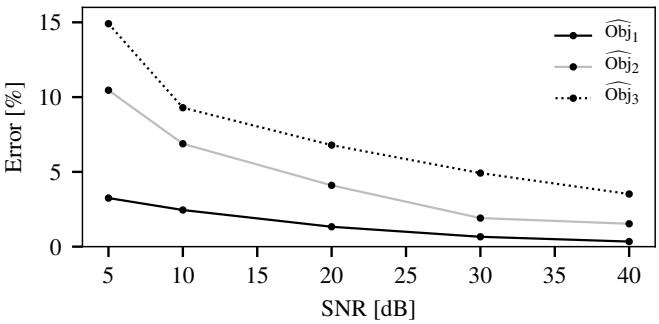


Fig. 13. Influence of the SNR on the reconstruction.

It is generally impossible to acquire spectroscopic data for large areas because the fields of view of spectrometers are generally very limited and much smaller than those of imagers. Moreover, the spectral coverage of observations becomes nowadays more and more extended. Therefore our algorithm, which has demonstrated its effectiveness to recover the spectroscopic information contained within wideband images and to reconstruct hyperspectral images with a homogenized spatial resolution at all wavelengths, appears a very promising tool that could be used for many scientific fields.

Many perspectives are possible. First the hyper parameters are fixed by hand, and their automatic estimation from data is a challenging question. Secondly, the spectral components

are not always available and data with more spectral details must be used. Finally, the method supposes that all images are acquired with the same spatial sampling step, and without shifts between the images.

APPENDIX A
COMPUTATION OF THE MULTICHANNEL QUADRATIC SOLUTION

In this section we present the computation of the multichannel quadratic solution given by

$$\underbrace{\begin{bmatrix}\widehat{\boldsymbol{f}}^1\\ \vdots \\ \widehat{\boldsymbol{f}}^M\end{bmatrix}}_{\widehat{\boldsymbol{f}}} = \underbrace{\begin{bmatrix}\boldsymbol{Q}^{1,1} & \dots & \boldsymbol{Q}^{1,M}\\ \vdots & \ddots & \vdots \\ \boldsymbol{Q}^{M,1} & \dots & \boldsymbol{Q}^{M,M}\end{bmatrix}^{-1}}_{\boldsymbol{Q}} \underbrace{\begin{bmatrix}\boldsymbol{q}^1\\ \vdots \\ \boldsymbol{q}^M\end{bmatrix}}_{\boldsymbol{q}},$$

where $\boldsymbol{Q} \in \mathbb{R}^{MN_kN_l \times MN_kN_l}$ is a non-circulant block circulant matrix and $\boldsymbol{q} \in \mathbb{R}^{MN_kN_l}$ is a multichannel vector. The computation of $\widehat{\boldsymbol{f}}$ relies on the inversion of the Hessian matrix $\boldsymbol{Q}$ inspired from [7], [51]. Inverting $\boldsymbol{Q}$ and storage of the inverse is possible by performing diagonalization of its circulant blocks $\boldsymbol{Q}^{i,j}$, resulting in a set of diagonal blocks $\boldsymbol{\Lambda}^{i,j}$ through the transfer equation

$$\boldsymbol{Q}^{i,j} = \boldsymbol{F}^\dagger \boldsymbol{\Lambda}^{i,j} \boldsymbol{F}, \quad i,j \in [1,\dots,M]^2, \tag{26}$$

where $\boldsymbol{F}$ and $\boldsymbol{F}^\dagger$ are the discrete Fourier matrix and its conjugate, respectively. This yields

$$\boldsymbol{Q} = \overline{\boldsymbol{F}}^\dagger \boldsymbol{\Lambda}_{\boldsymbol{Q}} \overline{\boldsymbol{F}}, \quad \text{and} \quad \boldsymbol{Q}^{-1} = \overline{\boldsymbol{F}}^\dagger \boldsymbol{\Lambda}_{\boldsymbol{Q}}^{-1} \overline{\boldsymbol{F}}, \tag{27}$$

with

$$\boldsymbol{\Lambda}_{\boldsymbol{Q}} = \begin{bmatrix}\boldsymbol{\Lambda}^{1,1} & \dots & \boldsymbol{\Lambda}^{1,M}\\ \vdots & \ddots & \vdots \\ \boldsymbol{\Lambda}^{M,1} & \dots & \boldsymbol{\Lambda}^{M,M}\end{bmatrix}, \overline{\boldsymbol{F}} = \begin{bmatrix}\boldsymbol{F} & & \\ & \ddots & \\ & & \boldsymbol{F}\end{bmatrix}. \tag{28}$$

The matrix $\boldsymbol{\Lambda}_{\boldsymbol{Q}}$ is a Non-Diagonal Block Diagonal (NDBD) matrix. Thanks to the permutation matrices $\boldsymbol{P}$, the NDBD matrix can be written as

$$\boldsymbol{\Lambda}_{\boldsymbol{Q}} = \boldsymbol{P}\boldsymbol{R}\boldsymbol{P}, \tag{29}$$

where $\boldsymbol{R} = \operatorname{diag}\left(\boldsymbol{R}^p\right), p = 1,\dots,N_kN_l$ is a matrix with full blocks on the diagonal. Each block $\boldsymbol{R}^p$ is an $M \times M$ full matrix, invertible in our case, with permutations that writes

$$\left(\boldsymbol{R}^p\right)_{i,j} = \left(\boldsymbol{\Lambda}^{i,j}\right)_{p,p} \tag{30}$$

for $i,j \in [1,\dots,M]^2$. Therefore, since the inversion of a block diagonal matrix is also a block diagonal matrix, the inverse of $\boldsymbol{\Lambda}_{\boldsymbol{Q}}$ can be written as

$$\boldsymbol{\Upsilon} := \boldsymbol{\Lambda}_{\boldsymbol{Q}}^{-1} = \boldsymbol{P}^T \boldsymbol{R}^{-1} \boldsymbol{P}^T \tag{31}$$

where $\boldsymbol{R}^{-1} = \operatorname{diag}\left(\left(\boldsymbol{R}^p\right)^{-1}\right)$, and $\boldsymbol{\Upsilon}$ is also a NDBD matrix, having a diagonal block $\boldsymbol{\Upsilon}^{i,j}$ given by

$$\left(\boldsymbol{\Upsilon}^{i,j}\right)_{p,p} = \left(\left(\boldsymbol{R}^p\right)^{-1}\right)_{i,j}. \tag{32}$$

In conclusion, $\boldsymbol{\Upsilon}$ is tractable and sparse, allowing pre computation and direct application in the iterative Algorithm 1. Thanks to such properties, the multichannel quadratic solution can be computed with the DFT and N_kN_l inversions of square matrices of size M. In the context of this work, N_kN_l are the number of pixels and M is the number of spectral components. In addition, each block $\boldsymbol{R}^p$ can be inverted in parallel. Finally, $\widehat{\boldsymbol{f}}$ is computed just by applying the matrix $\boldsymbol{\Upsilon}$ as

$$\widehat{\boldsymbol{f}} = \overline{\boldsymbol{F}}^\dagger \boldsymbol{\Lambda}_{\boldsymbol{Q}}^{-1} \mathring{\boldsymbol{q}} = \overline{\boldsymbol{F}}^\dagger \boldsymbol{\Upsilon} \overline{\boldsymbol{F}} \boldsymbol{q} \tag{33}$$

with $\mathring{\boldsymbol{q}} = \overline{\boldsymbol{F}}\boldsymbol{q}$.

APPENDIX B
UPDATE OF THE AUXILIARY VARIABLES

The update of all auxiliary variables is independent and can be calculated in parallel. The solution is given by the minimization

$$\widehat{\boldsymbol{b}}^m = \underset{\boldsymbol{b}^m}{\operatorname{argmin}} \underbrace{\frac{1}{2}\left([\boldsymbol{D}\,\boldsymbol{f}^m]_{k,l} - [\boldsymbol{b}^m]_{k,l}\right)^2 + \xi\left([\boldsymbol{b}^m]_{k,l}\right)}_{\psi\left([\boldsymbol{b}^m]_{k,l}\right)}$$

where ψ is a convex, differentiable, and separable function with respect to the minimizer. For the Huber potential φ, the auxiliary function is $\xi(b) = 2s|b|$. Since the criterion is convex differentiable, a sufficient condition is $\psi'\left(\left[\widehat{\boldsymbol{b}}^m\right]_{k,l}\right) = 0$, $\forall\ k \in [1,N_k]$ and $\forall\ l \in [1,N_l]$. The derivative function ψ' is computed by substituting the auxiliary function in ψ. This yields for the Huber potential

$$\psi'\left(\left[\widehat{\boldsymbol{b}}^m\right]_{k,l}\right) = \left(\left[\widehat{\boldsymbol{b}}^m\right]_{k,l} - [\boldsymbol{D}\boldsymbol{f}^m]_{k,l}\right) + s\,\operatorname{sign}\left(\left[\widehat{\boldsymbol{b}}^m\right]_{k,l}\right).$$

Thus, the obtained auxiliary variables are

$$\begin{aligned}\left[\widehat{\boldsymbol{b}}^m\right]_{k,l} &= \begin{cases}[\boldsymbol{D}\boldsymbol{f}^m]_{k,l} - [\boldsymbol{D}\boldsymbol{f}^m]_{k,l}, & \text{if } \left|[\boldsymbol{D}\boldsymbol{f}^m]_{k,l}\right| < s,\\ [\boldsymbol{D}\boldsymbol{f}^m]_{k,l} - s\,\operatorname{sign}\left([\boldsymbol{D}\boldsymbol{f}^m]_{k,l}\right), & \text{otherwise.}\end{cases}\\ &= [\boldsymbol{D}\boldsymbol{f}^m]_{k,l} - \frac{1}{2}\varphi'\left([\boldsymbol{D}\boldsymbol{f}^m]_{k,l}\right).\end{aligned} \tag{34}$$

Finally, we obtain

$$\widehat{\boldsymbol{b}} = \boldsymbol{D}\boldsymbol{f} - \frac{1}{2}\varphi'\left(\overline{\boldsymbol{D}}\boldsymbol{f}\right) \tag{35}$$

where φ' is the first derivative of Huber function

$$\varphi'(x) = \begin{cases}2x, & \text{if } |x| < s,\\ 2s\,\operatorname{sign}(x), & \text{otherwise.}\end{cases}$$

ACKNOWLEDGMENT

The authors would like to thank Nathalie Ysard for kindly providing the simulation of the hyperspectral object *Horse-Head*.

REFERENCES

[1] R. Lionello, J. A. Linker, and Z. Mikić, "Multispectral emission of the sun during the first whole sun month: Magnetohydrodynamic simulations," *The Astrophysical Journal*, vol. 690, no. 1, p. 902, 2008.

[2] D. A. Landgrebe, *Signal theory methods in multispectral remote sensing*. John Wiley & Sons, 2005, vol. 29.

[3] T. O. McBride, B. W. Pogue, S. Poplack, S. Soho, W. A. Wells, S. Jiang, K. D. Paulsen *et al.*, "Multispectral near-infrared tomography: a case study in compensating for water and lipid content in hemoglobin imaging of the breast," *Journal of biomedical optics*, vol. 7, no. 1, pp. 72–79, 2002.

[4] M. Dickinson, G. Bearman, S. Tille, R. Lansford, and S. Fraser, "Multispectral imaging and linear unmixing add a whole new dimension to laser scanning fluorescence microscopy," *Biotechniques*, vol. 31, no. 6, pp. 1272–1279, 2001.

[5] J. W. Goodman, *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[6] P. Guillard, T. Rodet, S. Ronayette, J. Amiaux, A. Abergel, V. Moreau, J. Augueres, A. Bensalem, T. Orduna, C. Nehmé *et al.*, "Optical performance of the jwst/miri flight model: characterization of the point spread function at high resolution," in *SPIE Astronomical Telescopes+ Instrumentation*. International Society for Optics and Photonics, 2010, pp. 77 310J–77 310J.

[7] N. P. Galatsanos and R. T. Chin, "Digital restoration of multichannel images," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 415–421, 1989.

[8] L. Denis, É. Thiébaut, and F. Soulez, "Fast model of space-variant blurring and its application to deconvolution in astronomy," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 2817–2820.

[9] É. Thiébaut, L. Dénis, F. Soulez, and R. Mourya, "Spatially variant psf modeling and image deblurring," in *SPIE Astronomical Telescopes+ Instrumentation*. International Society for Optics and Photonics, 2016, pp. 99 097N–99 097N.

[10] G. Aniano, B. Draine, K. Gordon, and K. Sandstrom, "Common-resolution convolution kernels for space-and ground-based telescopes," *Publications of the Astronomical Society of the Pacific*, vol. 123, no. 908, p. 1218, 2011.

[11] A. Boucaud, M. Bocchio, A. Abergel, F. Orieux, H. Dole, and M. A. Hadj-Youcef, "Convolution kernels for multi-wavelength imaging," *Astronomy & Astrophysics*, vol. 596, p. A63, 2016.

[12] N. Keshava and J. F. Mustard, "Spectral unmixing," *IEEE signal processing magazine*, vol. 19, no. 1, pp. 44–57, 2002.

[13] A. Plaza, P. Martínez, R. Pérez, and J. Plaza, "A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data," *IEEE transactions on geoscience and remote sensing*, vol. 42, no. 3, pp. 650–663, 2004.

[14] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 5, no. 2, pp. 354–379, 2012.

[15] J. B. Adams, M. O. Smith, and P. E. Johnson, "Spectral mixture modeling: A new analysis of rock and soil types at the viking lander 1 site," *Journal of Geophysical Research: Solid Earth*, vol. 91, no. B8, pp. 8098–8112, 1986.

[16] J. Settle and N. Drake, "Linear mixing and the estimation of ground cover proportions," *International Journal of Remote Sensing*, vol. 14, no. 6, pp. 1159–1177, 1993.

[17] V. Haertel and Y. E. Shimabukuro, "Spectral linear mixing model in low spatial resolution image data," in *Geoscience and Remote Sensing Symposium, 2004. IGARSS'04. Proceedings. 2004 IEEE International*, vol. 4. IEEE, 2004, pp. 2546–2549.

[18] O. Berne, C. Joblin, Y. Deville, J. Smith, M. Rapacioli, J. Bernard, J. Thomas, W. Reach, and A. Abergel, "Analysis of the emission of very small dust particles from spitzer spectro-imagery data using blind signal separation methods," *Astronomy & Astrophysics*, vol. 469, no. 2, pp. 575–586, 2007.

[19] N. Dobigeon, S. Moussaoui, M. Coulon, J.-Y. Tourneret, and A. O. Hero, "Joint bayesian endmember extraction and linear unmixing for hyperspectral imagery," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4355–4368, 2009.

[20] Y. Tarabalka, J. A. Benediktsson, and J. Chanussot, "Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 8, pp. 2973–2987, 2009.

[21] Z. Guo, T. Wittman, and S. Osher, "L1 unmixing and its application to hyperspectral image enhancement," in *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XV*, vol. 7334. International Society for Optics and Photonics, 2009, p. 73341M.

[22] L. Loncan, L. B. Almeida, J. M. Bioucas-Dias, X. Briottet, J. Chanussot, N. Dobigeon, S. Fabre, W. Liao, G. A. Licciardi, M. Simoes *et al.*, "Hyperspectral pansharpening: A review," *arXiv preprint arXiv:1504.04531*, 2015.

[23] F. Orieux, J.-F. Giovannelli, and T. Rodet, "Bayesian estimation of regularization and point spread function parameters for wiener–hunt deconvolution," *JOSA A*, vol. 27, no. 7, pp. 1593–1607, 2010.

[24] J. Yang, Y. Zhang, and W. Yin, "An efficient tvl1 algorithm for deblurring multichannel images corrupted by impulsive noise," *SIAM Journal on Scientific Computing*, vol. 31, no. 4, pp. 2842–2865, 2009.

[25] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, pp. 120–145, 2011.

[26] P. Bouchet, M. García-Marín, P.-O. Lagage, J. Amiaux, J.-L. Auguéres, E. Bauwens, J. Blommaert, C. Chen, Ö. Detre, D. Dicken *et al.*, "The Mid-Infrared Instrument for the James Webb Space Telescope, III: MIRIM, The MIRI Imager," *Publications of the Astronomical Society of the Pacific*, vol. 127, no. 953, p. 612, 2015.

[27] J. P. Gardner, J. C. Mather, M. Clampin, R. Doyon, M. A. Greenhouse, H. B. Hammel, J. B. Hutchings, P. Jakobsen, S. J. Lilly, K. S. Long *et al.*, "The james webb space telescope," *Space Science Reviews*, vol. 123, no. 4, pp. 485–606, 2006.

[28] M. A. Hadj-Youcef, F. Orieux, A. Fraysse, and A. Abergel, "Spatio-spectral multichannel reconstruction from few low-resolution multispectral data," in *2018 26th European Signal Processing Conference (EUSIPCO)*, Sep 2018.

[29] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*. Springer, 1986, pp. 115–128.

[30] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[31] S. Bongard, F. Soulez, É. Thiébaut, and É. Pecontal, "3d deconvolution of hyper-spectral astronomical data," *Monthly Notices of the Royal Astronomical Society*, vol. 418, no. 1, pp. 258–270, 2011.

[32] B. Hunt, "A matrix theory proof of the discrete convolution theorem," *IEEE Transactions on Audio and Electroacoustics*, vol. 19, no. 4, pp. 285–288, 1971.

[33] G. Demoment, "Image reconstruction and restoration: Overview of common estimation structures and problems," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 12, pp. 2024–2036, 1989.

[34] A. N. Tikhonov, A. Goncharsky, V. Stepanov, and A. G. Yagola, *Numerical methods for the solution of ill-posed problems*. Springer Science & Business Media, 2013, vol. 328.

[35] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[36] I. Daubechies and G. Teschke, "Wavelet based image decomposition by variational functionals," in *Proc. SPIE Vol*, vol. 5266, 2004, pp. 94–105.

[37] M. A. Figueiredo and R. D. Nowak, "An em algorithm for wavelet-based image restoration," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 906–916, 2003.

[38] D. Geman and G. Reynolds, "Constrained restoration and the recovery of discontinuities," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 14, no. 3, pp. 367–383, 1992.

[39] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Transactions on Image Processing*, vol. 4, no. 7, pp. 932–946, 1995.

[40] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[41] A. Chambolle, "An algorithm for total variation minimization and applications," *Journal of Mathematical imaging and vision*, vol. 20, no. 1-2, pp. 89–97, 2004.

[42] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Transactions on image processing*, vol. 6, no. 2, pp. 298–311, 1997.

[43] E. Chouzenoux, A. Jezierska, J.-C. Pesquet, and H. Talbot, "A majorize-minimize subspace approach for ℓ_2-ℓ_0 image regularization," *SIAM Journal on Imaging Sciences*, vol. 6, no. 1, pp. 563–591, 2013.

[44] J. Idier, "Convex half-quadratic criteria and interacting auxiliary variables for image restoration," *IEEE Transactions on Image Processing*, vol. 10, no. 7, pp. 1001–1009, 2001.

[45] M. D. Perrin, R. Soummer, E. M. Elliott, M. D. Lallo, and A. Sivaramakrishnan, "Simulating point spread functions for the james webb space telescope with webbpsf," in *Space Telescopes and Instrumentation 2012: Optical, Infrared, and Millimeter Wave*, vol. 8442. International Society for Optics and Photonics, 2012, p. 84423D.

[46] M. D. Perrin, A. Sivaramakrishnan, C.-P. Lajoie, E. Elliott, L. Pueyo, S. Ravindranath, and L. Albert, "Updated point spread function simulations for jwst with webbpsf," in *Space Telescopes and Instrumentation 2014: Optical, Infrared, and Millimeter Wave*, vol. 9143. International Society for Optics and Photonics, 2014, p. 91433X.
[47] A. Glasse, G. Rieke, E. Bauwens, M. García-Marín, M. Ressler, S. Rost, T. V. Tikkanen, B. Vandenbussche, and G. Wright, "The Mid-Infrared Instrument for the James Webb Space Telescope, IX: Predicted Sensitivity," *Publications of the Astronomical Society of the Pacific*, vol. 127, no. 953, p. 686, 2015.
[48] U. of Arizona, "Table of integration window for each filter," 2018, accessed 2018-31-10. [Online]. Available: http://ircamera.as.arizona.edu/MIRI/ImPCE_TN-00072-ATC-Iss2.xlsx
[49] A. Abergel, D. Teyssier, J. Bernard, F. Boulanger, A. Coulais, D. Fosse, E. Falgarone, M. Gerin, M. Perault, J.-L. Puget *et al.*, "Isocam and molecular observations of the edge of the horsehead nebula," *Astronomy & Astrophysics*, vol. 410, no. 2, pp. 577–585, 2003.
[50] J. R. Houck, T. L. Roellig, van Cleve, and *et al.*, "The Infrared Spectrograph (IRS) on the Spitzer Space Telescope," *Astrophysical Journal Supplement Series*, vol. 154, no. 1, pp. 18–24, Sep 2004.
[51] N. P. Galatsanos, M. N. Wernick, A. K. Katsaggelos, and R. Molina, "Multichannel image recovery," *Handbook of Image and Video Processing*, vol. 12, pp. 155–168, 2000.

M. Amine Hadj-Youcef was born in Blida, Algeria, in 1989. He holds a MSc degree and a Ph.D from the University of Bordeaux, France, and the University of Paris-Saclay, France, in 2015 and 2018, respectively.

Since then, he is working in Research and development for industry, in various fields such as artificial intelligence based computer vision, machine learning, and data science. He recently joined the R&D department at TAG Heuer Connected, Paris, France.

François Orieux is Assistant Professor at the Université Paris-Sud and a researcher in the Laboratoire des Signaux et Systèmes, Groupe Problèmes Inverses (Université Paris-Saclay, CNRS, CentraleSupélec), France. He is also an associate researcher with the Institut d'Astrophysique Spatiale (Univ. Paris-Saclay, CNRS).

He received his Ph.D. degree in signal processing at Université Paris-Sud, Orsay, France. His research focuses on Bayesian methodological approaches for ill-posed inverse problems resolution with examples of applications in astrophysics or biological microscopy.

Alain Abergel is professor of Physics and Astrophysics at the Paris-Saclay University. He is astrophysicist of the interstellar medium at the Institut d'Astrophysique Spatiale (Paris-Saclay University, CNRS), and is involved in the scientific exploitation of several space missions for astrophysics (ISO, Spitzer, Herschel, Planck, JWST, ...).

Aurélia Fraysse was born in Creteil, France, in 1978. She graduated from University Paris 12 in 2001. She received her Ph. D. degree in mathematics at University Paris 12 in 2005.

In 2006, she was a research assistant at Ecole Nationale Supérieure des Télécommunications (Telecom Paris). She is presently assistant professor at IUT Cachan and researcher with the Laboratoire des Signaux et Systèmes (CNRS-CentraleSupelec-Univ. Paris-Sud).

Journal of Real-Time Image Processing manuscript No.
(will be inserted by the editor)

Real-Time Optical Flow Processing on Embedded GPU: an Hardware-Aware Algorithm to Implementation Strategy

Mickaël Seznec · Nicolas Gac · François Orieux · Alvin Sashala Naik

Received: date / Accepted: date

Abstract Determining the optical flow of a video is a compute-intensive task essential for computer vision. For achieving this processing in real-time, the whole algorithm deployment chain must be thought of for efficiency first. The development is usually divided into two parts: first, designing an algorithm that meets precision constraints, then, implementing and optimizing its execution on the targeted platform. We argue that unifying those operations enhances performance on the embedded processor.

This paper is based on an industrial use case of computer vision. The objective is to determine dense optical flow in real-time on an embedded GPU platform: the Nvidia AGX Xavier. The CLG (Combined Local-Global) optical flow method, initially chosen, is analyzed to understand the convergence speed of its underlying optimization problem. The *Jacobi* solver is selected for implementation because of its parallel nature. The whole multi-level processing is then ported to the GPU, using several specific optimization strategies. In particular, we analyze the impact of fusing the solver's iterations with the roofline model.

As a result, with a 30W power budget, our implementation runs at 60FPS, on 640×512 images, with a four-level processing. Hopefully, this example should provide feedback on the issues that arise when trying to port a method to a parallel platform and serve for further implementations of computer vision algorithms on specialized hardware.

Mickaël Seznec · Alvin Sashala Naik
Thales Research and Technology. Palaiseau, France

Mickaël Seznec · Nicolas Gac · François Orieux
Laboratoire des Signaux et Systèmes, Université Paris-Saclay, CNRS, CentraleSupélec. Gif-Sur-Yvette, France

Keywords Algorithm design, Optical Flow, GPU Optimization, Linear Solvers, Image Processing

1 Introduction

Computer vision has become an essential aspect of widely adopted electronic devices in various fields: medicine [8], unmanned flight [15], or autonomous driving [6], for instance. The constant progress of these applications is driven by more sophisticated algorithms and more efficient hardware architectures. As both of these fields continue to progress, the difficulty of finding an optimal match between the two increases.

On the one hand, the algorithm design space of image processing methods is broad. New techniques are constantly developed that often depend on hyper-

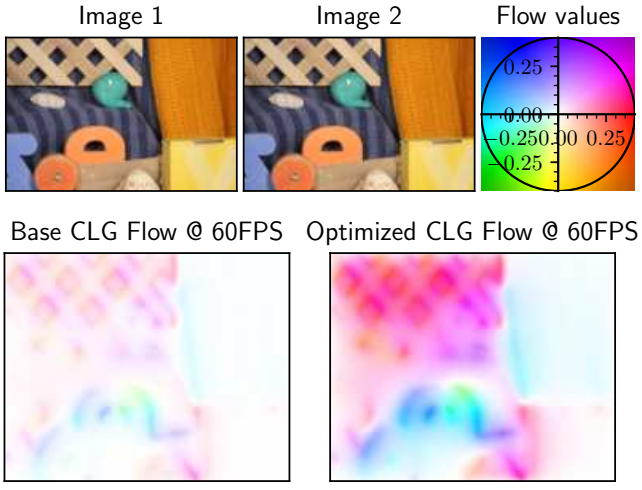


Fig. 1: For the same framerate on Jetson Xavier, our GPU-optimized multi-scale CLG Optical Flow converges further than the initial implementation.

parameters to control a trade-off between the speed and accuracy of the results. On the other hand, modern hardware architectures such as GPUs (Graphics Processing Units), FPGAs (Field-Programmable Gate Array), or SIMD (Single-Instruction Multiple-Data) processors have successfully improved the execution of vision algorithms. The increasing complexity in both of these domains calls for expertise that keeps being more and more specific. It is then challenging to combine these two skills to find an optimal match between the algorithm and the target.

In this article, we focus on the optical flow problem. The goal is, given two successive frames of a video, to find a per-pixel displacement vector. First numerical methods to solve it have been found by Horn and Schunk in the 1980s [12] and numerous refinements have been developed since [9,3]. For our analysis, we consider the CLG (Combined Local-Global) method [4], because it serves in an industrial application we use.

Our analysis then serves two goals. First, finding the impact of the solver choice and the values of hyper-parameters on the speed of the CLG method. This initial study gives rise to an initial implementation on the NVIDIA Jetson AGX Xavier, an embedded GPU SOC (System On Chip). Doing optical flow processing on this kind of lightweight device is crucial for power constrained systems, like drones [15]. The other goal is finding efficient optimization procedures for the CLG algorithm to achieve maximum performance. Overall, the study aims at finding algorithm-implementation synergies through the perspective of optical flow processing.

The main novelties brought by this article are listed below.

- It extends previous work [20] on the influence on speed and accuracy of the hyper-parameters of the CLG optical flow. Notably, the spectral radiuses of splitting solvers are provided, and new performance results on the Xavier GPU are presented.
- It introduces a complete algorithm implementation on the Jetson AGX Xavier, optimized in-depth with diverse techniques: buffer re-utilization, solver iteration fusion, and kernel launches batching.
- It analyses the impact of the multi-scale scheme on the performance of our implementation.

The rest of this article is structured as follows: section 2 outlines related work on optical flow processing for real-time systems and optimization strategies for parallel systems. Section 3 introduces mathematical notations for optical flow and analyzes solvers and hyper-parameters on the convergence speed. Section 4 deals with the implementation optimizations on GPU and focuses on arithmetic intensity to explain achieved performance. Section 5 concludes this paper and gives direction for further work.

2 Related Work

Optical flow has received a lot of attention since pioneering numerical methods introduced by Horn and Schunk [12] and Lucas & Kanade [14]. From there, many refinements have been incorporated on top of these frameworks. Review papers [2,22] explore comprehensively the different strategies used for computing optical flow.

In this article, we focus on a differential method, a family introduced by Horn & Schunk. It consists of minimizing a penalization function usually composed of two types of terms: model attach and regularization. On top of the original penalization function found in [12], Farnebäck *et al.* replace the linear interpolation with a quadratic one for better accuracy [9]. Brox *et al.* add a gradient conservation term [3] while Zach *et al.* use a L1-norm penalization instead of a quadratic one [26] to obtain better-defined object boundaries. The selected algorithm for our study is the CLG (Combined Local-Global) method, as defined by Bruhn *et al.* [4]. This method adds a neighboring condition to the model attach term, similar to the one found in [14]. This unifying model is less sensitive to noise, as the local information is averaged over multiple pixels. We consider this algorithm fixed in our work, so our results match the original implementation's accuracy. A detailed comparison with other related methods is made in [22].

Efficient implementation has always been key to an attractive optical flow method. For CLG, a CPU implementation has been described in [13] and Moussu [16] detailed its GPU counterpart. With respect to this previous work, our article details how to choose the right solver and hyper-parameters of CLG for fast convergence. It is completed by GPU optimizations, especially for the Jacobi solver.

There is plenty of literature about GPU optimization for linear algebra. Kernel fusion is a frequent technique manually applied to a sparse CG (Conjugate Gradient) solver in [1], or BCG (Biconjugate-CG) in [23]. Filipovic *et al.* propose a source-to-source compiler to perform fusion at the compilation stage [10]. Regarding the Jacobi solver specifically, Aslam *et al.* have benchmarked many computations and synchronization techniques. We differ from this work by not relying on sparse matrices to implement the Jacobi solver but by directly implementing the operators defined by those matrices. An implemenation on massively parallel processors of this solver is studied in [18] and also analyse the trade-off between memory locality and overhead computa-

tions. In [17], Nguyen *et al.* compare several GPU solvers for fastest convergence and comes to similar conclusions as ours: more iterations on simpler solvers are more efficient on GPUs.

To guide our optimization strategy, we rely on the roofline model, as introduced by Williams in [24]. It analyses a program's run based on its arithmetic intensity and the performance limits of the hardware in terms of memory and compute throughput. It is a general and powerful tool to find bottlenecks in an application that has already been used for GPUs [7].

3 Method-level approach

In this section, we examine the CLG algorithm from a mathematical perspective. This first analysis serves our optimization method by highlighting the degrees of freedom allowed by our application. After giving some mathematical context, we then explore the implications of the solver choice and the tuning of hyper-parameters.

3.1 Modeling optical flow

A category of optical flow algorithms provides a result by finding the solution to an optimization problem. In this section, we introduce the mathematical notations associated with this optimization problem.

The variables are named using the following convention: the lowercase $a \in \mathbb{R}$ is a coefficient, the bold lowercase $\boldsymbol{a} \in \mathbb{R}^n$ is a vector, the uppercase $A \in \mathbb{R}^{n\times m}$ is a matrix. The over-line symbol $\bar{a} \in \mathbb{R}^{I_h \times I_w}$ represents the field a over a two-dimensional image. Likewise, $\bar{\boldsymbol{a}}$ is a vector field, $\bar{A}$ is a matrix field. Finally, the double bar notation introduces *flattened* fields: $\bar{\bar{a}}$ is a two-dimensional field represented as a vector with a row-major convention.

In a sequence of images at time t, the optical flow at (x, y) is noted $\boldsymbol{w}_{x,y,t} = (u_{x,y,t}, v_{x,y,t}, 1)^T$. It has three components: $u_{x,y,t}$ and $v_{x,y,t}$ are the displacement in the x and y axis, respectively, with the time displacement equals 1. We also introduce $\boldsymbol{w}^* = (u_{x,y,t}, v_{x,y,t})^T$, for ease of notation.

Finally, $f_{x,y,t}$ represents the pixel intensity of the frame at time t, at coordinates x, y. Images are grayscale, so $f_{x,y,t}$ is a scalar. Later in the article, and for the sake of brevity, we may omit the x, y, t indices, so $\boldsymbol{w}_{x,y,t}$ becomes $\boldsymbol{w}$, for example.

With the variables now set, we present an energy definition that serves as a framework for many variational methods

$$E(\bar{\bar{\boldsymbol{w}}}) = \int_\Omega D(\boldsymbol{w}, \bar{f}) + R(\boldsymbol{w}) \,\mathrm{d}x\,\mathrm{d}y \tag{1}$$

where D is the data-fitting term while R plays the role of regularization, and Ω represents the 2D image domain.

For example, in [12], Horn & Schunck set D and R to

$$D_{\mathrm{HS}}(\boldsymbol{w}, \bar{f}) := \boldsymbol{w}^T J_0 \boldsymbol{w}, \text{ and} \tag{2}$$

$$R_{\mathrm{HS}}(\boldsymbol{w}) := \alpha\Big(\|\nabla_{x,y} u_{x,y}\|^2 + \|\nabla_{x,y} v_{x,y}\|^2\Big), \tag{3}$$

where $\nabla_{x,y}$ is a two-dimensional spatial gradient and $\alpha \in \mathbb{R}^+$ is the trade-off between data fitting and regularization penalization. $\bar{J}_0$ is a matrix field and corresponds to a quadratic penalization of the image intensity conservation, eq. (4), with a linear approximation, eq. (5) of the image's values

$$\|\bar{f}(x+u, y+v, t+1) - \bar{f}(x, y, t)\|^2 \tag{4}$$

$$\approx \|\bar{f}(x, y, t) + \nabla\bar{f}(x, y, t)^T \boldsymbol{w} - \bar{f}(x, y, t)\|^2 \tag{5}$$

$$= \|\nabla\bar{f}(x, y, t)^T \boldsymbol{w}\|^2 \tag{6}$$

$$= \boldsymbol{w}^T \nabla\bar{f}(x, y, t) \nabla\bar{f}(x, y, t)^T \boldsymbol{w} \tag{7}$$

$$= \boldsymbol{w}^T J_0 \boldsymbol{w}. \tag{8}$$

With this definition, the data-fitting term only incorporates pixel-wise intensity conservation. In [4], Bruhn *et al.* leverage the energy penalization found in [14] to average the intensity conservation over the pixel's neighborhood.

$$D_{\mathrm{Bruhn}}(\boldsymbol{w}, \bar{f}) := \boldsymbol{w}^T J_\rho \boldsymbol{w}, \text{ with} \tag{9}$$

$$J_\rho = (K_\rho \circledast \bar{J}_0)(x, y, t) \text{ and } \rho \in \mathbb{R}^+. \tag{10}$$

Here, K_ρ is a 2D Gaussian kernel with a standard deviation ρ, and $\circledast$ is a per-channel 2D-convolution operator applied to the matrix field $\bar{J}_0$. It means that the solution $\boldsymbol{w}$ should solve its intensity conservation equation and its neighbors'.

By replacing D by eq. (10) and R by eq. (3) in eq. (1), we have the CLG (Combined Local-Global) model, as defined in [4]

$$E_{\mathrm{CLG}}(\bar{\bar{\boldsymbol{w}}}) := \int_\Omega \boldsymbol{w}^T J_\rho \boldsymbol{w} + \alpha(\|\nabla_{x,y} u\|^2 + \|\nabla_{x,y} v\|^2)\,\mathrm{d}x\,\mathrm{d}y. \tag{11}$$

The convex optimization problem is now entirely defined. It is usually solved with an iterative gradient descent technique: each step yields a new approximate solution by displacing the current solution towards the opposite direction of the gradient. Two methods exist to compute the gradient of $E(\bar{\bar{\boldsymbol{w}}})$: the first one considers $\bar{f}$ and $\bar{\bar{\boldsymbol{w}}}$ to be continuous functions and employs the Euler-Lagrange equations. The second one discretizes

$\bar{f}$ and $\bar{\boldsymbol{w}}$ over the two-dimensional pixel grid first. This version is detailed in this article, with

$$E_{CLG}(\bar{\bar{\boldsymbol{w}}}^*) = \bar{\bar{\boldsymbol{w}}}^T H \bar{\bar{\boldsymbol{w}}} + \alpha\Big(\|D_x S_u \bar{\bar{\boldsymbol{w}}}^*\|^2 + \|D_y S_u \bar{\bar{\boldsymbol{w}}}^*\|^2 + \|D_x S_v \bar{\bar{\boldsymbol{w}}}^*\|^2 + \|D_y S_v \bar{\bar{\boldsymbol{w}}}^*\|^2\Big). \quad (12)$$

Equation (12) introduces $\bar{\bar{\boldsymbol{w}}}$, a $3 \times I_h \times I_w$ vector, such that $\bar{\bar{\boldsymbol{w}}}^T = \begin{bmatrix}\bar{u}^T & \bar{v}^T & \bar{\bar{1}}^T\end{bmatrix}$, similarly, $\bar{\bar{\boldsymbol{w}}}^{*T} = \begin{bmatrix}\bar{u}^T & \bar{v}^T\end{bmatrix}$. S_u and S_v are diagonal matrices that respectively select $\bar{u}$ and $\bar{v}$ parts of $\bar{\bar{\boldsymbol{w}}}$. D_x and D_y are discrete partial derivative operators along the x and y axes.

H is composed of diagonal matrices

$$H = \begin{bmatrix} \operatorname{diag}\bar{\bar{j}}_{\rho,0,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,0,1} & \operatorname{diag}\bar{\bar{j}}_{\rho,0,2} \\ \operatorname{diag}\bar{\bar{j}}_{\rho,1,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,1,1} & \operatorname{diag}\bar{\bar{j}}_{\rho,1,2} \\ \operatorname{diag}\bar{\bar{j}}_{\rho,2,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,2,1} & \operatorname{diag}\bar{\bar{j}}_{\rho,2,2} \end{bmatrix}, \quad (13)$$

with

$$J_\rho = \begin{bmatrix} j_{\rho,0,0} & j_{\rho,0,1} & j_{\rho,0,2} \\ j_{\rho,1,0} & j_{\rho,1,1} & j_{\rho,1,2} \\ j_{\rho,2,0} & j_{\rho,2,1} & j_{\rho,2,2} \end{bmatrix}. \quad (14)$$

Let us now compute the derivative of eq. (12) with respect to $\bar{\bar{w}}^*$

$$\nabla_{\bar{\bar{\boldsymbol{w}}}^*} E_{\mathrm{CLG}}(\bar{\bar{\boldsymbol{w}}}^*) = 2 S_{u,v} H \bar{\bar{\boldsymbol{w}}} + 2\alpha\Big(S_u^T (D_x^T D_x + D_y^T D_y) S_u \bar{\bar{\boldsymbol{w}}}^* + S_v^T (D_x^T D_x + D_y^T D_y) S_v \bar{\bar{\boldsymbol{w}}}^*\Big) \quad (15)$$

$$S_{u,v} H \bar{\bar{\boldsymbol{w}}} = \begin{bmatrix} \operatorname{diag}\bar{\bar{j}}_{\rho,0,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,0,1} \\ \operatorname{diag}\bar{\bar{j}}_{\rho,1,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,1,1} \end{bmatrix} \bar{\bar{\boldsymbol{w}}}^* + \begin{bmatrix} \bar{\bar{j}}_{\rho,0,2} \\ \bar{\bar{j}}_{\rho,1,2} \end{bmatrix}. \quad (16)$$

The selection matrix $S_{u,v}$ is necessary as H is applied to the vector $\bar{\bar{\boldsymbol{w}}}$ that contains ones in addition to u's and v's.

Equation (15) should now be set to zero to find a minimizer of E_{CLG}. By doing so, we obtain an equation of the generic form $A\boldsymbol{x} = \boldsymbol{b}$ where

$$A = \begin{bmatrix} \operatorname{diag}\bar{\bar{j}}_{\rho,0,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,0,1} \\ \operatorname{diag}\bar{\bar{j}}_{\rho,1,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,1,1} \end{bmatrix} - \alpha \begin{bmatrix} L & 0 \\ 0 & L \end{bmatrix} \quad (17)$$

with $L = D_x^T D_x + D_y^T D_y$, $\boldsymbol{b}^T = -[\bar{\bar{j}}_{\rho,0,2}, \bar{\bar{j}}_{\rho,1,2}]^T$, and $\boldsymbol{x} = \bar{\bar{\boldsymbol{w}}}^*$.

3.2 Solver Overview

The linear system of equations $A\boldsymbol{x} = \boldsymbol{b}$ can be solved in various ways. However, the characteristics of the optical flow setting restrict the choice of possible solvers. In a typical environment, with an HD image stream of dimensions 1280×480, there are over $2 \cdot 10^9$ coefficients in the matrix A. As is, an embedded system would never be able to store the whole matrix. Hopefully, the matrix is sparse, with more than 99.999% of its coefficients being zeros. It is then crucial to find a solver that takes advantage of this sparsity to make the computation possible on embedded devices.

The two following sections present two principal families of solvers for the optical flow. First, matrix splitting methods have been chosen in seminal work on flow estimation [12] and remain widely used to solve these linear systems [13]. Second, Krylov methods are often used for numerical simulations and benefit from a well-supplied scientific corpus [19].

3.2.1 Matrix Splitting

The matrix splitting methods partition the matrix A into two: $A = B + C$. Using this equality in $A\boldsymbol{x} = \boldsymbol{b}$ yields $B\boldsymbol{x} = \boldsymbol{b} - C\boldsymbol{x}$. Assuming B is invertible, an iterative scheme is constructed

$$\boldsymbol{x}^{k+1} = B^{-1}(\boldsymbol{b} - C\boldsymbol{x}^k) \quad (18)$$

$$\boldsymbol{x}^{k+1} = (I - B^{-1}A)\boldsymbol{x}^k + B^{-1}\boldsymbol{b}. \quad (19)$$

The choice of B leads to different methods. For example, choosing B to hold the diagonal of A: $B_{\mathrm{J}} := D_A$ is the Jacobi solver, while $B_{\mathrm{GS}} := D_A + L_A$ is the Gauss-Seidel method (with L_A, the lower triangular part of A).

In the case of optical flow problems, we can craft custom B matrices based on the structure of A. These variants contain the four non-empty diagonals of A

$$B_{\mathrm{J}}^{(\mathrm{diags})} := \begin{bmatrix} \operatorname{diag}\bar{\bar{j}}_{\rho,0,0} - \alpha D_L & \operatorname{diag}\bar{\bar{j}}_{\rho,0,1} \\ \operatorname{diag}\bar{\bar{j}}_{\rho,1,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,1,1} - \alpha D_L \end{bmatrix} \quad (20)$$

$$B_{\mathrm{GS}}^{(\mathrm{diags})} := \begin{bmatrix} \operatorname{diag}\bar{\bar{j}}_{\rho,0,0} - \alpha L_L & \operatorname{diag}\bar{\bar{j}}_{\rho,0,1} \\ \operatorname{diag}\bar{\bar{j}}_{\rho,1,0} & \operatorname{diag}\bar{\bar{j}}_{\rho,1,1} - \alpha L_L \end{bmatrix}. \quad (21)$$

This construction of B matrices is called the pointwise-coupled method in [13], as these matrices update $u_{x,y}$ and $v_{x,y}$ simultaneously. Later in this article, we call these versions "preconditioned" by analogy with the Krylov methods.

The spectral radius ρ_{SR} of $I - B^{-1}A$ must be studied to show how the specially designed matrices compare to the traditional ones. At each iteration of the solver, the error's norm $\|\boldsymbol{e}^k\|_2 = \|\boldsymbol{b} - A\boldsymbol{x}^k\|_2$ is multiplied by a factor ρ_{SR}. The end goal is then to find a matrix B such that the corresponding ρ_{SR} is as close to zero as possible.

Figure 2 presents results for the Jacobi and Gauss-Seidel solvers with their derived pointwise-coupled methods. The optical flow is analyzed under two parameter settings, with $\alpha = 5e^{-3}$ or $\alpha = 5e^{-6}$. The plot shows $-\log(\rho_{\mathrm{SR}})$ for easier comparison between solvers. Note

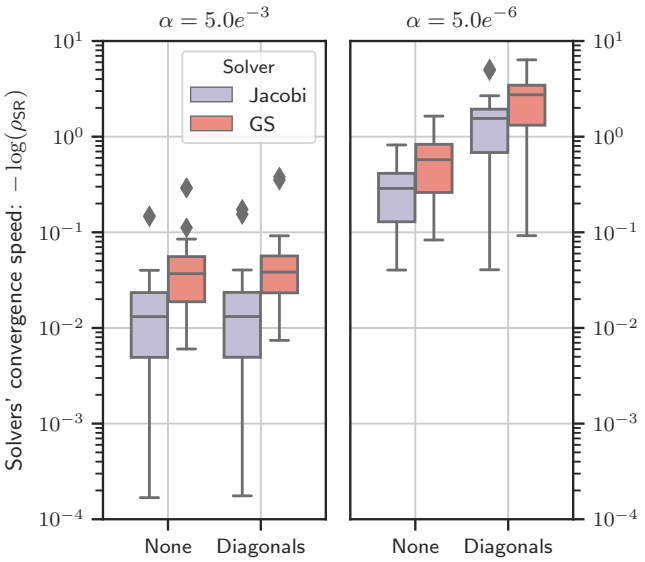


Fig. 2: Comparison of the theoretical convergence speed of the Jacobi and Gauss-Seidel (GS) methods. Standard solvers are refered by *None* and precontioned, or pointwise-coupled, variations with *Diagonals*.

that finding ρ_{SR} is a computationally heavy task, so this numerical analysis is done on cropped images from the Middlebury dataset [2].

We can draw three conclusions from fig. 2. First, a low alpha dramatically increases the convergence speed for all types of solvers by factors of $20 \sim 100$. Second, with the same flow parameters, Gauss-Seidel is about three times faster than Jacobi. Last, the pointwise-coupled method is useful only in a low alpha setting where a $5\times$ speedup is achieved.

3.2.2 Krylov's methods

Krylov solvers all emerge from the same premise: at each iteration, increase the possible solutions' space's dimension. Such spaces, called Krylov spaces, are defined by

$$\mathcal{K}_n(A, \boldsymbol{b}) = \text{span}\{\boldsymbol{b}, A\boldsymbol{b}, A^2\boldsymbol{b}, \dots, A^{n-1}\boldsymbol{b}\},\ n \in \mathbb{N}^*. \quad (22)$$

The choice of the solution in these subspaces leads to different methods: Conjugate Gradient (CG), MINimal RESidual (MINRES), or Generalized Minimal RESidual (GMRES), for example.

The speed of Krylov's methods depends on the matrix condition number κ of the matrix A. This characteristic quantifies how much our model's result changes with a small perturbation in the input data. A low condition number reflects a robust modelization of our problem. It also hints that Krylov solvers should converge rapidly [21].

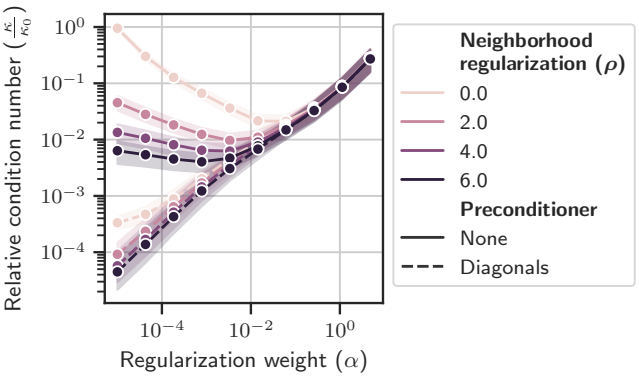


Fig. 3: Normalised condition number of the problem versus parameters' value.

Sometimes, the system can be enhanced by the use of a preconditioner M. With M being an invertible matrix, the linear system to solve becomes

$$M^{-1}A\boldsymbol{x} = M^{-1}\boldsymbol{b}. \quad (23)$$

Solving the system eq. (23) is equivalent to solve $A\boldsymbol{x} = \boldsymbol{b}$ with a change of variables $A' = M^{-1}A$ and $b' = M^{-1}\boldsymbol{b}$. This system should be faster to solve if $\kappa(A') < \kappa(A)$.

Similarly to the pointwise-coupled matrices defined for splitting solvers, a natural preconditioner for the optical flow is

$$M = \begin{bmatrix} \text{diag}\,\bar{\bar{j}}_{\rho,0,0} - \alpha D_L & \text{diag}\,\bar{\bar{j}}_{\rho,0,1} \\ \text{diag}\,\bar{\bar{j}}_{\rho,1,0} & \text{diag}\,\bar{\bar{j}}_{\rho,1,1} - \alpha D_L \end{bmatrix}. \quad (24)$$

Figure 3 summarizes the value of κ for several model parametrizations: with ρ, the local radius parameter, ranging from 0 to 6 and α, the global regularization weight, from $1e^{-5}$ to 10. The values shown are the ratio κ by κ_0 with κ_0 the value of κ taken with $\alpha = 0$ and $\rho = 0$. Just like in section 3.2.1, images have been cropped to compute κ.

We can now conduct a similar analysis for fig. 3 as we did for fig. 2. First, without preconditioning, κ follows a V-shape with respect to α. However, with a preconditioner M defined as in eq. (24), κ always increases with α. This difference is important, as, for low values of α, the preconditioner decreases κ by orders of magnitudes. With higher values of α, though, the effect of M is barely noticeable. Finally, we can assert that increasing ρ is significant with low α and no preconditioning.

Figure 3 confirms the results of fig. 2: solvers are the fastest when preconditioned and with low α values. The effect of α can be analyzed by looking back at eq. (11): with α close to zero, most of the penalization comes from $\boldsymbol{w}^T J_\rho \boldsymbol{w}$. This term is directly sensitive to the value ρ. Moreover, the preconditioner M "targets"

this term. It is then not a surprise to see how effective it is with low α.

With high values of α, the term $\|\nabla_{x,y} u\|^2 + \|\nabla_{x,y} v\|^2$ dominates. Then, the influences of ρ and M are negligible. Conversely, κ increases because the solution is solely determined by having a zero derivative so that any constant field would be a potential solution.

3.3 A coarse solver benchmark

While sections 3.2.1 and 3.2.2 presented theoretical results for solver convergence on small images, the actual performance is yet to be measured. In this section, we are interested in two indicators: convergence vs. iterations and convergence vs. time. The data is averaged over 30 images from several databases [2,5,11].

Since convergence vs. iterations is platform-independent, we can rely on it as an initial filter for limiting the number of solvers to test on the target hardware.

Then comes an implementation on target for the actual solvers' performance. In fact, a performant solver under the convergence vs. iterations measure may become less attractive if the time to perform one iteration is too slow on the targeted hardware.

3.3.1 Convergence vs. iterations

For fig. 4, we chose two sets of parameters to compare the convergence of the solvers mentioned above. We tried several Krylov solvers from the *sparse* module of Scipy but only reported Conjugate-Gradient (CG) as it was the most relevant. We developed two splitting methods: Jacobi and Red-Black Gauss-Seidel (Red-Black GS). The more traditional Gauss-Seidel solver has been discarded from the benchmark. It requires all pixels to be treated sequentially and thus is not appropriate for a parallel implementation. Red-Black GS is a variation on Gauss-Seidel that updates half of the pixels simultaneously [19].

The results differ greatly depending on α. On fig. 4, when α is low, preconditioned method converges quickly (up to 10^{-9} in 100 iterations). The CG method is the fastest, but splitting methods are not far behind. On the contrary, when α is higher, all solvers converge slowly ($\sim 10^{-5}$ in 100 iterations), and splitting methods are still slower.

Consistently with the results found in section 3.2, the effects of the preconditioner are less visible with higher α. On the left graph of fig. 4, the preconditioned helps the convergence of CG a little, but not as much as when $\alpha = 5e^{-6}$. Regarding splitting solvers, the results with or without preconditioning overlap.

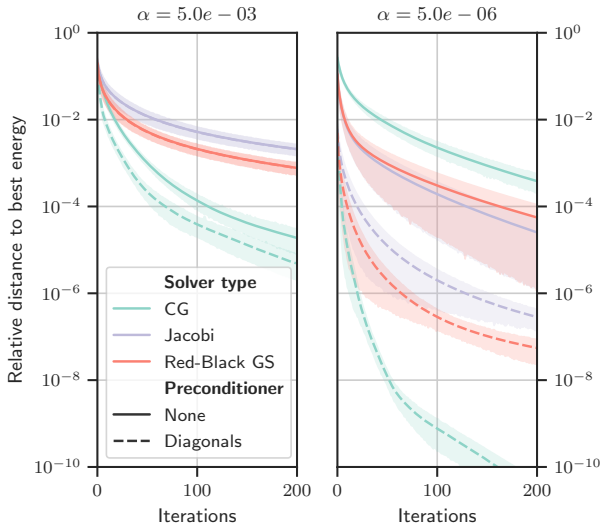


Fig. 4: Convergence vs. iterations with $\rho = 2.5$. On the lest $\alpha = 5e^{-3}$, on the right $\alpha = 5e^{-6}$

3.3.2 Convergence vs. time

This subsection presents solvers' results as implemented on the embedded target GPU: the Jetson AGX Xavier. The time spent on all implementations was roughly equal. Splitting solvers' implementation is relatively straightforward: all (Jacobi) or half (Red-Black GS) of the pixels are updated in parallel, in a "embarrassingly parallel" fashion.

For the Conjugate-Gradient method, one difficulty is to compute a vector's norm. This operation is not so well adapted to GPUs. Then, we leveraged the *CUB* library (CUDA UnBound) for state-of-the-art GPU reduction performance. We, moreover, took extra care to keep all intermediate results on GPU to avoid expensive latency in CPU-GPU communication.

Figure 5 shows convergence timings for different solvers on GPU until they reach a runtime of 200ms. Globally, the curves follow the same trend as fig. 4 and the order of the curves is respected. Splitting solvers are, however, catching up with CG's performance.

With a low alpha (left-hand side), Jacobi and Red-Black GS are faster than CG in the very first iterations and stay close to CG's performance for a longer time. With a high alpha (right-hand side), all preconditioned methods are on par.

An important finding of the benchmark is that the Conjugate-Gradient method is sensitive to numerical precision. On the right-hand side graph of fig. 5, the method diverges after about 100ms of compute. While arithmetic is done in FP32 (IEEE 754 *binary32*) precision, we observed identical behavior in FP64 [20]. This phenomenon also happened with $\alpha = 5.0e^{-3}$, after a

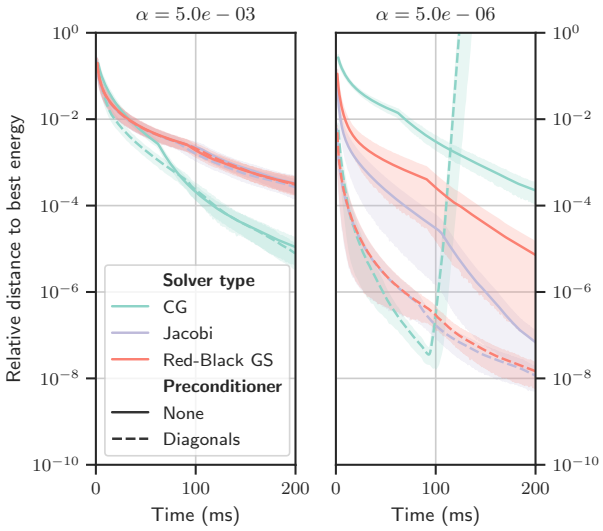


Fig. 5: Convergence vs. time on a Jetson AGX Xavier. Parameters are the same as on fig. 4.

vast number of iterations, though. We attribute this divergence to the sensitivity of the Conjugate-Gradient method to roundoff errors [25].

For further optimization, Jacobi was chosen as it is the simplest to implement, with adequate performance and good numerical stability. As described later in the article, it is also possible to fuse iterations of Jacobi.

4 Implementation-level approach

This section extends the analysis done in section 3. As a starting point, the solver is now considered fixed. That choice is possible thanks to the initial benchmark on the actual target.

An initial implementation of the CLG method on GPU is done, including the underlying solver and the multi-scale strategy, as detailed in [13]. First, we find sources of optimizations for the solver or elsewhere in the method. Then, we analyze the effects of multi-scale processing by measuring the performance of working on a particular level and the computational cost of changing scale.

4.1 Framework and optimizations

When optimizing the code, it is essential to follow a consistent strategy. One must profile the application first to find its main bottlenecks, then try to solve these hotspots, and always check that the application provides the same results. On the Jetson AGX Xavier platform, Nsight Systems and Nsight Compute are two NVIDIA-provided tools that profile executions of programs.

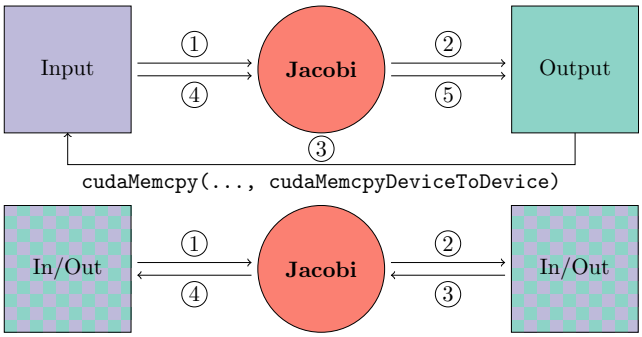


Fig. 6: Two iterations of Jacobi without (top) or with (bottom) buffer reuse. Top version reads from (1), writes to (2), copies output to input (3) and does a new iteration (4) and (5). Bottom version avoids copying by using buffers as both input or output.

The first one analyzes the whole system and provides CPU and GPU execution traces. This information highlight which kernels would benefit the most from optimization.

The second one dives deeper into a single kernel execution. It provides multiple metrics such as GPU cores occupancy, bandwidth, or a roofline model plot. These lower-level indicators facilitate the discovery of bottlenecks within the kernel.

4.1.1 Optimizations' overview

In this sub-section, we detail the different optimizations that we added to our CLG GPU implementation. They are standard techniques known in the literature, but their application for optical flow is original, as well as their analysis in this context. We present them in their order of importance: after each optimization, a new hotspot is selected until speed gains become marginal.

Buffer Reuse: this optimization acts on the Jacobi solver. At the k-th iteration, the program needs one location in memory for the input $x^{(k)}$ and one for the output $x^{(k+1)}$. An initial approach is to fix the memory position of inputs and outputs. This strategy then rely on a copy of the previous output to the current iteration's input: $x^{(k)} \leftarrow x^{(k-1)}$. The memory operation can be avoided by changing the input and output locations at each solver iteration, in a back-and-forth fashion. Figure 6 illustrates this technique.

Jacobi Fusion: the Jacobi solver consumes a lot of memory bandwidth: for each pixel, it fetches a neighborhood of values to compute the Laplacian in addition to coefficients from $\boldsymbol{b}$. All this data is processed with few operations: the solver is bandwidth limited. Our solution is to combine the computation of several iterations within a single kernel launch. This optimization is

probably the most important one so that section 4.1.2 extends its analysis.

Batched convolutions: the multi-scale processing of CLG relies on up and down-sampling the image to solve the problem at different scales. This change of resolution uses a Gaussian kernel convolution on images to preserve the down-sampling for high-frequency artifacts. Rather than launching a CUDA kernel for each convolution, we prefer to launch a single kernel that performs convolutions on many images at once. This means that the kernel launch overhead is limited and that the Gaussian filter weights are loaded once and reused for all images.

4.1.2 Fusing iterations of Jacobi

As mentioned previously, the main issue of the Jacobi solver on GPU is its high demand for memory resources. As is, the implementation saturates the VRAM bandwidth, and GPU compute units are starving. To quantify the phenomenon, let us introduce the arithmetic intensity of a program defined by the ratio between the number of floating-point operations (FLOPs) performed by a computed unit over the number of bytes moved to do these operations

$$\text{AI} = \frac{\text{FLOPs}}{\text{Bytes loaded}}. \tag{25}$$

A low AI is symptomatic of over-used memory bandwidth. Conversely, if AI is too high, the program requests so many FLOPs that the compute units cannot process them fast enough. Further analysis of the role of AI on a program's execution may be found in [24].

In our initial case, the CUDA kernel is programmed to do a single Jacobi iteration. This approach is straightforward but has several limitations: it requires one kernel launch per iteration so that the call overhead might become an issue. Moreover, each iteration output is written back to main memory, but this is not strictly needed. Combining several iterations within the same kernel would allow direct reuse of intermediate iterations in addition to load coefficients of $\boldsymbol{b}$ only once.

Bottom fig. 7 exposes a fusion of two iterations of Jacobi within a single kernel launch. Static parameters are loaded once and serve for both iterations. The output of the first Jacobi iteration is immediately reused for the second one. The two-iteration scheme requires loading a larger neighborhood of $\boldsymbol{x}$ values to satisfy all further dependencies.

Another important aspect of this implementation is shared memory. In the CUDA model, GPU threads are partitioned into Thread Blocks (TB). Threads of a common TB are executed on a single processing unit,

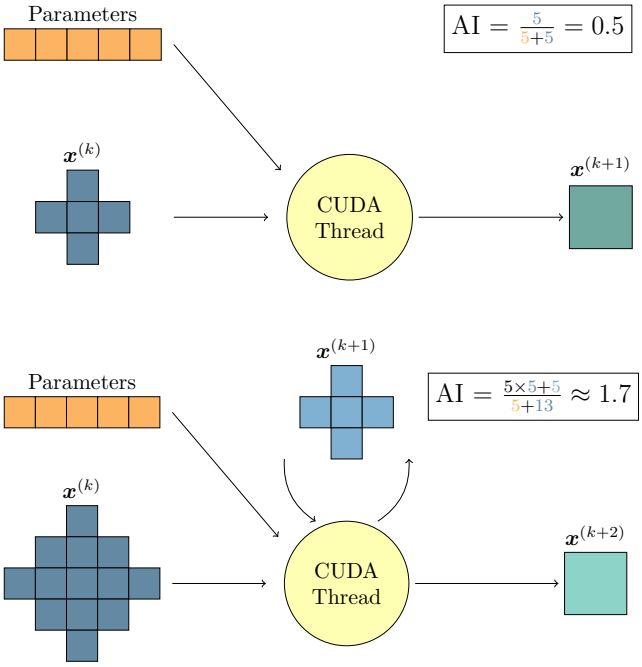


Fig. 7: Top: an iteration of Jacobi for a single output. Bottom: fusion of two Jacobi iterations. Arithmetic intensity is given for reference only, assuming one operation per $\boldsymbol{x}^{(k)}$.

the streaming multiprocessor, and have access to shared memory. This location is used to share the coefficient of $\boldsymbol{x}$ between GPU threads, leveraging the pixels' neighborhoods' spatial redundancies.

For now, let us set the TB size to 32×32. Initially, each thread of the TB load one coefficient of $\boldsymbol{x}^{(k)}$ from the main memory to the shared memory. Then, threads compute a first Jacobi iteration and wait for the TB to have finished thanks to the synchronization primitive `__syncthreads`. With the $\boldsymbol{x}^{(k+1)}$ coefficients being computed, the TB computes the subsequent Jacobi iteration.

Let us now find the approximate value of AI based on an implementation that fuses j iterations. At each new iteration, the size of the computed area decreases because of spatial dependencies. At the i-th iteration, $i \leq j$, the footprint's size is $(32 - 2i) \times (32 - 2i)$. We can now express AI as a function of j, the number of fused iteration

$$\text{AI}(j) = \frac{\alpha \sum_{i=1}^{j} (32 - 2i)^2}{\beta(32 \times 32)} \tag{26}$$

α is the number of FLOPs needed per pixel and per iteration and β is the number of bytes to load per pixel.

While the AI expressed in eq. (26) increases with j and then seems to benefit the implementation, it is important to understand that the total FLOPs required

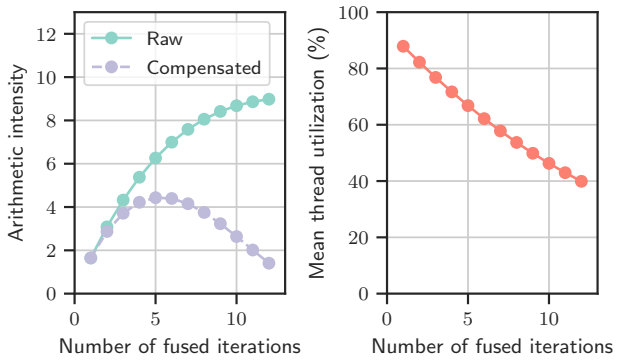


Fig. 8: Arithmetic intensity and mean thread utilization w.r.t. the number of fused iterations.

by the algorithm are not constant with j. A single non-fused Jacobi needs

$$W_{\text{no fusion}} = \alpha HW. \tag{27}$$

operations. With H and W the dimensions on the processed image. In comparison, in a j-fused implementation, each TB computes a patch of $(32-2j)^2$ pixels. To compute the entire image, we have

$$W_{\text{j-fusion}} = \lceil \frac{H}{32-2j} \rceil \lceil \frac{W}{32-2j} \rceil \alpha \sum_{i=1}^{j} (32-2i)^2. \tag{28}$$

Some operations are redundant with the fused iterations technique to handle patch borders and avoid inter-TB communication.

The ratio between $W_{\text{j-fusion}}$ and $j \cdot W_{\text{no fusion}}$ expresses the overhead of operations due to the fusion of operations

$$\frac{W_{\text{j-fusion}}}{jW_{\text{no fusion}}} \approx \frac{1}{j(32-2j)^2} \sum_{i=1}^{j} (32-2i)^2 \tag{29}$$

The left-hand side graph of fig. 8 shows the AI for different choices of j, the number of fused iteration. The solid curve represents the AI computed by the formula in eq. (26). The dashed curve is arithmetic intensity divided by the *compute* overhead, as expressed in eq. (29).

The *raw* AI is an increasing function of j: by looking at this metric only, it would make sense to choose j as large a possible to reduce the memory pressure. Conversely, the refined metric, *compensated* AI, indicates that because higher values of j induce too much redundant work, it is better to choose j close to 5.

The right-hand side of fig. 8 shows the percentage of active threads during the entire fused iteration. With each supplementary fused iteration, the footprint of computable coefficients shrinks, so fewer threads are operating.

This study of the Jacobi iteration fusion has exhibited the pros and cons of using many fused iterations. While done in a theoretical setting, it should help to analyze GPU execution performance.

4.2 Results

To measure the effects of the various optimizations presented in section 4.1.1, we have taken measurements on two GPU cards. The first one, an NVIDIA Titan V, is used in PCs and computing servers. We use it as the baseline of our development process. The second one, a Jetson AGX Xavier, is the actual target of our industrial application. After initial implementation and verification on Titan V, we deploy on Xavier, and we check if the optimization has the expected effect.

In our method, the optimizations' order is guided by results on the Jetson Xavier. For example, fig. 9 shows us that once the buffer reuse optimization is implemented, the time spent in memory transfers is still relatively high on Titan V but not on Xavier. Since we ultimately focus on this embedded target, we will not dwell on further optimization for memory transfers.

All the details about the hardware used for our experiments are available on table 1.

	Machine #1: desktop	Machine #2: embedded (Jetson AGX Xavier)
OS	Ubuntu 16.04	Ubuntu 18.04
Linux Kernel	4.15.0	4.9.140
CUDA	11.0	10.2
NVIDIA Driver	450	JetPack 4.4
CPU	Intel i7-3820	8-core ARM 64bits
GPU	Titan V (arch. 7.0)	Xavier (arch. 7.2)
TDP	∼500W	∼30W

Table 1: Environments of the experiments.

Buffer Reuse: On the initial runtime bar of fig. 9, we can see that a good part of the computation time spent on GPU is dedicated to memory transfers. The effects of the buffer reuse optimization are pretty different depending on the platform.

On Jetson Xavier, we can see that the time spent in memory operations goes from about 3ms to 0.5ms. The remaining memory time is spent uploading the input images and downloading the output flow. A further optimization could lead to marginal gains by using Unified Memory. This makes buffer transfers with zero-copy because the GPU and the CPU share the same memory.

On Titan V, we can see that those memory operations still require a lot of time (∼33% of the computation time). This is explained by the fact that the CPU and GPU memory are disjoint, so it takes more time

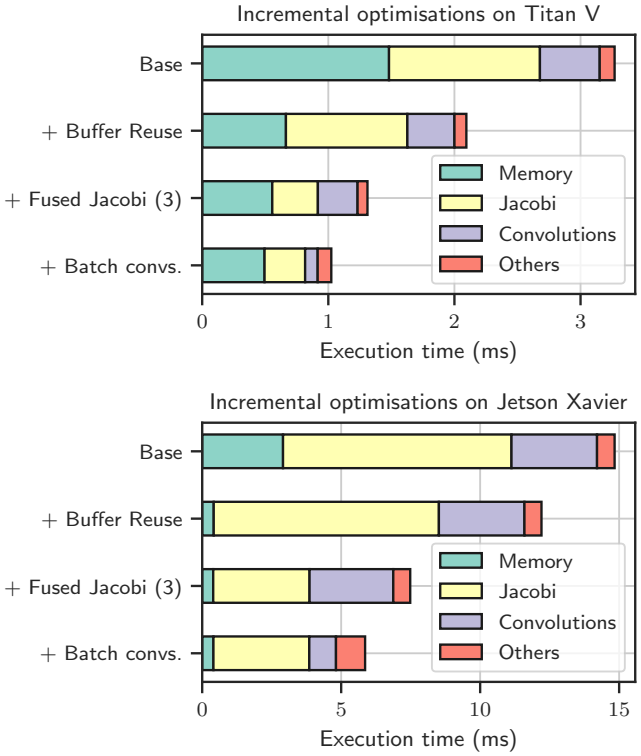


Fig. 9: Effects of cumulative optimizations on Titan V (top) and Jetson Xavier (bottom).

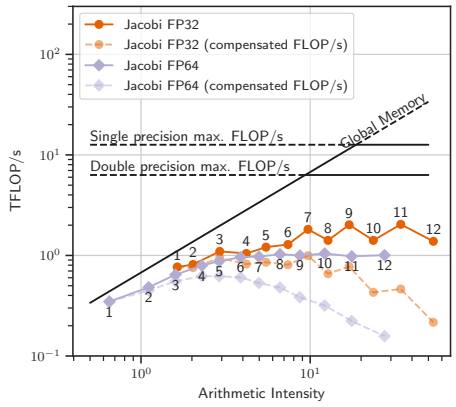


Fig. 10: A roofline model analysis of Jacobi iteration fusion on Titan V.

(proportionally to the power of the machine) to transfer the inputs and outputs.

Jacobi Fusion: tests shown on figs. 10 and 11 evaluates the performance of different number of fused Jacobi iterations, as explained in section 4.1.2. Figure 10 plots the achieved TFLOP/s (Tera Floating-Point Operations per second) with respect to the measured arithmetic intensity. This figure first shows that, for Titan V, the FP64 machine balance is reached for an AI of 10, while 20 is needed for FP32 operations. This value exposes the minimum number of operations per byte

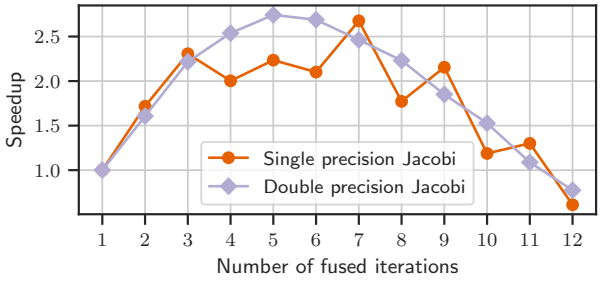


Fig. 11: Speedup vs. number of fused iterations on Titan V.

to compute to benefit from maximal hardware performance.

Without any fusion, it is clear that both FP32 and FP64 implementations are bandwidth-limited. As expected, the arithmetic intensity of increases with the number of fused iterations. With an ideal execution, the points should appear close to the roofline. Here, after few fusions, it is clear that the progression stalls. While the FLOP/s continue to increase, this growth is not sufficient to stay close to the roofline. We explain this behavior by the low number of active threads within a TB, as modeled on fig. 8. At each new fused iteration, the size of the region of interest of a TB decreases, then, more of its threads are idle.

Comparing raw performance on fig. 10 is complex. The *FLOP/s* metric, given by Nvidia Nsight Compute, directly measures the activity of computing units. As explained in section 4.1.2, some computations are redundant from the method point-of-view. To correct the *FLOP/s* metric, we divide it by the work overhead defined in eq. (29). This compensated curve draws a different conclusion than the initial one. For example, in FP32, the *raw FLOP/s* is highest for nine fused iterations. In the compensated model, the best fusion is lower: around three iterations.

This difference highlight a drawback of the analysis based on the roofline model only. When the total number of operations changes from one implementation to another, the achieved *FLOP/s* is not comparable. In our case, fig. 11 is more straightforward: it shows the execution time gain for different numbers of fused iterations.

The maximum performance is achieved with seven fused iterations in FP32 and five in FP64. These results tend to confirm the analysis of the *compensated* roofline model made on fig. 10.

We now choose a value of three fused iterations in FP32 for two reasons: it achieves good speedup both in FP32 and FP64 precisions, and it is more convenient to have a total number of iterations that is a multiple

of three, than seven, for example. On fig. 9, the time spend for Jacobi iterations is almost divided by three on Titan V and about by two on Jetson Xavier. The additional speedup, especially on Titan V, is explained by reducing kernel launch overhead.

Batched convolutions: after optimizing the Jacobi solver, fig. 9 shows that on Jetson Xavier, almost half of the runtime is spent doing convolutions. As explained in section 4.1, convolutions have been re-expressed to be run in a single CUDA function. Instead of launching a kernel per convolution, the batch computation reduces the overhead and lets the convolution filter's coefficients in the CUDA thread registers. This makes the runtime of convolutions decrease by a factor of two on Xavier.

		Ours (Jetson Xavier)		
	Moussu [16]	CPU	GPU (baseline)	GPU (optimized)
FPS	4.2*	0.4	4.5	**15.0**

Table 2: Throughput of 3000 Jacobi iterations. Same conditions as [16]. *scaled results.

To show the benefits of these optimizations, we compare our results with another Jacobi GPU implementation in table 2. This reference [16] was measured on a Geforce 9400 GT, released in 2008. We scaled that result to account for the 14.3× larger memory bandwidth in the Jetson Xavier. We chose to multiply according to this metric as memory requests are the bottleneck of non-optimized Jacobi implementations (see fig. 10). Our original GPU implementation is then on-par with results extrapolated from Moussu. The optimized version, however, benefits from a large speed-up (3.5× vs. [16]) thanks to the buffer reuse and iteration fusion strategies. For reference, we included a CPU version in the comparison. This implementation leverages OpenMP parallelization but under-performs compared to GPU versions.

4.3 Multi-level analysis

As detailed in [4], multi-level processing aims at finding optical flows at different problem scales. The technique helps in finding large displacements and iterates quickly on higher levels due to the reduced problem size. Consequently, we measure the actual performance of our GPU implementation on different image sizes. Those results should guide decisions back at the algorithm level to set the number of iterations per level that fits a given time frame.

Table 3 presents results for the Jacobi solver on various image sizes. Below 320 × 256, there is not enough processing to saturate the GPU, so the time does not vary vastly between different sizes. However, with larger dimensions we see that the processing time grows linearly with the number of pixels in the image. We can now use this knowledge to choose the number of iterations per scale of the problem.

Size	20 × 16	40 × 32	80 × 64	160 × 128
Time (ms)	6.34	6.39	10.4	13.1
Size	320 × 256	640 × 512	1280 × 1024	2560 × 2048
Time (ms)	26.3	83.6	341	1,360

Table 3: Time to perform 1000 Jacobi iterations on Jetson Xavier vs. image size.

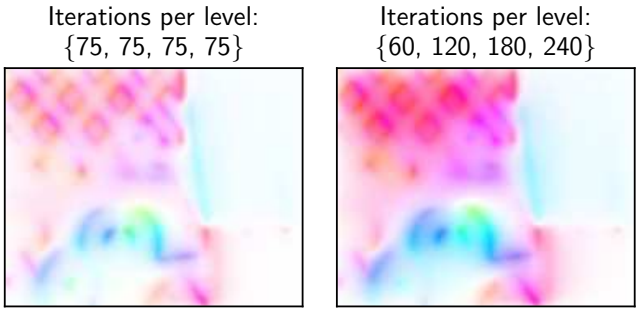


Fig. 12: Results on 640 × 512 images at 60 FPS. Doing more iterations at higher levels converges faster.

On fig. 12, the optical flow for two choices of parameters is displayed. The left-hand flow was obtained by doing 75 iterations on each scale, from the 640 × 512 level, to the upper ones: 320×256, 160×128, and 80×64. The right-hand flow sets 60 iterations at the highest-resolution level and 120, 180, and 240 iterations at the lower ones. While both configurations run at the same speed, 60 FPS on the Jetson Xavier with 640 × 512 images, the configuration using more iterations on the higher levels yields a smoother flow. It has converged more on less textured regions and seems better for practical use.

5 Conclusion

This article has shown the interest in combining analyses at the algorithm and implementation levels to obtain the best performance.

Initially, we pre-selected candidate GPU solvers for a subsequent GPU optimization. This first analysis also provided an understanding of the hyper-parameters on the convergence speed. Then, the multi-scale CLG algorithm was ported on the embedded Jetson AGX Xavier GPU. Several optimizations have enhanced the algorithm's run time: re-utilization of intermediate Jacobi buffers, solver iteration fusion, and batching of convo-

lution. Overall, these techniques decreased the runtime of the algorithm by more than 2×.

The multi-scale behavior of the method has also been studied. Results have shown that higher levels are processed faster but that the speedup plateaus for images smaller than 80×64. This result allowed us to choose the right parameters for the best possible convergence within a limited time frame.

In the end, our GPU implementation of the CLG optical flow method runs at 60 frames per second on 640×512 images with a 30W power budget. Tuning the number of iterations per level set allowed us to produce a smoother flow in the same time frame. Overall, this implementation opens up the use of CLG optical flow for embedded applications like drones or robotics.

Further work could analyze the performance of multigrid solvers and their optimal configuration on GPUs or apply this optimization method to other optical flow methods.

References

1. Aliaga, J.I., Pérez, J., Quintana-Ortí, E.S.: Systematic Fusion of CUDA Kernels for Iterative Sparse Linear System Solvers. In: J.L. Träff, S. Hunold, F. Versaci (eds.) Euro-Par 2015: Parallel Processing. Springer. DOI: 10.1007/978-3-662-48096-0_52
2. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A Database and Evaluation Methodology for Optical Flow **92**(1), 1–31. DOI: 10.1007/s11263-010-0390-2
3. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High Accuracy Optical Flow Estimation Based on a Theory for Warping. In: T. Pajdla, J. Matas (eds.) Computer Vision - ECCV 2004. Springer. DOI: 10.1007/978-3-540-24673-2_3
4. Bruhn, A., Weickert, J., Schnörr, C.: Lucas/Kanade Meets Horn/Schunck: Combining Local and Global Optic Flow Methods **61**(3), 1–21. DOI: 10.1023/B:VISI.0000045324.43199.43
5. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: A. Fitzgibbon et al. (Eds.) (ed.) European Conf. on Computer Vision (ECCV), Part IV, LNCS 7577, pp. 611–625. Springer-Verlag
6. Capito, L., Ozguner, U., Redmill, K.: Optical Flow based Visual Potential Field for Autonomous Driving. In: 2020 IEEE Intelligent Vehicles Symposium (IV), pp. 885–891. DOI: 10.1109/IV47402.2020.9304777
7. Ding, N., Williams, S.: An Instruction Roofline Model for GPUs. In: 2019 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS), pp. 7–18. DOI: 10.1109/PMBS49563.2019.00007
8. Dougherty, L., Asmuth, J.C., Gefter, W.B.: Alignment of CT Lung Volumes with an Optical Flow Method **10**(3), 249–254. DOI: 10.1016/S1076-6332(03)80098-3
9. Farnebäck, G.: Two-Frame Motion Estimation Based on Polynomial Expansion. In: J. Bigun, T. Gustavsson (eds.) Image Analysis, Lecture Notes in Computer Science, pp. 363–370. Springer. DOI: 10.1007/3-540-45103-X_50
10. Filipovič, J., Madzin, M., Fousek, J., Matyska, L.: Optimizing CUDA Code By Kernel Fusion—Application on BLAS **71**(10), 3934–3957. DOI: 10.1007/s11227-015-1483-z
11. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset **32**(11), 1231–1237. DOI: 10.1177/0278364913491297
12. Horn, B.K.P., Schunck, B.G.: Determining optical flow **17**(1), 185–203. DOI: 10.1016/0004-3702(81)90024-2
13. Jara-Wilde, J., Cerda, M., Delpiano, J., Härtel, S.: An Implementation of Combined Local-Global Optical Flow **5**, 139–158. DOI: 10.5201/ipol.2015.44
14. Lucas, B.D., Kanade, T.: An Iterative Image Registration Technique with an Application to Stereo Vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence. Morgan Kaufmann. URL http://dl.acm.org/citation.cfm?id=1623264.1623280
15. McGuire, K., de Croon, G., De Wagter, C., Tuyls, K., Kappen, H.: Efficient Optical Flow and Stereo Vision for Velocity Estimation and Obstacle Avoidance on an Autonomous Pocket Drone **2**(2), 1070–1076. DOI: 10.1109/LRA.2017.2658940
16. Moussu, C.: GPU based real-time optical Flow computation. p. 108. Imperial College London
17. Nguyen, M.T., Castonguay, P., Laurendeau, E.: GPU parallelization of multigrid RANS solver for three-dimensional aerodynamic simulations on multiblock grids **75**(5), 2562–2583. DOI: 10.1007/s11227-018-2653-6
18. Podestá, E., do Nascimento, B.M., Castro, M.: Energy Efficient Stencil Computations on the Low-Power Many-core MPPA-256 Processor. In: M. Aldinucci, L. Padovani, M. Torquati (eds.) Euro-Par 2018: Parallel Processing. Springer. DOI: 10.1007/978-3-319-96983-1_46
19. Saad, Y.: Iterative Methods for Sparse Linear Systems. SIAM
20. Seznec, M., Gac, N., Orieux, F., Naik, A.S.: An Efficiency-Driven Approach For Real-Time Optical Flow Processing On Parallel Hardware. In: 2020 IEEE International Conference on Image Processing (ICIP), pp. 3055–3059. DOI: 10.1109/ICIP40778.2020.9191164
21. Shewchuk, J.R.: An Introduction to the Conjugate Gradient Method without the Agonizing Pain. Carnegie-Mellon University. Department of Computer Science
22. Sun, D., Roth, S., Black, M.J.: A Quantitative Analysis of Current Practices in Optical Flow Estimation and the Principles Behind Them **106**(2), 115–137. DOI: 10.1007/s11263-013-0644-x
23. Tabik, S., Ortega, G., Garzón, E.M.: Performance evaluation of kernel fusion BLAS routines on the GPU: Iterative solvers as case study **70**(2), 577–587. DOI: 10.1007/s11227-014-1102-4
24. Williams, S.W.: Auto-Tuning Performance on Multicore Computers. EECS Department University of California
25. Woźniakowski, H.: Roundoff-error analysis of a new class of conjugate-gradient algorithms **29**, 507–529. DOI: 10.1016/0024-3795(80)90259-1
26. Zach, C., Pock, T., Bischof, H.: A Duality Based Approach for Realtime TV-L 1 Optical Flow. In: F.A. Hamprecht, C. Schnörr, B. Jähne (eds.) Pattern Recognition, vol. 4713, pp. 214–223. Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-74936-3_22

Fast myopic 2D-SIM Super Resolution Microscopy with Joint Modulation Pattern Estimation

François Orieux[1], Vincent Loriette[2], Jean-Christophe Olivo-Marin[3], Eduardo Sepulveda[4], Alexandra Fragola[2]

[1] Laboratoire des Signaux et Systèmes (Univ. Paris-Sud – CNRS – CentraleSupélec – Université Paris-Saclay), 3 rue Joliot-Curie, 91 192 Gif-sur-Yvette, France
[2] Laboratoire de physique et d'étude des matériaux (CNRS – UPMC – ESPCI), 75 005 Paris, France
[3] Unité d'Analyse d'Images Biologiques (Institut Pasteur – CNRS), 75 015 Paris, France
[4] Laboratoire de Physique Nucléaire et des Hautes Énergies (IN2P3 – CNRS – UPMC – UPD), 75 252 Paris, France

E-mail: `orieux@l2s.centralesupelec.fr`

Abstract. Super-resolution in Structured Illumination Microscopy (SIM) is obtained through de-aliasing of modulated raw images, in which high frequencies are measured indirectly inside the optical transfer function. Usual approaches that use 9 or 15 images are often too slow for dynamic studies. Moreover, as experimental conditions change with time, modulation parameters must be estimated within the images. This paper tackles the problem of image reconstruction for fast super resolution in SIM, where the number of available raw images is reduced to *four* instead of nine or fifteen. Within an optimization framework, the solution is inferred *via* a joint myopic criterion for image and modulation (or acquisition) parameters, leading to what is frequently called a myopic or semi-blind inversion problem. The estimate is chosen as the minimizer of the nonlinear criterion, numerically calculated by means of a block coordinate optimization algorithm. The effectiveness of the proposed method is demonstrated for simulated and experimental examples. The results show precise estimation of the modulation parameters jointly with the reconstruction of the super resolution image. The method also shows its effectiveness for thick biological samples.

1. Introduction

Fluorescence microscopy, where only molecules marked by fluorophores are visible, is a fundamental tool of biology, but remains fundamentally limited in resolution by diffraction (often modeled by a filter in Fourier space). The last twenty years have seen numerous developments in super resolution fluorescence microscopy, enabling resolutions in the 10 to 100 nm using these techniques for dynamic studies remains a challenge as they require sample scanning or multiple image acquisitions. Localization super resolution methods, such as photo-activated localization microscopy (PALM) or stochastic optical reconstruction microscopy (STORM) [1], are based on the localization of individual, and supposedly separate, photoactivable fluorophores. Thousands of exposures are necessary to build the final high-resolution image, which strongly limits the usefulness of these techniques for live imaging. Stimulated emission depletion (STED) microscopy [2] provides nanometer resolution by reducing the diffraction spot size through stimulated emission. The reconstructed image is obtained by scanning the sample. Even though a recent publication has demonstrated the capability of parallelizing 2000 STED spots, 100 acquired images are still required to build the super-resolution image [3].

An alternative approach used in structured illumination microscopy (SIM) consists of illuminating and imaging the entire field of view and using a limited amount of raw data acquisitions. Illumination by a sinusoidal fringe pattern makes high spatial frequencies of the sample response, previously filtered, appear inside the support of the transfer function. An algorithm, after measurement, reconstructs a high-resolution image by demodulation and Wiener filtering.

SIM has provided [4, 5, 6] resolution down to 100 nm at a rate of $\approx$11 Hz and an $8\,\mu m^2$ field of view [7]. Recently, Betzig and coworkers [4] demonstrated nonlinear (with the presence of high-order modulation) SIM capability to reach a resolution of 62 nm in living cells. These performances paved the way towards high-resolution imaging in living samples where numerous biological processes require a subsecond temporal resolution. For living cells studies, TIRF illumination, where only a hundred nanometer thick slice of the sample is observed, is the most common in the literature [4, 5, 6] but forbids observation inside the cell. Although SIM can be used for 3D samples, it requires at least 15 images [4, 5, 6, 7] per optical section.

Furthermore, the reconstruction may generate artifacts that often appear as residual modulation, especially with thick samples, if sufficient attention is not paid to estimation of the modulation parameters [8, 9, 5]. Schaefer et al. [8] present a detailed and precise analysis of possible artifacts as well as an algorithm based on the *post* analysis of the result to minimize it. Wicker et al. [5] propose an algorithm based on the weighted cross-correlation between the central spectrum and the replication due to the modulation in Fourier space. On the contrary, the proposed approach does not depend on the possible patterns of artifacts but rather on data simulation and criterion fitting.

A major alternative to harmonic modulation in classical SIM is the Blind-SIM

approach [10, 11, 12, 13]. With the blind-SIM approach, a large number of images is generated with different illumination patterns created by speckle. The stochastic variation of the modulation allows to retrieve the high-resolution image. Very good results have been obtained and the technique is promising for robust reconstruction, despite the great number of images required. Ayuk et al. [12] adapted the technique to distorted modulations. However, currently, blind-SIM is not known to be able to perform nonlinear reconstruction. Unfortunately, Blind-SIM has not been validated on living biological samples yet. In our case, the distortion made by the sample does not seem to be the major limitation, and we have not further investigated this possibility, especially as we consider the existing problem of harmonic modulation.

A paper by Dong et al. [14] shows an algorithm to reconstruct the image with only four raw images. In comparison to the proposed approach, the reconstruction is defined by the algorithm without clear argumentation about which image is reconstructed. Criterion based methods instead model explicitly the information as presented in Sec. 3, and optimization algorithm is a tool to reach the minimum. As a consequence, the reconstruction is not really well-defined, and the stability of the alternative reconstruction step with the modulation will be hard to guarantee. Finally, contrary to the previous work [15], the proposed approach allows to simplify the optical setup and estimate the modulation parameters that were fixed in the above cited work. Another important distinction is the nature of the estimator and algorithm that were the posterior mean and Monte Carlo Markov Chain (MCMC) in [15].

This paper describes a new methodological framework and an original algorithm for 2D joint myopic estimation based on a criterion to minimize. Compared to previous techniques, the proposed approach focuses on harmonic modulation estimation within the original Gustafsson framework and small number of raw images. The image reconstructed cannot be better than that obtained with perfectly reconstructed classical SIM, or Blind-SIM or [15]. However, even if the image reconstructed compares directly to other approaches, the proposed approach introduces several improvements in the acquisition and reconstruction steps. The algorithm is derived from the estimator and criterion, whereas the usual approaches define their solutions by the algorithms. The first advantage of our proposed method, as already mentioned in [15], is that it only needs *four* raw images to reconstruct one super resolution image in linear SIM and eight raw images in nonlinear SIM. In addition, there is no limitation to the modulation pattern that can be arbitrary if a *parameterized pattern* is available (we study only harmonic patterns here). Second, our algorithm can estimate, along with the image, the modulation parameters that need to be estimated from the data, even with four images. The Shroff *et al* method [9] is the only other method, to the best of our knowledge, which can provide a phase parameter value when only four raw images are used. The cross-correlation proposed by Wicker et al. [5] provides a very good result but requires component separation and is presented for phase estimation only. The proposed algorithm also estimates frequency (or orientation) and contrast modulation, and with a much better accuracy than [9].

Finally, when fringes are generated by the interference of the first orders of a diffracted beam, our method allows the use of the zero (or more) order without constraints. In such cases, standard methods need at least fifteen images [5, 6, 16]. In contrast, our method only needs four images. This point is interesting for two more reasons: interference with zero order introduces a modulation at half the frequency of the main fringe, leading to a better SNR for parameter estimation, and the optical setup does not need to block the zero order.

The remainder of this paper is organized as follows. Section 2 shows the underlying principles of SIM and our proposed approach for image number reduction. Section 3 presents our data model and specific constraints on the image and modulation parameters allowing us to jointly estimate all the unknowns, that is all modulation parameters and pixels of the image. Section 4 is devoted to the joint optimizer algorithm and the final section 5 presents the results. Simulated examples are used to quantify the effectiveness of the proposed algorithm and results for experimental data with thick samples are presented.

2. Amplitude modulation and redundancy

The diffraction theory states that incoherent optical systems can be described in Fourier space with the optical transfer function (OTF) $H(\nu_x, \nu_y)$, which is theoretically equal to zero for all frequencies beyond the cut-off frequency ν_c. All information outside this bound is lost, and the underlying idea of SIM is injecting high frequencies inside the support of the OTF, owing to amplitude modulation, before filtering.

Let us denote the original image $f(x, y) \in L_2$ and its continuous Fourier transform $F(\nu_x, \nu_y)$. In SIM the illumination pattern is considered to be, up to an amplitude factor,

$$m(x,y) = 1 + \beta \cos\left(\pi k_x x + \pi k_y y + \phi\right) + \alpha \cos\left(2\pi k_x x + 2\pi k_y y + 2\phi\right). \quad (1)$$

The fringe at the $k_x/2$ frequency comes from the interference between the zero order and the two orders ± 1, and the fringe at the k_x frequency from the interference between the orders -1 and $+1$. The parameters α and β are contrast parameters between 0 and 1.

After modulation, the image in the Fourier space becomes

$$\begin{aligned} G(\nu_x, \nu_y) = F(\nu_x, \nu_y) + & \\ & \frac{\alpha}{2} e^{-2i\pi\phi} \left[F(\nu_x - k_x, \nu_y - k_y) + F(\nu_x + k_x, \nu_y + k_y)\right] + \\ & \qquad \frac{\beta}{2} e^{-i\pi\phi} \left[F\left(\nu_x - \frac{k_x}{2}, \nu_y - \frac{k_y}{2}\right) + F\left(\nu_x + \frac{k_x}{2}, \nu_y + \frac{k_y}{2}\right)\right]. \end{aligned} \quad (2)$$

and it is finally filtered by the OTF as $H(\nu_x, \nu_y)G(\nu_x, \nu_y)$. This result is illustrated in figure 1 for a 1D signal.

To the best of our knowledge, all existing approaches [4, 5, 6, 7, 9] state that U_0, U_1, and U_2 (see figure 1) are three unknowns that can be recovered by a linear combination of at least three or more modulated images with the same frequency modulation but different phases ϕ. Unfortunately, this approach completely disregards that redundancy is introduced, as mentioned by Heintzmann in [17]. It needs at least *five* images when the zero order is present for one grid orientation, and *fifteen* raw images for one 2D image.

Indeed, because of the Fourier Hermitian symmetry of the real image, U_0 and U_2 are complex conjugates of each other. Moreover, the part U_1 introduced by the zero order is not an additional unknown. In essence, irrespective of the sinusoidal pattern, only two unknowns can be present inside the raw data: the low frequencies between 0 and ν_c, labeled LF in figure 1 and the high frequencies between ν_c and $2\nu_c$, labeled HF. The raw data is then a mix of several copies of different frequencies in the range of these two unknowns (LF and HF). Based on this observation, we demonstrate that estimating the high-resolution image still involves the resolution of a linear system but with only *four* datasets for four unknowns in a full 2D dataset. Moreover, every extra modulation with frequency $|\boldsymbol{k}| < \nu_c$ does not introduce an extra unknown.

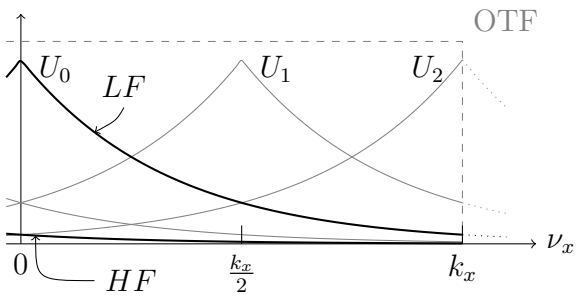


Figure 1: Illustration of amplitude modulation for SIM with zero order. The original spectrum LF is replicated around $\pm\frac{k_x}{2}$ and $\pm k_x$. The OTF removes all frequencies beyond ν_c, equal to k_x in this illustration. The unknown HF part comes from the $-k_x$ replication while the redundant spectra are in gray.

3. Data and image model

3.1. Forward model

The unknown image is modeled as N pixels collected in a vector $\boldsymbol{f} \in \mathbb{R}^N$. The pixel size is defined to be sufficient to represent all possible frequencies that can be reconstructed from the raw images. The image is illuminated by a light field $\boldsymbol{m}$ modeled as $\boldsymbol{M}\boldsymbol{f}$, where the matrix $\boldsymbol{M}$ is diagonal with $\text{diag}(\boldsymbol{M}) = \boldsymbol{m} \in \mathbb{R}^N$. The illumination is structured to produce modulations as described by equation (1). Therefore, the vector $\boldsymbol{m}$, or the diagonal of $\boldsymbol{M}$, depends on the five unknown parameters $\boldsymbol{\theta} = [k_x, k_y, \phi, \alpha, \beta]$ and corresponds to the fringe pattern.

The optics and microscope response are supposed to be linearly spatial invariant. The output is obtained by the convolution of the input image $\boldsymbol{Mf}$ with a known point spread function (PSF) thanks to simulation of the Airy disc or measurement. Being linear, the convolution operation can be written as $\boldsymbol{CMf}$, where $\boldsymbol{C}$ is a block circulant with circulant block (BCCB) convolution matrix. The BCCB nature of $\boldsymbol{C}$ makes the application of the convolution manageable in Fourier space

$$\boldsymbol{CM_\theta f} = \boldsymbol{F}^\dagger \boldsymbol{\Lambda_C F M_\theta f} \tag{3}$$

where $\boldsymbol{\Lambda_C} \in \mathbb{C}^{N\times N}$ is a diagonal complex matrix with the OTF as the diagonal and $\boldsymbol{F}, \boldsymbol{F}^\dagger$ are the unitary forward and reverse Fourier transform matrices, respectively. Equation (3) states that the convolution can be easily computed by filtering in the Fourier domain.

The third element is camera detector integration. We assume the detector as being squared with perfect integration on its surface. The camera integration is modeled as a convolution then sampling at the detector resolution. Next, camera convolution is integrated with the optic convolution, and only sampling remains. This operation is written as a matrix $\boldsymbol{S} \in \{0,1\}^{N\times M}$ with only one 1 per line, 0 otherwise. Concretely $\boldsymbol{S}$ is identity if $N = M$, or make a subsampling by a factor of two if $N = 2M$. In the latter case, the image resolution is doubled with respect to the raw data resolution, and the camera convolution is a 2×2 square response.

Finally, the full model writes

$$\boldsymbol{g}_i = \boldsymbol{SCM}_{\boldsymbol{\theta}_i}\boldsymbol{f} + \boldsymbol{n}_i = \boldsymbol{H}_i\boldsymbol{f} + \boldsymbol{n}_i \tag{4}$$

where $\boldsymbol{g}_i \in \mathbb{R}^M$ are i-th raw data and $\boldsymbol{n}_i \in \mathbb{R}^M$ an unknown noise. Then, I raw data are acquired, stacked, and the complete model is written as

$$\boldsymbol{g} = \boldsymbol{Hf} + \boldsymbol{n} \tag{5}$$

where

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{H}_1 & & \\ & \ddots & \\ & & \boldsymbol{H}_I \end{pmatrix} \begin{pmatrix} \boldsymbol{I} \\ \vdots \\ \boldsymbol{I} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{n} = \begin{pmatrix} \boldsymbol{n}_1 \\ \vdots \\ \boldsymbol{n}_I \end{pmatrix}. \tag{6}$$

However, the naive use of this forward model is ill posed because the least square solution of the system

$$\left(\boldsymbol{H}^t\boldsymbol{H}\right)\boldsymbol{f} = \boldsymbol{H}^t\boldsymbol{g}$$

may have several solutions (without additional choice), or is unstable otherwise [18]. This characteristic comes from the sub-sampling $\boldsymbol{S}$ and the convolution $\boldsymbol{C}$, and it is the main reason for the use of Wiener filtering in classical methods [16, 6]. We propose an alternative approach where the image and the modulation parameter models are defined jointly with the forward model owing to a *regularized criterion*.

3.2. Image model

To eliminate the ill-posed feature of the least square criterion, we introduce a regularization term leading to the mixed criterion

$$J(\boldsymbol{x}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{f}\|^2. \tag{7}$$

This kind of criterion is well known and has been widely studied and used in various fields and applications [19, 20, 21, 22, 23, 18]. The prior fidelity term $\|\boldsymbol{D}\boldsymbol{f}\|^2$ stabilizes the problem and further particular solutions. To counterbalance noise amplification arising from the data fidelity term, $\boldsymbol{D}$ is chosen as a differential operator or high-pass filter. We then choose the Laplacian second-order differential operator along the lines and columns implemented as the 2D impulse response

$$d = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}.$$

Moreover, we implement filtering in Fourier space $\boldsymbol{D} = \boldsymbol{F}^\dagger \boldsymbol{\Lambda_D} \boldsymbol{F}$, where $\boldsymbol{\Lambda_D}$ is diagonal and the corresponding high-pass filter. Owing to Parseval equality, the criterion can then be written as

$$J(\boldsymbol{x}) = \|\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}\|^2 + \lambda\|\boldsymbol{\Lambda_D}\mathring{\boldsymbol{f}}\|^2.$$

where $\mathring{\boldsymbol{f}} = \boldsymbol{F}\boldsymbol{f}$ is the discrete Fourier transform (DFT) of the image. The regularization parameter λ determines the balance between the data fidelity term, that tends to introduce high frequencies with noise amplification and instability, and the prior fidelity term that tends to favor image smoothness. Owing to this Tikhonov regularization, the criterion remains quadratic with a fast algorithm such as a conjugate gradient [24] to optimize it. Note that this prior corresponds to the classical Wiener filter. Other prior are possible, such as TV or $\ell_2 - \ell_1$, which would lead to a nonlinear estimator for better image reconstruction with the price of a more complex and slower algorithm.

The choice of the λ parameter value is a difficult question, and a significant part of the literature is dedicated to it [25, 21, 26, 27, 15, 28, 29]. In this work, the value is considered known and fixed by hand to reconstruct satisfactory images. This possibility relies on the Bayesian framework such as in [15].

3.3. Modulation parameters model

The modulation parameters must be estimated from the data jointly with the image. With I raw data acquired, $5 \times I$ parameters must be estimated. Nevertheless, little information is available about them except some interval constraints.

- Concerning the modulation frequencies $\boldsymbol{k}$, the nearest discrete frequencies from the DFT of data can be easily located [9] at indices (n, m). This allows stating that the true frequency $\boldsymbol{k}^* = (k_x^*, k_y^*)$ is inside the interval

$$k_x^* \in \left[\left(n - \frac{1}{2}\right)\Delta_x, \left(n + \frac{1}{2}\right)\Delta_x\right]$$

and

$$k_y^* \in \left[\left(m - \frac{1}{2}\right)\Delta_y, \left(m + \frac{1}{2}\right)\Delta_y\right]$$

where Δ_x, and Δ_y are the spectral resolution of the DFT.

- The phase value is unknown. Shroff *et al* [9] guess the phase from the data spectrum phase value at (n, m) frequency. These approaches assume that phases from other replications are negligible and that frequency modulation is exactly at the discrete frequency $(n\Delta_x, m\Delta_y)$. Unfortunately, such hypotheses are not available for 2D super resolution SIM since frequency modulation slightly change with experimental conditions, especially with significant noise. We consider here that the true phase value is located inside the interval

$$\phi^* \in [0, 2\pi].$$

- Knowledge about the contrast parameters is also limited. As for the phases, we only use physical information saying that the contrast parameters are inside the interval

$$\alpha^*, \beta^* \in [0, 1]^2.$$

3.4. Complete model

The solution to the problem of joint estimation of the image $\boldsymbol{f}$ and the full set $\boldsymbol{\theta} = [\boldsymbol{k}, \boldsymbol{\phi}, \boldsymbol{\alpha}, \boldsymbol{\beta}]$ of $5 \times I$ modulation parameters are chosen as the joint minimizer of constrained regularized least square criterion

$$\begin{aligned} \hat{\boldsymbol{f}}, \hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{f},\boldsymbol{\theta}} \|\boldsymbol{g} - \boldsymbol{H}_{\boldsymbol{\theta}}\boldsymbol{f}\|^2 + \lambda\|\boldsymbol{D}\boldsymbol{f}\|^2 \quad \text{subject to} \\ \boldsymbol{k} \in [\boldsymbol{k}_m, \boldsymbol{k}_M], \\ \boldsymbol{\phi} \in [0, 2\pi]^I, \\ \boldsymbol{\alpha}, \boldsymbol{\beta} \in [0, 1]^{2\times I}. \end{aligned} \tag{8}$$

The problem is well posed; the regularization factor $\lambda \boldsymbol{D}^t\boldsymbol{D}$ ensures the uniqueness of the solution, stabilizes the inversion, and circumvents noise amplification. The defined image and parameters solution correspond to the values that best reproduce the data $\boldsymbol{g}$ while satisfying the constraints. The criterion is, however, globally non-convex with several local minimizers. Sec. 4 is devoted to an optimisation algorithm that try to reach a satisfactory solution.

From this point, two main questions arise: First, how to compute the solution defined by equation (8)? Second, does this solution solve the initial problem of joint estimation of modulation parameters and image reconstruction ? The next two sections are devoted to the proposed algorithm that computes the solution and to showing results that illustrate the effectiveness of the proposition.

4. Myopic SIM image reconstruction

Several computational difficulties arise in resolving the problem defined by equation (8).

(i) The solution cannot be expressed explicitly. This point is addressed with an alternate block optimization algorithm on $\boldsymbol{\theta}$ and $\boldsymbol{f}$.

(ii) The criterion, w.r.t. the modulation parameters $\boldsymbol{\theta}$, is nonlinear because of the cosine and can have several local minima. Hopefully, the number of parameters is small and, as demonstrated here, global optimization is feasible.

(iii) The dimension, w.r.t. the image $\boldsymbol{f}$, is very large and non-stationary because of the presence of the forward model $\boldsymbol{H}$. Despite the fact that the regularized least square solution is explicit,

$$\hat{\boldsymbol{f}} = \left(\boldsymbol{H}^t\boldsymbol{H} + \lambda \boldsymbol{D}\boldsymbol{D}\right)^{-1} \boldsymbol{H}^t\boldsymbol{g}, \tag{9}$$

the problem is too large to invert the matrix and make $\boldsymbol{f}$ directly tractable with (9). The proposed solution relies on the well-known conjugate gradient algorithm to solve large linear systems.

To solve the joint problem, we propose an alternate minimization algorithm 1 that we describe in more detail in the next section. Nevertheless, the quality of the initial image $\boldsymbol{f}^{(0)}$ is an important point. The non-linearity and multi-modality of the criterion make the convergence of line 1.5 dependent on the initial point. In section 5.5.1, we propose a robust and easily feasible initial image.

Algorithm 1 SIM image reconstruction

1: **procedure** SIM($\boldsymbol{g}, \boldsymbol{H}, \boldsymbol{D}, \lambda, \boldsymbol{f}^{(0)}$)
2: $k \leftarrow 0$
3: **repeat**
4: $k \leftarrow k + 1$
5: $\boldsymbol{\theta}^{(k)} \leftarrow \arg\min_{\boldsymbol{\theta}} J\left(\boldsymbol{f}^{(k)}, \boldsymbol{\theta}\right)$
6: $\boldsymbol{f}^{(k)} \leftarrow \arg\min_{\boldsymbol{f}} J\left(\boldsymbol{f}, \boldsymbol{\theta}^{(k)}\right)$ ▷ Done with conjugate gradient
7: **until** Stopping criterion is met
8: **return** $\boldsymbol{f}^{(k)}$

4.1. Modulation parameters optimization

The optimization of the modulation parameters is not straightforward, even more so when only four raw data images are available. This step requires finding the parameters $\boldsymbol{\theta}$ that best reproduce the data with a fixed image $\boldsymbol{f}$ with respect to equation (8). However, minor simplifications are possible. First, the regularization term can be removed because it does not depend on $\boldsymbol{\theta}$. Second, the criterion can be split into one criterion for each raw data $\boldsymbol{g}_i$ because the modulation parameter model is independent

when $\boldsymbol{f}$ is fixed. Consequently, the line 5 of the algorithm 1 corresponds to I's parallel optimization of the constrained nonlinear least square criterion

$$\begin{aligned} J_i(\boldsymbol{\theta}_i) = \|\boldsymbol{g}_i - \boldsymbol{S}\boldsymbol{H}\boldsymbol{M}_{\boldsymbol{\theta}_i}\boldsymbol{f}\|^2 \quad \text{subject to} \\ \boldsymbol{k}_i \in [\boldsymbol{k}_m, \boldsymbol{k}_M], \quad \phi_i \in [0, 2\pi], \quad \alpha_i, \beta_i \in]0, 1[^2 \end{aligned} \tag{10}$$

with $\boldsymbol{\theta}_i = [k_{xi}, k_{y_i}, \phi_i, \alpha_i, \beta_i]$ and the diagonal elements of $\boldsymbol{M}_{\boldsymbol{\theta}_i}$, with $q \in [1, \ldots, N]$,

$$\begin{aligned} \boldsymbol{M}_i[q] = 1 + \beta_i \cos\left(\pi k_{x,i} x[q] + \pi k_{y,i} y[q] + \phi_i\right) + \\ \alpha_i \cos\left(2\pi k_{x,i} x[q] + 2\pi k_{y,i} y[q] + 2\phi_i\right). \end{aligned} \tag{11}$$

This criterion is only the least squares between a "simulator output" and the data, when the modulation parameters are varied.

The optimization scheme of criterion (10) has been established owing to practical observation. First, the contrast parameters (α, β) and cosine parameters $(\boldsymbol{k}, \phi)$ form two distinct families of parameters. The estimated value for the cosine parameters can be good enough even if the contrast parameter is wrong but with a significant value such as 0,5. On the contrary, if the cosine parameters are not good enough, the estimation of the contrast can be extremely low, yielding to 0 value. In other words, if the cosine shape is incorrectly estimated, the best residual is obtained without any modulation. These two observations led us to first estimate the cosine parameters with ad hoc values for the contrast in the first loop of algorithm 1, and to use the contrast parameter as an empirical diagnostic tool, as already suggested in [5, 6].

The cosine parameters are more difficult: the criterion is nonlinear because of the cosine, can have several local *minima*, and the three parameters (k_x, k_y, and ϕ) are strongly correlated. Therefore, we must exhaustively search the full volume in three directions using a global optimization algorithm

$$\widehat{\boldsymbol{\theta}} = \left\{\boldsymbol{\theta}^* \in \mathcal{S} \middle| \|\boldsymbol{g} - \boldsymbol{H}_{\boldsymbol{\theta}^*}\boldsymbol{f}\|^2 \leq \|\boldsymbol{g} - \boldsymbol{H}_{\boldsymbol{\theta}}\boldsymbol{f}\|^2, \forall \boldsymbol{\theta} \in \mathcal{S}\right\} \tag{12}$$

where $\mathcal{S}$ is a finite set of evaluated points inside $[\boldsymbol{k}_m, \boldsymbol{k}_M] \times [0, 2\pi]$. All other tested algorithms (nonlinear conjugate gradient, newton method, and Nealder-Mead for instance), most often local ones, were unable to find the global minimum necessary for good image reconstruction. The advantage of this algorithm is the parallel evaluation of each tested point in $\mathcal{S}$ and the criterion evaluation $J(\boldsymbol{\theta})$ is only required at the current point.

The contrast parameters lead to an easily solved linear least squares with the details explained in Appendix A.

4.2. Image optimization

Unlike the modulation parameters, the image optimization step (line 6 of algorithm 1) involves a large but linear and strictly convex problem. The solution equation (9) can

then be approximated owing to a preconditioned conjugate gradient algorithm that solves the system using the gradient vector

$$\nabla J(\boldsymbol{f}) = 2\boldsymbol{H}^t(\boldsymbol{H}\boldsymbol{f} - \boldsymbol{g}) + 2\lambda \boldsymbol{D}^t\boldsymbol{D}\boldsymbol{f}$$

that does not involve large matrix inversion. As this step is inside a larger alternate minimization strategy, this subpart can be stopped early. We do not describe in more detail the conjugate gradient because a significant amount of literature is devoted to it [24, 30]. Appendix B describes a preconditioning matrix that largely diminishes the number of iterations.

5. Results

This section shows the results of our *myopic SIM* approach for experimental and simulated data with only four raw images, with zero order present in the fringe ($\beta \neq 0$). We compare our results with the classical approach where at least 9 or 15 raw images are needed and show that we provide the same quality as we already demonstrated in [15]. With our method, we can also estimate the modulation parameters. We compare our results to Shroff [9], which is the sole method, to the best of our knowledge, that can estimate only the phase parameters with four images (modulation depth are fixed to the true value in the Shroff tests). This method estimates the frequencies modulation by locating the maximum value pixel, if visible, in Fourier space. Thus, the phase estimate is the phase value at this frequency component and therefore neglects wrong location and noise corruption. The amplitude is also estimated by ratio with the null frequency, corrected by the OTF value. This method is very fast but is not precise and inefficient when the frequency modulation is near the Abbe limit.

The algorithm is initialized with $\boldsymbol{f}^{(0)}$ being the wiener deconvolution [31] of a widefield (image obtained using a uniform illumination) image and $\alpha^{(0)} = \beta^{(0)} = 0{,}5$. The widefield can be a real one if available, or the mean of several modulated images. In the latter case, nine classical images are needed to avoid incorrect convergence, and thus, it is not advisable for our method. In our tests, a widefield image is available. The initial exhaustive grid search is performed with $10 \times 10 \times 40$ points for $(k_x, k_y, and\phi)$. The search interval is initialized with prior constraints as explained in section 3.3.3 and are reduced by a factor 0,9 after each loop for refinement. The implementation is done with Python and standard library (numpy, ...) and typical computing time is between 2 and 10 minutes.

5.1. Results on simulated data

The simulations are conducted using two test images: `mire`, and `boat`. For all simulated data, the acquisition scheme is

- one widefield image ($N = M = 256 \times 256$).

- three modulated images ($I = 3$) with respective grid orientation $\left[0, \frac{2\pi}{3}, -\frac{2\pi}{3}\right] + \frac{\pi}{4}$, with contrast $\alpha = 0{,}2$, $\beta = 0{,}8$, and modulation frequency $|\boldsymbol{k}| = 0{,}2$. Figure 2 shows the dataset for the `mire` test.
- an Airy theoretic OTF with reduced frequency cut $|\nu_c| \approx 0{,}26$
- white gaussian noise with different levels ($\gamma_{\boldsymbol{n}} = 10$ and $0{,}1$).
- the frequency modulation is 0.23 in reduced frequency.
- all images are in the same displayed with same dynamics with original image values between 0 and 256.

The algorithm settings are always the same as described in section 3 and the process takes a few minutes with a 256×256 pixels image on a 4×1.9 GHz CPU with 8 GB of memory. The tests were conducted with the images `mire`, and `boat`, and two noise levels with `mire` and `mire-hn`. All known methods [4, 5, 6, 32] rely on the cross-correlation of extracted spectra with the classical 9 or 15 images. They are thus inefficient with only four images. We compare the proposed *myopic SIM* method to Shroff *et al* [9], which is the sole method that can also provide phase values.

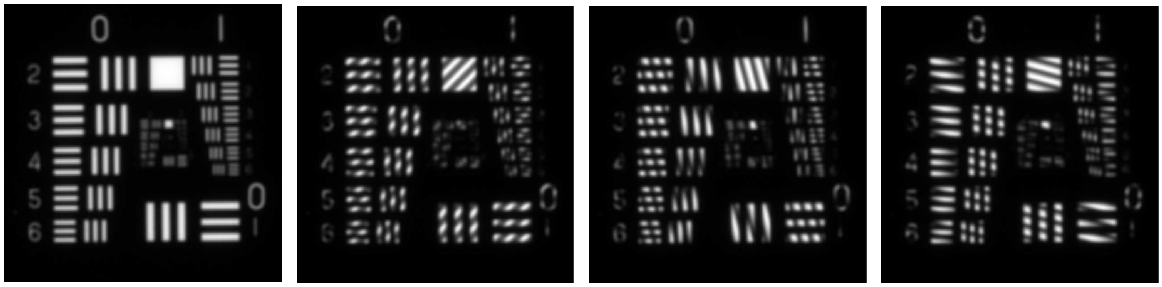
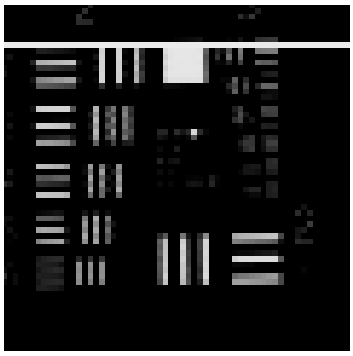


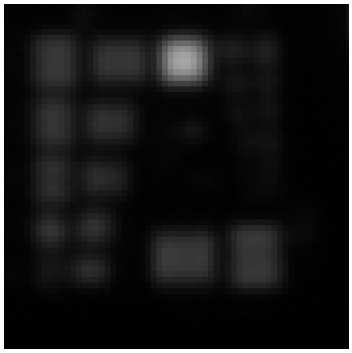
Figure 2: The four raw images used for the myopic SIM reconstruction in the `mire` test. The first image on the left is the widefield image, and the three other are modulated. The main visible fringes come from the interference between the zero order and the ± 1 order, with $k_x/2$ frequency.

The figure 3 shows a zoom of the `mire` results with 3a being the true image. The figure 3c is the reconstruction where modulation parameters are estimated with the [9] method. Artifacts are present and make the image difficult to analyze. The power spectral density (PSD, or squared absolute value of the Fourier transform) is illustrated in figure 4c. The PSD exhibits classical interference in SIM because of the wrong estimation of the modulation parameters. On the contrary, figures 3d and 4d show the results of our *myopic SIM* approach. The results show an image without visual artifacts and more high frequencies than widefield. The PSD figure 4d shows only very limited effect on high frequency. This effect is so small that no differences are visible in comparison with the reconstruction made with true parameter values.

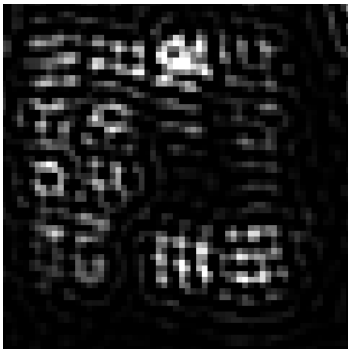
To appreciate the gain in resolution, the figure 5 shows a slice of the image for the widefield and the *myopic SIM* reconstruction. More details and high frequencies are visible.



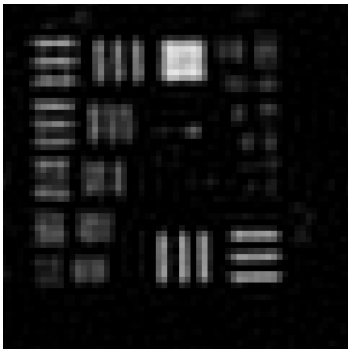
(a) True image



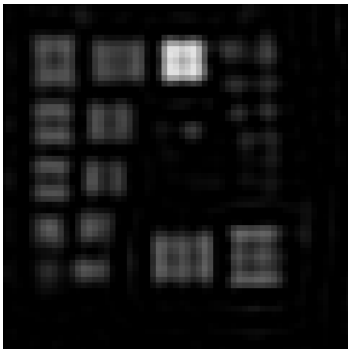
(b) Widefield



(c) SIM w. [9]

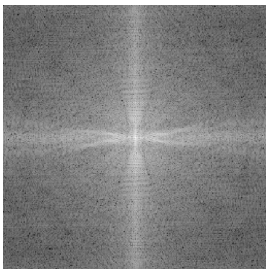


(d) Myopic SIM w. 4 images

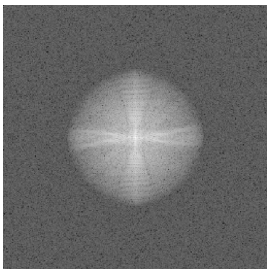


(e) Deconvolution

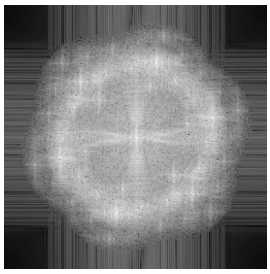
Figure 3: Zoom on results for the `mire` test. Figures 3a and 3b show the true image and widefield, respectively. Figure 3c shows the results with four images with modulation parameters estimated with [9]. Artifacts are strongly present. Figure 3d shows the results of our proposed method. Almost no artifacts are visible. Horizontal and vertical lines are caused by the periodic hypothesis of the DFT algorithm. Figure 3e is the deconvolution of the widefield (Figure 3b) as done in [15] and [31].



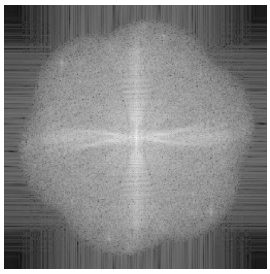
(a) True image



(b) Widefield

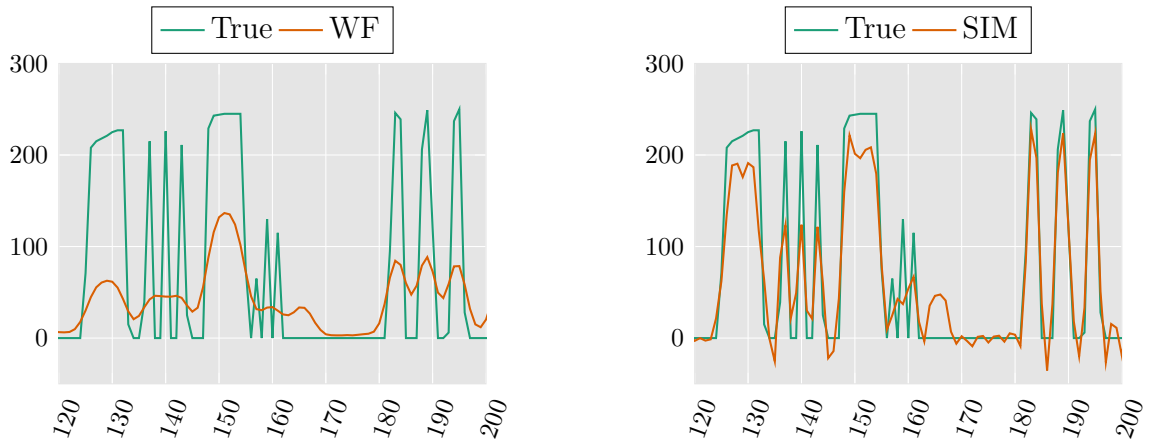


(c) SIM w. [9]



(d) Myopic SIM

Figure 4: Results for the `mire` test. Figures 4a and 4b show the true image and widefield module spectrum, respectively. Figure 4c shows the results with four images with modulation parameters estimated with [9]. Complex interference patterns are clearly visible. Figure 4d shows the results of our proposed method. Almost no artifacts are visible.



(a) The widefield (WF) comparison

(b) The myopic SIM comparison

Figure 5: Slice comparison for `mire` test between the true image, the widefield, and the myopic SIM reconstruction. Figure 5a shows the widefield and the true image. We see that the finer details are no longer visible such as around 140. Figure 5b shows that myopic SIM can estimate these originally lost frequencies.

Table 1 shows the error of the estimated parameter values w.r.t. the true values, for the Shroff *et al* [9] method and our *myopic SIM* method. We see that, both in absolute and relative error, our method has better estimation by between one and two orders of magnitude. This explains the quality of the reconstructed images. The level of required precision illustrates the difficulty of the problem.

Figure 6 shows the comparison between two reconstructions with two different levels of noise. The `mire-hn` test has been simulated with 10 times more noise (that is $\gamma_{\boldsymbol{n}} = 0.1$ and $\gamma_{\boldsymbol{n}} = 100$) and the myopic method is robust and can reconstruct a good high-resolution image without artifacts. However, the reconstructed image quality for `mire-hn` is obviously slightly lower than expected because more noise affects the measurements.

Finally, the *myopic SIM* method has been tested with `boat` image. The figure 7 shows the original image on the left, then the widefield, the SIM reconstruction with parameters estimated with [9], and our *myopic SIM* method. Again, the proposed approach is effective with good parameter estimation and good image reconstruction, without visible artifacts. The figure 8 shows a slice of the image and illustrates the gain in detail.

5.2. Results on experimental data

Tests have been conducted on real microscopic biological data. The setup has already been described in several references such as [33] or [15]. Images are obtained on living HeLa cells where mitochondria are labeled with Mitotracker green, and observed with a 96X 1,2 NA objective lens ($N = 1014 \times 1024$, $M = 512 \times 512$) and $I = 3$). The

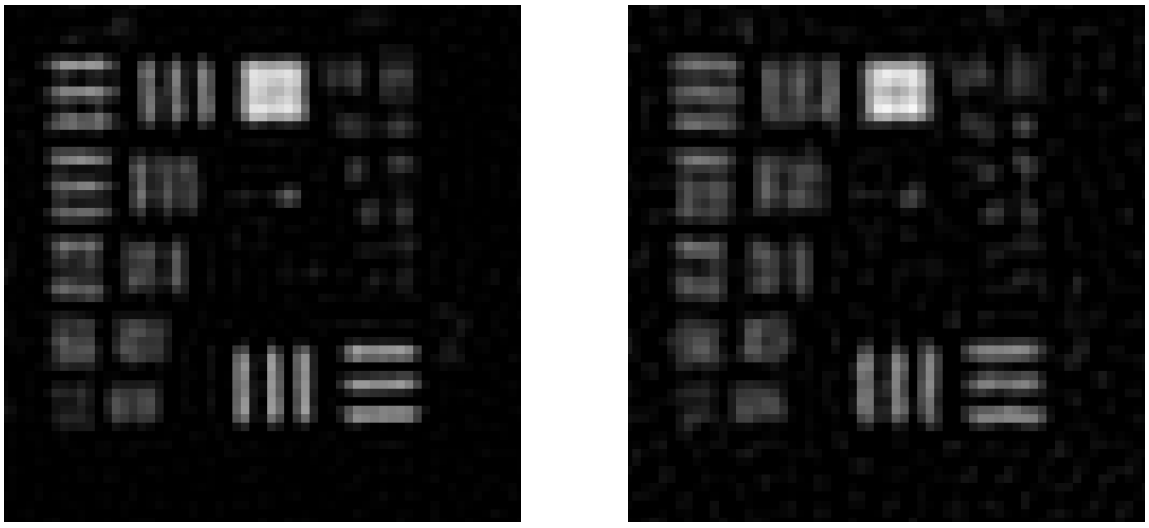
k_x abs. error		k_y abs. error		ϕ abs. error	
[9]	Myopic	[9]	Myopic	[9]	Myopic
$8 \cdot 10^{-4}$	$1 \cdot 10^{-5}$	$8 \cdot 10^{-4}$	$2 \cdot 10^{-5}$	1	$1 \cdot 10^{-2}$
$1 \cdot 10^{-3}$	$1 \cdot 10^{-6}$	$2 \cdot 10^{-3}$	$2 \cdot 10^{-6}$	$6 \cdot 10^{-1}$	$2 \cdot 10^{-3}$
$2 \cdot 10^{-3}$	$2 \cdot 10^{-5}$	$1 \cdot 10^{-3}$	$1 \cdot 10^{-6}$	$7 \cdot 10^{-1}$	$7 \cdot 10^{-3}$

(a) Absolute error of modulation parameters for mire test

k_x rel. error (%)		k_y rel. error (%)		ϕ rel. error (%)	
[9]	Myopic	[9]	Myopic	[9]	Myopic
$6 \cdot 10^{-1}$	$9 \cdot 10^{-3}$	$6 \cdot 10^{-1}$	$1 \cdot 10^{-2}$		
2	$2 \cdot 10^{-3}$	$9 \cdot 10^{-1}$	$1 \cdot 10^{-3}$	27	$1 \cdot 10^{-1}$
$9 \cdot 10^{-1}$	$1 \cdot 10^{-2}$	2	$2 \cdot 10^{-3}$	33	$4 \cdot 10^{-1}$

(b) Relative error of modulation parameters for mire test

Table 1: Absolute and relative errors of modulation parameters for the `mire` test and the three modulated images. Two estimations are compared: the method of Shroff *et al* [9] and the proposed *myopic*. The results show that our method improves the reconstruction of between one and two order of magnitude w.r.t. [9]

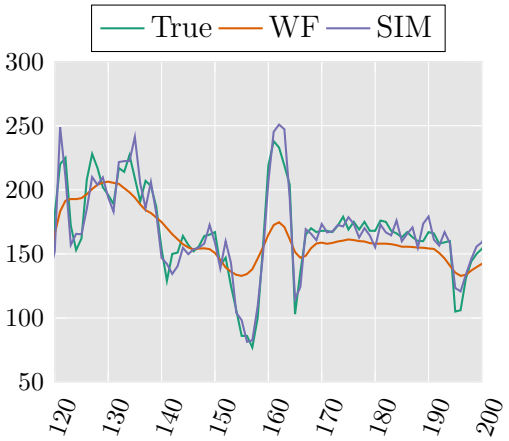


(a) `mire` reconstruction (b) `mire-hn` reconstruction

Figure 6: Image reconstruction results with the proposed myopic method for `mire` and `mire-hn` tests. Figures show good reconstruction results for `mire-hn` even with 10 times more noise.

(a) Boat (b) Widefield (c) SIM w. [9] (d) Myopic SIM

Figure 7: The result for `boat` tests.



(a) `boat` slice

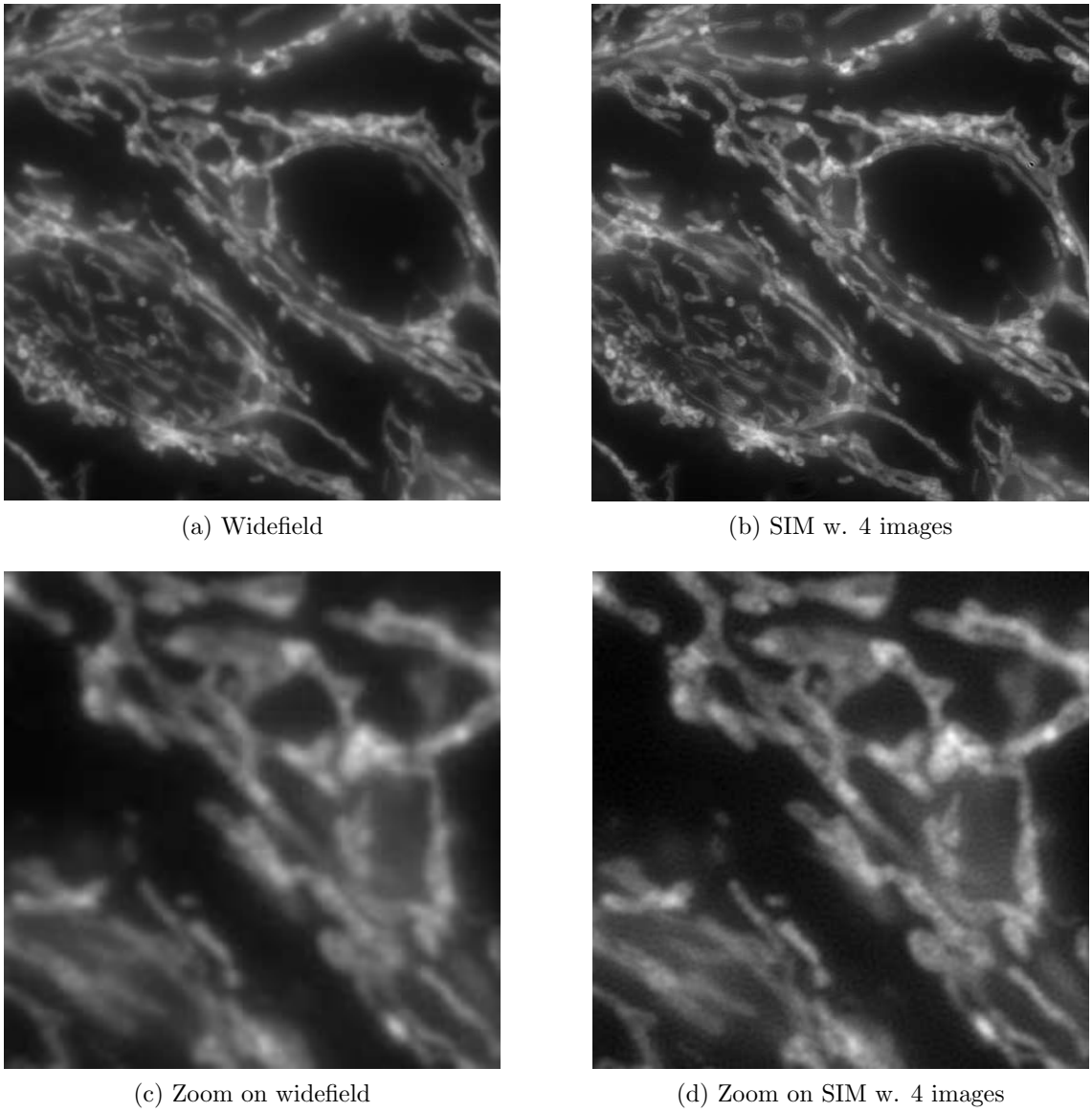
Figure 8: Slice comparison for `boat` test between the true image, the widefield, and the myopic SIM reconstruction. The results clearly show that high frequencies are recovered w.r.t. to the widefield.

illumination pattern was generated by three-beam interference and the SIM image was obtained from four raw images. The exposure time is 500 ms per raw image. Figures 9b and 9a, respectively, show the classical uniform image and the four-image ($I = 3$) SIM reconstruction using the myopic SIM method. The final obtained resolution is related to the optical setup and the chosen frequency modulation $\boldsymbol{k}$, but an improvement in spatial resolution is clearly visible without visual artifacts while we observe a thick sample.

6. Conclusion

We propose a new *myopic SIM* approach to jointly estimate the modulation parameters with significant precision with a high-resolution image, allowing reconstruction without visible artifacts even with only four raw images. The demonstration has been shown in simulated data and biological samples.

We demonstrated that inverse methodological data processing methods have a great impact on imaging capability in SIM by increasing the frame rate by a factor of 225 %



(a) Widefield

(b) SIM w. 4 images

(c) Zoom on widefield

(d) Zoom on SIM w. 4 images

Figure 9: Thick sample data with 96X lens, and marked mitochondria. The exposure time of the widefield is equivalent to that for the four SIM raw data, in the presence of the zero order for the modulation. Spatial resolution enhancement is clearly visible.

(from 9 to 4 images) up to 375 % (from 15 to 4 images), or by the reduction of the number of raw images needed for one reconstruction by a factor of two. Moreover, our estimator allows optical setup simplification by removing the need to block the zero order. Furthermore, the algorithm is independent of the fringe pattern.

Perspective applications are numerous, such as fast 3D imaging. One limitation is the nonstationary noise because of the Poisson nature of the flux. We also might consider out-of-focus background estimation with work first published in [33]. Strong out-of-focus light may modulation parameters to diffucult to estimate without proper handle. Hyper-

parameter estimation, as in [15], is a possibility with modulation parameters estimation done with a Metropolis-Hastings step for instance. Finally, the use of optimized or more specialized optimization algorithm and implementation could drastically reduce the computation time.

Acknowledgments

The authors would like to sincerely thank Stéphanie Bonneau, assistant professor of Université Pierre et Marie Curie (UPMC), associated with the Laboratoire Jean Perrin (CNRS – UMPC), for providing the HeLa cells.

Appendix A. Contrast estimation

The contrast estimation, for one raw image and without the constraints, leads to the minimization of

$$J(\theta_i) = \|\boldsymbol{g}_i - \boldsymbol{SCM}_{\theta_i}\boldsymbol{f}\|^2 \tag{A.1}$$

where $\theta_i = \alpha_i$ or β_i and the other parameters are fixed. The modulation matrix $\boldsymbol{M}_{\theta_i}$ can be written as

$$\boldsymbol{M}_{\theta_i} = \boldsymbol{I} + \alpha_i \boldsymbol{K}_{k_x,k_y,\phi} + \beta_i \boldsymbol{K}'_{k_x,k_y,\phi} \tag{A.2}$$

where $\boldsymbol{I}$ is the identity matrix and

$$\boldsymbol{K} = \cos(\pi k_x x + \pi k_y y + \phi) \tag{A.3}$$

$$\boldsymbol{K}' = \cos(2\pi k_x x + 2\pi k_y y + \phi). \tag{A.4}$$

This simple observation of linearity w.r.t. α or β allows obtaining an explicit minimizer of

$$\begin{aligned} J(\theta_i) &= \|\boldsymbol{g}_i - \boldsymbol{SCM}_{\theta_i}\boldsymbol{f}\|^2 \\ &= \|\boldsymbol{g}_i - \boldsymbol{SCf} - \alpha_i \boldsymbol{SCKf} - \beta_i \boldsymbol{SCK}'\boldsymbol{f}\|^2 \end{aligned} \tag{A.5}$$

and therefore,

$$\begin{gathered} \hat{\alpha}_i = \arg\min_{\alpha_i} J(\alpha_i) = \frac{\tilde{\boldsymbol{g}}_i{}^t \tilde{\boldsymbol{f}}}{\|\tilde{\boldsymbol{f}}\|^2} \\ \text{with} \quad \tilde{\boldsymbol{f}} = \boldsymbol{SCKf} \quad \text{and} \quad \tilde{\boldsymbol{g}}_i = \boldsymbol{g}_i - \boldsymbol{SCf} - \beta_i \boldsymbol{SCK}'\boldsymbol{f}, \end{gathered} \tag{A.6}$$

and

$$\begin{gathered} \hat{\beta}_i = \arg\min_{\beta_i} J(\beta_i) = \frac{\tilde{\boldsymbol{g}}_i{}^t \tilde{\boldsymbol{f}}}{\|\tilde{\boldsymbol{f}}\|^2} \\ \text{with} \quad \tilde{\boldsymbol{f}} = \boldsymbol{SCK}'\boldsymbol{f} \quad \text{and} \quad \tilde{\boldsymbol{g}}_i = \boldsymbol{g}_i - \boldsymbol{SCf} - \alpha_i \boldsymbol{SCKf}. \end{gathered} \tag{A.7}$$

Appendix B. Preconditionner for conjugate gradient

The goal of a preconditioning matrix is to be the best possible approximation of the Hessian matrix while being easily invertible. The Hessian matrix of problem (7) is

$$\boldsymbol{H}_e = \boldsymbol{H}^t\boldsymbol{H} + \lambda \boldsymbol{D}^t\boldsymbol{D}. \tag{B.1}$$

The $\boldsymbol{H}^t\boldsymbol{H}$ matrix is a block matrix composed of blocks such as

$$\boldsymbol{H}_c = \boldsymbol{M}_{\boldsymbol{\theta}_i}^t \boldsymbol{C}^t\boldsymbol{S}^t\boldsymbol{S}\boldsymbol{C}\boldsymbol{M}_{\boldsymbol{\theta}_i} \tag{B.2}$$

that forbids the inversion of the full matrix, as well as factorization in Fourier space. However, the $\boldsymbol{M}_{\boldsymbol{\theta}}$ matrix is diagonal with fluctuations around 1 coming from the modulation and can thus be approximated by the identity matrix. The Hessian matrix blocks can then be approximated by

$$\boldsymbol{H}_c \approx \boldsymbol{C}^t\boldsymbol{S}^t\boldsymbol{S}\boldsymbol{C}. \tag{B.3}$$

Moreover, it appears that the sub-sampling matrix $\boldsymbol{S}^t\boldsymbol{S}$ is composed of 1 or 0 on the diagonal. An additional approximation is feasible

$$\boldsymbol{H}_c \approx \boldsymbol{C}^t\boldsymbol{C} \tag{B.4}$$

where only the convolution part remains. Finally, the Hessian matrix can be approximated by

$$\boldsymbol{H}_e \approx \begin{pmatrix} \boldsymbol{C}^t\boldsymbol{C} & & \\ & \ddots & \\ & & \boldsymbol{C}^t\boldsymbol{C} \end{pmatrix} + \lambda \boldsymbol{D}^t\boldsymbol{D}, \tag{B.5}$$

easily factorized in Fourier space

$$\boldsymbol{H}_e \approx \boldsymbol{F}^\dagger \left(|\boldsymbol{\Lambda}_{\boldsymbol{C}}|^2 + \lambda \, |\boldsymbol{\Lambda}_{\boldsymbol{D}}|^2 \right) \boldsymbol{F} \tag{B.6}$$

where $\boldsymbol{F}$ is the matrix Fourier transform, with the property $\boldsymbol{F}^{-1} = \boldsymbol{F}^\dagger$, with$\boldsymbol{\Lambda}_{\boldsymbol{C}}$ and $\boldsymbol{\Lambda}_{\boldsymbol{D}}$ as complex diagonal matrices with the transfer function of the optics and the regularization operator, respectively, on the diagonal.

This approximation has an inverse easily tractable

$$\boldsymbol{M} \approx \boldsymbol{F}^\dagger \left(|\boldsymbol{\Lambda}_{\boldsymbol{C}}|^2 + \lambda \, |\boldsymbol{\Lambda}_{\boldsymbol{D}}|^2 \right)^{-1} \boldsymbol{F} \tag{B.7}$$

and can be used in a preconditioned conjugate algorithm. In practice, a strong improvement in the number of iterations to reach the optimum is observed.

References

References

[1] Vivien Marx. “Is super-resolution microscopy right for you?” In: *Nature Methods* 10.12 (2013), pp. 1157–1163. ISSN: 1548-7091.

[2] Stefan W. Hell. "Microscopy and its Focal Switch." In: *Nature methods* 6.1 (2009), pp. 24–32. ISSN: 1548-7105.

[3] Fabian Bergermann et al. "2000-fold parallelized dual-color STED fluorescence nanoscopy." EN. In: *Optics express* 23.1 (2015), pp. 211–23. ISSN: 1094-4087.

[4] D. Li et al. "Extended-resolution structured illumination imaging of endocytic and cytoskeletal dynamics". In: *Science* 349.6251 (2015). ISSN: 0036-8075.

[5] Kai Wicker et al. "Phase optimisation for structured illumination microscopy". In: *Optics Express* 21.2 (2013), p. 2032. ISSN: 1094-4087.

[6] Liisa M Hirvonen et al. "Structured illumination microscopy of a living cell." In: *European biophysics journal : EBJ* 38.6 (2009), pp. 807–12. ISSN: 1432-1017.

[7] Lin Shao et al. "Super-resolution 3D microscopy of live whole cells using structured illumination". In: *Nature Methods* october (2011), pp. 2–6. ISSN: 1548-7091.

[8] L. H. Schaefer, D. Schuster, and J. Schaffer. "Structured illumination microscopy: artefact analysis and reduction utilizing a parameter optimization approach". In: *Journal of Microscopy* 216.May (2004), pp. 165–174.

[9] Sapna A. Shroff, James R. Fienup, and David R. Williams. "Phase-shift Estimation in Sinusoidally Illuminated Images for Lateral Superresolution". In: *Journal of the Optical Society of America. sr. A.* 26.2 (2009), pp. 413–424.

[10] E. Mudry et al. "Structured illumination microscopy using unknown speckle patterns". In: *Nature Photonics* 6.5 (2012), pp. 312–315. ISSN: 1749-4885.

[11] Aurélie Jost et al. "Optical Sectioning and High Resolution in Single-Slice Structured Illumination Microscopy by Thick Slice Blind-SIM Reconstruction". In: *PLOS ONE* 10.7 (2015). Ed. by Vadim E. Degtyar. ISSN: 1932-6203.

[12] R. Ayuk et al. "Structured illumination fluorescence microscopy with distorted excitations using a filtered blind-SIM algorithm". In: *Optics Letters* 38.22 (2013), p. 4723. ISSN: 0146-9592.

[13] Simon Labouesse et al. "Blind fluorescence structured illumination microscopy: A new reconstruction strategy". In: *GRETSI*. Lyon, 2015, pp. 2–5.

[14] Siyuan Dong et al. "Resolution doubling with a reduced number of image acquisitions." EN. In: *Biomedical optics express* 6.8 (2015), pp. 2946–52. ISSN: 2156-7085.

[15] F. Orieux et al. "Bayesian Estimation for Optimized Structured Illumination Microscopy". In: *Image Processing, IEEE Trans. on* 21.2 (2012), pp. 601–614.

[16] M. G. L. Gustafsson. "Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy". In: *Journal of Microscopy* 198.2 (2000), pp. 82–87. ISSN: 00222720.

[17] Rainer Heintzmann. "Saturated patterned excitation microscopy with two-dimensional excitation patterns". In: *Elsevier* 34 (2003), pp. 283–291.

[18] Jérôme Idier, ed. *Bayesian Approach to Inverse Problems*. ISTE Ltd and John Wiley & Sons Inc., 2008.

[19] G. Demoment. "Image reconstruction and restoration: overview of common estimation structures and problems". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.12 (1989), pp. 2024–2036. ISSN: 00963518.

[20] F. Soulez, É. Thiébaut, and L. Denis. "Restoration of Hyperspectral Astronomical Data with Spectrally Varying Blur". In: *EAS Publications Series* 59 (2013), pp. 403–416. ISSN: 1633-4760.

[21] Clément Gilavert, Saïd Moussaoui, and Jérôme Idier. "Efficient Gaussian Sampling for Solving Large-Scale Inverse Problems Using MCMC". In: *IEEE Transactions on Signal Processing* 63.1 (2015), pp. 70–80.

[22] Christian Labat and Jérôme Idier. *Convergence of truncated half-quadratic algorithms using conjugate gradient*. Tech. rep. 2006.

[23] Jérôme Idier. "Convex Half-Quadratic Criteria and Interacting Auxiliary Variables for Image Restoration". In: *IEEE Transactions on Image Processing* 10.7 (2001), pp. 1001–1009.

[24] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2000. ISBN: 9780387303031.

[25] F. Orieux, O. Féron, and J.-F. Giovannelli. "Gradient scan gibbs sampler : an efficient high-dimensional sampler application in inverse problems". In: *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE Int. Conf. on*. 2015.

[26] Yuling Zheng. "Algorithmes bayésiens variationnels accélérés et applications aux problèmes inverses de grande taille". PhD thesis. Université Paris-Sud, 2014.

[27] Orieux, F. et al. "Estimating hyperparameters and instrument parameters in regularized inversion Illustration for Herschel/SPIRE map making". In: *A&A* 549.A83 (2013).

[28] F. Orieux, O. Féron, and J.-F Giovannelli. "Sampling high-dimensional Gaussian distributions for general linear inverse problems". In: *Signal Processing Letters, IEEE* 19.5 (2012), pp. 251–254.

[29] S. Derin Babacan, Rafael Molina, and Aggelos K. Katsaggelos. "Variational Bayesian Super Resolution". In: *IEEE Transactions on Image Processing* 20.4 (2011), pp. 984–999.

[30] Jonathan Richard Shewchuk. *An Introduction to the Conjugate Gradient Method Without the Agonizing Pain*. Tech. rep. Carnegie Mellon University, 1994.

[31] F. Orieux, J.-F. Giovannelli, and T. Rodet. "Bayesian estimation of regularization and PSF parameters for Wiener-Hunt deconvolution". In: *Journal of the Optical Society of America A* 27.7 (2010), pp. 1593–1607.

[32] Mats G. L. Gustafsson et al. "Three-dimensional resolution doubling in wide-field fluorescence microscopy by structured illumination." In: *Biophysical journal* 94.12 (2008), pp. 4957–70. ISSN: 1542-0086.

[33] P. Vermeulen et al. "Out of focus background subtraction for fast structured illumination super resolution microscopy of optically thick samples". In: *Journal of Microscopy* (2015).

Gradient Scan Gibbs Sampler: an efficient algorithm for high-dimensional Gaussian distributions

O. Féron*, F. Orieux and J.-F. Giovannelli

***Abstract*—This paper deals with Gibbs samplers that include high dimensional conditional Gaussian distributions. It proposes an efficient algorithm that avoids the high dimensional Gaussian sampling and relies on a random excursion along a small set of directions. The algorithm is proved to converge, *i.e.* the drawn samples are asymptotically distributed according to the target distribution. Our main motivation is in inverse problems related to general linear observation models and their solution in a hierarchical Bayesian framework implemented through sampling algorithms. It finds direct applications in semi-blind / unsupervised methods as well as in some non-Gaussian methods. The paper provides an illustration focused on the unsupervised estimation for super-resolution methods.**

I. Introduction

A. Context and problem statement

Gaussian distributions are common throughout signal and image processing, machine learning, statistics,... being convenient from both theoretical and numerical standpoints. Moreover, they are versatile enough to describe very diverse situations. Nevertheless, efficient sampling including these distributions is a cumbersome problem in high dimensions and this paper deals with this question.

Our main motivation is in inverse problems [1], [2] and the methodology resorts to a hierarchical Bayesian strategy, numerically implemented through Monte-Carlo Markov Chain algorithms and more specifically the Gibbs Sampler (GS). Indeed, consider the general linear direct model $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$, where $\mathbf{y}$, $\mathbf{n}$ and $\mathbf{x}$ are the observation, the noise and the unknown image and $\mathbf{A}$ is a given linear operator. Consider, again, two independent prior distributions for $\mathbf{n}$ and $\mathbf{x}$ that are Gaussian conditionally to a vector $\boldsymbol{\theta}$, namely the hyper-parameter vector. The estimation of both $\mathbf{x}$ and $\boldsymbol{\theta}$ relies on the sampling of the joint posterior $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$, and this is the core question of the paper. It commonly requires the handling of the high dimensional conditional posterior $p(\mathbf{x}|\boldsymbol{\theta}, \mathbf{y})$ that is Gaussian with given mean $\mathbf{m}$ and precision $\mathbf{Q}$.

The framework considered in this paper directly covers non-stationary and inhomogeneous Gaussian models for image and noise. The paper also has fallouts for non-Gaussian models based on conditionally Gaussian ones involving auxiliary / latent variables[1] (*e.g.*, location or scale mixtures of Gaussian) for edge preserving [3]–[5] and for sparse signals [6], [7]. It also includes other hierarchical models [8], [9] involving labels for inversion-segmentation. This framework also includes linear variant direct models and some non-linear direct models, based on conditional linear ones, *e.g.* bilinear or multilinear. In addition, it covers a majority of current inverse problems, *e.g.* unsupervised [5] and semi-blind [10], by including hyperparameters and acquisition parameters in the vector $\boldsymbol{\theta}$.

Large scale Gaussian distributions are also useful for Internet data processing, *e.g.* to model social networks and to develop recommender systems [11]. They are also widely used in epidemiology and disease mapping [12], [13] as they provide a simple way to include spatial correlations. The question is also in relation to spatial linear regression with (smooth) spatially varying parameters [14]. In these cases the question of efficient sampling including Gaussian distributions in high dimensions becomes crucial and it is all the more true in the "Big Data" context.

In the following we address the general problem of sampling from a joint distribution $p(\mathbf{x}, \boldsymbol{\theta})$ where the conditional distribution $p(\mathbf{x}|\boldsymbol{\theta})$ is a high-dimensional Gaussian distribution.

B. Existing approaches

The difficulty is directly related to handling the high-dimensional precision $\mathbf{Q}$. The factorization (Cholesky, square root,...), diagonalization and inversion of $\mathbf{Q}$ could be used but they are generally unfeasible in high dimensions due to both computational cost and memory footprint. Nevertheless, such solutions are practicable in two famous cases.

- If $\mathbf{Q}$ is circulant or circulant-block-circulant an efficient strategy [15], [16] relies on its diagonalization computed by FFT. More generally, an efficient strategy exists if $\mathbf{Q}$ is diagonalizable by a fast transform, *e.g.* discrete cosine transform for Neumann boundary conditions [17], [18].
- When $\mathbf{Q}$ is sparse, a possible strategy [13], [19], [20] relies on a Cholesky decomposition and a linear system resolution. Another strategy is a GS [21] that simultaneously updates large blocks of variables.

O. Féron is with EDF Research & Developments, 92140 Clamart, France and with Univ. Paris Dauphine, FiME, 75116 Paris, France, olivier-2.feron@edf.fr. F. Orieux is with Univ. Paris-Sud 11, L2S, UMR 8506, 91190 Gif-sur-Yvette, France, orieux@l2s.centralesupelec.fr. J.-F. Giovannelli is with Univ. Bordeaux, IMS, UMR 5218, F-33400 Talence, France, Giova@IMS-Bordeaux.fr.

[1] It is based on the fact that for a couple of random variables (U, V), the conditional law for $U|V$ is Gaussian and the marginal law for U is non-Gaussian. A famous example is a Gaussian variable with precision under a Gamma distribution: the resulting marginal follow a Student distribution.

In order to address more general cases, solutions founded on iterative algorithms for objective optimization or linear system resolution have recently been proposed.

1) An efficient algorithm has been proposed by several authors [6], [17], [18], [22], [23] (previously used in applications [8], [10]). It is founded on a Perturbation-Optimization (PO) principle: adequate stochastic perturbation of a quadratic criterion and optimization of the perturbed criterion. However, in order to obtain a sample from the right distribution, an exact optimization is needed, but in practice an empirical truncation of the iterations is implemented, leading to an approximate sample. [24] introduces a Metropolis step in order to asymptotically retrieve an exact sample and then to ensure, in a global MCMC procedure, the convergence to the correct invariant distribution.
2) In [25], [26] the authors propose a Conjugate Direction Sampler (CDS) based on two crucial properties: (i) a Gaussian distribution admits Gaussian conditional distributions and (ii) a set of mutually conjugate directions w.r.t. $\mathbf{Q}$ is available. The key point of the algorithm is to sample along these mutually conjugate directions instead of optimizing as in the classical Conjugate Gradient optimization algorithm.

In the first case, the only constraint on $\mathbf{Q}$ is that a sample from $\mathcal{N}(0, \mathbf{Q})$ must be accessible, which is often the case in inverse problem applications. In the second case, $\mathbf{Q}$ must have only distinct eigenvalues to make the CDS give an exact sample. Otherwise it leads to an approximate sample as described in [26].

The proposed algorithm uses the same approach as the CDS and extends the efficiency to, theoretically, any matrix $\mathbf{Q}$.

C. Contribution

The existing methods described above and the proposed one are both founded on a Gibbs sampler. However, the existing ones attempt to sample the high dimensional Gaussian component $\mathbf{x} \in \mathbb{R}^N$ whereas the proposed method does not. Our main contribution is to avoid the high dimensional sampling and only requires small dimensional sampling. More precisely, given a subspace $D \subset \mathbb{R}^N$, the objective is to sample the sub-component of $\mathbf{x}$ according to the subspace D. It must be sampled under the appropriate conditional distribution $\pi(\mathbf{x}_D|\mathbf{x}_{\setminus D}, \boldsymbol{\theta})$, with the decomposition $\mathbf{x} = (\mathbf{x}_D, \mathbf{x}_{\setminus D})$. The algorithm takes advantage of the ease of calculating the conditional pdf of a multivariate Gaussian distribution, when D is appropriately built, as explained in section II. These ideas are strongly related to other existing works.

- If the subset D is composed of only one direction in the canonical coordinates, the algorithm amounts to a pixel-by-pixel GS [3].
- The marginal chain $\mathbf{x}^{(t)}$ can also be viewed as the one produced by a specific random scan sampler [27]–[29]. The random scans are related to the random choice of D, depending on the current value $\boldsymbol{\theta}^{(t)}$.
- Other algorithms based on optimization principles [26], [30] aim at producing a complete optimization. On the other hand, in essence, the proposed approach only requires a few steps of the optimization process.
- A similar idea is at work in Hamiltonian (or Langevin) Monte Carlo [31]–[34] (see also [35]): the proposed distribution takes advantage of an ascent direction of the target to increase the acceptation probability. Here, the exact distribution is sampled, so the proposal is always accepted.

However, to our knowledge, the proposed algorithm does not directly join the class of existing strategies. One contribution of this paper is to give sufficient assumptions for convergence, *i.e.* the samples are asymptotically distributed according to the joint pdf $p(\mathbf{x}, \boldsymbol{\theta})$.

D. Outline

Subsequently, Section II presents the proposed algorithm and section III gives an illustration through an academic problem in super-resolution. Section IV presents conclusions and perspectives.

II. Gradient Scan Gibbs Sampler

In this section we describe the proposed algorithm: a GS with a high dimensional conditional Gaussian distribution. The objective is to generate samples from a joint distribution $p(\mathbf{x}, \boldsymbol{\theta})$, where $\mathbf{x} \in \mathbb{R}^N$ is highly dimensional and $p(\mathbf{x}|\boldsymbol{\theta})$ is a Gaussian distribution $\mathcal{N}(\mathbf{m}_{\boldsymbol{\theta}}, \mathbf{Q}_{\boldsymbol{\theta}}^{-1})$:

$$p(\mathbf{x}|\boldsymbol{\theta}) = (2\pi)^{-N/2} (\det \mathbf{Q}_{\boldsymbol{\theta}})^{1/2} \exp -J_{\boldsymbol{\theta}}(\mathbf{x}) \tag{1}$$

with the potential $J_{\boldsymbol{\theta}}$ defined as:

$$J_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{2} (\mathbf{x} - \mathbf{m}_{\boldsymbol{\theta}})^{\mathrm{t}} \mathbf{Q}_{\boldsymbol{\theta}} (\mathbf{x} - \mathbf{m}_{\boldsymbol{\theta}}). \tag{2}$$

All the other variables of the problem are grouped into $\boldsymbol{\theta} \in \Theta$ and we assume that the sampling from $p(\boldsymbol{\theta}|\mathbf{x})$ is tractable (directly or with several steps of the GS, including Metropolis-Hastings steps).

A. Preliminary results

This section presents classical definitions and results, mostly based on [25], needed to provide convergence proof and links between matrix factorization and optimization / sampling procedures.

Definition 1. *Consider $\mathbf{Q}$ a $N \times N$ symmetric definite positive matrix. A set $\{\mathbf{d}_n, n = 1, \ldots, N\}$ of non-zero vectors in $\mathbb{R}^N$ such that: $\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{d}_m = 0$ for $n, m = 1, \ldots, N$, $n \neq m$ is said mutually conjugate w.r.t. $\mathbf{Q}$.* △

A mutually conjugate set $\{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$ w.r.t. $\mathbf{Q}$ is a basis of $\mathbb{R}^N$, then, for all $\mathbf{x} \in \mathbb{R}^N$:

$$\mathbf{x} = \sum_{n=1}^{N} \alpha_n \mathbf{d}_n \quad \text{with} \quad \alpha_n = \frac{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{x}}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{d}_n}.$$

So, if $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{Q}^{-1})$ is a Gaussian random vector with mean $\mathbf{m}$ and precision $\mathbf{Q}$, then the α_n are also Gaussian:

$$\alpha_n \sim \mathcal{N}\left(\frac{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{m}}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{d}_n} \; ; \; \frac{1}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{d}_n} \right) \tag{3}$$

and reciprocally if the α_n are distributed under (3) then $\mathbf{x} \sim \mathcal{N}(\mathbf{m}, \mathbf{Q}^{-1})$.

In particular, let $\mathbf{x}^0 \in \mathbb{R}^N$ be a "current" point and $\mathbf{d}_1 \in \mathbb{R}^N$ a given "direction". One can find $\mathbf{d}_2, \ldots, \mathbf{d}_N$ such that $\{\mathbf{d}_1, \ldots, \mathbf{d}_N\}$ is mutually conjugate w.r.t. $\mathbf{Q}$ and $\mathbf{x}^0$ writes:

$$\mathbf{x}^0 = \sum_{n=1}^{N} \alpha_n^0 \, \mathbf{d}_n.$$

Consider now the N_D-dimensional subset

$$D(\mathbf{x}^0) = \left\{ \sum_{n=1}^{N} \alpha_n \mathbf{d}_n, \alpha_n \in \mathbb{R},\ n \leq N_D, \alpha_n = \alpha_n^0,\ n > N_D \right\}$$
$$= \left\{ \mathbf{x}^0 + \sum_{n=1}^{N_D} (\alpha_n - \alpha_n^0)\mathbf{d}_n,\ (\alpha_1, \ldots, \alpha_{N_D}) \in \mathbb{R}^{N_D} \right\}$$

We are interested in the conditional pdf $p(\mathbf{x}|\mathbf{x} \in D(\mathbf{x}^0))$. The following result and its proof can be found in [25].

Proposition 1. *A sample $\widetilde{\mathbf{x}}$ according to $p(\mathbf{x}|\mathbf{x} \in D(\mathbf{x}^0))$ can be obtained by:*

1) sample independently the set $(\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_{N_D})$ with:

$$\widetilde{\alpha}_n \sim \mathcal{N}\left(\frac{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q}(\mathbf{x}^0 - \mathbf{m})}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{d}_n} \ ; \ \frac{1}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q} \mathbf{d}_n} \right), n = 1, \ldots, N_D$$

2) compute $\widetilde{\mathbf{x}} = \mathbf{x}^0 - \sum_{n=1}^{N_D} \widetilde{\alpha}_n \, \mathbf{d}_n$

B. Gradient Scan Gibbs Sampler (GSGS)

In the following we propose a GS in order to sample the joint probability $p(\mathbf{x}, \boldsymbol{\theta})$. The principle is to sample, at each iteration of the GS, only N_D directions of $\mathbf{x}$ instead of sampling the whole high dimensional variable. The chosen first direction of the set D will be the gradient of the potential of $p(\mathbf{x}|\boldsymbol{\theta})$, with a stochastic perturbation to ensure, in the general case, the convergence of the resulting Markov chain. The following directions are chosen so as to get a mutually conjugate subset with respect to the precision of $p(\mathbf{x}|\boldsymbol{\theta})$.

We call our proposed algorithm the Gradient Scan Gibbs Sampler (GSGS) which is described by Algorithm 1. In this algorithm the chosen first sampling direction $\mathbf{d}_1$ is given by the gradient of the potential of $p(\mathbf{x}|\boldsymbol{\theta})$, with an additional random perturbation $\tilde{\boldsymbol{\varepsilon}}$ that follows a probability density $p(\boldsymbol{\varepsilon})$. In fact, we expect the gradient to be a good direction towards regions of high probabilities. Also, the gradient is easily computable and so gives an easy rule to sample from any current point $\mathbf{x}$. Moreover, the other conjugate directions are iteratively computable as described in the Conjugate Direction Sampling (CDS) algorithm [25] used to get an approximated sample from a Gaussian distribution. In fact, the GSGS is embedding steps of the CDS in a global GS.

The objective is now to study the convergence properties of the GSGS. We begin with two classical results.

- If the Markov chain is aperiodic, $\phi-$irreducible for some nonzero measure ϕ[2], and has an invariant probability π, then it converges to π from π-almost every starting point (cf. Theorem 4.4 of [36]).
- Moreover, if the Markov chain is Harris recurrent, then it converges to π from all starting point [36], [37].

[2]In all the paper we will consider ϕ as the Lebesgue measure and we will omit it for simplicity.

Algorithm 1 : Gradient scan Gibbs sampler (GSGS).

Define an initial point $\mathbf{x}^{(0)}$, a number N_D and a stopping criterion. Iterate .

1: sample $\boldsymbol{\theta}^{(t)} \sim p(\boldsymbol{\theta}|\mathbf{x}^{(t-1)})$

2: set $\mathbf{Q}_t = \mathbf{Q}_{\boldsymbol{\theta}^{(t)}}$ and $\mathbf{m}_t = \mathbf{m}_{\boldsymbol{\theta}^{(t)}}$, and compute the gradient $\mathbf{g} = \nabla J_{\boldsymbol{\theta}}(\mathbf{x}^{(t-1)}) = \mathbf{Q}_t(\mathbf{x}^{(t-1)} - \mathbf{m}_t)$

3: sample a perturbation $\tilde{\boldsymbol{\varepsilon}} \sim p(\boldsymbol{\varepsilon})$

4: compute a set of N_D mutually conjugate directions $(\mathbf{d}_1, \ldots, \mathbf{d}_{N_D})$ w.r.t. $\mathbf{Q}_t$ such that

$$\mathbf{d}_1 = \mathbf{g} + \tilde{\boldsymbol{\varepsilon}}$$

5: sample independently the set $(\widetilde{\alpha}_1, \ldots, \widetilde{\alpha}_{N_D})$ with:

$$\widetilde{\alpha}_n \sim \mathcal{N}\left(\frac{\mathbf{d}_n^{\mathrm{t}} \mathbf{g}}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q}_t \mathbf{d}_n} \ ; \ \frac{1}{\mathbf{d}_n^{\mathrm{t}} \mathbf{Q}_t \mathbf{d}_n} \right), \ n \leq N_D$$

6: compute $\mathbf{x}^{(t)} = \mathbf{x}^{(t-1)} - \sum_{n=1}^{N_D} \widetilde{\alpha}_n \, \mathbf{d}_n$

7: $t \leftarrow t + 1$.

until the stopping criterion is reached.

The Harris recurrence of GS, or more generally Metropolis-within-Gibbs samplers is well studied in [37]. In particular, the Theorem 12 and Corollary 13 of [37] ensures that if the Markov chain produced by the GSGS is irreducible then it is Harris recurrent. Consequently, in the following we focus on showing that the Markov chain is aperiodic, irreducible and with stationary distribution $p(\mathbf{x}, \boldsymbol{\theta})$.

It is trivial to see that the Markov chain $(\mathbf{x}^{(t)}, \boldsymbol{\theta}^{(t)})_{t \geq 0}$, produced by the GSGS, is aperiodic since for any non-negligible subset $A \in \mathbb{R}^N$ including $\mathbf{x}^{(t-1)}$, $\mathbb{P}(\mathbf{x}^{(t)} \in A) > 0$. The existence of an invariant probability and the irreducibility can be shown by thinking of a random scan GS for the marginal component $(\mathbf{x}^{(t)})_{t \geq 0}$.

Proposition 2. *The Markov chain produced by Algorithm 1 admits $p(\mathbf{x}, \boldsymbol{\theta})$ as an invariant distribution, even without perturbations of the gradient direction (i.e. $\tilde{\boldsymbol{\varepsilon}} = 0$).*
Moreover, if the density $p(\boldsymbol{\varepsilon})$ is supported on $\mathbb{R}^N$, the Markov chain produced by Algorithm 1 is irreducible, and therefore its law converges to $p(\mathbf{x}, \boldsymbol{\theta})$.

Proof. see appendix A. □

Proposition 2 then shows that the joint probability $p(\mathbf{x}, \boldsymbol{\theta})$ remains an invariant distribution in the limit case where the first direction $\mathbf{d}_1$ is exactly the gradient of $p(\mathbf{x}|\boldsymbol{\theta})$, without random perturbation. However the perturbation is needed to ensure the irreducibility (and then the convergence) of the chain.

If the gradient is not perturbed, the mutually conjugate set D is then given by a deterministic function of $\boldsymbol{\theta}^{(t)}$ and

$\mathbf{x}^{(t-1)}$. In this case, we need more assumptions to ensure the Markov chain to be irreducible. For example, we can have the following result.

Proposition 3. *Suppose the following conditions are satisfied:*

H-1 The function $\boldsymbol{\theta} \mapsto \mathbf{Q}_{\boldsymbol{\theta}}$ *is continuous*

H-2 $\forall(\mathbf{x}, \boldsymbol{\theta}) \in \mathbb{R}^N \times \Theta$ *and* $\forall r > 0$, $\mathbb{P}(\mathcal{B}(\boldsymbol{\theta}, r)|\mathbf{x}) > 0$, *with* $\mathcal{B}(\boldsymbol{\theta}, r)$ *the ball in* Θ, *centered in* $\boldsymbol{\theta}$, *of radius* r.

H-3 $\forall \mathbf{x} \in \mathbb{R}^N$, $\exists \boldsymbol{\theta} \in \Theta$ *such as:*

H-3.1 $\mathbf{Q}_{\boldsymbol{\theta}}$ *has* N *distinct eigenvalues,*

H-3.2 $\mathbf{x} - \mathbf{m}_{\boldsymbol{\theta}}$ *is not orthogonal to any eigenvector of* $\mathbf{Q}_{\boldsymbol{\theta}}$,

Then the Markov chain produced by Algorithm 1 without the perturbation step 3 ($\tilde{\boldsymbol{\varepsilon}} = 0$*) is irreducible.*

Proof. see appendix B □

The conditions described in Proposition 3 are very restrictive and, in particular, condition H-3.1 is difficult, if not impossible, to prove in practice. This condition ensures that every non-negligible subset of $\mathbb{R}^N$ can be reached with a non-zero probability. It can be interpreted in the framework of Krylov spaces as in [26]. For example, if there is t such as the Krylov space

$$\mathcal{K}^N(\mathbf{Q}_{\boldsymbol{\theta}^{(t)}}, \mathbf{x}^{(t)}) := \text{span}\left(\mathbf{x}^{(t)}, \mathbf{Q}_{\boldsymbol{\theta}^{(t)}}\mathbf{x}^{(t)}, \ldots, \mathbf{Q}^N_{\boldsymbol{\theta}^{(t)}}\mathbf{x}^{(t)}\right)$$

is of rank N then the Markov chain is irreducible. This condition can be weakened in our case because the Gaussian parameters $\mathbf{m}_{\boldsymbol{\theta}^{(t)}}$ and $\mathbf{Q}_{\boldsymbol{\theta}^{(t)}}$ are changing since $\boldsymbol{\theta}$ is changing at each iteration of the GS. Therefore a sufficient condition to ensure the irreducibility of the chain can be expressed as follows:

Proposition 4. *If there is* $T > N$ *such as the union of Krylov spaces*

$$\cup_{t=1}^{T} \mathcal{K}^N(\mathbf{Q}^{(t)}_{\boldsymbol{\theta}}, \mathbf{x}^{(t)}) \cup \mathcal{K}^N(\mathbf{Q}^{(t)}_{\boldsymbol{\theta}}, \mathbf{m}^{(t)}_{\boldsymbol{\theta}})$$

is of rank N *then the Markov chain built by the GSGS without perturbation of the gradient is irreducible.*

Proof. The condition implies that for any non-negligible subset $A \subset \mathbb{R}^N$, $\mathbb{P}\left(\mathbf{x}^{(T)} \in A|\mathbf{x}^{(0)}\right) > 0$, which ensures the irreducibility. □

The issue of determining general conditions, as in Proposition 3, is an open problem at this time. The fact that the condition described in Proposition 4 is satisfied, highly depends on the model's characteristics. That is why the GSGS (with the random perturbation step 3) is the one that ensures, in all cases, the convergence of the Markov chain to the joint distribution $p(\mathbf{x}, \boldsymbol{\theta})$.

The above results do not allow us to get any convergence rate of the Markov chain. The latter is, in fact, very important to ensure in practice the efficiency of the estimators produced by simulations in finite time. In particular, the geometric ergodicity [38] is a very well known property that gives a Central Limit Theorem and ensures the Markov chain to quickly converge and give estimations of standard errors. However the Algorithm 1 aims to be general while the precise study of geometric convergence (especially to quantify the convergence rate) would need to specify the distributions on the parameters $\boldsymbol{\theta}$ and on the perturbation $\boldsymbol{\varepsilon}$. At this time, only weak assumptions are considered on these probabilities and the next section discusses the different choices of $p(\boldsymbol{\varepsilon})$ from a feasibility point of view.

C. Choice of $p(\boldsymbol{\varepsilon})$

As previously specified, the only condition to ensure the convergence of the GSGS in the general case, is to choose a distribution $p(\boldsymbol{\varepsilon})$ supported in $\mathbb{R}^N$. In practice we also expect a sample from $p(\boldsymbol{\varepsilon})$ to be easily accessible. A natural choice is the Gaussian iid distribution $\mathcal{N}(0, \mathbf{I}_N)$, $\mathbf{I}_N$ being the $N \times N$ identity matrix. This was already studied in [39] in the case of only sampling from a Gaussian distribution $p(\mathbf{x})$ and where results are shown in small dimensions.

Our empirical studies in high dimension (one example is shown in section III) incited us to choose the Gaussian distribution $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$, when it is possible. The sampling from this distribution may actually be easily computable, provided that $\mathbf{Q}_{\boldsymbol{\theta}}$ has, for example, the specific factorization form described in [30]:

$$\mathbf{Q}_{\boldsymbol{\theta}} = \sum_{k=1}^{K} \mathbf{M}_k^{\text{t}} \mathbf{R}_k^{-1} \mathbf{M}_k$$

In this case, the sampling from $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$ is easily computable by using the Perturbation Optimization (PO) algorithm [30]. The latter consists in (i) randomly modifying the potential $J_{\boldsymbol{\theta}}(\mathbf{x})$ to get a perturbed potential $\tilde{J}_{\boldsymbol{\theta}}$ and (ii) optimizing $\tilde{J}_{\boldsymbol{\theta}}$. The first step of this optimization procedure consists in computing the gradient $\nabla \tilde{J}_{\boldsymbol{\theta}}$ and it is trivial to show that it can be decomposed: $\nabla \tilde{J}_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla J_{\boldsymbol{\theta}}(\mathbf{x}) + \boldsymbol{\varepsilon}$, with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$. Therefore, the perturbed gradient $\mathbf{d}_1$ of the GSGS, with a random perturbation $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$, can be obtained by using the PO algorithm truncated to one step of the optimization procedure.

Although, at this time, this choice is empirical we may have some intuition to recommend, when it is possible, the distribution $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$. The first direction $\mathbf{d}_1$ is related to the gradient of $J_{\boldsymbol{\theta}}$, in accordance with the objective to get a direction towards regions of high probability. This gradient is mostly driven by the highest eigenvalues of $\mathbf{Q}_{\boldsymbol{\theta}}$. The perturbation $\boldsymbol{\varepsilon}$ is only needed to ensure the GSGS convergence, but the objective is to keep a direction towards high probability regions. The sampling from $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$ seems to be a good compromise: it gives values of $\boldsymbol{\varepsilon}$ mostly driven by the highest eigenvalues of $\mathbf{Q}_{\boldsymbol{\theta}}$ and then the resulting direction $\mathbf{d}_1$ still continues to encourage the exploration space of high probability.

We may also notice that some relaxations of the GSGS are possible, following classical arguments of a random scan GS. For example, it is not necessary to sample the perturbation from $p(\boldsymbol{\varepsilon})$ at each iteration, it is sufficient to do this an infinite number of times to ensure the chain to be irreducible.[3] As we will see in section III, a low frequency sampling of $\boldsymbol{\varepsilon}$ can improve the algorithm's efficiency.

[3]From any point $\left(\mathbf{x}^{(t)}, \boldsymbol{\theta}^{(t)}\right)$, let $s > t$ be the closest next time where $\boldsymbol{\varepsilon}$ is sampled, then for any non-negligible subset $A \in \mathbb{R}^N \times \boldsymbol{\Theta}$, we have $P(\mathbf{x}^{(t)}, A) > 0$.

III. Unsupervised super resolution as a large scale problem

A. Problem statement

The paper details an application of the proposed GSGS to a super-resolution problem (identical to the one presented in [30], [40]): several blurred, noisy and down-sampled (low resolution) observations of a scene are available to retrieve the original (high resolution) scene [41], [42].

The usual direct model reads: $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} = \mathbf{S}\mathbf{H}\mathbf{x} + \mathbf{n}$. In this equation, $\mathbf{y} \in \mathbb{R}^M$ collects the pixels of the low resolution images (five 128×128 images, *i.e.* $M = 81920$) and $\mathbf{x} \in \mathbb{R}^N$ collects the pixels of the original image (one 256×256 image, *i.e.* $N = 65536$). The noise $\mathbf{n} \in \mathbb{R}^M$ accounts for measurement and modeling errors. $\mathbf{H}$ is a $N \times N$ circulant-block-circulant convolution matrix accounting for the optical and the sensor parts of the observation system. Here it is a square window of 5-pixel-width. $\mathbf{S}$ is a $M \times N$ matrix modeling motion (here translation) and decimation: it is a down-sampling binary matrix indicating which pixel of the blurred image is observed.

The noise is chosen to be $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \gamma_{\mathbf{n}}^{-1}\mathbf{I})$. Regarding the object, the chosen prior accounts for smoothness: $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \gamma_{\mathbf{x}}^{-1}\mathbf{D}^{\mathrm{t}}\mathbf{D})$ where $\mathbf{D}$ is the $N \times N$ circulant convolution matrix of the Laplacian filter. The hyperparameters $\gamma_{\mathbf{n}}$ and $\gamma_{\mathbf{x}}$ are unknown and the assigned priors are conjugate : Gamma distributions $\gamma_{\mathbf{n}} \sim \mathcal{G}(\alpha_{\mathbf{n}}; \beta_{\mathbf{n}})$ and $\gamma_{\mathbf{x}} \sim \mathcal{G}(\alpha_{\mathbf{x}}; \beta_{\mathbf{x}})$. They are weakly informative for large variances and uninformative Jeffreys' prior when the $(\alpha_{\mathbf{x}}, \beta_{\mathbf{x}})$ tends to $(0, 0)$. As a consequence, the full posterior pdf writes

$$\begin{aligned} p(\mathbf{x}, \gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x}, \gamma_{\mathbf{n}}) p(\mathbf{x}|\gamma_{\mathbf{x}}) p(\gamma_{\mathbf{x}}) p(\gamma_{\mathbf{n}}) \\ &\propto \gamma_{\mathbf{n}}^{\alpha_{\mathbf{n}}+N/2-1} \gamma_{\mathbf{x}}^{\alpha_{\mathbf{x}}+(M-1)/2-1} \\ &\quad \exp\left[-\gamma_{\mathbf{n}}\|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{x}\|^2/2\right] \exp\left[-\beta_{\mathbf{n}}\gamma_{\mathbf{n}}\right] \\ &\quad \exp\left[-\gamma_{\mathbf{x}}\|\mathbf{D}\mathbf{x}\|^2/2\right] \exp\left[-\beta_{\mathbf{x}}\gamma_{\mathbf{x}}\right]. \end{aligned} \tag{4}$$

The conditional law of the image writes

$$p(\mathbf{x}|\mathbf{y}, \gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}) \propto \exp\left[-\frac{\gamma_{\mathbf{n}}}{2}\|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{x}\|^2 - \frac{\gamma_{\mathbf{x}}}{2}\|\mathbf{D}\mathbf{x}\|^2\right].$$

Accordingly the negative logarithm gives the criterion

$$J_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}(\mathbf{x}) = \frac{\gamma_{\mathbf{n}}}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2 + \frac{\gamma_{\mathbf{x}}}{2}\|\mathbf{D}\mathbf{x}\|^2$$

and the gradient

$$\begin{aligned} \nabla J_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}(\mathbf{x}) &= \gamma_{\mathbf{n}}\mathbf{A}^{\mathrm{t}}(\mathbf{A}\mathbf{x} - \mathbf{y}) + \gamma_{\mathbf{x}}\mathbf{D}^{\mathrm{t}}\mathbf{D}\mathbf{x} \\ &= \mathbf{Q}(\mathbf{x} - \mathbf{m}) \end{aligned}$$

with $\mathbf{m} = \mathbf{Q}_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}^{-1}\gamma_{\mathbf{n}}\mathbf{A}^{\mathrm{t}}\mathbf{y}$, and the Hessian

$$\mathbf{Q}_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}} = \nabla^2 J_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}(\mathbf{x}) = \gamma_{\mathbf{n}}\mathbf{A}^t\mathbf{A} + \gamma_{\mathbf{x}}\mathbf{D}^{\mathrm{t}}\mathbf{D}$$

B. Gibbs sampler

The posterior pdf is explored by the proposed GS in Algorithm 2, based on the GSGS, that iteratively updates $\gamma_{\mathbf{n}}$, $\gamma_{\mathbf{x}}$ and a sub-component of $\mathbf{x}$. Regarding the hyperparameters, the conditional pdf are Gamma and their parameters are easy to compute.

Algorithm 2 : GSGS for super-resolution.

Set $t = 1$, define an initial point $\mathbf{x}^{(0)}$, and repeat

1: Sample $\gamma_{\mathbf{n}}^{(t)} \sim p\left(\gamma_{\mathbf{n}}|\mathbf{y}, \mathbf{x}^{(t-1)}\right)$ as

$$\mathcal{G}\left(\frac{N}{2}; \frac{2}{\|\mathbf{y} - \mathbf{S}\mathbf{H}\mathbf{x}^{(t-1)}\|^2}\right).$$

and $\gamma_{\mathbf{x}}^{(t)} \sim p\left(\gamma_{\mathbf{x}}|\mathbf{y}, \mathbf{x}^{(t-1)}\right)$ as

$$\mathcal{G}\left(\frac{M-1}{2}; \frac{2}{\|\mathbf{D}\mathbf{x}^{(t-1)}\|^2}\right).$$

2: Set $\mathbf{Q}_t = \mathbf{Q}_{\gamma_{\mathbf{x}}^{(t)}, \gamma_{\mathbf{n}}^{(t)}}$ and compute the gradient

$$\mathbf{g}^{(t)} = \nabla J_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}\left(\mathbf{x}^{(t-1)}\right) = \mathbf{Q}_t(\mathbf{x}^{(t-1)} - \mathbf{m})$$

3: Sample a perturbation $\boldsymbol{\varepsilon}^{(t)} \sim \mathcal{N}(0, \mathbf{Q}_t)$

4: Compute a set of N_D mutually conjugate directions $\{\mathbf{d}_1, \ldots, \mathbf{d}_{N_D}\}$ with the first being $\mathbf{d}_1 = \mathbf{g}^{(t)} + \boldsymbol{\varepsilon}^{(t)}$.

5: Sample independently the set $(\widetilde{\alpha}_n)_{n=1,\ldots,N_D}$ with:

$$\widetilde{\alpha}_n \sim \mathcal{N}\left(\frac{\mathbf{d}_n\mathbf{g}^{(t)}}{\mathbf{d}_n\mathbf{Q}_t\mathbf{d}_n}; \frac{1}{\mathbf{d}_n\mathbf{Q}_t\mathbf{d}_n}\right)$$

6: Compute $\mathbf{x}^{(t)} \leftarrow \mathbf{x}^{(t-1)} - \sum_{n=1}^{N_D} \widetilde{\alpha}_n \mathbf{d}_n$.

7: $t \leftarrow t + 1$.

until the stopping criterion is reached.

The set of mutually conjugate directions w.r.t. $\mathbf{Q}_{\gamma_{\mathbf{x}}, \gamma_{\mathbf{n}}}$, at step 4 of Algorithm 2, is computed by the Gram-Schmidt process applied to gradient, as usually found in conjugated gradient optimization algorithm. The procedure is similar to the algorithm described in [26]. Finally the estimator is the posterior mean computed as the empirical mean of the samples.

Despite the convergence proof with almost any law for the perturbation $\boldsymbol{\varepsilon}$ (provided that the density $p(\boldsymbol{\varepsilon})$ is supported in $\mathbb{R}^N$), some tuning is necessary to practically obtain a good space's exploration. In practice, Step 3 has a major influence and, as already discussed in section II-C, we observe that a working perturbation corresponds to those of the PO algorithm [30]

$$\boldsymbol{\varepsilon}^{(t)} = {\gamma_{\mathbf{n}}^{(t)}}^{-1/2}\mathbf{A}^{\mathrm{t}}\boldsymbol{\varepsilon}_{\mathbf{n}} + {\gamma_{\mathbf{x}}^{(t)}}^{-1/2}\mathbf{D}^{\mathrm{t}}\boldsymbol{\varepsilon}_{\mathbf{x}}$$

where $\boldsymbol{\varepsilon}_{\times}$ are two Gaussian normalized random vectors, leading to a Gaussian perturbation $\boldsymbol{\varepsilon}^{(t)}$ of covariance $\mathbf{Q}_t$. However, the proposed algorithm has numerous advantages over the PO algorithm. First the proposed algorithm has a convergence proof because it does not suffer from truncation, even in the extreme case with $N_D = 1$. Second the perturbation has the sole constraint of having $\mathbb{R}^N$ as support. Moreover a perturbation is not required at each iteration.

C. Numerical results

The posterior law (4) has been explored with the following four algorithms or settings.

- The adaptive RJ-PO algorithm [40], directly tuned with the acceptance probability, here chosen to be 0.9. This acceptance probability leads to an average number of

	PO	RJ-PO	GSGS(150)	GSGS(20)	GSGS(2)
$\widehat{\gamma_{\mathbf{n}}}$	0.9725	0.9718	0.9694	0.9694	0.7078
$\widehat{\sigma_{\gamma_{\mathbf{n}}}}$	0.0061	0.0063	0.0061	0.0063	0.0062
$\widehat{\gamma_{\mathbf{x}}}$	1.05 e-03	1.07e-03	1.06e-03	1.29e-03	9.62e-03
$\widehat{\sigma_{\gamma_{\mathbf{x}}}}$	1.5e-05	3.7e-05	1.7e-05	2.4e-05	6.2e-03
loop [s.]	3.4	2.4	2.4	0.5	0.1
total [s.]	515	362	353	72	9

Table I: Hyperparameter estimates and estimation variances for $\gamma_{\mathbf{n}} = 1$.

around 150 iterations of the conjugate gradient algorithm to compute the proposal, and with 6% of rejected samples.

- The PO algorithm [30] with a number of 150 iterations for the optimization.
- Algorithm 2 with $N_D = 150$. The idea is to build an algorithm close to RJ-PO's computing time.
- Algorithm 2 with $N_D = 20$. The idea is to show that our algorithm offers the possibility to reduce the number of iterations while still offering a good exploration and with guaranteed convergence.
- Algorithm 2 with $N_D = 2$. The idea is to show a very fast algorithm that offers a partially correct exploration. This case is particular in the sense that the perturbation is done only once for the whole algorithm.

The posterior mean (PM) estimations of the high-resolution image are given in Fig. 1 as well as the posterior standard deviation (PSD). From these results we can say that all algorithms provide similar quality for the image estimation. The same statement can be made for the standard deviation. However the posterior standard deviation with $N_D = 2$ seems incorrect. A possible interpretation is that the perturbation vector $\boldsymbol{\varepsilon}$ is simulated only once during the whole algorithm. Thus, the space is surely not sufficiently explored and the covariance estimation is severely biased. Indeed, since $\boldsymbol{\varepsilon}_{\times}$ are drawn only once, the stochastic explorations are limited to the conjugate direction plus the two directions $\boldsymbol{\varepsilon}_{\mathbf{x}}$ and $\boldsymbol{\varepsilon}_{\mathbf{n}}$. However the mean estimation does not seem to be affected and this algorithm is able to provide very quickly a good estimation of the image and hyperparameter values. We must notice that in our test with $N_D = 10$ the chain converged to a close, but wrong distribution, giving good results in the image but an slightly underestimation of $\gamma_{\mathbf{n}}$.

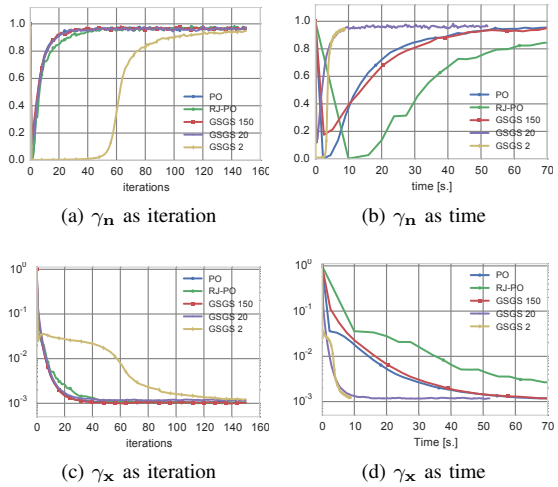
The chains of the hyperparameters are illustrated in Fig. 2. Figs. 2a and 2c represent the samples as a function of the iterations. We observe that, except for $N_D = 2$, all the chains have the same behavior with the same convergence period. The $N_D = 2$ has slower (in terms of the number of iterations) convergence but reaches the same stationary distribution.

Figs. 2b and 2d represent the samples as a function of time (in seconds). The chain behavior of algorithms PO, RJ-PO and GSGS(150) is very similar. This result is obvious since these algorithms compute almost the same number of gradients per iteration. That said, we see that for $N_D = 20$ and $N_D = 2$, the impact on the convergence time is significant. Table I shows some quantitative results. In particular the case $N_D = 20$ is five times faster than RJ-PO.

	PO	RJ-PO	GSGS(20)	GSGS(2)
$\widehat{\gamma_{\mathbf{n}}}$	9.9e-03	9.9e-03	9.9e-03	9.9e-03
$\widehat{\sigma_{\gamma_{\mathbf{n}}}}$	6.0e-05	6.05e-05	4.8e-05	5.5e-05
$\widehat{\gamma_{\mathbf{x}}}$	1.86e-03	1.84e-03	4.86e-03	2.29e-03
$\widehat{\sigma_{\gamma_{\mathbf{x}}}}$	3.2e-04	3.2e-04	7.2e-04	3.4e-05

Table II: Hyperparameter estimates and estimation variances for $\gamma_{\mathbf{n}} = 1e-02$.

In addition, Table II shows the estimated values of the hyperparameters with a higher noise level. Again the results are close with a good estimation of $\gamma_{\mathbf{n}}$.



(a) $\gamma_{\mathbf{n}}$ as iteration (b) $\gamma_{\mathbf{n}}$ as time

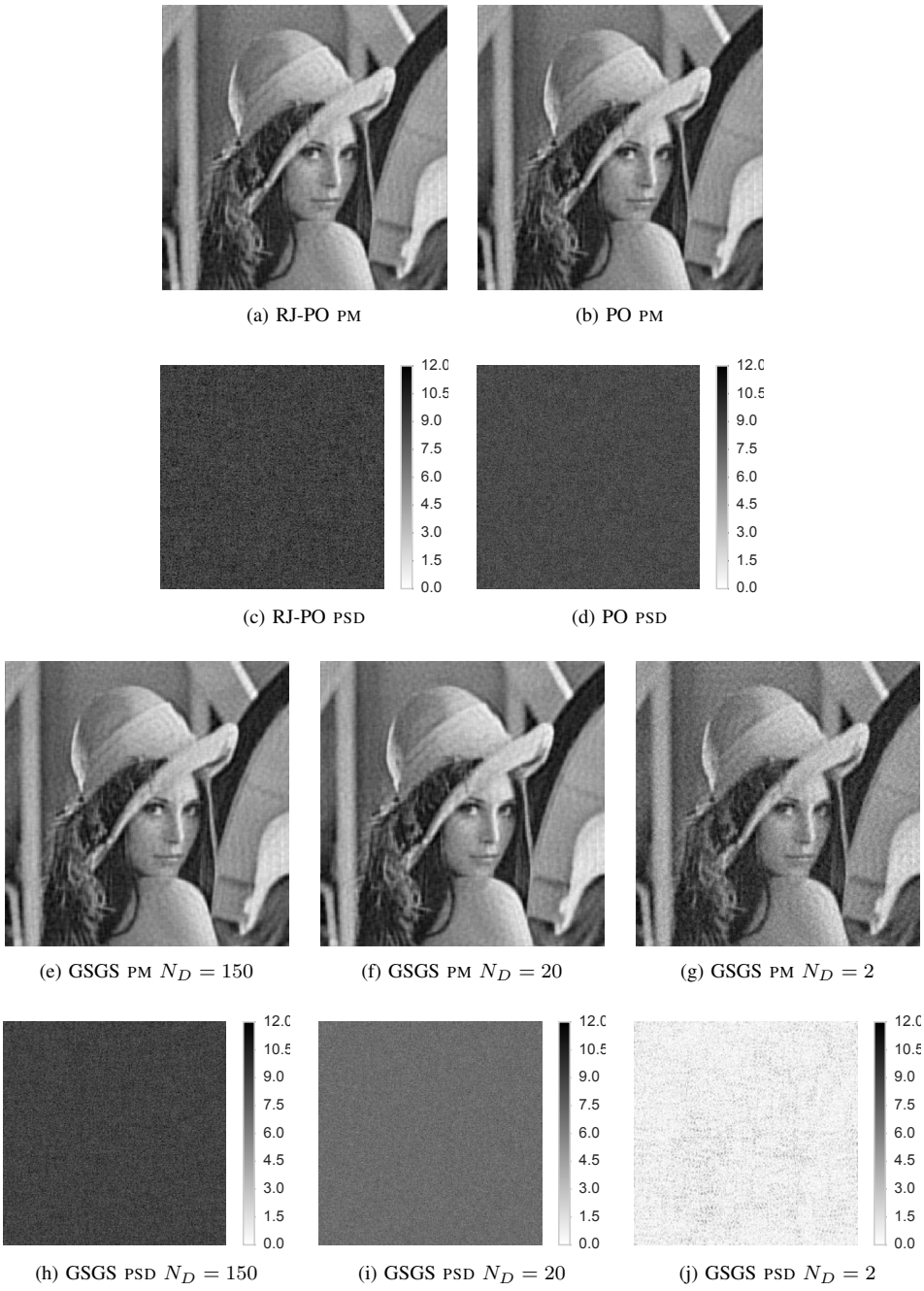
(c) $\gamma_{\mathbf{x}}$ as iteration (d) $\gamma_{\mathbf{x}}$ as time

Figure 2: Chains of hyper parameters $\gamma_{\mathbf{x}}$ and $\gamma_{\mathbf{n}}$.

To illustrate the effect of the perturbation for good space exploration, Fig. 3 shows the results when no perturbations $\boldsymbol{\varepsilon}^{(t)}$ are introduced and with $N_D = 10$. In this case, the hypotheses of Proposition 2 are no longer verified and those of Proposition 3 cannot be verified in practice. Moreover, the results show that both the covariance and the hyperparameters are wrongly estimated. This effect leads to an over-regularized image. A possible explanation is that the conjugate directions of the GSGS explore in a privileged way the directions of small variance (highest eigenvalues of $\mathbf{Q}$).

Regarding the computational cost, all the presented algorithms are dominated by the cost of the matrix-vector product $\mathbf{Qx}$. The cost thus depends on the specific problems and the structure of $\mathbf{Q}$ in the same way as for the conjugate gradient algorithm. For super-resolution problems, the cost of the matrix-vector product is almost equal to two discrete Fourier transforms of images. That said, the total number of matrix-vector products is related to N_D and the number of Gibbs iterations. Moreover, the computational cost is linear with respect to N_D.

The main concluding comment is that the proposed algorithm allows a great improvement in the convergence time of the GS. However the speed improvement can come with a bad covariance estimation if the number N_D of directions for the image $\mathbf{x}$ is not sufficient.



(a) RJ-PO PM (b) PO PM

(c) RJ-PO PSD (d) PO PSD

(e) GSGS PM $N_D = 150$ (f) GSGS PM $N_D = 20$ (g) GSGS PM $N_D = 2$

(h) GSGS PSD $N_D = 150$ (i) GSGS PSD $N_D = 20$ (j) GSGS PSD $N_D = 2$

Figure 1: Image results.

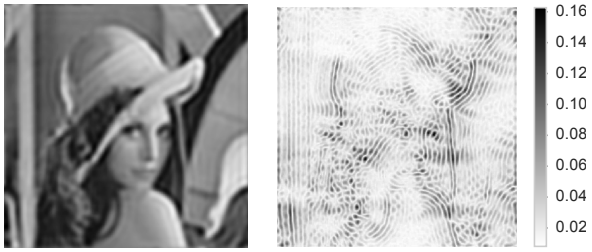
IV. Conclusion

The handling of high-dimensional distribution, especially Gaussian, appears in many linear inverse and estimation problems. With growing interest in "Big Data" and non stationary problems this task becomes critical. Moreover, the uncertainty around the estimated values, or the confidence interval, remains one of the difficult points combined with the hyperparameter estimation for automatic method designs.

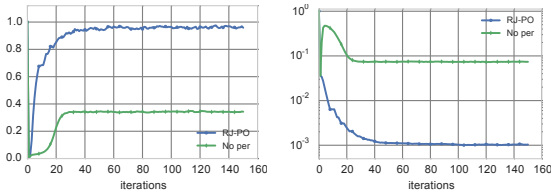
The main contributions of this paper are (i) the proposition of a new algorithm in the class of the Gibbs samplers, able to address the case of high-dimensional Gaussian conditional distributions, and (ii) the convergence proof of the algorithm. It relies on a random excursion along a small set of directions instead of working with high dimensional distributions. The directions are appropriately chosen according to the gradient of the potential of the distribution.

This new algorithm is shown to be an efficient alternative to existing work like the PO-type algorithms: we ensure the theoretical convergence of the algorithm and, in some cases, we can show a drastic computing-time improvement.

The convergence of the algorithm is proved, provided



(a) PM (b) PSD



(c) $\gamma_{\mathbf{n}}$ (d) $\gamma_{\mathbf{x}}$

Figure 3: Results without perturbation and $N_D = 20$.

that a random perturbation around the gradient direction is introduced. Even if in theory the only condition to ensure convergence is to choose a perturbation distribution supported in the whole space, it appears in practice that the results are sensitive to the choice of the distribution. Moreover, the choice of the Gaussian distribution $\mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$ is the only case where the algorithm is more efficient than the PO and RJ-PO algorithms. The objective of further work will be to better understand this sensitivity and the open problem of the choice of the perturbation's distribution.

In further work the objective will be to study the convergence rate of the GSGS. In particular, the geometric ergodicity is an important property that ensures a fast convergence and allows us to give estimations of standard errors. The geometric ergodicity of Gibbs samplers has long been studied [43] and a lot of results are shown in the Gaussian case [44], as well as for applications in Bayesian hierarchical models [45], also in the case of joint Gaussian and Gamma distributions [46], [47], the latter being close to our illustration example.

Also, one has to choose the number N_D of mutually conjugate directions to sample at each iteration of the algorithm. In theory, this does not affect the convergence properties of the algorithm. As a perspective, one can propose an automatic choice of N_D, following the work in [40] for the RJ-PO. A research field could be the study of the algorithm's efficiency with respect to the eigenvalues of Q in the high dimensional case.

The proposed algorithm is somewhat independent of the chosen direction. The use of a preconditioner to compute direction, as in preconditioned conjugate gradient, should improve the computational cost by an N_D parameter smaller than at the present time. It depends, however, on each problem addressed.

From an experimental standpoint an additional assessment of the proposed method could rely on a numerical comparison with other existing approaches, for instance Hamltonian or Langevin algorithm [31]–[34].

This paper is focused on linear conditionally Gaussian models. By use of hidden variables, the algorithm should also be able to work with non Gaussian models that are still conditionally Gaussian.

APPENDIX

A. Proof of Proposition 2

This appendix is devoted to prove Proposition 2. It is mainly inspired by the proofs presented in [28] (see also [27], [29]) for different random scan strategies in order to sample $p(\mathbf{x}|\boldsymbol{\theta})$. The only difference is that the random choice is not according to a set of coordinates of $\mathbf{x}$ in the canonical basis, but according to a mutually conjugate set with respect to a current matrix $\mathbf{Q}_{\boldsymbol{\theta}}$. Therefore the same arguments as detailed in [28] can be used to prove the irreducibility: if the support of the density $p(\boldsymbol{\varepsilon})$ is $\mathbb{R}^N$, all the directions can be explored in one step of the algorithm. Therefore any $\mathbf{y} \in \mathbb{R}^N$ can be reached in one step by taking, for example, $\mathbf{d}_1 = \mathbf{x}^{(t-1)} - \mathbf{y}$, $\widetilde{\alpha}_1 = 1$, $\widetilde{\alpha}_n = 0$, $n = 2, \ldots, N_D$. Using classical continuity arguments, we can deduce that the probability of reaching any open ball $\mathcal{B}(\mathbf{y}, r)$, centered in $\mathbf{y}$ of radius r, conditional to any current point $\mathbf{x}^{(t)}$, is strictly positive, which ensures the chain to be irreducible.

The rest of the proof focuses on the fact that $p(\mathbf{x}, \boldsymbol{\theta})$ is an invariant probability of the chain. We use the same arguments and notations of [28]. Let $\mathbf{x} \in \mathbb{R}^N$ and a set D of mutually conjugate directions with respect to a definite positive matrix $\mathbf{Q}$. We decompose $\mathbf{x} = (\mathbf{x}_D, \mathbf{x}_{\setminus D})$ which is always possible as explained in section II-A.

Define $(\mathbf{x}', \boldsymbol{\theta}') \in \mathbb{R}^N \times \Theta$ a current point and $(\mathbf{x}', \boldsymbol{\theta}') \in \mathbb{R}^N \times \Theta$ the point obtained by Algorithm 1 with the transition Kernel:

$$P(\mathbf{x}, \boldsymbol{\theta} | \mathbf{x}', \boldsymbol{\theta}') = \pi(\boldsymbol{\theta} | \mathbf{x}', \boldsymbol{\theta}') \pi(\mathbf{x}_D | \mathbf{x}_{\setminus D}, \mathbf{x}', \boldsymbol{\theta}) \delta(\mathbf{x}_{\setminus D} - \mathbf{x}'_{\setminus D})$$

with π denoting any conditional probability and δ is the Dirac function. The objective is to show that if $(\mathbf{x}', \boldsymbol{\theta}')$ is distributed according to the joint distribution p, then $(\mathbf{x}, \boldsymbol{\theta})$ is also distributed according to p.

Let $A \subset \mathbb{R}^N$ be a measurable set. The following lines are the result of the definition of the transition Kernel, the use of the general product rule, and of sequential integration with

respect to $\boldsymbol{\theta}'$, $\mathbf{x}'_D$ and $\mathbf{x}'_{\setminus D}$:

$$\begin{aligned}
&\mathbb{P}((\mathbf{x},\boldsymbol{\theta}) \in A)\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta}) P(\mathbf{x},\boldsymbol{\theta}|\mathbf{x}',\boldsymbol{\theta}') p(\mathbf{x}',\boldsymbol{\theta}') \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{x}'\mathrm{d}\boldsymbol{\theta}'\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|\mathbf{x}',\boldsymbol{\theta}') \pi(\mathbf{x}_D|\mathbf{x}_{\setminus D},\mathbf{x}',\boldsymbol{\theta}) \ldots\\
&\qquad \ldots \delta(\mathbf{x}_{\setminus D} - \mathbf{x}'_{\setminus D}) p(\mathbf{x}',\boldsymbol{\theta}') \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{x}'\mathrm{d}\boldsymbol{\theta}'\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta}) p(\mathbf{x}',\boldsymbol{\theta}) \pi(\mathbf{x}_D|\mathbf{x}_{\setminus D},\mathbf{x}',\boldsymbol{\theta}) \ldots\\
&\qquad \ldots \delta(\mathbf{x}_{\setminus D} - \mathbf{x}'_{\setminus D}) \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{x}'\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta}) p(\mathbf{x}'_{\setminus D},\boldsymbol{\theta}) \pi(\mathbf{x}_D|\mathbf{x}_{\setminus D},\mathbf{x}'_{\setminus D},\boldsymbol{\theta}) \ldots\\
&\qquad \ldots \delta(\mathbf{x}_{\setminus D} - \mathbf{x}'_{\setminus D}) \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{x}'_{\setminus D}\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta}) p(\mathbf{x}_{\setminus D},\boldsymbol{\theta}) \pi(\mathbf{x}_D|\mathbf{x}_{\setminus D},\boldsymbol{\theta}) \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}\\
&= \int \mathbb{1}_A(\mathbf{x},\boldsymbol{\theta}) p(\mathbf{x},\boldsymbol{\theta}) \mathrm{d}\mathbf{x}\mathrm{d}\boldsymbol{\theta}
\end{aligned}$$

Hence the joint probability $p(\mathbf{x},\boldsymbol{\theta})$ is an invariant probability of the Markov chain produced by Algorithm 1.

B. Proof of Proposition 3

This appendix is dedicated to prove Proposition 3. Let $(\mathbf{x}^{(0)},\boldsymbol{\theta}^{(0)}) \in \mathbb{R}^N \times \Theta$ be a current point and $(\mathbf{x}^{(t)},\boldsymbol{\theta}^{(t)})$ the point produced by the chain of Algorithm 1 at iteration t. The objective is to prove that for any non-negligible subset $A \subset \mathbb{R}^N \times \Theta$, there is $T \geq 0$ such as $\mathbb{P}((\mathbf{x}^{(T)},\boldsymbol{\theta}^{(T)}) \in A|\mathbf{x}^{(0)},\boldsymbol{\theta}^{(0)}) > 0$. Using the hypothesis H-2, it is sufficient to prove that for any non-negligible subset $A_x \in \mathbb{R}^N$, there is $T \geq 0$ such as:

$$\mathbb{P}(\mathbf{x}^{(T)} \in A_x|\mathbf{x}^{(0)},\boldsymbol{\theta}^{(0)}) > 0 \tag{5}$$

Given $\mathbf{x}^{(0)}$, we denote by $\boldsymbol{\theta}$ the corresponding element that respects conditions H-3. It is sufficient to prove the Proposition in the following framework:
F-1 $\boldsymbol{\theta}^{(N+1)} = \boldsymbol{\theta}^{(N)} = \ldots = \boldsymbol{\theta}^{(0)} = \boldsymbol{\theta}$,
F-2 $\mathbf{m}_{\boldsymbol{\theta}} = 0$,
F-3 $\mathbf{Q}_{\boldsymbol{\theta}} = \mathrm{diag}(q_1,\ldots,q_N)$ is diagonal.
Indeed, if we prove the inequality (5) with fixed $\boldsymbol{\theta}$ for $N+1$ iterations, continuity arguments using conditions H-1 and H-2 will end the proof of the Proposition. The simplifications F-2 and F-3 can be assumed by a change of variable $\mathbf{y}^{(t)} = \mathbf{x}^{(t)} - \mathbf{m}_{\boldsymbol{\theta}}$ and by considering the basis of $\mathbb{R}^N$ formed by the eigenvectors of $\mathbf{Q}_{\boldsymbol{\theta}}$.

In this simplified framework, the chain of Algorithm 1 produces $\mathbf{x}^{(t)},\ t = 1,\ldots,N+1$, such as:

$$\mathbf{x}^{(t)} = (\mathbf{I} - \alpha^{(t)}\mathbf{Q}_{\boldsymbol{\theta}})(\mathbf{I} - \alpha^{(t-1)}\mathbf{Q}_{\boldsymbol{\theta}}) \ldots (\mathbf{I} - \alpha^{(1)}\mathbf{Q}_{\boldsymbol{\theta}})\mathbf{x}^{(0)},$$

with $\mathbf{I}$ the identity matrix in $\mathbb{R}^N$ and, noting $\mathbf{x} = (x_1,\ldots,x_N)^{\mathrm{t}}$, we have, for $n = 1,\ldots,N$:

$$\mathbf{x}_n^{(t)} = (1 - \alpha^{(t)}q_n)(1 - \alpha^{(t-1)}q_n) \ldots (1 - \alpha^{(1)}q_n)\mathbf{x}_n^{(0)}. \tag{6}$$

The hypothesis H-3.2 ensures that $\mathbf{x}_n^{(0)} \neq 0,\ n = 1,\ldots,N$, therefore we can assume without loss of generality that $\mathbf{x}_n^{(0)} = 1,\ n = 1,\ldots,N$, and equation (6) is, in this case:

$$\mathbf{x}_n^{(t)} = (1 - \alpha^{(t)}q_n)(1 - \alpha^{(t-1)}q_n) \ldots (1 - \alpha^{(1)}q_n). \tag{7}$$

The following Lemma proves that any point in $\mathbb{R}^N$ can be reached by the chain in $N+1$ iterations.

Lemma 1. *For any $\mathbf{y} \in \mathbb{R}^N$, there is $\alpha = (\alpha^{(1)},\ldots,\alpha^{(N+1)})$ such as $\mathbf{x}^{(N+1)} = \mathbf{y}$, where $\mathbf{x}^{(N+1)}$ is defined by (7) with $t = N+1$.*

Proof. This can be done by interpreting it as an interpolation problem: given $\mathbf{y} \in \mathbb{R}^N$, the objective is to show that there is a polynomial P_α^{N+1} such as:

$$\begin{aligned}
P_\alpha^{N+1}(q_n) &= y_n, \quad n = 1,\ldots,N &(8)\\
P_\alpha^{N+1}(0) &= 1 &(9)
\end{aligned}$$

with P_α^{N+1} defined by the right hand side of (7) with $t = N+1$. The constraint (9) is due to the specific form of P_α^{N+1}. Also the fact that the parameters $\alpha^{(n)}$ must be real, implies that the polynomial P_α^{N+1} must have only real roots. It is well known that there is a polynomial of degree N that respects (8) and (9). Let us denote by Q such a polynomial. But the roots of Q may be complex. However we can show that there is a polynomial of degree $N+1$ with real roots that respects the conditions (8) and (9). Indeed, let us consider the polynomial Q and a polynomial R of degree $N+1$ such as $R(q_1) = R(q_2) = \ldots = R(q_N) = R(0) = 0$. Therefore any polynomial $P_\tau = Q + \tau R$, $\tau \in \mathbb{R}$, respects conditions (8) and (9), and it is trivial to show that for τ^* sufficiently large, the polynomial P_{τ^*} has all its roots $r_n^* \in \mathbb{R},\ n = 1,\ldots,N$. Therefore, taking $P_\alpha^{N+1} = P_{\tau^*}$, *i.e.* $\alpha^{(n)} = 1/r_n^*$ ends the proof of the lemma. □

Using this lemma and the continuity of P_α^{N+1} with respect to α, it is trivial to prove (5) and then the Proposition.

REFERENCES

[1] J. Idier, Ed., *Bayesian Approach to Inverse Problems*. London: ISTE Ltd and John Wiley & Sons Inc., 2008.

[2] J.-F. Giovannelli and J. Idier, Eds., *Regularization and Bayesian Methods for Inverse Problems in Signal and Image Processing*. London: ISTE Ltd and John Wiley & Sons Inc., 2015.

[3] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[4] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Processing*, vol. 4, no. 7, pp. 932–946, July 1995.

[5] J.-F. Giovannelli, "Unsupervised Bayesian convex deconvolution based on a field with an explicit partition function," *IEEE Trans. Image Processing*, vol. 17, no. 1, pp. 16–26, Jan. 2008.

[6] X. Tan, J. Li, and P. Stoica, "Efficient sparse Bayesian learning via Gibbs sampling," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 3634 –3637.

[7] G. Kail, J.-Y. Tourneret, F. Hlawatsch, and N. Dobigeon, "Blind deconvolution of sparse pulse sequences under a minimum distance constraint: a partially collapsed Gibbs sampler method," *IEEE Trans. Signal Processing*, vol. 60, no. 6, pp. 2727–2743, June 2012.

[8] O. Féron, B. Duchêne, and A. Mohammad-Djafari, "Microwave imaging of piecewise constant objects in a 2D-TE configuration," *International Journal of Applied Electromagnetics and Mechanics*, vol. 26, no. 6, pp. 167–174, IOS Press 2007.

[9] H. Ayasso and A. Mohammad-Djafari, "Joint NDT image restoration and segmentation using Gauss-Markov-Potts prior models and variational Bayesian computation," *IEEE Trans. Image Processing*, vol. 19, no. 9, pp. 2265–2277, 2010.

[10] F. Orieux, J.-F. Giovannelli, and T. Rodet, "Bayesian estimation of regularization and point spread function parameters for Wiener–Hunt deconvolution," *J. Opt. Soc. Amer.*, vol. 27, no. 7, pp. 1593–1607, July 2010.

[11] Q. Liu, E. Chen, B. Xiang, C. H. Q. Ding, and L. He, "Gaussian process for recommender systems," in *Lecture Notes in Computer Science, Knowledge Science, Engineering and Management*, vol. 7091, 2011, pp. 56–67.

[12] J. E. Besag, "On the correlation structure of some two-dimentional stationary processes," *Biometrika*, vol. 59, no. 1, pp. 43–48, 1972.

[13] H. Rue, "Fast sampling of Gaussian Markov random fields," *J. R. Statist. Soc. B*, vol. 63, no. 2, pp. 325–338, 2001.

[14] A. E. Gelfand, H.-J. Kim, C. F. Sirmans, and S. Banerjee, "Spatial modeling with spatially varying coefficient processes," *J. Amer. Statist. Assoc.*, vol. 98, no. 462, pp. 387–396, 2003.

[15] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 33, no. 4, pp. 959–963, Aug. 1985.

[16] R. Chellappa and A. Jain, *Markov Random Fields: Theory and Application*. Academic Press Inc, 1992.

[17] J. M. Bardsley, "MCMC-based image reconstruction with uncertainty quantification," *SIAM Journal of Scientific Computation*, vol. 34, no. 3, pp. A1316–A1332, 2012.

[18] J. M. Bardsley, M. Howard, and J. G. Nagy, "Efficient MCMC-based image deblurring with Neumann boundary conditions," *Electronic Transactions on Numerical Analysis*, vol. 40, pp. 476–488, 2013.

[19] H. Rue and L. Held, *Gaussian Markov Random Fields: Theory and Applications*, ser. Monographs on Statistics and Applied Probability. Chapman & Hall, 2005, vol. 104.

[20] P. Lalanne, D. Prévost, and P. Chavel, "Stochastic artificial retinas: algorithm, optoelectronic circuits, and implementation," *Applied Optics*, vol. 40, no. 23, pp. 3861–3876, 2001.

[21] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer Verlag, Berlin Germany, 2003.

[22] G. Papandreou and A. Yuille, "Gaussian sampling by local perturbations," in *Proc. Int. Conf. on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, Dec. 2010, pp. 1858–1866.

[23] F. Orieux, J.-F. Giovannelli, T. Rodet, H. Ayasso, and A. Abergel, "Super-resolution in map-making based on a physical instrument model and regularized inversion. Application to SPIRE/Herschel." *Astron. Astrophys.*, vol. 539, Mar. 2012.

[24] C. Gilavert, S. Moussaoui, and J. Idier, "Rééchantillonnage gaussien en grande dimension pour les problèmes inverses," in *Actes 24$^{\text{e}}$ coll. GRETSI*, Brest, France, Sep. 2013.

[25] C. Fox, "A conjugate direction sampler for normal distributions with a few computed examples," Electronics Technical Report No. 2008-1, University of Otago, Dunedin, New Zealand, Tech. Rep., 2008.

[26] A. Parker and C. Fox, "Sampling Gaussian distributions in Krylov spaces with conjugate gradients," *SIAM J. Sci. Comput.*, vol. 34, no. 3, pp. B312–B334, 2012.

[27] R. A. Levine, Z. Yu, W. G. Hanley, and J. J. Nitao, "Implementing random scan Gibbs samplers," *Computational Statistics*, vol. 20, no. 1, pp. 177–196, 2005.

[28] R. A. Levine and G. Casella, "Optimizing random scan Gibbs samplers," *Journal of Multivariate Analysis*, vol. 97, no. 10, pp. 2071–2100, 2006.

[29] K. Latuszynski, G. O. Roberts, and J. Rosenthal, "Adaptive Gibbs samplers and related MCMC methods," *Annals of Applied Probability*, vol. 23, no. 1, pp. 66–98, 2013.

[30] F. Orieux, O. Féron, and J.-F. Giovannelli, "Sampling high-dimensional Gaussian fields for general linear inverse problem," *IEEE Signal Proc. Let.*, vol. 19, no. 5, pp. 251–254, May 2012.

[31] O. Stramer and R. L. Tweedie, "Langevin-type models i: Diffusions with given stationary distributions, and their discretizations," *Methodology and Computing in Applied Probability*, vol. 1, no. 3, pp. 283–306, 1999.

[32] ——, "Langevin-type models ii: Self-targeting candidates for MCMC algorithms," *Methodology and Computing in Applied Probability*, vol. 1, no. 3, pp. 307–328, 1999.

[33] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth, "Hybrid Monte Carlo," *Physics Letters B*, vol. 195, no. 2, pp. 216–222, 1987.

[34] R. M. Neal, "MCMC using Hamiltonian dynamics," in *Handbook of Markov Chain Monte Carlo*, G. J. S. Brooks, A. Gelman and X.-L. Meng, Eds. Chapman & Hall, 2010, ch. 5, pp. 113–162.

[35] C. Vacar, J.-F. Giovannelli, and Y. Berthoumieu, "Langevin and Hessian with Fisher approximation stochastic sampling for parameter estimation of structured covariance," in *Proc. IEEE ICASSP*, Prague, Czech Republic, May 2011, pp. 3964–3967.

[36] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in practice*. Boca Raton, USA: Chapman & Hall/CRC, 1996.

[37] G. O. Roberts and S. Rosenthal, "Harris recurrence of Metropolis-within-Gibbs and trans-dimensional Markov chains," *The Annals of Applied Probability*, vol. 16, no. 4, pp. 2123–2139, 2006.

[38] S. Meyn and R. Tweedie, *Markov chains and stochastic stability*. Springer-Verlag, London, 1993.

[39] F. Orieux, O. Féron, and J.-F. Giovannelli, "Gradient scan Gibbs sampler: An efficient high-dimensional sampler. Application in inverse problems," in *ICASSP*, Apr. 2015.

[40] C. Gilavert, S. Moussaoui, and J. Idier, "Efficient Gaussian sampling for solving large-scale inverse problems using MCMC," *IEEE Trans. Image Processing*, vol. 63, no. 1, pp. 70–80, Jan. 2015.

[41] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Proc. Mag.*, pp. 21–36, May 2003.

[42] G. Rochefort, F. Champagnat, G. Le Besnerais, and J.-F. Giovannelli, "An improved observation model for super-resolution under affine motion," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3325–3337, Nov. 2006.

[43] G. O. Roberts and N. G. Polson, "On the geometric convergence of the Gibbs sampler," *J. R. Statist. Soc. B*, vol. 56, no. 2, pp. 377–384, 1994.

[44] G. O. Roberts and S. K. Sahu, "updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler," *J. R. Statist. Soc. B*, vol. 59, no. 2, pp. 291–317, 1997.

[45] P. O. and G. O. Roberts, "Stability of the Gibbs sampler for Bayesian hierarchical models," *Ann. Statist.*, vol. 36, no. 1, pp. 95–117, 2008.

[46] J. P. Hobert and C. J. Geyer, "Geometric ergodicity of Gibbs and block samplers for a hierarchical random effects model," *Journal of Multivariate Analysis*, vol. 67, no. 2, pp. 414–430, 1998.

[47] A. Johnson and O. Burbank, "Geometric ergodicity and scanning strategies for two-component Gibbs samplers," *Communications in Statistics - Theory and Methods*, vol. 44, no. 15, pp. 3125–3145, 2015.

A&A 549, A83 (2013)
DOI: 10.1051/0004-6361/201219950

Astronomy & Astrophysics

Estimating hyperparameters and instrument parameters in regularized inversion
Illustration for *Herschel*/SPIRE map making

F. Orieux[1,3], J.-F. Giovannelli[2,3], T. Rodet[3], and A. Abergel[4]

[1] Institut d'Astrophysique de Paris (CNRS – Univ. Paris 6), 75014 Paris, France
e-mail: orieux@iap.fr
[2] Univ. Bordeaux, IMS, UMR 5218, 33400 Talence, France
e-mail: Giova@IMS-Bordeaux.fr
[3] Laboratoire des Signaux et Systèmes (CNRS – Supélec – Univ. Paris-Sud 11), 91192 Gif-sur-Yvette, France
e-mail: orieux,rodet@lss.supelec.fr
[4] Institut d'Astrophysique Spatiale (CNRS – Univ. Paris-Sud 11), 91405 Orsay, France
e-mail: abergel@ias.u-psud.fr

Received 5 July 2012 / Accepted 19 October 2012

ABSTRACT

We describe regularized methods for image reconstruction and focus on the question of hyperparameter and instrument parameter estimation, i.e. unsupervised and myopic problems. We developed a Bayesian framework that is based on the posterior density for all unknown quantities, given the observations. This density is explored by a Markov chain Monte-Carlo sampling technique based on a Gibbs loop and including a Metropolis-Hastings step. The numerical evaluation relies on the SPIRE instrument of the *Herschel* observatory. Using simulated and real observations, we show that the hyperparameters and instrument parameters are correctly estimated, which opens up many perspectives for imaging in astrophysics.

Key words. methods: data analysis – methods: statistical – methods: numerical – techniques: image processing

1. Unsupervised myopic inversion

The agreement of physical models and observations is a crucial question in astrophysics, however, observation instruments inevitably have defects and limitations (limited pass-band, non-zero response time, attenuation, error and uncertainty, etc.). Their inversion by numerical processing must, as far as possible, be based on an instrument model that includes a description of these defects and limitations. The difficulties of such inverse problems, and notably their often ill-posedness, were well identified several decades ago in various communities: signal and image processing and statistics, and also mathematical physics and astrophysics. It seems pertinent to take advantage of the knowledge amassed by these communities concerning both the analysis of the problems and their solutions.

The ill-posedness comes from a deficit of available information (and not only from a "simple numerical problem"), which becomes all the more marked as resolution requirements increase. The inversion methods must therefore take other information into account to compensate for the deficits in the observations: this is known as regularization. Each reconstruction method is thus specialised for a certain class of objects (point sources, diffuse emission, superposition of the two, etc.) according to the information accounted for. Consequently, in as much as it relies on various sources of information, each method is based on a trade-off, which usually requires the setting of hyperparameters, denoted by $\boldsymbol{\xi}$ in the following. The question of their automatic tuning, namely unsupervised inversion, has been extensively studied and numerous attempts investigate statistical approaches: approximated, pseudo or marginal likelihood, in a Bayesian or non-Bayesian sense, EM, SEM and SAEM algorithms, etc. The reader may consult papers such as (Zhou et al. 1997; de Figueiredo & Leitao 1997; Saquib et al. 1998; Descombes et al. 1999; Molina et al. 1999; Lanterman et al. 2000; Pascazio & Ferraiuolo 2003; Blanc et al. 2003; Chantas et al. 2007; Giovannelli 2008; Babacan et al. 2010; Orieux et al. 2010a) and reference books such as (Winkler 2003, Part.VI), (Li 2001, Chap. 7) or (Idier 2008, Chap. 8). Alternative methods are based on the L-curve (Hansen 1992; Wiegelmann & Inhester 2003) or on generalised cross-validation (Golub et al. 1979; Fortier et al. 1993; Ocvirk et al. 2006).

The construction of maps of high resolution and accuracy relies on increasingly complex instruments. So, inversion methods require instrument models that faithfully reflect the physical reality to distinguish, in the observations, between what is caused by the instrument and what is due to the actual sky. Then, a second set of parameters comes into play: the instrument parameters, denoted by $\boldsymbol{\eta}$ in the following, such as lobe width, amplitude of secondary lobes, response time, or gain. Their values are of prime importance and their settings are generally based on dedicated observation and rely on models and/or calibrations that inevitably contain errors. For example, the lobe widths are usually determined from a specific observation in a spectral band of non-zero width; consequently the result depends on the source spectrum. Correction factors can be applied but, naturally, they also contain errors when the source spectrum is poorly known or unknown. In contrast, our aim is to achieve myopic inversion, i.e. to estimate the instrument

parameters without dedicated observation. The question arises in various fields: optical imaging (Pankajakshani et al. 2009), interferometry (Thiébaut 2008), satellite observation (Jalobeanu et al. 2002), magnetic resonance force microscopy (Dobigeon et al. 2009), fluorescence microscopy (Zhang et al. 2007), deconvolution (Orieux et al. 2010b), etc. A similar problem deals with non-parametric intrument response (blind inversion), for which the literature is also very abundant: (Mugnier et al. 2004; Thiébaut & Conan 1995; Fusco et al. 1999; Conan et al. 1998) in astronomy and (Lam & Goodman 2000; Likas & Galatsanos 2004; Molina et al. 2006; Bishop et al. 2008; Xu & Lam 2009) in the signal-image literature represent examples. The present paper is devoted to parameter estimation for the instrument model developed in our previous paper (Orieux et al. 2012b), based on an accurate instrument model.

A threefold problem has to be solved: from a unique observation, estimate the hyperparameters the instrument parameters and the map. This is referred to as unsupervised and myopic inversion. From the methodological point of view, the proposed inversion method comes within a Bayesian approach (Idier 2008). In this family, we find the classic Wiener and Kalman methods that calculate the expectation or the maximizer of a posterior density. In an equivalent way, the Phillips-Twomey-Tikhonov methods calculate the minimizer of a least-squares criterion with quadratic penalization. These methods are based on a second-order analysis (Gaussian models, quadratic criteria) and lead to linear processing. The work proposed here is in a similar methodological vein as far as estimating the map goes; however, the contribution concerns the estimation of the hyperparameters and instrument parameters. We resort to an entirely Bayesian approach (also called full-Bayes) that models the information for each variable (observations, unknown map as well as hyperparameters and instrument parameters) through a probability density. Based on an a posteriori distribution for all the unknown variables, the proposed method jointly estimates the instrument parameters, the hyperparameters, and the map of interest. Regarding experimental data processing, the present paper follows (Orieux et al. 2012b) on inversion for the SPIRE instrument onboard *Herschel*, which requires the hyperparameters to be fixed by hand and the instrument parameters to be known. The proposed method can automatically tune these parameters and may permit the systematic and automatic processing of large information streams coming from present and future space-based instruments (e.g. *Herschel*, *Planck*, JWST, etc.).

The paper is structured as follows. Section 2 introduces the notation and sets out the problem. Section 3 presents the inversion method: it introduces the prior densities and leads to the posterior density. Section 4 describes the computing method based on Markov chain Monte-Carlo (Gibbs) stochastic sampling algorithms. The work is essentially evaluated on simulated observations and on a first set of real observations in the context of the SPIRE instrument onboard *Herschel*. The results are presented in Sect. 5. Finally, some conclusions and perspectives are provided in Sect. 6.

2. Notation, instrument, and map models

To produce accurate and reliable maps, the inversion must exploit a description that represents the acquisition process as faithfully as possible. In this sense, the instrument model

- is based on a map of the sky noted $\mathcal{X}$, which is naturally a function of continuous spatial variables $(\alpha,\beta) \in \mathbb{R}^2$ (and possibly a spectral variable $\lambda \in \mathbb{R}_+$);
- and accurately describes the formation of a set of N discrete observations grouped together in a vector $\boldsymbol{y} \in \mathbb{R}^N$.

A general description of the map of the sky as a function of continuous spatial variables can be written starting from a basic function ψ by combination and regular shifting:

$$\mathcal{X}(\alpha,\beta) = \sum_{ij} x_{ij}\, \psi\left(\alpha - i\,\delta_\alpha, \beta - j\,\delta_\beta\right). \quad (1)$$

The function ψ must be chosen so that this decomposition can describe the maps of interest and is easy to handle. It may be, among other choices, a pixel indicator, a cardinal sine function, or a wavelet (although in the last case the function ψ and the coefficients also depend on a scaling parameter). Whatever the choice, the map of interest is finally represented by its coefficients x_{ij}, the number of which is arbitrarily large and collected in $\boldsymbol{x} \in \mathbb{R}^M$ in what follows. In practice, we choose the Gaussian family as this greatly simplifies the (theoretical and numerical) calculations of the model outputs, including for complex models (Orieux et al. 2012b; Rodet et al. 2008).

The presented work is quite generic in the sense that it is not a priori attached to a specific instrument. It deals with a general linear instrument model that describes, at least to a fair approximation, the physics of the processes in play: optics, electrics, and thermodynamics. It also includes the passage from a continuous physical reality to a finite number of discrete observations. The instrument is then described by

$$\boldsymbol{y} = \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{x} + \boldsymbol{n}, \quad (2)$$

i.e. a general linear model w.r.t. $\boldsymbol{x}$ (a special case of which is the convolutive model). This model shows the instrument parameters $\boldsymbol{\eta} \in \mathbb{R}^K$ that define the form of the instrument response. The component $\boldsymbol{n} = \boldsymbol{y} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{x}$ represents the measuring and modelling errors additively. For the SPIRE instrument (Griffin et al. 2010) of the *Herschel* Space Observatory (Pilbratt et al. 2010) launched in May 2009, the paper of Orieux et al. (2012b) gives the details of the instrument model construction. The results of Sect. 5 are based on this instrument.

3. Probabilistic models and inversion

The proposed inversion is developed in the framework of Bayesian statistics. It relies on the posterior density $p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}|\boldsymbol{y})$ for the unknown quantities $\boldsymbol{x}$ (image), $\boldsymbol{\xi}$ (hyperparameters), and $\boldsymbol{\eta}$ (instrument parameters) given $\boldsymbol{y}$ (observations). This density brings together the information about the unknowns in the sense that it attaches more or less confidence to each value of the triplet $(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta})$. A summary of this density in the form of a mean and a standard deviation will provide (1) a point estimate (the posterior mean) for the map of interest and the parameters, and (2) an indication of the associated uncertainty (the posterior standard deviation).

Remark 1. In statistical terms (Robert 2005), the posterior mean is an optimal estimator. More precisely, of all the possible estimators (whether Bayesian or not, empirical or not, a computation code, etc.), the posterior mean yields the minimum mean square error (MMSE)[1]. Regarding first-order statistics, this estimator has, moreover, a zero mean bias.

[1] The mean square error is the expected value of the squared norm of the difference between estimated value and true value. The expectation is under the distribution of the observation and the unknown. The MSE is the sum of the variance and the squared bias of the estimator (under the distribution of the observation and the unknown).

The posterior density is deduced as the ratio of the joint density for all considered quantities $p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{y})$ and the marginal density for the observations $p(\boldsymbol{y})$ by application of Bayes' rule

$$p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}|\boldsymbol{y}) = \frac{p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{y})}{p(\boldsymbol{y})}. \tag{3}$$

Seen as a function of the unknowns $(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta})$, this posterior density is proportional to the joint density:

$$p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}|\boldsymbol{y}) \propto p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{y}\right). \tag{4}$$

This joint density is essential as all the other densities (marginal, conditional, prior, posterior, etc.) can be deduced from it. It can be factorised in various forms and, in preparatio for the developments to follow, we write

$$\begin{aligned} p(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta},\boldsymbol{y}) &= p\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}\right) p\left(\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}\right) \\ &= p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}) p\left(\boldsymbol{x}|\boldsymbol{\xi},\boldsymbol{\eta}\right) p\left(\boldsymbol{\xi},\boldsymbol{\eta}\right) \\ &= p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}) p\left(\boldsymbol{x}|\boldsymbol{\xi}\right) p(\boldsymbol{\xi}) p(\boldsymbol{\eta}), \end{aligned} \tag{5}$$

including the fact that (1) the hyperparameters $\boldsymbol{\xi}$ and the instrument parameters $\boldsymbol{\eta}$ are a priori independent and (2) the object $\boldsymbol{x}$ and the instrument parameters $\boldsymbol{\eta}$ are also a priori independent.

The different probability densities will be defined in the following sections according to the information available on each set of variables and according to practical concerns about dealing with the probability densities and numerical computation time.

3.1. Modelling of errors and likelihood

The factor $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta})$ in Eq. (5) is the density for the observations $\boldsymbol{y}$ given the map $\boldsymbol{x}$, the instrument parameters $\boldsymbol{\eta}$, and the hyperparameters $\boldsymbol{\xi}$, i.e. the likelihood of the unknowns attached to the observations.

Given Eq. (2), the construction of this likelihood is based on the model for the error $\boldsymbol{n}$. The analysis developed in this paper is essentially founded on its mean $m_{\boldsymbol{n}}$ and its covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{n}}$, and the proposed model is Gaussian:

$$\boldsymbol{n} \sim \mathcal{N}(m_{\boldsymbol{n}}, \boldsymbol{\Sigma}_{\boldsymbol{n}}). \tag{6}$$

The choice of the Gaussian model is also justified via information property: based on the sole information of finite mean and covariance, the Gaussian density is the model that introduces the least information (Kass & Wasserman 1996). This property is also mentioned as a maximum entropy property of the Gaussian density.

$m_{\boldsymbol{n}}$ is a scalar that models a possible non-zero mean for the noise, such as an offset (in the proposed numerical evaluation of Sect. 5, one offset for each bolometer is introduced). Regarding the covariance matrix, to lighten the notation, we set $\boldsymbol{\Sigma}_{\boldsymbol{n}}^{-1} = \gamma_{\boldsymbol{n}}\boldsymbol{\Pi}_{\boldsymbol{n}}$: $\gamma_{\boldsymbol{n}}$ is a scale factor (called precision, homogeneous to an inverse variance) and $\boldsymbol{\Pi}_{\boldsymbol{n}}$ contains the structure itself. For a stationary model $\boldsymbol{\Pi}_{\boldsymbol{n}}^{-1}$ has a Tœplitz structure, for an auto-regressive model $\boldsymbol{\Pi}_{\boldsymbol{n}}$ is a band matrix, for a independent model $\boldsymbol{\Pi}_{\boldsymbol{n}}$ is diagonal, for a white and stationary model $\boldsymbol{\Pi}_{\boldsymbol{n}} = \boldsymbol{I}$, the identity matrix. In the developments below, the structure of $\boldsymbol{\Pi}_{\boldsymbol{n}}$ is given while the scale factor $\gamma_{\boldsymbol{n}}$ and the mean $m_{\boldsymbol{n}}$ are unknown and included in the vector $\boldsymbol{\xi}$. The results in Sect. 5 are presented for the case $\boldsymbol{\Pi}_{\boldsymbol{n}} = \boldsymbol{I}$; hence $\gamma_{\boldsymbol{n}}$ is the inverse of the noise power.

Remark 2. The proposed developments account for characteristics of the error $\boldsymbol{n}$ that may differ from channel to channel, sensor to sensor, etc. This will be the case in Sect. 5: a mean and power of the noise will be assigned and estimated for each bolometer.

As the error $\boldsymbol{n}$ is Gaussian and additive (Eqs. (6) and (2)), the vector of observations $\boldsymbol{y}$, given $\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}$, is also Gaussian

$$\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta} \sim \mathcal{N}\left(\boldsymbol{m}_{\boldsymbol{y}|*}, \boldsymbol{\Sigma}_{\boldsymbol{y}|*}\right)$$

with mean

$$\boldsymbol{m}_{\boldsymbol{y}|*} = \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{x} + m_{\boldsymbol{n}} \tag{7}$$

and with the same covariance as $\boldsymbol{n}$: $\boldsymbol{\Sigma}_{\boldsymbol{y}|*} = \boldsymbol{\Sigma}_{\boldsymbol{n}}$. So, the likelihood of the unknowns attached to the observations reads

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\xi},\boldsymbol{\eta}) \;=\; & (2\pi)^{-N/2}\gamma_{\boldsymbol{n}}^{N/2}\det\left[\boldsymbol{\Pi}_{\boldsymbol{n}}\right]^{1/2} \\ & \times\exp\left[-\frac{1}{2}\gamma_{\boldsymbol{n}}\left(\boldsymbol{y}-\boldsymbol{m}_{\boldsymbol{y}|*}\right)^{t}\boldsymbol{\Pi}_{\boldsymbol{n}}\left(\boldsymbol{y}-\boldsymbol{m}_{\boldsymbol{y}|*}\right)\right]. \end{aligned} \tag{8}$$

It includes the information provided by the observations as the transform of a map $\boldsymbol{x}$ by the instrument, taking its parameters $\boldsymbol{\eta}$ and the noise parameters $\gamma_{\boldsymbol{n}}$ and $m_{\boldsymbol{n}}$ into consideration.

3.2. Prior density for the map and spatial regularity

The aim of this section is to introduce a prior density $p(\boldsymbol{x}|\boldsymbol{\xi})$ for the unknown map coefficients $\boldsymbol{x}$ based upon available information about the map $\mathcal{X}$. The present work is mainly devoted to extended emissions. From a spatial standpoint, such maps are relatively regular, i.e. they involve positive correlation. From the spectral standpoint, the power is mainly located at relatively low frequencies. The Gaussian density includes these second-order properties in a simple way. This choice can also be justified based on a maximum entropy principle. Its main interest here is to result in a linear processing method. It is written in the form

$$p(\boldsymbol{x}|\gamma_{\boldsymbol{x}}) = (2\pi)^{-M/2}\gamma_{\boldsymbol{x}}^{M/2}\det\left[\boldsymbol{\Pi}_{\boldsymbol{x}}\right]^{1/2}\exp\left[-\frac{1}{2}\gamma_{\boldsymbol{x}}\boldsymbol{x}^{t}\boldsymbol{\Pi}_{\boldsymbol{x}}\boldsymbol{x}\right], \tag{9}$$

where $\gamma_{\boldsymbol{x}}$ is a precision parameter (homogeneous to an inverse-variance that controls the regularity strength) and $\boldsymbol{\Pi}_{\boldsymbol{x}}$ is a precision matrix (homogeneous to an inverse-covariance matrix that controls the regularity structure). When the precision $\gamma_{\boldsymbol{x}}$ is low (strong prior variance), the regularity information is weakly taken into account. Conversely, when the precision $\gamma_{\boldsymbol{x}}$ is high (weak prior variance), the penalization of non-regular maps is high, i.e. the regularity is strongly imposed.

The subsequent developments are devoted to the design of $\boldsymbol{\Pi}_{\boldsymbol{x}}$ to account for the desired regularity of the map. A simple regularity measure $R_{\mathrm{c}}[\mathcal{X}]$ of the map $\mathcal{X}$ is the energy of some of its derivatives. These derivatives can address the spatial variables (α,β) separately, can rely on cross derivatives and can intervene at various orders. This is the classical Philipps-Twomey-Thikonov penalization idea (Tikhonov & Arsenin 1977). It can also embed directional derivatives or any differential operator (Mallat 2008). In the simplest and natural case, we choose

$$R_{\mathrm{c}}[\mathcal{X}] = \left\|\frac{\partial\mathcal{X}}{\partial\alpha}\right\|^{2} + \left\|\frac{\partial\mathcal{X}}{\partial\beta}\right\|^{2},$$

where $\|u\|$ is the standard function norm [2]. Given the decomposition (1), it is easy to establish the partial derivatives of $\mathcal{X}$ from the derivatives of ψ. In the direction α, by noting $\psi'_{\alpha} = \partial\psi/\partial\alpha$, we have

$$\begin{aligned} \left\|\frac{\partial\mathcal{X}}{\partial\alpha}\right\|^{2} \;=\; & \sum_{ij\,i'j'} x_{ij}\,x_{i'j'}\int_{\mathbb{R}^2}\psi'_{\alpha}\left(\alpha - i'\,\delta_{\alpha},\beta - j'\,\delta_{\beta}\right) \\ & \times\psi'_{\alpha}\left(\alpha - i\,\delta_{\alpha},\beta - j\,\delta_{\beta}\right)\mathrm{d}\alpha\,\mathrm{d}\beta, \end{aligned}$$

[2] The function squared norm is defined by $\|u\|^2 = \iint u(\alpha,\beta)^2\,\mathrm{d}\alpha\,\mathrm{d}\beta$.

which brings out the autocorrelation $\Psi_\alpha = \psi'_\alpha \star \psi'_\alpha$ of the derivative of ψ. We then have a quadratic form in $\boldsymbol{x}$

$$\left\|\frac{\partial \mathcal{X}}{\partial \alpha}\right\|^2 = \sum_{ij\ i'j'} x_{ij}\, x_{i'j'} \Psi_\alpha \left[(i'-i)\,\delta_\alpha, (j'-j)\,\delta_\beta\right] = \boldsymbol{x}^t \boldsymbol{\Psi}_\alpha \boldsymbol{x}.$$

As the coefficients $\Psi_\alpha \left[(i'-i)\,\delta_\alpha, (j'-j)\,\delta_\beta\right]$ depend only on the difference between indices, the matrix $\boldsymbol{\Psi}_\alpha$ has a Tœplitz structure and the computations amounts to a discrete convolution that can be efficiently implemented by the use of fast Fourier transform (FFT). Finally, by performing the same development in the β dimension, a global quadratic norm appears: $R_c[\mathcal{X}] = \boldsymbol{x}^t(\boldsymbol{\Psi}_\alpha + \boldsymbol{\Psi}_\beta)\boldsymbol{x}$ and designs the precision matrix $\boldsymbol{\Pi}_{\boldsymbol{x}} = \boldsymbol{\Psi}_\alpha + \boldsymbol{\Psi}_\beta$. For more details and for a spectral interpretation, see Sect. 2.1 and Appendix A of (Orieux et al. 2012b).

3.3. Prior distribution for hyperparameters (hyperprior)

The hyperparameters are the unknown parameters of the densities for the error and for map Eqs. (8), (9) and they are collected in the vector $\boldsymbol{\xi} = [m_{\boldsymbol{n}}, \gamma_{\boldsymbol{n}}, \gamma_{\boldsymbol{x}}]$. It has been said that $\gamma_{\boldsymbol{x}}$ and $\gamma_{\boldsymbol{n}}$ are the precisions (scale parameters) and $m_{\boldsymbol{n}}$ is a mean (position parameter) of Gaussian densities.

The choice of the prior distributions for these hyperparameters is driven by two requirements: (i) little information is available a priori on their values and their relations and (ii) the chosen distributions must lead to efficient algorithms (see Sect. 4.2). Following this line of thought, we choose a prior distribution determined by Jeffreys' principle[3]: $p(\gamma) = 1/\gamma$ for $\gamma_{\boldsymbol{x}}$ and $\gamma_{\boldsymbol{n}}$ and $p(m_{\boldsymbol{n}}) = 1$. Moreover, regarding the triplet of hyperparameters $\boldsymbol{\xi} = [m_{\boldsymbol{n}}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{n}}]$, they are modelled as independent variables, since no information is available about their eventual relations. Finally

$$p(m_{\boldsymbol{n}}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{n}}) = 1/\gamma_{\boldsymbol{x}}\gamma_{\boldsymbol{n}}, \tag{10}$$

has two advantages

1. First of all, the posterior conditional densities for $\gamma_{\boldsymbol{x}}$ and $\gamma_{\boldsymbol{n}}$ (resp. for $m_{\boldsymbol{n}}$), as shown in Sect. 4.2, will be gamma densities (resp. Gaussian density), which will make the implementation easier.
2. This prior distribution is non-informative (which introduces a minimum of information on the value of the hyperparameters) in the sense that it is invariant by certain parameterization changes (Robert 2005; Kass & Wasserman 1996).

3.4. Prior density for the instrument parameters

The instrument parameter $\boldsymbol{\eta}$ operates in a complex nonlinear way in the description of the observations. In consequence, whatever the prior density, the conditional posterior density for $\boldsymbol{\eta}$ (see Sect. 4.3) will not have a standard form. The choice is thus purely oriented by the information on the instruments and the question that arises concerns the encoding of the available information in the form of a probability density. If we have no information except a minimum and a maximum value for a given parameter, the choice of a uniform density over the interval is a reasonable one. If we have a nominal value with an associated uncertainty and no other information, the most suitable choice is a Gaussian density. The rest of the development is valid whatever the choice, and we consider the Gaussian case in the following.

[3] It yields a non-informative prior distribution based on a key feature that it is invariant under reparameterization. It is deduced as the determinant of the Fisher information matrix.

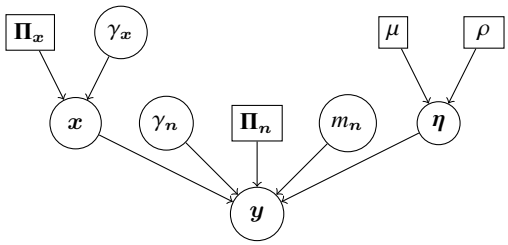


Fig. 1. Graphical dependency representation (hierarchical structure). The round (square) nodes correspond to unknown (fixed) quantities. The directions of the arrows indicate the dependencies.

In addition, having no information available about possible links among the various parameters, we take it that the parameters are, a priori, independent and thus

$$p(\boldsymbol{\eta}) = \prod_{k=1}^{K} (2\pi\rho_k)^{-1/2} \exp\left[-\frac{(\eta_k - \mu_k)^2}{2\rho_k}\right]. \tag{11}$$

In practice and for the first results presented in Sect. 5, the means μ_k and variances ρ_k were taken from the SPIRE observer manual or were fixed ad-hoc at plausible values.

3.5. Posterior density: histograms, mean and standard deviation

The posterior density (3) for all the unknowns $\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\eta}$, is deduced from the joint density (5) for all quantities concerned as follows:

$$p(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\eta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\eta}) p(\boldsymbol{x}|\boldsymbol{\xi}) p(\boldsymbol{\xi}) p(\boldsymbol{\eta}).$$

In this expression,

- the density $p(\boldsymbol{x}|\gamma_{\boldsymbol{x}})$ for the unknown map $\boldsymbol{x} \in \mathbb{R}^M$ is Gaussian (Eq. (9));
- the distribution $p(\boldsymbol{\xi})$ (for $\boldsymbol{\xi} = [m_{\boldsymbol{n}}, \gamma_{\boldsymbol{n}}, \gamma_{\boldsymbol{x}}]$) is a Jeffreys' distribution (Eq. (10));
- the instrument parameters $\boldsymbol{\eta} \in \mathbb{R}^K$ is modelled by a Gaussian density $p(\boldsymbol{\eta})$ (Eq. (11));
- the density $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\eta})$ for the observations $\boldsymbol{y} \in \mathbb{R}^N$ given the rest of the variables (i.e. likelihood) is a Gaussian density (Eq. (8)) and it is a function of $\boldsymbol{x}$ through $\boldsymbol{m}_{\boldsymbol{y}|*}$ given by Eq. (7).

Finally, from Eqs. (8), (9), (10), and (11), the posterior density can be written

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\eta}|\boldsymbol{y}) \propto\ & \gamma_{\boldsymbol{x}}^{M/2-1}\, \gamma_{\boldsymbol{n}}^{N/2-1} \prod_{k=1}^{K} (2\pi\rho_k)^{-1} \\ & \times \exp\left[-\frac{1}{2}\frac{(\eta_k - \mu_k)^2}{\rho_k}\right] \exp\left[-\frac{1}{2}\,\gamma_{\boldsymbol{x}}\, \boldsymbol{x}^t \boldsymbol{\Pi}_{\boldsymbol{x}} \boldsymbol{x}\right] \\ & \times \exp\left[-\frac{1}{2}\gamma_{\boldsymbol{n}} (\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{y}|*})^t \boldsymbol{\Pi}_{\boldsymbol{n}} (\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{y}|*})\right]. \end{aligned} \tag{12}$$

This density brings together all information about the unknowns, and the estimators and algorithms presented below are entirely based on it. However, it is too complex to be analyzed directly as a whole and the difficulty stems from (i) the dimension of $\boldsymbol{x}$ (of size $M \sim 10^5$ in practice) and (ii) the joint presence of other parameters (hyperparameters $\boldsymbol{\xi}$ and instrument parameters $\boldsymbol{\eta}$). Moreover, for the latter in particular, the dependence

Table 1. Gibbs algorithm.

Initialize $\boldsymbol{x}^{(0)}, \boldsymbol{\xi}^{(0)}, \boldsymbol{\eta}^{(0)}$
for $q = 1, 2, \ldots$ **do**
(1) sample $\boldsymbol{x}^{(q)}$ under $p(\boldsymbol{x}|\boldsymbol{\xi}^{(q-1)}, \boldsymbol{\eta}^{(q-1)}, \boldsymbol{y})$
(2) sample $\boldsymbol{\xi}^{(q)}$ under $p(\boldsymbol{\xi}|\boldsymbol{x}^{(q)}, \boldsymbol{\eta}^{(q-1)}, \boldsymbol{y})$
(3) sample $\boldsymbol{\eta}^{(q)}$ under $p(\boldsymbol{\eta}|\boldsymbol{x}^{(q)}, \boldsymbol{\xi}^{(q)}, \boldsymbol{y})$
end for

is complicated and cannot be identified with a standard form. The proposed approach is to explore the posterior density by means of stochastic sampling (Robert & Casella 2004; Gilks et al. 1996). The idea is to produce a set of samples $\boldsymbol{x}^{(q)}, \boldsymbol{\xi}^{(q)}, \boldsymbol{\eta}^{(q)}$, for $q = 1, 2, \ldots, Q$, drawn at random under the posterior density. It is then possible, for example, to deduce histograms that approximate marginal densities, together with means and standard deviations. This strategy is by no means new but interest in its practical use has revived in recent years as new forms of algorithms have been developed and computer power has increased.

Concerning the estimates themselves for the map, the hyperparameters and the instrument parameters, we choose the posterior mean (PM), as indicated at the beginning of Sect. 3 (see also Remark 1). We will also look at the dispersion around the mean through the posterior standard deviation (PSD) and the links among components through posterior correlations. Using the set of samples $\boldsymbol{x}^{(q)}, \boldsymbol{\xi}^{(q)}, \boldsymbol{\eta}^{(q)}$, for $q = 1, 2, \ldots, Q$, the posterior mean $\boldsymbol{\mu}_{\mathrm{P}}$ and the posterior covariance matrix $\boldsymbol{\Gamma}_{\mathrm{P}}$ are computed by

$$\boldsymbol{\mu}_{\mathrm{P}} \approx \frac{1}{Q} \sum_{q=1}^{Q} \bar{\boldsymbol{x}}^{(q)} \tag{13}$$

$$\boldsymbol{\Gamma}_{\mathrm{P}} \approx \frac{1}{Q} \sum_{q=1}^{Q} \left(\bar{\boldsymbol{x}}^{(q)} - \boldsymbol{\mu}_{\mathrm{P}}\right)\left(\bar{\boldsymbol{x}}^{(q)} - \boldsymbol{\mu}_{\mathrm{P}}\right)^{t}, \tag{14}$$

where $\bar{\boldsymbol{x}}$ denotes the column concatenation $\bar{\boldsymbol{x}} = [\boldsymbol{x}; \boldsymbol{\xi}; \boldsymbol{\eta}]$. Practically, it is not possible to compute the entire covariance $\boldsymbol{\Gamma}_{\mathrm{P}}$, but it is possible to compute its diagonal elements to characterize the marginal errors for each component (each pixel, hyperparameters, instrument parameters) and to compute a few nondiagonal elements to measure the correlations between components.

4. Exploration of posterior density and the computation algorithm

We have introduced an instrument model and various probability densities to define the posterior density that brings together the information on the map and the parameters (hyperparameters and instrument parameters). We have also defined the posterior mean (PM) as an estimate and the posterior standard deviation (PSD) as a measure of the uncertainty. We have then introduced the idea of computations via stochastic sampling. The developments in the present section concern the algorithm for computing these samples.

The production of samples of the posterior density for the set $(\boldsymbol{x}, \boldsymbol{\xi}, \boldsymbol{\eta})$ is not possible directly because of the complexity of the density. We therefore use a Gibbs algorithm (Robert & Casella 2004; Gilks et al. 1996), which breaks the problem down into three simpler subproblems: sampling $\boldsymbol{x}$, $\boldsymbol{\xi}$, and $\boldsymbol{\eta}$ separately. This is an iterative algorithm, described in Table 1: each variable $\boldsymbol{x}$, $\boldsymbol{\xi}$, and $\boldsymbol{\eta}$ is drawn under its conditional posterior density given the current value of the other two variables. For each of the three steps, this conditional posterior density can be deduced directly (up to a multiplying factor) from the posterior density (12): all we have to do is to keep only the factors depending on the variable of interest. This algorithm is a Markov chain Monte-Carlo (MCMC) algorithm (Robert & Casella 2004; Gilks et al. 1996) and is known to give (after a certain time, called the burn-in time) samples under the posterior density.

The conditional density of the map coefficients $\boldsymbol{x}$ (Step (1), Table 1) is Gaussian (Sect. 4.1). For the precisions $\boldsymbol{\xi}$ (Step (2), Table 1), the conditional densities are gamma densities (Sect. 4.2). They will be sampled using standard existing numerical routines (e.g. in Matlab). In contrast, the conditional density of the instrument parameters $\boldsymbol{\eta}$ (Step (3), Table 1) has a much more complex nonstandard form, so that it cannot be directly sampled by existing routines. To overcome this difficulty, sampling was carried out by means of a Metropolis-Hastings step (Sect. 4.3).

4.1. Map sampling

The density for the map $\boldsymbol{x}$ conditionally on the other variables is deduced from (12) by extracting the factors depending on $\boldsymbol{x}$:

$$p(\boldsymbol{x}|\boldsymbol{y}, \gamma_{\boldsymbol{x}}, \gamma_{\boldsymbol{n}}, \boldsymbol{\eta}) \propto \exp\left[\frac{1}{2}\gamma_{\boldsymbol{x}} \boldsymbol{x}^{t} \boldsymbol{\Pi}_{\boldsymbol{x}} \boldsymbol{x} + \gamma_{\boldsymbol{n}} (\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{y}|*})^{t} \boldsymbol{\Pi}_{\boldsymbol{n}} (\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{y}|*})\right]. \tag{15}$$

Considering the expression for $\boldsymbol{m}_{\boldsymbol{y}|*}$ given by Eq. (7), the argument of this exponential is quadratic in $\boldsymbol{x}$. We deduce that we have a Gaussian density and, by rearranging the argument, we can determine the covariance and the mean

$$\boldsymbol{\Sigma}_{\boldsymbol{x}|*} = \left(\gamma_{\boldsymbol{n}} \boldsymbol{A}_{\boldsymbol{\eta}}^{t} \boldsymbol{\Pi}_{\boldsymbol{n}} \boldsymbol{A}_{\boldsymbol{\eta}} + \gamma_{\boldsymbol{x}} \boldsymbol{\Pi}_{\boldsymbol{x}}\right)^{-1} \tag{16}$$

$$\boldsymbol{m}_{\boldsymbol{x}|*} = \gamma_{\boldsymbol{n}} \boldsymbol{\Sigma}_{\boldsymbol{x}|*} \boldsymbol{A}_{\boldsymbol{\eta}}^{t} \boldsymbol{\Pi}_{\boldsymbol{n}} \boldsymbol{y}, \tag{17}$$

Remark 3. For a fixed value of the hyperparameters and the instrument parameters, the map $\boldsymbol{m}_{\boldsymbol{x}|*}$ defined by Eqs. (16), (17) is the maximizer (and the mean) of the conditional posterior density (15). This is the regularized least-squares solution denoted $\widehat{\boldsymbol{x}}(\mu)$ parameterized by $\mu = \gamma_{\boldsymbol{x}}/\gamma_{\boldsymbol{n}}$. This corresponds to the solution defined in our previous paper (Orieux et al. 2012b). For a convolutive instrument model, it is the Wiener solution (also called Wiener-Hunt solution; Orieux et al. 2010b).

Step (1) of Table 1 consists of sampling this Gaussian but this operation is made very difficult by three elements: (i) the large size of the map; (ii) the correlation introduced by the instrument model and the prior density; and (iii) the absence of structure of the instrument model (invariance, sparse nature). This problem can be solved using several approaches. For a convolutive instrument model (2), it is possible to approximately diagonalise the correlation matrix by FFT, thus producing a sample for the cost of an FFT (Chellappa & Chatterjee 1985; Chellappa & Jain 1992; Geman & Yang 1995; Giovannelli 2008; Orieux et al. 2010b). If where the inverse of the correlation matrix is sparse, a partially parallel Gibbs sampler may be particularly efficient (Winkler 2003, Chap. 8). In the present case, neither the correlation nor its inverse possess the required properties. A general solution relies on factorizing the correlation matrix (Cholesky decomposition, diagonalization, etc.) but the large size of the matrix ($M \times M$ with $M \sim 10^5$) does not permit the required calculations to be performed here.

The proposed solution consists of constructing a criterion such that its minimizer is a sample under the desired posterior conditional density. To do this, we perturb the means of the noise

component and of the map component by an additive component with covariance $\gamma_{\boldsymbol{x}}\boldsymbol{\Pi}_{\boldsymbol{x}}$ and $\gamma_{\boldsymbol{n}}\boldsymbol{\Pi}_{\boldsymbol{n}}$. A perturbed regularized least squares criterion is then introduced

$$J(\boldsymbol{x}) = \gamma_{\boldsymbol{n}} \left(\widetilde{\boldsymbol{y}} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{x}\right)^t \boldsymbol{\Pi}_{\boldsymbol{n}} \left(\widetilde{\boldsymbol{y}} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{x}\right) + \gamma_{\boldsymbol{x}} \left(\boldsymbol{x} - \widetilde{\boldsymbol{m}}_x\right)^t \boldsymbol{\Pi}_{\boldsymbol{x}} (\boldsymbol{x} - \widetilde{\boldsymbol{m}}_x),$$

and it can be shown (see Orieux et al. 2012a) that its minimizer

$$\widetilde{\boldsymbol{x}} = \left(\gamma_{\boldsymbol{n}} \boldsymbol{A}_{\boldsymbol{\eta}}^t \boldsymbol{\Pi}_{\boldsymbol{n}} \boldsymbol{A}_{\boldsymbol{\eta}} + \gamma_{\boldsymbol{x}} \boldsymbol{\Pi}_{\boldsymbol{x}}\right)^{-1} \times \left(\gamma_{\boldsymbol{n}} \boldsymbol{A}_{\boldsymbol{\eta}}^t \boldsymbol{\Pi}_{\boldsymbol{n}} \widetilde{\boldsymbol{y}} + \gamma_{\boldsymbol{x}} \boldsymbol{\Pi}_{\boldsymbol{x}} \widetilde{\boldsymbol{m}}_x\right) \quad (18)$$

is Gaussian and does indeed have the correlation and mean defined by (16) and (17). This very powerful result has already been used by (Féron 2006; Orieux et al. 2012b). In different forms, similar ideas have been introduced and used by (Rue 2001; Rue & Held 2005; Lalanne et al. 2001; Tan et al. 2010).

Remark 4. For the non perturbed criterion ($\widetilde{\boldsymbol{y}} = \boldsymbol{y}$ and $\widetilde{\boldsymbol{m}}_x = 0$), we have the regularized least-squares solution Eqs. (16), (17), that was mentioned in Remark 3.

Remark 5. The approach described in Orieux et al. (2012a) involves the sampling of the prior density (9) that is not properly defined here: the matrix $\boldsymbol{\Pi}_{\boldsymbol{x}}$ does not penalise the mean of the map (it is of deficient rank). But, for the same reason, the solution (18) does not depend on the mean of the realization of the prior density. Therefore, the simulated sample can have an arbitrary mean value.

4.2. Hyperparameter sampling

To determine the posterior conditional density for $\gamma_{\boldsymbol{x}}$, we examine the posterior density (12), and only keep the factors where $\gamma_{\boldsymbol{x}}$ appears, which gives

$$p(\gamma_{\boldsymbol{x}}|\boldsymbol{y}, \boldsymbol{x}, \gamma_{\boldsymbol{n}}, \boldsymbol{\eta}) \propto \gamma_{\boldsymbol{x}}^{M/2-1} \exp\left[-\frac{\gamma_{\boldsymbol{x}}}{2}\|\boldsymbol{x}\|^2_{\boldsymbol{\Pi}_{\boldsymbol{x}}}\right],$$

and we recognize a Gamma density (see Appendix A)

$$\gamma_{\boldsymbol{x}} \sim \mathcal{G}\left(M/2\ ,\ 2/\|\boldsymbol{x}\|^2_{\boldsymbol{\Pi}_{\boldsymbol{x}}}\right). \quad (19)$$

Concerning $\gamma_{\boldsymbol{n}}$, we also refer to the posterior density (12) and find

$$p(\gamma_{\boldsymbol{n}}|\boldsymbol{y}, \boldsymbol{x}, \gamma_{\boldsymbol{x}}, \boldsymbol{\eta}) \propto \gamma_{\boldsymbol{n}}^{N/2-1} \exp\left[-\frac{1}{2}\gamma_{\boldsymbol{n}}\|\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{y}|*}\|^2_{\boldsymbol{\Pi}_{\boldsymbol{n}}}\right],$$

which is also a Gamma density

$$\gamma_{\boldsymbol{n}} \sim \mathcal{G}\left(N/2\ ,\ 2/\|\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{y}|*}\|^2_{\boldsymbol{\Pi}_{\boldsymbol{n}}}\right). \quad (20)$$

For both $\gamma_{\boldsymbol{x}}$ and $\gamma_{\boldsymbol{n}}$ the second parameter of the Gamma density introduces a quadratic norm (regularity of the map in Eq. (19), and goodness-of-fit in Eq. (21)), which can be easily computed.

Remark 6. An intuitive interpretation can be given to these results starting from the fact that the mean of the Gamma density is equal to the product of its parameters (see Appendix A), here $N/\left\|\boldsymbol{y} - \boldsymbol{m}_{\boldsymbol{y}|*}\right\|^2_{\boldsymbol{\Pi}_{\boldsymbol{n}}}$ for (21). In this sense, the conditional posterior mean is the inverse of the empirical variance of the residuals. Consequently, when the goodness-of-fit term is small, the mean of the density is large and so the sampled value of $\gamma_{\boldsymbol{n}}$ is also high reporting a high precision, i.e. a weak variance (and vice versa). The same holds for the map regularity in relation with the mean of the density (19) given by $M/\|\boldsymbol{x}\|^2_{\boldsymbol{\Pi}_{\boldsymbol{x}}}$. These observations support the coherence of the model and reinforce the prior choice for these hyperparameters as a convenient one.

Table 2. Step of the Metropolis-Hastings sampler, which replaces Step (3) of Table 1. The current sample at step q is $\boldsymbol{\eta}^{(q)}$ and it is either replaced or not by the proposed sample $\boldsymbol{\eta}^{\mathrm{p}}$.

(a) Draw a sample $\boldsymbol{\eta}^{\mathrm{p}}$ under a proposal density.
(b) Compute the acceptation ratio ρ by Eq. (23).
(c) Replace $\boldsymbol{\eta}^{(q)}$ by $\boldsymbol{\eta}^{\mathrm{p}}$ (i.e. $\boldsymbol{\eta}^{(q+1)} = \boldsymbol{\eta}^{\mathrm{p}}$) with the probability $\min(1, \rho)$, otherwise keep $\boldsymbol{\eta}^{(q)}$ (i.e. $\boldsymbol{\eta}^{(q+1)} = \boldsymbol{\eta}^{(q)}$).

Regarding the mean of the noise, $m_{\boldsymbol{n}}$, it is a scalar whose posterior conditional density is also deduced from the posterior density (12) and from (7)

$$p(m_{\boldsymbol{n}}|\boldsymbol{y}, \boldsymbol{x}, \gamma_{\boldsymbol{n}}, \gamma_{\boldsymbol{x}}, \boldsymbol{\eta}) \propto \exp\left[-\frac{1}{2}\gamma_{\boldsymbol{n}}(m_{\boldsymbol{n}} - m_r)^2\right],$$

where m_r is the empirical mean of the residuals $\boldsymbol{y} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{x}$. We then have a Gaussian density

$$m_{\boldsymbol{n}} \sim \mathcal{N}(m_r, \gamma_{\boldsymbol{n}}). \quad (21)$$

In the numerical evaluation of Sect. 5, one such mean is estimated for each bolometer.

Remark 7. If we examine the relationships above, we see that $p(\gamma_{\boldsymbol{x}}|\boldsymbol{y}, \boldsymbol{x}, \gamma_{\boldsymbol{n}}, \boldsymbol{\eta}) = p(\gamma_{\boldsymbol{x}}|\boldsymbol{x})$, in other words, $\gamma_{\boldsymbol{x}}$ and $(\boldsymbol{y}, \gamma_{\boldsymbol{n}}, \boldsymbol{\eta})$ are independent conditionally on $\boldsymbol{x}$. Similarly, we note that $p(\gamma_{\boldsymbol{n}}|\boldsymbol{y}, \boldsymbol{x}, \gamma_{\boldsymbol{x}}, \boldsymbol{\eta}) = p(\gamma_{\boldsymbol{n}}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\eta})$, which means that $\gamma_{\boldsymbol{n}}$ and $\gamma_{\boldsymbol{x}}$ are independent conditionally on $(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\eta})$. In addition, $m_{\boldsymbol{n}}$ is independant of $\gamma_{\boldsymbol{x}}$, given $\boldsymbol{y}, \boldsymbol{x}, \gamma_{\boldsymbol{n}}, \boldsymbol{\eta}$.

4.3. Instrument parameter sampling

The last step (Step (3) of Table 1) is more complex. As for the other variables, the posterior conditional density can be deduced from the posterior density (12) by keeping the factors that bring in $\boldsymbol{\eta}$. There are two of these: the likelihood and the prior density, and we thus have

$$p(\boldsymbol{\eta}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\xi}) \propto \prod_{k=1}^{K} \exp\left[-\frac{1}{2}\frac{(\eta_k - \mu_k)^2}{\rho_k}\right] \times \exp\left[-\frac{1}{2}\gamma_{\boldsymbol{n}}(\boldsymbol{y} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{x})^t \boldsymbol{\Pi}_{\boldsymbol{n}} (\boldsymbol{y} - \boldsymbol{A}_{\boldsymbol{\eta}}\boldsymbol{x})\right]. \quad (22)$$

However, this is not a usual density, notably because there is no simple mathematical form to represent the dependence of the observation w.r.t. $\boldsymbol{\eta}$. Thus, Step (3) of Table 1 cannot be carried out directly with standard sampling routines and we resort to a Metropolis-Hastings step in a random-walk version (Robert & Casella 2004; Gilks et al. 1996) described in Table 2. It can be briefly explained as follows. Because it is impossible to draw a sample directly under the conditional posterior density (22), a sample is drawn under another density (namely the proposal density), but is not systematically accepted. Acceptance or rejection is also random with a precisely defined probability (see Eq. (23)) to ensure that, at convergence, we have samples under the target density (Robert & Casella 2004; Gilks et al. 1996). The algorithm is divided into three sub-steps summarized in Table 2 and detailed here.

(a) Draw a proposal $\boldsymbol{\eta}^{\mathrm{p}}$ as a perturbation of the current value: $\boldsymbol{\eta}^{\mathrm{p}} = \boldsymbol{\eta}^{(q)} + \boldsymbol{\varepsilon}$, deduce the instrument matrix $\boldsymbol{A}_{\boldsymbol{\eta}^{\mathrm{p}}}$, and the corresponding model output $\boldsymbol{m}_{\boldsymbol{y}}^{\mathrm{p}} = \boldsymbol{A}_{\boldsymbol{\eta}^{\mathrm{p}}}\boldsymbol{x}^{(q)} + m_{\boldsymbol{n}}$.

(b) Compute the acceptation ratio

$$\rho = \frac{p(\boldsymbol{\eta}^{\mathrm{p}}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\xi})}{p(\boldsymbol{\eta}^{(q)}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\xi})}, \tag{23}$$

based on the conditionnal posterior law ratio that compares the goodness-of-fit for the current parameter and the proposed one.

(c) Accept or reject the proposal, at random, with probability $\min(1, \rho)$. To do so, draw u uniformly in $[0, 1]$ and take

$$\boldsymbol{\eta}^{(q+1)} = \begin{cases} \boldsymbol{\eta}^{\mathrm{p}} & \text{if } u < \min\{1, \rho\} \\ \boldsymbol{\eta}^{(q)} & \text{otherwise.} \end{cases}$$

These three substeps are inserted instead of Step (3) of Table 1.

The algorithm can be explained as follows. Starting with a current value $\boldsymbol{\eta}^{(q)}$, the algorithm proposes a new value $\boldsymbol{\eta}^{\mathrm{p}}$ and compares the goodness-of-fit for the two values. When the proposed value improves the fit, $\rho > 1$ and $\boldsymbol{\eta}^{\mathrm{p}}$ is accepted. When the proposed value degrades the fit, $\boldsymbol{\eta}^{\mathrm{p}}$ can be accepted or rejected, with a probability that is higher or lower, depending on how weak the degradation is.

Remark 8. There are other more complex (and potentially more efficient) approaches for Metropolis-Hastings sampling. In particular, the proposal density can be adapted, e.g. directional random walk (Vacar et al. 2011). They are not exploited here but are considered in the development perspectives.

5. Numerical results

The previous sections presented the approach for building the posterior density and for its exploration by stochastic sampling using a Gibbs algorithm including a Metropolis-Hastings step. The mean and standard deviation (SD) of the posterior density are numerically computed as empirical averages based on simulated samples, from relations (13) and (14). The developments below show the practicability of the proposed method (models, estimate and algorithm), and provide a first numerical evaluation.

5.1. Evaluation methodology

The evaluation is based on the SPIRE instrument (Griffin et al. 2010) of the *Herschel* Space Observatory (Pilbratt et al. 2010) launched in May 2009. It focuses on the PMW channel (centred around 350 μm) and the Large Map protocol in the nominal operating conditions: scan back and forth with constant speed (30″/s) over two almost perpendicular directions (88°). The scans are associated with a high sampling frequency ($F_{\mathrm{s}} \approx 30$ Hz) providing spatially redundant observation and Fig. 2 shows the corresponding redundancy/pointing map. The spatial shift between basis functions (see Eq. (1)) is fixed at $\delta_\alpha = \delta_\beta = 2''$, based on our earlier work (Orieux et al. 2009, 2012b), to obtain the best gain in resolution without important increase of the computational cost. The angular size of the reconstructed map is $20' \times 20'$, i.e. a map of 600×600 coefficients. The associated direct model, including the whole acquisition chain (scanning strategy, mirror, horns, wavelength filters, bolometers, and electronics) is detailed in our previous paper (Orieux et al. 2012b) and represented by Eq. (2) of the present paper.

The unsupervised method is assessed based on two synthetic maps of extended emission (the Galactic Cirrus (Fig. 3e) and a realization of the prior density Eq. (9)) as well as based on a real observation (reflection nebula NGC 7023, Fig. 7). The paper also proposes a first assessment of the unsupervised and myopic approach based on a synthetic map with broader spectral content (the Galactic Cirrus with point sources (Fig. 3e)).

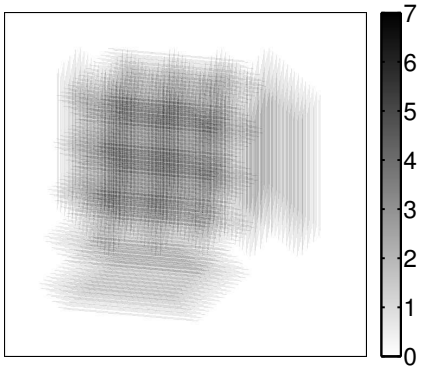


Fig. 2. Redundancy/pointing map associated with our experiment of five crossed scans.

In the simulated cases, a zero-mean white Gaussian noise is added to the model output. Moreover, in these cases, since the original map (the "sky truth" denoted by $\boldsymbol{x}^*$) is known, the quality of the reconstruction (denoted by $\widehat{\boldsymbol{x}}$) can be quantified through an error:

$$\mathcal{E} = \sum_{i,j} |x^*_{ij} - \widehat{x}_{ij}|^2 \Big/ \sum_{i,j} |x^*_{ij}|^2, \tag{24}$$

where only coefficients in the observed area are taken into account, allowing assessments and comparisons between methods.

5.2. Algorithm behaviour and general comments

As explained in Sect. 4, the algorithm provides a series of samples that form a Markov chain for hyperparameters, instrument parameter and map. The MCMC theory then ensures that it correctly explores the parameter space and produces a density of samples reflecting the posterior density. Practically, the algorithm has been executed for the unsupervised problem as well as for the unsupervised and myopic problem. It has been run several times (1) using identical initial conditions and (2) using different initial conditions. In both cases, the same qualitative and quantitative behaviour as presented here has been systematically observed.

The computation time takes about one hour for the unsupervised (nonmyopic) case and about ten hours for the unsupervised and myopic case. The main computational cost is due to computating the instrument model output given by Eq. (2).

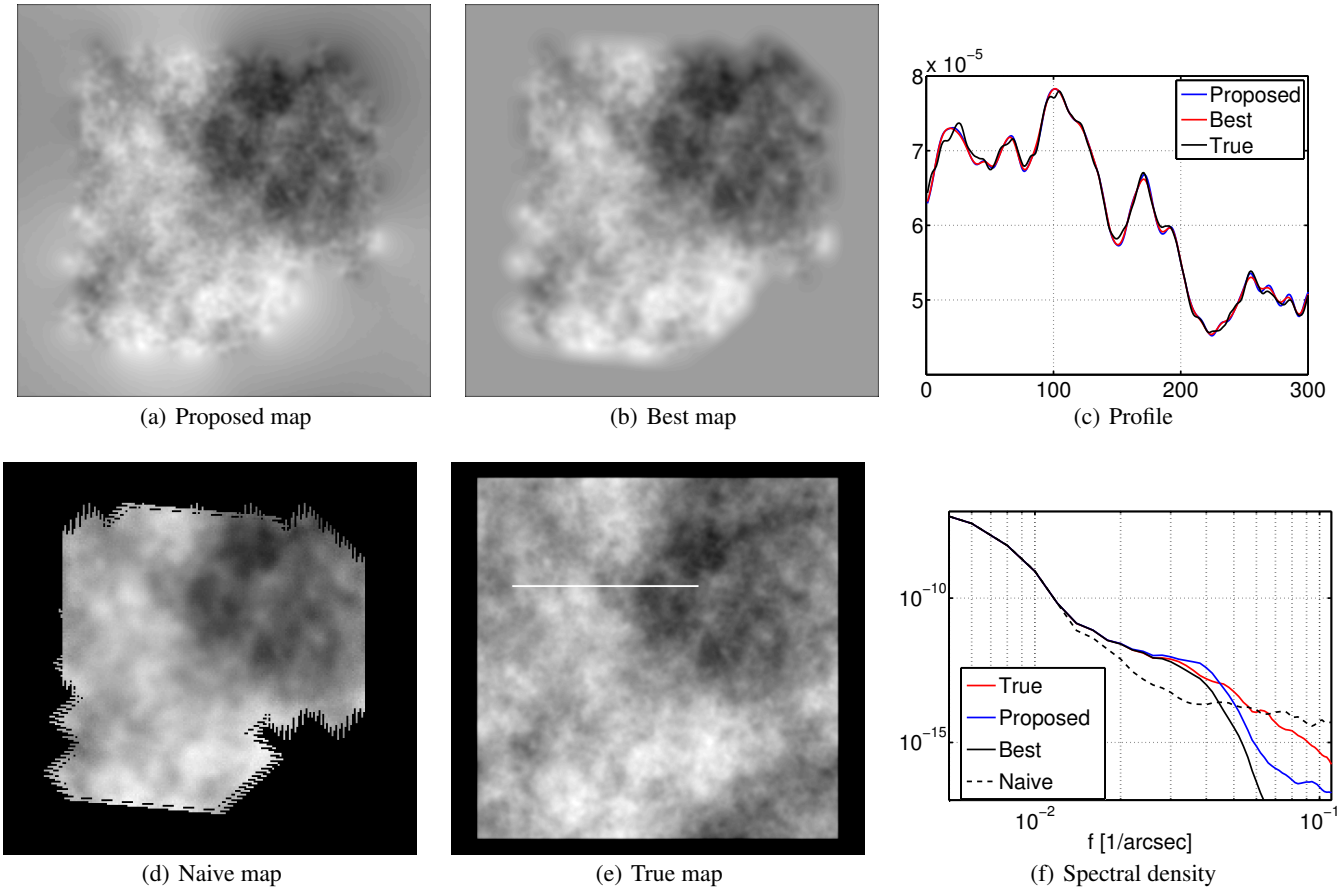
Figures 6–8 present some typical elements of the algorithm operation: visualization of progression, convergence phase (burn-in period), stable phase, etc. The evolution of the chain is shown for hyperparameters (Figs. 6 and 7) and for instrument parameter (Fig. 8). It is thus possible to grasp how the parameter space is explored.

5.3. Unsupervised approach

5.3.1. Assessment of map estimation

The qualitative and quantitative assessment of the reconstructed maps is presented here for the Galactic Cirrus; the first results are shown in Fig. 3.

– The unsupervised method (proposed method) is outlined in Fig. 3a. The hyperparameters are automatically set (without knowing of the sky truth).



(a) Proposed map (b) Best map (c) Profile

(d) Naive map (e) True map (f) Spectral density

Fig. 3. Comparison of reconstructed map for the Galactic Cirrus: proposed map (Fig. 3a), best map (Fig. 3b), naive map (Fig. 3d), and true map (Fig. 3e). Figure 3c shows a profile (marked by the white line in Fig. 3e) and Fig. 3f the spectrum (circular means of power spectra). Uncertainties are given in Fig. 4 and quantitative results are given in Table 3. Comments are given in Sect. 5.3.1.

- The best-supervised method is outlined in Fig. 3b. The hyperparameters are set by hand to minimize the error $\mathcal{E}$ (knowing the sky truth). It is referred to as the best map and was previously presented in our paper (Orieux et al. 2012b).
- The naive map (coaddition) and the true map are shown in Figs. 3d and e.
- In addition, Fig. 3c gives spatial profiles (vertical in the middle of the map) and Fig. 3f gives the spectral[4] profiles.

As expected and shown in (Orieux et al. 2012b), the inversion based on an accurate instrument model considerably improves the quality of the map: see the proposed map and the best map compared to the naive map and to the true map. The proposed map is visually very similar to the true map. In particular, our method restores details of small spatial scales (with spectral extension from null to high frequency) that are invisible on the naive map but are present on the true map (see Fig. 3c). In addition, our method correctly restores the structures of large spatial scales (low frequencies) and also the mean level of the map (null frequency), i.e. the photometry.

To assess the pertinence of estimating γ_n and γ_x in terms of map quality, we compared the proposed map with the best map. They are visually very similar (see Figs. 3a and b). In the quantitative terms given by Table 3, the best map produces an error $\mathcal{E}$ of 0.0129% and the proposed map produces an error $\mathcal{E}$ of 0.016%, which is only slightly higher. In other words, the proposed unsupervised method automatically (without knowing the sky truth) determines hyperparameters that produce a map almost as good as the best map (which requires knowing the sky truth).

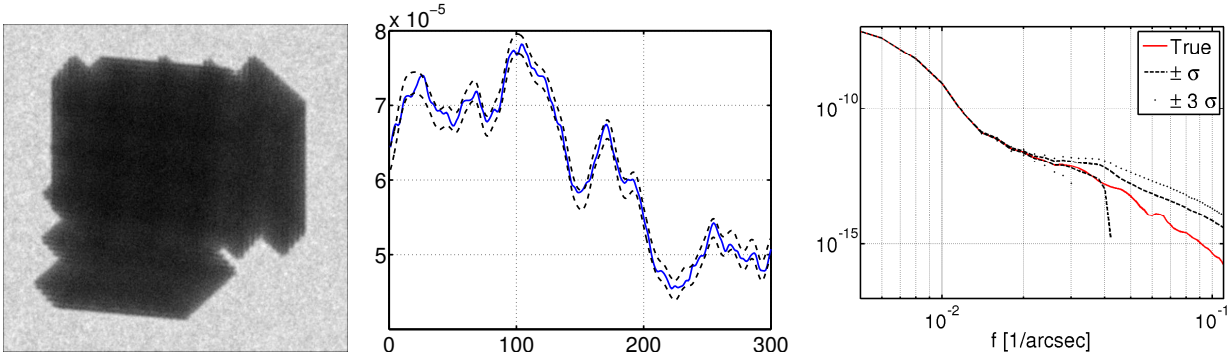
However, the proposed map shows a fine grainy texture that is visible neither in the true nor in the best map. This feature is also visible on the residual map of the coefficiens (Fig. 5). This is also observable in Fig. 3f: the spectrum of the proposed map passes above the spectrum of the true map in the spectral band 0.025–0.035 arsec^{-1}. This defect is related to a slight overevaluation of the observation contribution with respect to the prior contribution. It is referred to as under-regularization and yields an overamplification of the observation in this spectral

Table 3. Comparison of reconstruction error $\mathcal{E}$ (see Eq. (24)) for the Galactic Cirrus (the error $\mathcal{E}$ only accounts for the observed area of the map).

	Reconstruction error $\mathcal{E}$
Unsupervised ($\widehat{\gamma_{\boldsymbol{x}}}$)	0.016%
Best supervised ($\gamma_{\boldsymbol{x}}^{\text{best}}$)	0.0129%
Naive map	0.0435%

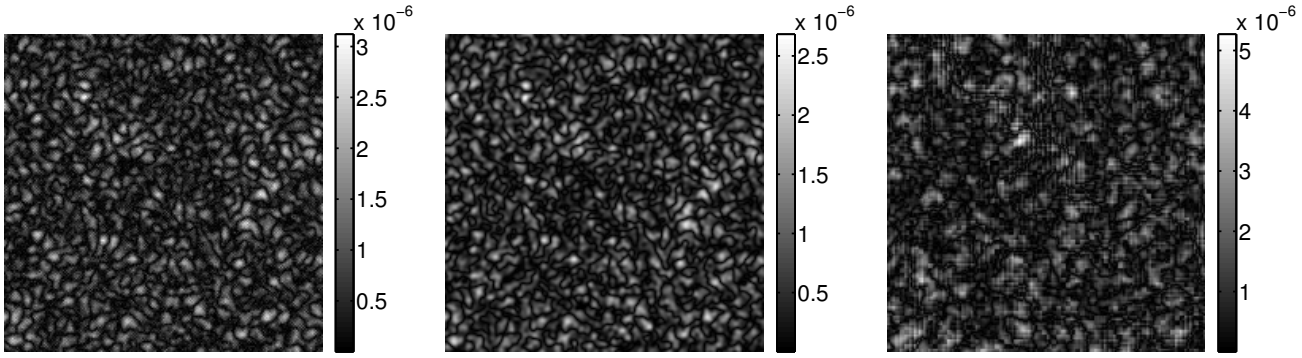
Notes. The proposed approach (which does not require knowing the true map) produces an error only very slightly higher than the best map (which does require knowing true map).

[4] This spectrum is computed from the FFT-2D of the map by averaging the coefficients in regularly spaced concentric rings. This gives a 1D spectrum containing the isotropic approximation of the spectral map properties.



(a) Map of PSD (b) True map and PM ± PSD (c) True spectrum and PM ± PSD

Fig. 4. PSD and quantification of uncertainties. Fig. 4a shows the map of the PSD and Figs. 4b and c show the interval around the estimated map ±PSD and the true map. In the spatial domain, Fig. 4b is a profile (marked by the white line on 3e) and in the spectral domain Fig. 4c is the circular means of the power spectra.



(a) Proposed map residuals (b) Best map residuals (c) Naive map residuals

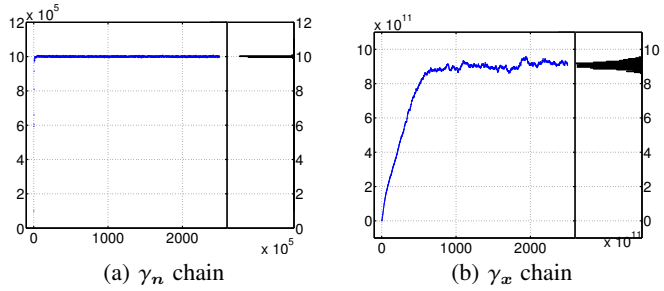
Fig. 5. Residuals of maps in the central part of measured coefficiens. This illustrates the grainy structure of the proposed map wrt. best map. The naive map residuals suffer twice more errors and a squared feature caused by the pixel model.

band. This confirms the behaviour previously observed in deconvolution (Orieux et al. 2010b; Giovannelli 2008) or noted for the maximum likelihood (Fortier et al. 1993). Nevertheless, it is remarkable that this defect is correctly notified by the PSD, as explained in the next paragraph.

Indeed, the approach naturally provides a measure of reliability through the PSD shown in Fig. 4. Two zones can be seen in Fig. 4a, in accordance with Fig. 2: the central zone (where observations are available) and the peripheral zone (extrapolated from observations of the central zone and based on the prior regularity). The boundary between the two zones also exhibits the variation of the observation hit and scanning strategy notably well. In addition, the posterior standard deviation also illustrates the difference between the zones observed with our without cross scan. From a spatial standpoint, Fig. 4b shows an interval around the estimated map with plus/minus PSD. The main result is that the true map is clearly within the interval. In a similar way, from a spectral standpoint the results are given by Fig. 4c (in relation with Fig. 3f): the true spectrum is also within the interval. More specifically, incorrectly reconstructed in the spectral band (above 0.025 arcsec^{-1}), the stronger PSD clearly shows that the estimated spectrum is certainly submitted to unsatisfactory errors or confidence.

5.3.2. Assessment of hyperparameter estimation

This section assesses the unsupervised capabilities through evaluating the hyperparameter estimation using the Galactic



(a) γ_n chain (b) γ_x chain

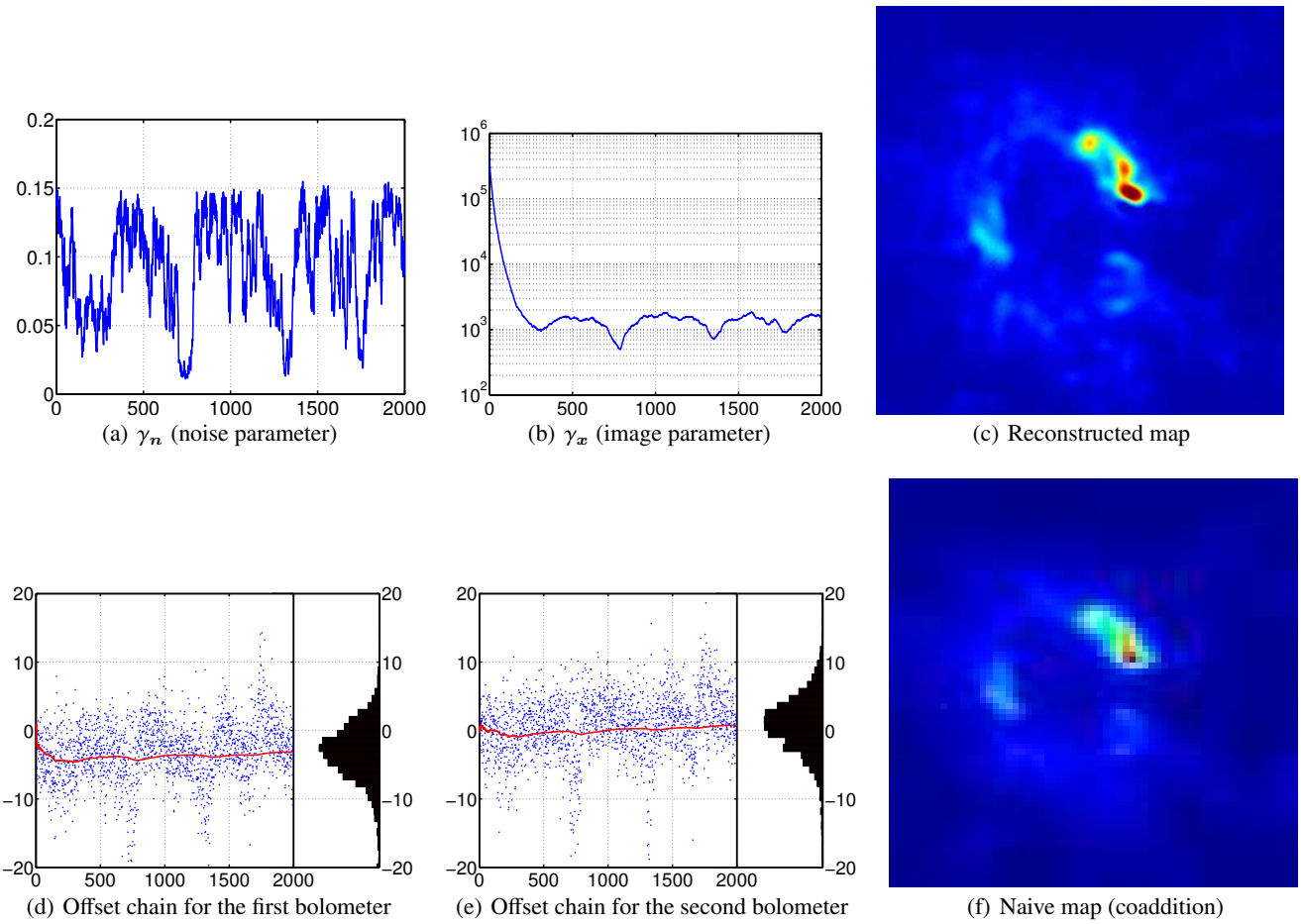
Fig. 6. Chains and histograms for γ_n, Fig. 6a, and γ_x, Fig. 6b, for the Galactic Cirrus. The chains show the burn-in period (about 1000 iterations) and the steady state. The corresponding histograms are computed on steady state only.

Cirrus and a realization of the prior for the map (which makes a true value γ_x^* available). Fig. 6 shows the chains and the histograms that approximate the marginal posterior densities $p(\gamma_n|\boldsymbol{y})$ and $p(\gamma_x|\boldsymbol{y})$. In both cases, the histogram is relatively narrow although the prior is a wide non-informative Jeffreys' distribution (see Eq. (10)). In other words, the observations are sufficiently informative to quantify noise and regularity level and the method is able to capture this information.

From a quantitative standpoint, results are given in Table 4. For the Galactic Cirrus and prior realization, the estimated values $\widehat{\gamma_n}$ are very similar to the true value γ_n^* (error is less

Table 4. Hyperparameter estimation: true values, estimates, PSD and best values.

	γ_n^*	$\widehat{\gamma_n}$	PSD	γ_x^*	$\widehat{\gamma_x}$	PSD	γ_x^{best}
Cirrus	10^6	1.009×10^6	4.07×10^3	–	8.99×10^{11}	2.46×10^{10}	2.47×10^{12}
Prior	10^6	1.003×10^6	4.05×10^3	4×10^{11}	3.28×10^{11}	1.07×10^{10}	8.37×10^{11}



(a) γ_n (noise parameter) (b) γ_x (image parameter) (c) Reconstructed map

(d) Offset chain for the first bolometer (e) Offset chain for the second bolometer (f) Naive map (coaddition)

Fig. 7. Results for real observation processing (reflection nebula NGC 7023). Chains for the noise parameter (γ_n) and for the image parameter (γ_x) in Figs. 7a and b. The stationary state is attained after a burn-in time of about 500 iterations. Figure 7c shows the corresponding map. Figures 7d and e illustrate chains and marginal histogram for two bolometer offsets.

than 1%). Moreover, the PSD are very low (0.40%). In the case of prior realization, the estimated value $\widehat{\gamma_x}$ is in the correct range but the error is larger (about 17%) and the PSD is 1.7%. This difference can be naturally explained by two elements: (i) the noise is added at the system output, so it is directly observed, whereas the map is at the system input, i.e. indirectly observed; (ii) the added noise is a realization of the prior density for the noise while the Galactic Cirrus is not a realization of the prior density for the map.

5.3.3. Real observation processing

This section proposes a first assessment for a real observation. It is based on the reflection nebula NGC 7023 acquired during the science demonstration phase of *Herschel*, which as been presented in (Abergel et al. 2010) and was processed in our previous paper (Orieux et al. 2012b). There and here, computations are made on the level-1 files processed using HIPE. In our previous paper (Orieux et al. 2012b), the offsets were removed in a pre-processing step and the regularization parameter was tuned by hand compromise between gain in resolution and overamplification of the observations. In contrast, here both are automatically tuned.

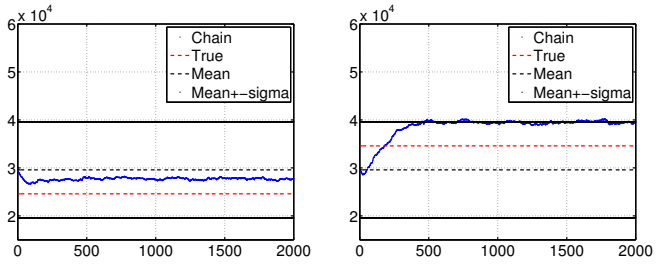
Figure 7 presents the evolution of the chains for the hyperparameters: Fig. 7a for the noise parameter γ_n and Fig. 7b for the image parameter γ_x. It is important to notice that the algorithm behaves in a very similar manner for the real observation and for the simulated observation (see Fig. 7 compared to Fig. 6). The figures also give an empirical indication of the algorithm operation: after a burn-in time (empirically less than about 500 iterations) the stationary state is attained and the chain remains in a steady state: the samples are drawn under the posterior density. Concerning the offsets, the chains begin in the steady state, thanks to a good initialization based on (Orieux et al. 2012b) results. All bolometer offsets behave in the same manner with two example illustrated Figs. 7d and e. The empircal mean of the offsets sample start to stabilize at approximately 500 samples.

Figure 7c shows the corresponding reconstructed map. Its quality is equivalent to the quality of the map restored by

Table 5. Quantitative evaluation of the estimation of the instrument parameter using the Galactic Cirrus with point sources.

Case	η^*	$\hat{\eta}$	$\hat{\eta} - \eta^*$	$(\hat{\eta} - \eta^*)/\eta^*$	$\bar{\sigma}$
1	2.46×10^4	2.29×10^4	-1.65×10^3	6.7%	2.2×10^2
2	3.46×10^4	3.27×10^4	-1.86×10^3	5.4%	2.9×10^2

Notes. Prior mean and standard deviation are $\mu = 2.96 \times 10^4$ and $\rho = 10^4$.



(a) Case 1, parameter chain (b) Case 2, parameter chain

Fig. 8. Instrument parameter chain (myopic and unsupervised approach) for the of Galactic Cirrus with point sources. *Left* (*right*) part of the figure deals with case 1, i.e. $\eta_1^* = 2.46 \times 10^4$ (case 2, i.e. $\eta_2^* = 3.46 \times 10^4$). The horizontal axis gives the iteration index and the vertical range is the prior interval in a two-standard-deviations sense. The true value is shown by the straight line.

empirically tuning the hyperparameter presented in Orieux et al. (2012b), Fig. 8. In other words, the proposed unsupervised method automatically determines hyperparameters (noise power and offsets as well as sky power) that produce a map almost as good as the map produced by a hand-made hypermarameter tuning. In addition, the map remains far better than the naive map shown in Fig. 7f.

5.4. Myopic and unsupervised approach

The myopic and unsupervised question is a threefold problem that is much more ambitious: estimate the instrument parameter, the hyperparameters, and the map itself from a unique observation. In addition, the instrument parameter intervenes in a complex way in the description of the observations, and moreover, the problem is stated in a context that is doubly delicate: ill-posedness and high-resolution.

In Orieux et al. (2012b), the equivalent PSF has a Gaussian shape whose standard deviation is proportional to the wavelength: $\sigma_o(\lambda) = \eta\lambda$. It is then integrated w.r.t. the wavelength (to include the spectral extend) and w.r.t. the time parameter (to account for the bolometer response) to form the global instrument response. To test the method, we consider the instrument parameter η to be poorly known and introduce elements of the feasibility to estimate it.

The prior is the Gaussian density given by Eq. (11), with $K = 1$. Its mean is taken from the SPIRE observer manual $\mu = 2.96 \times 10^4$ [″/m] and its standard deviation is set to $\rho = 10^4$, i.e. a relatively large uncertainty. It is about 33% of the mean and an equivalent prior interval is $[0.96 \times 10^4, 4.96 \times 10^4]$ in a two-standard-deviations sense. Two cases are investigated for the true value (used to simulated observations): $\eta_1^* = 2.46 \times 10^4$ and $\eta_2^* = 3.46 \times 10^4$. The conditional posterior for η (Sect. 4.3) does not have a standard form and its sampling (step (3) of Table 1) relies on a Metropolis-Hastings sampler, itself based on a random-walk with a Gaussian excursion. The size of the excursion was chosen so that the acceptation rate is around 50%.

Two maps are used for the observation: the Galactic Cirrus and the Galactic Cirrus with point sources. In each case, the algorithm was run several times from identical and different initializations, and shows similar qualitative and quantitative behaviours as those in Fig. 8.

Nevertheless, as expected, the spectral content of the Galactic Cirrus is not sufficiently extended towards high frequencies to provide an excitation that is adequate for instrument identification. In contrast, the Galactic Cirrus with point sources is more extended and estimations are more accurate. Table 5 presents quantitative assessments. The main result is that the estimation error is about 6%. It is a remarkable result given the difficulty of the problem (triple problem, complex relations, ill-possedness, and high resolution) and given that the prior uncertainty is about 33%. In other words, the method is able to capture information about instrument parameter, jointly with noise level, regularity level, and map from a unique observation. However, the parameter η seems to be slightly underestimated which, we explain as follows. The input map (with point sources) presents a broad spectral extent whereas the prior favours spatially extended maps (dominated by relatively low frequencies), so the posterior advocates a narrower PSF to compensate for this spectral discrepancy.

Figures 9a and b show the related maps. They must be compared to the map restored with the true instrument parameter and the best hyperparameter presented in Fig. 7b of Orieux et al. (2012b) and in Fig. 9d here. They must also be compared to the true map and the naive map also given in Orieux et al. (2012b) and in Figs. 9c–e here. As previously, the proposed maps show a fine grainy texture but despite this defect, they remain similar to the true map. The quality of the proposed maps is similar to the quality of the map restored with the true instrument parameter and the best hyperparameter. In addition, several point sources of the true map are visible on the proposed maps but not on the naive map. In other words, the proposed method automatically determines instrument parameter and hyperparameters that produce a map almost as good as the best one and better than the naive map.

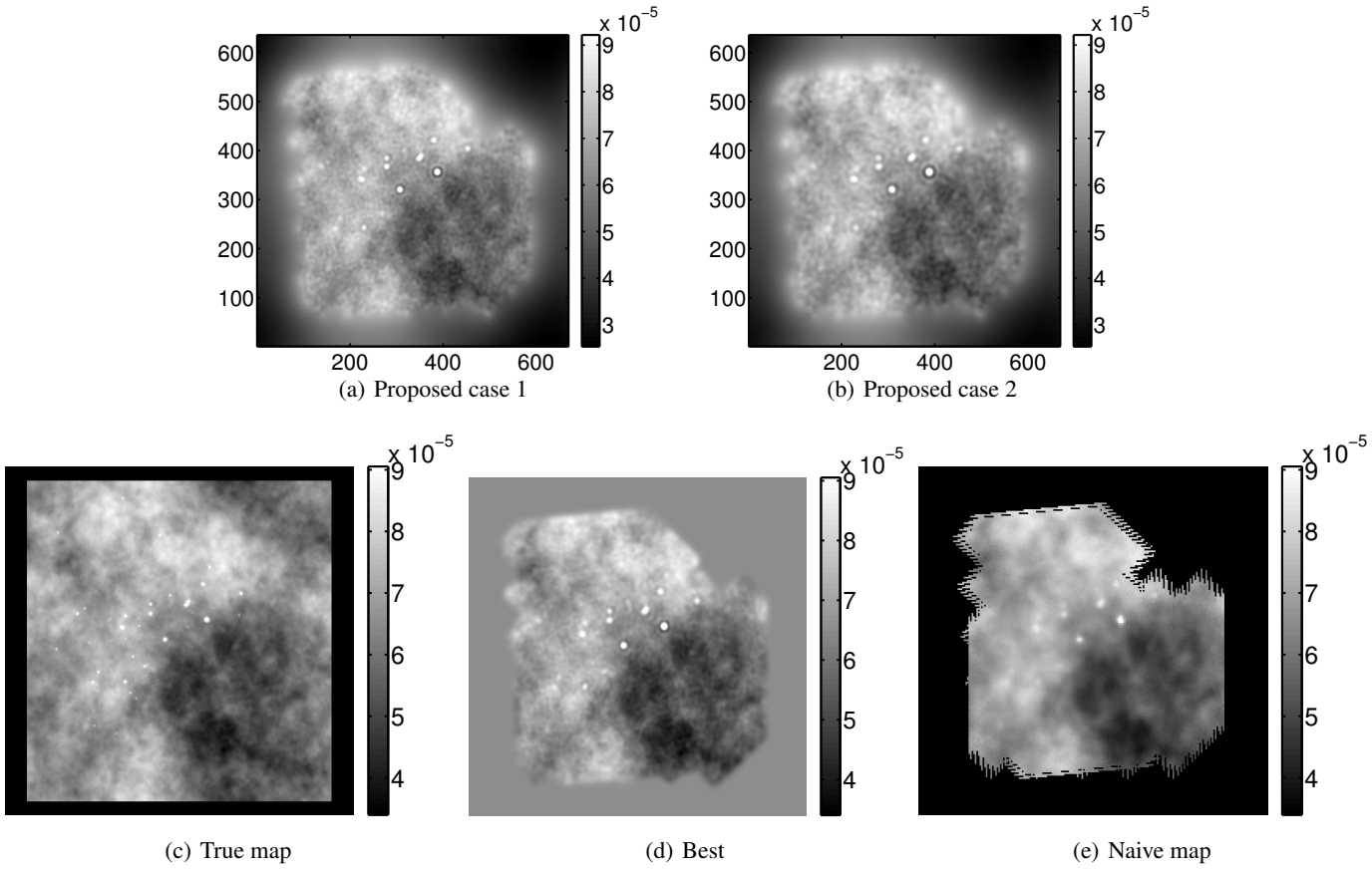
6. Conclusion

We described regularized methods for image reconstruction and focused on parameter estimation:

- hyperparameters, which guide the trade-off between prior-based and observation-based information;
- instrument parameter, which tunes the physical characteristics of the model of the acquisition system.

They were jointly estimated with the map of interest. We were therefore dealing with an unsupervised and myopic inverse problem.

The most delicate point is jointly handling the different types of variables and their interactions in direct terms but, above all, in inverse terms. From a methodology point of view, we worked in the framework of hierarchical full Bayes strategies that model



(a) Proposed case 1 (b) Proposed case 2

(c) True map (d) Best (e) Naive map

Fig. 9. Restoration of cirrus superimposed on point sources. The proposed maps must be compared to the maps restored with the true instrument parameter and the best hyperparameter and with the naive map.

the available information for each set of variables (map, hyperparameters, instrument parameter, and observations) under a probability density. We defined the posterior density, which gathers the information on the map of interest and the parameters, given the observations. We then defined the posterior mean as an estimate of the map and the posterior standard deviation as a measure of uncertainty, which gives an uncertainty map. This approach makes it possible to work in a global and consistent framework to solve the problem as a whole. It draws its inspiration from our earlier works on deconvolution (Orieux et al. 2010b) and adapts them to the case at hand.

The posterior density was explored by stochastic sampling using a Gibbs algorithm. The sampling of the map was difficult: we are dealing with a large-sized multivariate normal density for which classical techniques do not apply. We overcame this difficulty by constructing a sample as the minimizer of a well-chosen perturbated criterion (Orieux et al. 2012a). Another problematic point is the instrument parameter sampling: we are dealing with a very complex, nonstandard density. This difficulty was overcome by means of a Metropolis-Hastings step. The estimate of the map as well as the parameters (posterior mean) and the uncertainties (posterior standard deviation) were calculated numerically as empirical averages based on the simulated samples.

We presented a first application of the developments (Bayesian estimation method and stochastic sampling algorithm) in a real context: the SPIRE instrument of the *Herschel* Space Observatory. The study was essentially performed on simulated observations and has also yielded some initial results on real observations. We concluded that the approach is applicable and enables joint estimation of the map, the hyperparameters, and the instrument parameter from a unique observation. We showed, among other results, that the quality of the proposed map is similar to that obtained when the instrument parameter is known and the hyperparameters are fixed by hand in a supervised way (using the sky truth). The method shows remarkable results given the difficulty of the problem. It seems to us that these initial results are particularly promising and worth developing. They may open up many new perspectives for imaging in astrophysics in a myopic and unsupervised framework.

Appendix A: Gamma probability density

The gamma pdf for $\gamma > 0$, with given parameters $a > 0$ and $b > 0$, is written

$$\mathcal{G}(\gamma|a,b) = \frac{1}{b^a\Gamma(a)}\gamma^{a-1}\exp\left(-\gamma/b\right). \tag{A.1}$$

The following properties hold: mean is $\mathbb{E}_{\mathcal{G}}[\gamma] = ab$, variance is $\mathbb{V}_{\mathcal{G}}[\gamma] = ab^2$ and maximizer is $b(a-1)$ if and only if $a > 1$.

References

Abergel, A., Arab, H., Compiègne, M., et al. 2010, A&A, 518, L96
Babacan, S., Molina, R., & Katsaggelos, A. 2010, IEEE Trans. Image Process., 19, 53
Bishop, T., Molina, R., & Hopgood, J. 2008, in Proc. IEEE ICIP
Blanc, A., Mugnier, L., & Idier, J. 2003, J. Opt. Soc. Amer. (A), 20, 1035

Chantas, G. K., Galatsanos, N. P., & Woods, N. A. 2007, IEEE Trans. Image Process., 16, 1821
Chellappa, R., & Chatterjee, S. 1985, IEEE Trans. Acoust. Speech Signal Process., ASSP-33, 959
Chellappa, R., & Jain, A. 1992, Markov Random Fields: Theory and Application (Academic Press Inc)
Conan, J.-M., Mugnier, L., Fusco, T., Michau, V. & Rousset, G. 1998, App. Opt., 37, 4614
de Figueiredo, M. T., & Leitao, J. M. N. 1997, IEEE Trans. Image Process., 6, 1089
Descombes, X., Morris, R., Zerubia, J., & Berthod, M. 1999, IEEE Trans. Image Process., 8, 954
Dobigeon, N., Hero, A., & Tourneret, J.-Y. 2009, IEEE Trans. Image Process., 18
Féron, O. 2006, Ph.D. Thesis, Université de Paris-Sud, Orsay, France
Fortier, N., Demoment, G., & Goussard, Y. 1993, J. Visual Comm. Image Repres., 4, 157
Fusco, T., Véran, J.-P., Conan, J.-M., & Mugnier, L. M. 1999, Astron. Astrophys. Suppl. Ser., 134, 193
Geman, D., & Yang, C. 1995, IEEE Trans. Image Process., 4, 932
Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. 1996, Markov Chain Monte Carlo in practice (Boca Raton: Chapman & Hall/CRC)
Giovannelli, J.-F. 2008, IEEE Trans. Image Process., 17, 16
Golub, G. H., Heath, M., & Wahba, G. 1979, Technometrics, 21, 215
Griffin, M. J., Abergel, A., Abreu, A., et al. 2010, A&A, 518, L3
Hansen, P. 1992, SIAM Rev., 34, 561
Idier, J., ed. 2008, Bayesian Approach to Inverse Problems (London: ISTE Ltd and John Wiley & Sons Inc.)
Jalobeanu, A., Blanc-Feraud, L., & Zerubia, J. 2002, in Proc. IEEE ICASSP, 4, 3580
Kass, R. E., & Wasserman, L. 1996, J. Amer. Statist. Assoc., 91, 1343
Lalanne, P., Prévost, D., & Chavel, P. 2001, Appl. Opt., 40
Lam, E. Y., & Goodman, J. W. 2000, J. Opt. Soc. Am. A, 17, 1177
Lanterman, A. D., Grenander, U., & Miller, M. I. 2000, IEEE Trans. Pattern Anal. Mach. Intell., 22, 337
Li, S. Z. 2001, Markov Random Field Modeling in Image Analysis (Tokyo: Springer-Verlag)
Likas, A. C., & Galatsanos, N. P. 2004, IEEE Trans. Image Process., 52, 2222
Mallat, S. 2008, A Wavelet Tour of Signal Processing: The Sparse Way (Academic Press Inc.)
Molina, R., Katsaggelos, A. K., & Mateos, J. 1999, IEEE Trans. Image Process., 8, 231
Molina, R., Mateos, J., & Katsaggelos, A. K. 2006, IEEE Trans. Image Process., 15, 3715
Mugnier, L., Fusco, T., & Conan, J.-M. 2004, J. Opt. Soc. Amer., 21, 1841
Ocvirk, P., Pichon, C., Lançon, A., & Thiébaut, E. 2006, MNRAS, 365, 46
Orieux, F., Rodet, T., & Giovannelli, J.-F. 2009, in Proc. IEEE ICIP, Le Caire, Egypt
Orieux, F., Giovannelli, J.-F., & Rodet, T. 2010a, J. Opt. Soc. Amer., 27, 1593
Orieux, F., Giovannelli, J.-F., & Rodet, T. 2010b, in Proc. IEEE ICASSP, Dallas
Orieux, F., Féron, O., & Giovannelli, J.-F. 2012a, IEEE Signal Process. Lett.
Orieux, F., Giovannelli, J.-F., Rodet, T., et al. 2012b, A&A, 539, A38
Pankajakshani, P., Zhang, B., Blanc-Féraud, L., et al. 2009, Applied Optics, 48, 4437
Pascazio, V., & Ferraiuolo, G. 2003, IEEE Trans. Image Process., 12, 572
Pilbratt, G. L., Riedinger, J. R., Passvogel, T., et al. 2010, A&A, 518, L1
Robert, C. P. 2005, The Bayesian Choice, Statistiques et probabilités appliquées (Paris, France: Springer)
Robert, C. P., & Casella, G. 2004, Monte-Carlo Statistical Methods, Springer Texts in Statistics (New York: Springer)
Rodet, T., Orieux, F., Giovannelli, J.-F., & Abergel, A. 2008, IEEE J. of Selec. Topics in Signal Proc., 2, 802
Rue, H. 2001, J. R. Statist. Soc. B, 63
Rue, H., & Held, L. 2005, Monographs on Statistics and Applied Probability, 104, Gaussian Markov Random Fields: Theory and Applications (Chapman & Hall)
Saquib, S. S., Bouman, C. A., & Sauer, K. D. 1998, IEEE Trans. Image Process., 7, 1029
Tan, X., Li, J., & Stoica, P. 2010, in Proc. IEEE ICASSP, 3634
Thiébaut, E. 2008, in Proc. SPIE: Astronomical Telescopes and Instrumentation, 7013, 70131
Thiébaut, E., & Conan, J.-M. 1995, J. Opt. Soc. Amer. (A), 12, 485
Tikhonov, A., & Arsenin, V. 1977, Solutions of Ill-Posed Problems (Washington DC: Winston)
Vacar, C., Giovannelli, J.-F., & Berthoumieu, Y. 2011, in Proc. IEEE ICASSP, Prague, Czech Republic
Wiegelmann, T., & Inhester, B. 2003, Sol. Phys., 214, 287
Winkler, G. 2003, Image Analysis, Random Fields and Markov Chain Monte Carlo Methods (Springer Verlag, Berlin Germany)
Xu, Z., & Lam, E. Y. 2009, Opt. Lett., 34, 1453
Zhang, B., Zerubia, J., & Olivo-Marin, J.-C. 2007, App. Opt., 46, 1819
Zhou, Z., Leahy, R. M., & Qi, J. 1997, IEEE Trans. Image Process., 6, 844

Sampling high-dimensional Gaussian distributions for general linear inverse problems

F. Orieux*, O. Féron and J.-F. Giovannelli

***Abstract*—This paper is devoted to the problem of sampling Gaussian distributions in high dimension. Solutions exist for two specific structures of inverse covariance: sparse and circulant. The proposed algorithm is valid in a more general case especially as it emerges in linear inverse problems as well as in some hierarchical or latent Gaussian models. It relies on a perturbation-optimization principle: adequate stochastic perturbation of a criterion and optimization of the perturbed criterion. It is proved that the criterion optimizer is a sample of the target distribution. The main motivation is in inverse problems related to general (non-convolutive) linear observation models and their solution in a Bayesian framework implemented through sampling algorithms when existing samplers are infeasible. It finds a direct application in myopic / unsupervised inversion methods as well as in some non-Gaussian inversion methods. An illustration focused on hyperparameter estimation for super-resolution method shows the interest and the feasibility of the proposed algorithm.**

***Index Terms*—Stochastic sampling, high-dimensional sampling, inverse problem, Bayesian strategy, unsupervised, myopic**

I. Introduction

This work deals with simulation of high-dimensional Gaussian and conditional Gaussian distributions. The difficulty of the problem is directly related to handling high-dimensional covariances $\boldsymbol{R}$ and precision matrices $\boldsymbol{Q} = \boldsymbol{R}^{-1}$. The problem has already been investigated and solutions exist in two cases.

- When $\boldsymbol{Q}$ is sparse, two strategies are available. The first one [1, chap. 8], relies on a parallel Gibbs sampler based on a chessboard-like decomposition. It takes advantage of the sparsity of $\boldsymbol{Q}$ to update simultaneously large blocks of variables. The second strategy [2, 3] relies on a Cholesky decomposition $\boldsymbol{Q} = \boldsymbol{L}^{\mathrm{t}}\boldsymbol{L}$: a sample $\boldsymbol{x}$ is obtained by solving the linear system $\boldsymbol{L}\boldsymbol{x} = \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a zero-mean white Gaussian vector. The sparsity of $\boldsymbol{Q}$ ensures feasible numerical factorization and the sparsity of $\boldsymbol{L}$ ensures feasible numerical resolution of the linear system.
- [4, 5] propose a solution for circulant matrix $\boldsymbol{Q}$, even non-sparse. In this case, the covariance is diagonal in the Fourier domain: the sampling is based on independent sampling of the Fourier coefficients. Finally, the sample is computed by FFT and it has been used in [6–10].

To our knowledge there is no solution for more general structure in high dimension because factorization (Cholesky, QR, square root,...), diagonalization and inversion of $\boldsymbol{Q}$ and $\boldsymbol{R}$ are numerically infeasible. The obstacle is due to both computational cost and memory footprint. The proposed algorithm overcomes this obstacle when $\boldsymbol{Q}$ is of the form

$$\boldsymbol{Q} = \sum_{k=1}^{K} \boldsymbol{M}_k^{\mathrm{t}} \boldsymbol{R}_k^{-1} \boldsymbol{M}_k \tag{1}$$

as it appears in inverse problems [11]. Indeed, let us consider the general linear forward model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$, where $\boldsymbol{y}$, $\boldsymbol{n}$ and $\boldsymbol{x}$ are the observation, the noise and the unknown image and $\boldsymbol{A}$ is a linear operator. Consider, again, two prior distributions for $\boldsymbol{n}$ and $\boldsymbol{x}$ that are Gaussian conditionally on a parameter $\boldsymbol{\theta}$. This framework is very general: it includes linear inverse problems [11] as well as some hierarchical or latent Gaussian models [12] and it can be used in many applications. In image reconstruction, it covers a majority of current problems, e.g. unsupervised [8] or myopic (semi-blind) [9] inverse problems, by including acquisition parameters and hyperparameters in $\boldsymbol{\theta}$. Moreover, the framework also includes non-linear models, based on conditional linear models such as bilinear or multilinear ones (see Section III-B). The framework also covers some non-stationary or inhomogeneous Gaussian priors and non-Gaussian priors involving auxiliary / latent variables [6, 8, 13–15] (e.g., location or scale mixtures of Gaussian), by including these variables in $\boldsymbol{\theta}$.

Let us focus on the joint estimation of $\boldsymbol{x}$ and $\boldsymbol{\theta}$ from the posterior $p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. It commonly requires the handling of the conditional posterior $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ that is Gaussian with precision matrix $\boldsymbol{Q}$ of the form (1), as will be shown in section II-B. In the general case, $\boldsymbol{Q}$ is neither sparse nor circulant so existing sampling algorithms fail when the dimension of $\boldsymbol{x}$ is very large while the proposed one handles this case. It relies on a perturbation-optimization principle: adequate stochastic perturbation of a quadratic criterion and optimization of the perturbed criterion. A recent paper [16] briefly describes a similar algorithm for compressed sensing in signal processing. Our paper deepens and generalizes this contribution.

Subsequently, Section II presents the proposed algorithm and its direct application to linear inverse problems. Section III gives an illustration through an academic problem in super-resolution. Section IV presents conclusions and perspectives.

II. Perturbation-optimization algorithm

A. Description

We focus on the problem of sampling from a target Gaussian distribution whose precision matrix $\boldsymbol{Q}$ is in the form (1). When $\boldsymbol{Q}$ is neither sparse nor circulant, existing algorithms fail in high dimension because of an excessive memory footprint as illustrated in section III. We propose a solution based on the Perturbation-Optimization (PO) algorithm described hereafter, whose memory footprint is far smaller.

F. Orieux is with Pasteur Institute, 75015 Paris, France, orieux@pasteur.fr. O. Féron is with EDF Research & Developments, OSIRIS, 92140 Clamart, France, olivier-2.feron@edf.fr. J.-F. Giovannelli is with Univ. Bordeaux, IMS, 33400 Talence, France, Giova@IMS-Bordeaux.fr.

Proposition 1: The optimizer $\widehat{\boldsymbol{x}}$ of criterion (5) resulting from Algorithm 1 is Gaussian

$$\widehat{\boldsymbol{x}} \sim \mathcal{N}\left(\boldsymbol{Q}^{-1}\left(\sum_{k=1}^{K}\boldsymbol{M}_k^{\mathrm{t}}\boldsymbol{R}_k^{-1}\boldsymbol{m}_k\right), \boldsymbol{Q}^{-1}\right) . \tag{2}$$

Proof: The optimizer $\widehat{\boldsymbol{x}}$ of criterion (5) is explicit:

$$\begin{aligned}\widehat{\boldsymbol{x}} &= \left[\sum_{k=1}^{K}\boldsymbol{M}_k^{\mathrm{t}}\boldsymbol{R}_k^{-1}\boldsymbol{M}_k\right]^{-1}\left(\sum_{k=1}^{K}\boldsymbol{M}_k^{\mathrm{t}}\boldsymbol{R}_k^{-1}\boldsymbol{\eta}_k\right)\\ &= \boldsymbol{Q}^{-1}\left(\sum_{k=1}^{K}\boldsymbol{M}_k^{\mathrm{t}}\boldsymbol{R}_k^{-1}\boldsymbol{\eta}_k\right) .\end{aligned} \tag{3}$$

It is clearly a Gaussian vector as a linear combination of K Gaussian vectors. Its expectation and covariance are calculated below using elementary algebra: from (4) and (3), we have

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{x}}\right] &= \boldsymbol{Q}^{-1}\left(\sum_{k}\boldsymbol{M}_k^{\mathrm{t}}\boldsymbol{R}_k^{-1}\mathbb{E}\left[\boldsymbol{\eta}_k\right]\right)\\ &= \boldsymbol{Q}^{-1}\left(\sum_{k}\boldsymbol{M}_k^{\mathrm{t}}\boldsymbol{R}_k^{-1}\boldsymbol{m}_k\right)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}^{\mathrm{t}}\right] &= \boldsymbol{Q}^{-1}\left(\sum_{k,k'}\boldsymbol{M}_k^{\mathrm{t}}\boldsymbol{R}_k^{-1}\mathbb{E}\left[\boldsymbol{\eta}_k\boldsymbol{\eta}_{k'}^{\mathrm{t}}\right]\boldsymbol{R}_{k'}^{-1}\boldsymbol{M}_{k'}\right)\boldsymbol{Q}^{-1}\\ &= \boldsymbol{Q}^{-1}\left(\sum_{k}\boldsymbol{M}_k^{\mathrm{t}}\boldsymbol{R}_k^{-1}\left(\boldsymbol{R}_k+\boldsymbol{m}_k\boldsymbol{m}_k^{\mathrm{t}}\right)\boldsymbol{R}_k^{-1}\boldsymbol{M}_k\right)\boldsymbol{Q}^{-1}\\ &= \boldsymbol{Q}^{-1}+\mathbb{E}\left[\widehat{\boldsymbol{x}}\right]\mathbb{E}\left[\widehat{\boldsymbol{x}}\right]^{\mathrm{t}}\end{aligned}$$

that completes the proof. ∎

The feasibility of Step P clearly depends on the capability to sample from Gaussian distributions $\mathcal{N}(\boldsymbol{m}_k, \boldsymbol{R}_k)$. It is usually the case in inverse problems and it will be actually the case in super-resolution applications shown in section III-A and in other contributions shortly described in section III-B.

Regarding Step O, J being quadratic, a large literature [17] is available about its numerical optimization, e.g. gradient procedure (standard, corrected, conjugate, optimal step size…). Such algorithms require the computation of criterion (5) and its gradient. The feasibility of Step O clearly depends on the capability to compute that without the storage of large matrices. It is usually the case in inverse problems and it will be actually the case in applications shown in section III-A and described in section III-B.

However, the desired sample is the exact optimizer, so, Step O could require N iterations of a conjugate gradient algorithm for a problem of dimension N. Therefore the complexity could be $O(N^3)$ that is equivalent to the one of a Cholesky decomposition. However, the optimization procedure can be stopped earlier without practical loss of precision and the complexity falls down to $O(PN^2)$ for P iterations. In addition, for a band matrix, the complexity of the proposed algorithm becomes $O(MPN)$ and the one of the Cholesky decomposition becomes $O(MN^2)$. Anyway, the main advantage of the proposed algorithm is its reduced memory footprint: it avoids the storage of neither $\boldsymbol{Q}$ nor its (Cholesky, QR, square root,…) factors.

Algorithm 1 : Perturbation-Optimization algorithm.

1: **Step P (Perturbation)**: Generate independent vectors

$$\boldsymbol{\eta}_k \sim \mathcal{N}(\boldsymbol{m}_k, \boldsymbol{R}_k), \quad \text{for } k = 1, \dots K \tag{4}$$

2: **Step O (Optimization)**: Compute $\widehat{\boldsymbol{x}}$ as the minimizer of

$$J(\boldsymbol{x}) = \sum_{k=1}^{K}\left(\boldsymbol{\eta}_k - \boldsymbol{M}_k\boldsymbol{x}\right)^{\mathrm{t}}\boldsymbol{R}_k^{-1}\left(\boldsymbol{\eta}_k - \boldsymbol{M}_k\boldsymbol{x}\right) \tag{5}$$

Remark 1: Still regarding Step O, it would be awkward if $\boldsymbol{Q}$ was badly scaled, but it is not the case here for the following reason. In usual ill-conditioned inverse problems, $\boldsymbol{A}$ is badly-scaled but the aim of regularization is precisely to overcome this difficulty and to produce a well-scaled matrix $\boldsymbol{Q}$.

B. Application to inverse problems

The purpose is to solve an inverse problem, stated by the forward model $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n}$, in a Bayesian framework based on the following models:

- $\boldsymbol{A}$ describes any observation system that can depend on unknown acquisition parameters,
- priors for the noise $\boldsymbol{n}$ and the object $\boldsymbol{x}$ are Gaussian $\mathcal{N}(\boldsymbol{m}_n, \boldsymbol{R}_n)$ and $\mathcal{N}(\boldsymbol{m}_x, \boldsymbol{R}_x)$, conditionally on a set of hyperparameters and auxiliary variables.

In a general statement, acquisition parameters, hyperparameters and auxiliary variables are collected in $\boldsymbol{\theta}$. The general inverse problem then consists in estimating $\boldsymbol{x}$ and $\boldsymbol{\theta}$ through the posterior $p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})$. Its exploration can be achieved by means of a Gibbs sampler which iteratively samples from $p(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})$ and $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. The conditional posterior $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ is a correlated Gaussian distribution: $\mathcal{N}(\boldsymbol{m}_x^{\mathrm{post}}, \boldsymbol{R}_x^{\mathrm{post}})$ with

$$\begin{aligned}\boldsymbol{R}_x^{\mathrm{post}} &= \left(\boldsymbol{A}^t\boldsymbol{R}_n^{-1}\boldsymbol{A} + \boldsymbol{R}_x^{-1}\right)^{-1}\\ \boldsymbol{m}_x^{\mathrm{post}} &= \boldsymbol{R}_x^{\mathrm{post}}\left(\boldsymbol{A}^t\boldsymbol{R}_n^{-1}\left[\boldsymbol{y} - \boldsymbol{m}_n\right] + \boldsymbol{R}_x^{-1}\boldsymbol{m}_x\right)\end{aligned}$$

where $\boldsymbol{\theta}$ is embedded in $\boldsymbol{A}, \boldsymbol{R}_n$ and $\boldsymbol{R}_x$ for simpler notations.

If $\boldsymbol{A}$ has no particular properties, $\boldsymbol{Q} = (\boldsymbol{R}_x^{\mathrm{post}})^{-1}$ is neither sparse nor circulant, and existing sampling algorithms are not applicable. The PO algorithm makes it possible to sample from $\mathcal{N}(\boldsymbol{m}_x^{\mathrm{post}}, \boldsymbol{R}_x^{\mathrm{post}})$ by applying Algorithm 1 with $K = 2$, $\boldsymbol{M}_1 = \boldsymbol{A}$, $\boldsymbol{M}_2 = \boldsymbol{I}$, $\boldsymbol{R}_1 = \boldsymbol{R}_n$, $\boldsymbol{R}_2 = \boldsymbol{R}_x$, $\boldsymbol{m}_1 = \boldsymbol{m}_n$ and $\boldsymbol{m}_2 = \boldsymbol{m}_x$. In this context, it can be said that the optimization procedure converts prior samples into a posterior one.

III. Illustration

The proposed PO algorithm makes it possible to resort to stochastic sampling algorithms in inverse problems providing two main advances:

- capability to jointly estimate extra unknowns included in $\boldsymbol{\theta}$ (acquisition parameters, hyperparameters, …),
- access to the entire unknown distribution providing uncertainties (standard deviation, credibility interval,…).

These advances are illustrated in the present section.

A. Unsupervised super-resolution

We detail an application of the proposed PO algorithm to the super-resolution (SR) academic problem: several blurred and down-sampled (low resolution) images of a scene are available in order to retrieve the original (high resolution) scene [18, 19]. It is shown that the crucial novelty, enabled by the proposed PO algorithm, is to allow the use of sampling algorithms in SR methods and thus to provide hyperparameter estimation.

We resort to the standard forward model in SR: $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{n} = \boldsymbol{P}\boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}$. In this equation, $\boldsymbol{y} \in \mathbb{R}^M$ collects the pixels of the low resolution images, here 5 images of 128×128 pixels ($M = 81920$) and $\boldsymbol{x} \in \mathbb{R}^N$ collects the pixels of the original image, here 256×256 pixels ($N = 65536$). The noise $\boldsymbol{n} \in \mathbb{R}^M$ accounts for measurement and modeling errors. $\boldsymbol{H}$ is a $N \times N$ circulant convolution matrix that accounts for the convolution part of the observation system. Practically, the impulse response is a Laplace shape with FWHM of 4 pixels. $\boldsymbol{P}$ is a $M \times N$ decimation matrix: it is a binary matrix indicating which pixel is observed. Finally, $\boldsymbol{A}$ is a $M \times N$ matrix (that is to say 81920×65536). The prior distribution for $\boldsymbol{n}$ is $\mathcal{N}(\boldsymbol{0}, \gamma_n^{-1}\boldsymbol{I})$ and the one for $\boldsymbol{x}$ is $\mathcal{N}(\boldsymbol{0}, \gamma_x^{-1}\boldsymbol{D}^t\boldsymbol{D})$ where $\boldsymbol{D}$ is the $N \times N$ circulant convolution matrix of the Laplacian filter. The hyperparameters γ_n and γ_x are *unknown* and their prior law are Jeffreys'. The posterior [9] is

$$p(\boldsymbol{x}, \gamma_n, \gamma_x | \boldsymbol{y}) \propto \gamma_n^{M/2-1} \gamma_x^{(N-1)/2-1} \exp\left[-\gamma_n \|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{H}\boldsymbol{x}\|^2/2 - \gamma_x \|\boldsymbol{D}\boldsymbol{x}\|^2/2\right]. \quad (6)$$

It is explored by a Gibbs sampler: iteratively sampling γ_n, γ_x and $\boldsymbol{x}$ under their respective posterior conditional distribution

$$p(\gamma_n^{(k)} | \boldsymbol{x}, \gamma_x, \boldsymbol{y}) = \mathcal{G}\left(1 + M/2, 2/\left\|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{H}\boldsymbol{x}^{(k-1)}\right\|^2\right)$$

$$p(\gamma_x^{(k)} | \boldsymbol{x}, \gamma_n, \boldsymbol{y}) = \mathcal{G}\left(1 + (N-1)/2, 2/\left\|\boldsymbol{D}\boldsymbol{x}^{(k-1)}\right\|^2\right)$$

$$p(\boldsymbol{x}^{(k)} | \gamma_x, \gamma_n, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{m}_x^{\text{post}}, \boldsymbol{R}_x^{\text{post}})$$

$$\text{with} \quad \boldsymbol{R}_x^{\text{post}} = \left(\gamma_n^{(k)} \boldsymbol{H}^t \boldsymbol{P}^t \boldsymbol{P} \boldsymbol{H} + \gamma_x^{(k)} \boldsymbol{D}^t \boldsymbol{D}\right)^{-1}$$

$$\text{and} \quad \boldsymbol{m}_x^{\text{post}} = \gamma_n^{(k)} \boldsymbol{R}_x^{\text{post}} \boldsymbol{P}^t \boldsymbol{H}^t \boldsymbol{y}.$$

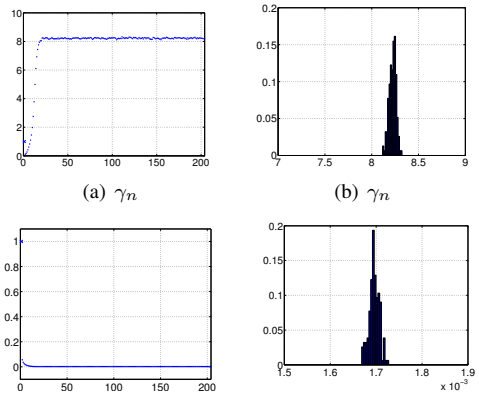
The conditional posteriors for the hyperparameters are Gamma distributions so they are easy to sample.

The conditional posterior for $\boldsymbol{x}$ is Gaussian, but the use of existing algorithms is impossible due to the structure and the size of $\boldsymbol{R}_x^{\text{post}}$. Regarding the structure, according to Section II-B, with $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{H}$: $\boldsymbol{A}$ is non-circulant due to the decimation and $\boldsymbol{A}$ is non-sparse due to large support of the impulse response. Regarding the size, $\boldsymbol{R}_x^{\text{post}}$ (and its Cholesky factor) is a huge $N \times N$ matrix, that is to say 65536×65536 and its footprint in memory would be 32 GB. As a consequence, neither the precision matrix nor its Cholesky factor can be stored on standard computers.

On the contrary, the proposed PO algorithm only requires the storage of four 256×256 matrices and its footprint in memory is only 2 MB that is easy to manage on standard computers. Regarding the computational cost:

- Step P requires a sample under each prior distribution: $\boldsymbol{x}$ is computed by FFT (see item 2 of Section I) and $\boldsymbol{n}$ is trivially computed since it is a white noise.
- Step O is achived by a conjugate gradient procedure with optimal step size. It only requires computations of convolutions (by FFT), decimation and zero-padding.

So, the proposed PO algorithm is feasible and it easily provides a desired sample. Practically, it takes[1] about one second (i.e. around $P = 50$ gradient iterations) to obtain one sample.



(a) γ_n (b) γ_n (c) γ_x (d) γ_x ($\times 10^{-3}$)

Fig. 1. Chains and histograms of hyperparameters γ_n and γ_x.

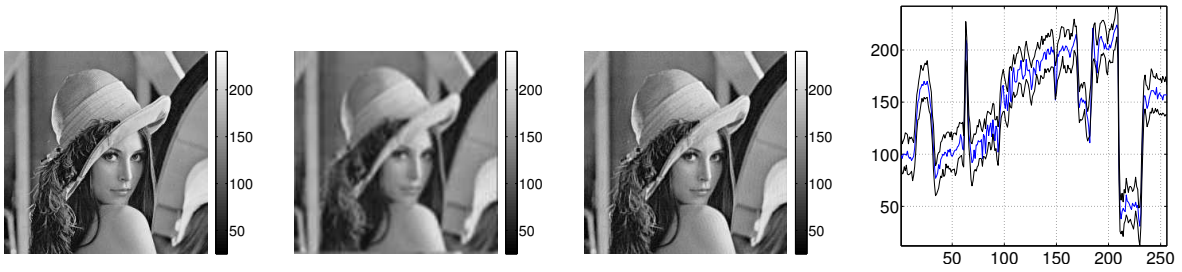
Fig. 1 shows the iterates and illustrates the operation and convergence. After a burn-in period of about 25 iterations, the algorithm is in its converged state and the total number of iterations is 59 to ensure a good exploration of the distribution. Histograms approximating marginal posteriors are also given and the posterior means are $\widehat{\gamma}_n \approx 7.7$ and $\widehat{\gamma}_x \approx 2.2 \times 10^{-3}$.

Concerning the images themselves, results are shown in Fig. 2: the estimated image in 2(c) clearly shows a better resolution than the data in 2(b) and it is visually close to the original image in 2(a). Nevertheless, it is important to keep in mind that, w.r.t. other SR methods, the proposed PO algorithm does not improve image quality itself but the crucial novelty is to allow for hyperparameter estimation. In this sense, it is clear that the approach produces correct hyperparameters i.e. correct balance between data and prior. Moreover, uncertainties are derived from the samples through the posterior standard deviation. It is illustrated in Fig. 2(d): the true image is inside the 99% credibility interval around the estimate. As a conclusion, the proposed PO algorithm makes it possible to resort to sampling algorithms in SR method whereas it was not possible before. It then enables hyperparameter estimation while other SR methods require hand-made hyperparameter tuning. In addition, it enables to compute uncertainties based on posterior standard deviation.

B. Three other examples

The PO algorithm has been used in three other contexts: electromagnetic inverse scattering [14], fluorescent microscopy trough structured illumination [20] and super-resolution from data provided the Herschel observatory in astronomy [**?**].

[1]The algorithm is implemented within the computing environment Matlab on a PC with a 3 GHz CPU and 3 GB of RAM.



(a) True (b) Data (c) Estimate (d) Uncertainty

Fig. 2. Image reconstruction: true image 2(a), one of the low resolution images 2(b) and the proposed estimate 2(c). The plot 2(d) is a true image slice inside the 99% credibility interval around the estimate.

The problems are tackled in a Bayesian framework and implemented by means of stochastic sampling. In these contexts, the distribution for the object given the other variables is Gaussian with large size precision matrix. Its structure is neither sparse nor circulant making the use of existing algorithms impossible. This is due to non-linearity and label variables in [14] and non-invariance of the observation model in [20, **?**]. Nevertheless, the precision matrix is in the form (1), so, the proposed PO is applicable.

IV. Conclusion

The paper presents an algorithm for sampling high-dimensional Gaussian distributions when existing algorithms are infeasible. It relies on a perturbation-optimization principle: adequate stochastic perturbation of a criterion and optimization of the perturbed criterion. It is shown that the criterion optimizer is a sample of the target distribution. The algorithm is applicable for a particular decomposition of the precision matrix that emerges in general linear inverse problems.

There is a wide class of applications, in particular any processing problem based on a conditional linear forward model and conditional Gaussian priors for noise and object. The interest and the feasibility of the proposed algorithm have been illustrated in [14, 20, **?**] and in this paper on a more academic super-resolution problem allowing automatic tuning of hyperparameters.

An interesting perspective deals with the case of stopped optimization procedure. It is a question under consderation to prove that, embedded in a Gibbs loop, a finite number (maybe one) of iteration of the optimization step is enough to guarantee convergence towards the target law.

V. Acknowledgment

The authors would like to thank J. Idier (IRRCyN) for inspiration of this work [21], T. Rodet and A. Djafari (L2S), for fruitful discussions, C. Vacar (IMS) and P. Szacherski (IMS and CEA-LETI) for carefully reading the paper.

References

[1] G. Winkler, *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods*. Springer Verlag, BerlinGermany, 2003.
[2] H. Rue, "Fast sampling of Gaussian Markov random fields," *J. R. Statist. Soc. B*, vol. 63, no. 2, 2001.
[3] P. Lalanne, D. Prévost, and P. Chavel, "Stochastic artificial retinas: algorithm, optoelectronic circuits, and implementation," *Applied Optics*, vol. 40, 2001.
[4] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. ASSP-33, pp. 959–963, Aug. 1985.
[5] R. Chellappa and A. Jain, *Markov Random Fields: Theory and Application*. Academic Press Inc, 1992.
[6] D. Geman and C. Yang, "Nonlinear image recovery with half-quadratic regularization," *IEEE Trans. Image Processing*, vol. 4, no. 7, pp. 932–946, July 1995.
[7] P. Charbonnier, L. Blanc-Féraud, G. Aubert, and M. Barlaud, "Deterministic edge-preserving regularization in computed imaging," *IEEE Trans. Image Processing*, vol. 6, no. 2, pp. 298–311, Feb. 1997.
[8] J.-F. Giovannelli, "Unsupervised Bayesian convex deconvolution based on a field with an explicit partition function," *IEEE Trans. Image Processing*, vol. 17, no. 1, pp. 16–26, Jan. 2008.
[9] F. Orieux, J.-F. Giovannelli, and T. Rodet, "Bayesian estimation of regularization and point spread function parameters for Wiener–Hunt deconvolution," *J. Opt. Soc. Amer.*, vol. 27, no. 7, pp. 1593–1607, 2010.
[10] H. Helgason, V. Pipiras, and P. Abry, "Fast and exact synthesis of stationary multivariate Gaussian time series using circulant embedding," *Signal Processing*, 2011.
[11] J. Idier, Ed., *Bayesian Approach to Inverse Problems*. London: ISTE Ltd and John Wiley & Sons Inc., 2008.
[12] H. Rue, S. Martino, and N. Chopin, "Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations," *J. R. Statist. Soc. B*, vol. 71, pp. 1–35, 2009.
[13] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.
[14] O. Féron, B. Duchêne, and A. Mohammad-Djafari, "Microwave imaging of piecewise constant objects in a 2D-TE configuration," *International Journal of Applied Electromagnetics and Mechanics*, vol. 26, no. 6, pp. 167–174, IOS Press 2007.
[15] H. Ayasso and A. Mohammad-Djafari, "Joint NDT image restoration and segmentation using Gauss-Markov-Potts prior models and variational Bayesian computation," *IEEE Trans. Image Processing*, vol. 19, no. 9, pp. 2265–2277, 2010.
[16] X. Tan, J. Li, and P. Stoica, "Efficient sparse Bayesian learning via Gibbs sampling," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 3634 –3637.
[17] J. Nocedal and S. J. Wright, *Numerical Optimization*, ser. Series in Operations Research. New York: Springer Verlag, 2000.
[18] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Trans. Signal Processing Mag.*, pp. 21–36, May 2003.
[19] G. Rochefort, F. Champagnat, G. Le Besnerais, and J.-F. Giovannelli, "An improved observation model for SR under affine motion," *IEEE Trans. Image Processing*, vol. 15, no. 11, pp. 3325–3337, Nov. 2006.
[20] F. Orieux, E. Sepulveda, V. Loriette, B. Dubertret, and J.-C. Olivo-Marin, "Bayesian estimation for optimized structured illumination microscopy," *IEEE Trans. Image Processing*, 2011, in press.
[21] J. Idier, "Optimization and sampling in the Gaussian case," Personnal communication, 2006.

université
PARIS-SACLAY

Titre : Contributions en imagerie computationnelle et problèmes inverses

Mots clés : Problèmes inverses, MCMC, deconvolution, hyperspectrale, multispectrale

Mes travaux portent sur les méthodes de résolution de problèmes inverses, notamment en imagerie computationnelle. Je présenterais tout d'abord des contributions sur la restauration et la fusion de données pour des problèmes d'imagerie multi et hyperspectrale résolus avec des approches variationnelles et des algorithmes de majorisation-minimisation. Je parlerais également de contributions sur des algorithmes MCMC pour les problèmes en grande dimension avec des problèmes de reconstruction d'images biologique. Enfin je mentionnerais également mes travaux sur la synthèse de Fourier et l'adéquation algorithme-architecture ainsi que des perspectives sur l'accélération d'algorithmes par préconditionneur et l'apprentissage machine.

Title : Contributions in Computationnal Imaging and Inverse Problems

Keywords : Inverse problems, MCMC, deconvolution, hyperspectrale, multispectrale

My work focuses on inverse problem solving methods, particularly in computational imaging. I will first present contributions on data restoration and fusion for multi- and hyperspectral imaging problems solved with variational approaches and majorization-minimization algorithms. I will also discuss contributions on MCMC algorithms for high-dimensional problems with biological image reconstruction problems. Finally, I will also mention my work on Fourier synthesis and algorithm-architecture matching, as well as perspectives on algorithm acceleration via preconditioning and machine learning.