Contributions en traitement du signal par méthodes d'apprentissage à noyaux

Paul Honeine

Habilitation à Diriger des Recherches

Contributions en traitement du signal par méthodes d'apprentissage à noyaux

présentée et soutenue publiquement le 9 décembre 2013

par

Paul HONEINE

Maître de conférences
à l'Université de technologie de Troyes

Devant la commission d'examen :

Rapporteurs :

Stéphane CANU	Professeur à l'INSA de Rouen
Thierry DENŒUX	Professeur à l'Université de technologie de Compiègne
Christian JUTTEN	Professeur à l'Université Joseph Fourier de Grenoble

Examinateurs :

Patrick FLANDRIN	Professeur à l'Ecole Normale Supérieure de Lyon
Pascal LARZABAL	Professeur à l'Université Paris Sud
Régis LENGELLE	Professeur à l'Université de technologie de Troyes
Cédric RICHARD	Professeur à l'Université de Nice Sophia-Antipolis

Ecole Doctorale de l'Université de technologie de Compiègne

A Celui qui a créé l'Homme, intelligent et libre.

A mes parents sans qui rien n'aurait été possible.

Table des matières

Première partie

Notice de titre et travaux

1

Curriculum Vitæ

Sommaire

Paul HONEINE
Né le 02/10/1977, à Beyrouth (Liban)
Nationalités libanaise et française
Situation familiale : célibataire
http://honeine.fr/paul/

Université de technologie de Troyes
12 rue Marie Curie, 10000 Troyes
Téléphone : +33 (3) 25 71 56 25
Fax : +33 (3) 25 71 56 99
paul.honeine@utt.fr

1.1 Parcours

1.1.1 Expérience professionnelle

2008–... **Maître de conférences** à l'Université de technologie de Troyes (UTT), Institut Charles Delaunay (UMR CNRS 6279), équipe du Laboratoire de Modélisation et Sûreté des Systèmes (LM2S) (CNU section 61)

2007–2008 **Post-doctorant** à l'Institut Charles Delaunay (ICD), UMR CNRS 6279, UTT
Projet ANR : Apprentissage et noyaux pour la représentation et la décision en traitement du signal
Collaborations : S. Canu (INSA Rouen, coordinateur), A. Rakotomamonjy (Univ. Rouen), M. Davy (LAGIS, Lille), O. Cappé (ENST Paris), C. Richard (ICD/LM2S)

2003–2007 **Ingénieur R&D** en contrôle acoustique et vibratoire, Sonalyse [1] (startup), Alés (Gard).

1.1.2 Formation

2003–2007 **Doctorant**, à l'Université de technologie de Troyes. Thèse sous la direction de Cédric Richard.
Financement : ANRT (convention CIFRE) et ANR
Intitulé : Méthodes à noyaux pour l'analyse et la décision en environnement non-stationnaire
Soutenue le 13 décembre 2007, devant le jury composé de P. Flandrin (président, ENS de Lyon), S. Canu (rapporteur, INSA de Rouen), B. Torrésani (rapporteur, Univ. de Provence), M. Davy (Ecole Centrale de Lille), R. Gribonval (INRIA), H. Snoussi (UTT), et C. Richard (UTT).

2002–2003 **Etudiant en DEA Contrôle Industriel**, à l'Université Libanaise, Faculté de Génie, Liban (classé 1er sur 10, mention Très Bien)
Stage de recherche au LM2S, UTT, financé par un contrat avec PSA Peugeot Citroën, centre DRIA/SARA/EMSA/PEFH (Perception et Facteurs Humains), sur la théorie de l'information pour l'analyse du typage sonore de véhicules

1997–2002 **Eléve-ingénieur** en Génie Mécanique, à l'Université Libanaise, Faculté de Génie, Liban (mention Très Bien)

1.1.3 Prix et distinctions

2010 **Prime d'Excellence Scientifique** (PES) pendant 4 ans (2010–2014)

2009 **Best Paper Award** au 19ème congrés IEEE MLSP pour le papier [Honeine and Richard, 2009] (premier auteur)

2008 Distinction ***Reviewer Appreciation*** de IEEE SP

2008 Finaliste pour le **prix de thèse** du rectorat de Reims, Région Champagne-Ardenne

1. J'ai été impliqué dans plusieurs projets industriels sur l'analyse vibratoire et acoustique de machines tournantes, avec Eurocopter (développement d'algorithmes pour des mesures vibratoires spécifiques aux turbines), SNR (contrôle vibratoire automatique des roulements en chaîne de production), et Renault (contrôle acoustique de soudure au laser).

1.2 Bilan statistique des publications

Ouvrage **4** chapitres de livres internationaux (2013, 2013, 2012, 2010)

Revue **16** articles de revues internationales, dont 11 IEEE
3 articles de revues francophones, invités (Traitement du Signal)

Congrés **50** communications dans des congrés internationaux à comité de lecture et actes
(4× EUSIPCO'13, 2× WHISPERS'13, SPAWC'13, ICASSP'13, MLSP'12, EUSIPCO'12, 2×SSP'12, ISIT'12, Simhydro'12, 2× IGARSS'12, Hydro'12, WHISPERS'12, 3× ICT'12, 2× ICASSP'12, 2× Asilomar'11, 2× SiPS'11, 2× EUSIPCO'11, WoSSPA'11, 3× SSP'11, Asilomar'10, EMBC'10, EUSIPCO'10, 2× IGARSS'10, WHISPERS'10, ICASSP'10, MLSP'09, SSP'09, ICASSP'09, Globecom'08, Asilomar'08, EUSIPCO'08, SAM'08, SSP'07, ISIT'07, ICASSP'06)
18 communications dans des congrés francophones à comité de lecture et actes (GRETSI et CIFA)
10 papiers dans des workshops (avec actes) et présentations dans des journées thématiques GdR

Prix 1 best paper award (IEEE MLSP'09), premier auteur

Brevet 1 brevet international (2010)

Indices Revues : *largest impact factor* ≈ 6, *5-year impact factor* ≈ 7 (papier publié en 2011, premier auteur)
Citations sur les 5 dernières années : h-index = 9 (selon Google scholar)

Bilan

année :			2013		2012		2011		2010		2009		2008		2007		2006		2005
articles de conférences[1] :	68	=	13	+	15	+	16	+	8	+	6	+	4	+	4	+	1	+	1
articles de revues, ouvrages :	23	=	10	+	3	+	4	+	3	+	1	+	1	+	1				

1.3 Principales collaborations (internationales et nationales)

Le premier de mes collaborateurs est Cédric Richard, avec qui j'ai eu le plaisir de faire mon stage de DEA et ma thèse de doctorat. Depuis, j'ai eu le plaisir de collaborer avec lui sur plusieurs projets. Mon appréhension de la recherche se nourrit des travaux que nous avons effectués ensemble. J'espère avoir hérité de lui cette rigueur scientifique et son exceptionnelle acuité scientifique.

J'entretiens également des collaborations privilégiées avec plusieurs équipes libanaises. En premier lieu, j'ai travaillé avec Clovis Francis de l'Université Libanaise, avec qui j'ai co-dirigé la thèse de Maya Kallas et plus récemment la thèse de Nisrine Ghadban. J'ai aussi collaboré avec Hassan Amoud du Centre Azm pour la Recherche en Biotechnologie et ses Applications. J'ai entretenu des collaborations avec Chafic Saïdé et Roger Achkar de l'American University of Science and Technology, dans le cadre des travaux de recherche de Chafic Saïdé. J'ai récemment collaboré avec Joumana Farah de l'Université Saint-Esprit de Kaslik sur les réseaux de capteurs sans fil.

Les collaborations les plus remarquables sont présentées dans la suite par niveau, et dans chaque niveau, par ordre chronologique.

Niveau international

2007–... **José C. M. Bermudez** de Federal University of Santa Catarina, Brésil,
sur le traitement adaptatif de l'information, depuis ma thèse de doctorat (12 articles).

2010–... **Clovis Francis** de l'Université Libanaise, Liban,
dans le cadre de la thèse de Maya Kallas et de la thèse de Nisrine Ghadban (17 articles).

2010–2013 **Hassan Amoud** de l'Université Libanaise, Liban,
dans le cadre d'une collaboration avec le centre Azm au Liban (21 articles).

2013–... **Joumana Farah** de l'Université Saint-Esprit de Kaslik, Liban,
sur les réseaux de capteurs sans fil, depuis 2013 (5 articles).

1. Conférences à comité de lecture et actes

Niveau national

2005–2012 **Patrick Flandrin** du Laboratoire de Physiques, ENS-Lyon,
sur l'analyse de la non-stationnarité, depuis ma thèse et dans le cadre du projet StaRAC (11 articles).

2009–2012 **Pierre Borgnat** du Laboratoire de Physiques, ENS-Lyon,
sur la détection de non-stationnarité par les substituts, dans le cadre du projet StaRAC (7 articles).

2010–... **Henri Lantéri** du Laboratoire Lagrange, Observatoire de la Côte d'Azur, Nice
sur l'étude de la contrainte de non-négativité (6 articles).

1.4 Activités scientifiques

Les thèmes de recherche que j'ai développés au sein de l'équipe LM2S à l'UTT sont identifiés par les mots-clés suivants :

- analyse et classification de signaux non stationnaires ; détection de non-stationnarité ; distributions temps-fréquence ;
- apprentissage statistique pour la reconnaissance des formes ; méthodes à noyaux ; représentations parcimonieuses ;
- apprentissage en ligne ; filtrage adaptatif non linéaire ; identification de systèmes non linéaires ;
- traitement du signal et traitement collaboratif de l'information dans les réseaux de capteurs ;
- analyse et classification en imagerie hyperspectrale ; démélange spectral, linéaire et non linéaire.

Mes activités de recherche émergent naturellement dans la présentation des projets de recherche auxquels j'ai participé.

Au cours de ces 5 dernières années, j'ai participé très activement à 5 projets ANR : KernSig, StaRAC, Vigirés'eau, SCALA et HYPANEMA. Par ailleurs, j'ai une activité de recherche importante sur la thématique des réseaux de capteurs : plateforme CAPSEC, deux projets Abondement Carnot, le projet Région WiDiD, etc.

1.4.1 Porteur/coordinateur de plusieurs projets

J'ai été porteur/coordinateur de 4 projets, selon différents niveaux de collaboration et de partenariat.

- Au niveau UTT-UTC, j'ai été porteur/coordinateur d'un projet dans le cadre du programme « Abondement Carnot » (2009–2012), sur le traitement de l'information dans les réseaux de capteurs sans fil. Ce projet a permis de développer et de consolider la collaboration entre trois équipes, deux de l'UTT, LM2S et ERA (Environnement de Réseaux Autonomes), et une équipe du laboratoire Heudiasyc de l'UTC.
- Au niveau de la Région, je suis porteur/coordinateur du projet WiDiD « Wireless Diffusion Detection » (2012–2015) dans le cadre de la plateforme CAPSEC, programme Essaimage-Région.
- Au niveau national, je suis coordinateur de l'équipe UTT dans le cadre du projet HYPANEMA (2012–2015), ANR programme blanc. Ce projet associe 4 laboratoires de recherche : Laboratoire Lagrange (Nice), IRIT (Toulouse), Gipsa-lab (Grenoble) et LM2S (Troyes).
- Au niveau international, j'ai été porteur/coordinateur d'un projet dans le cadre du programme franco-libanais « CEDRE » (2011–2012). Dans ce cadre, j'ai co-dirigé la thèse de Mlle Maya Kallas qui est actuellement Maître de Conférences à l'Université de Lorraine.

Par ailleurs, je suis porteur d'un projet de thèse CIFRE en cours de finalisation avec SNECMA. Un résumé de ces différents projets est présenté dans la suite, par ordre chronologique décroissant.

2012–2016 **HYPANEMA** (ANR, programme blanc) sur des *algorithmes de démélange non linéaire pour l'analyse de données hyperspectrales* (budget total 1 539 717 €, dont 215 323 € accordés à l'UTT)
Ce projet associe des équipe dirigées par C. Richard (coordinateur, Université de Nice Sophia-Antipolis, Laboratoire Lagrange), J.-Y. Tourneret (INP Toulouse, IRIT), J. Chanussot (Grenoble INP, Gipsa-lab) et moi-même (UTT)
↠ Je suis coordinateur de l'équipe UTT de ce projet.
↠ Je suis également coordinateur d'un projet du « Programme Excellence » (Région Champagne-Ardenne) qui complète le projet ANR HYPANEMA (subvention 41 600 €).

2012–2015 **WiDiD : Wireless Diffusion Detection** (Projet Essaimage-Région) sur la *détection et surveillance de la diffusion d'espèce biochimique nocive à base de réseaux de capteurs* (aide accordée 110 000 €)
Ce projet regroupe à l'UTT : P. Honeine, F. Mourad et H Snoussi.
↠ Je suis porteur/coordinateur de ce projet.
↠ Je co-dirige, avec Farah Mourad, la thèse de doctorat de Sandy Mahfouz dans ce cadre.

2011 **Projet CEDRE** du programme franco-libanais de Coopération pour l'Evaluation et le Développement de la Recherche) (aide accordée 6 000 €).
Ce projet, dont je suis porteur/coordinateur, associe l'UTT et l'Université de Nice Sophia-Antipolis du côté français, et l'Université Libanaise et le Centre AZM (C. Francis, M. Khalil et H. Amoud) du côté libanais. Il s'aligne parfaitement avec le projet de recherche de M. Kallas, doctorante en cotutelle UTT-UL sous la direction de C. Francis et de moi-même.
↠ J'ai été porteur/coordinateur de ce projet.
↠ J'ai co-dirigé, avec Clovis Francis, la thèse de doctorat de Maya Kallas dans ce cadre.

2009–2011 **Projet Abondement Carnot**, sur la *détection de changements par traitement de l'information dans les réseaux de capteurs collaboratifs* (aide accordée 36 000 €)
Ce projet associe les partenaires issus de 3 entités de recherches à l'ICD (UTT) et à l'Heudiasyc (UTC) : P. Honeine (resp., ICD/LM2S) C. Richard (ICD/LM2S), H. Snoussi (ICD/LM2S), G. Doyen (ICD/ERA), M. Esseghir (ICD/ERA), et F. Abdallah (Heudiasyc/DI).
↠ J'ai été porteur/coordinateur de ce projet.

1.4.2 Implication active dans plusieurs projets

J'ai participé activement dans plusieurs projets, avec une implication directe sur l'encadrement d'un doctorant ou post-doctorant dans chaque projet. Un résumé de ces projets est présenté dans la suite.

2012–2016 **SCALA** (ANR, programme CSOSG) sur la *Surveillance Continue d'Activité et Localisation d'Agression* dans les réseaux de systèmes cyber-physiques SCADA (1 178 000 €, dont 710 000 € à l'UTT).
Ce projet franco-allemand associe *Suez environnement (équipe Ondéo Systems), Diateam, Eurawasser (Fraunhofer Institute)* à l'UTT. Equipe UTT : I. Nikiforov (coordinateur), L. Fillatre, P. Honeine, P. Beauseroy, Ph. Cornu.
↠ Je co-dirige la thèse de Patric Nader, sur la « Détection de cyber-intrusions par apprentissage statistique »

2009–2012 **Vigirés'eau** (ANR, programme CSOSG) sur la *Surveillance en temps réel de la qualité de l'eau potable d'un réseau de distribution en vue de la détection d'intrusions* (1 124 000 €, dont 430 000 € à l'UTT).
Ce projet associe *Suez environnement* (équipe Ondéo Systems) à l'UTT, et regroupe d'une part des chercheurs de l'ICD, L. Fillatre (coordinateur), I. Nikiforov, H. Snoussi, P. Honeine, et C. Richard (Laboratoire Lagrange), et d'autre part une équipe de *Suez* avec Francis Campan (resp.), Stéphane Deveughéle, Hao-Nhiên Pham, Guillaume Gancel (Safége), Pierre-Antoine Jarrige (Safége).
Trois coordinateurs de l'UTT se sont succédés à ce projet : C. Richard, L. Fillatre et moi-même.
↠ J'ai co-dirigé, avec Cédric Richard, la thèse de Zineb Noumir dans le cadre de ce projet.

2007–2010 **StaRAC** (ANR, programme blanc) sur la *Stationnarité Relative et Approches Connexes* (225 000 €)
Ce projet associe P. Flandrin (coordinateur, ENS Lyon), P. Borgnat (ENS Lyon), P.-O. Amblard (GIPSA Grenoble), et C. Richard (UTT).
↠ J'ai participé activement à l'encadrement du post-doctorant Hassan Amoud dans ce projet.

2006–2009 **KernSig** (ANR, programme blanc) sur l'*Apprentissage et noyaux pour la représentation et la décision en traitement du signal* (193 000 €)
Ce projet associe S. Canu (coordinateur, INSA Rouen), A. Rakotomamonjy (Univ. Rouen), M. Davy (LAGIS, Lille), O. Cappé (ENST Paris), et C. Richard (UTT).
↠ J'ai été post-doctorant dans le cadre de ce projet.

1.4.3 Autres projets de recherche

2012–2014 **Risk-Perform** (Projet Région Emergence) sur la maitrise des risques en finances. Ce projet associe l'UTT à *Reims Management School* (équipe dirigée par S. Lleo).
Equipe UTT : M. Fouladirad (coordinatrice), Y. Dijoux, E. Deloux, et P. Honeine.

2012–2014 **Mobiloc** (Abondement Carnot) sur la Localisation et contrôle de capteurs embarqués mobile (42 000 €)
Ce projet qui constitue une suite au projet Carnot (dont j'ai été coordinateur) regroupe 4 entités de recherche, à l'UTT et à l'UTC : H. Snoussi (resp., ICD/LM2S), P. Honeine (ICD/LM2S), F. Yalaoui (ICD/LOSI), L. Amodeo (ICD/LOSI), F. Hnaien (ICD/LOSI), H. Chehade (ICD/LOSI), L. Khoukhi (ICD/ERA), M. Esseghir (ICD/ERA), F. Abdallah (Heudiasyc/DI), V. Frémont (Heudiasyc/DI)

2011–2012 **CoBISS** (ANR, programme Emergence) : *Compact Bidimensional Sampling Spectrometer* (337 000 €).
Ce projet, porté par l'équipe LNIO (Laboratoire de Nanotechnologie et d'Instrumentation Optique) de l'UTT, associe l'UTT et le Laboratoire des Technologies de la Microélectronique (CEA-Grenoble, UMR CNRS 5129)

2010–2013 **Plateforme SURECAP** (Contrat de Projets Etat-Région (CPER)) sur la *Fonction de surveillance dans les réseaux de capteurs sans fil* (301 000 €) du CPER axe ICOS (Information, Communication, Organisation et Sécurité des systèmes), thématique « S3 : Sécurité et Sûreté des Systèmes »

2010–2011 **WiCaN (Wireless Camera Network)** (Projet oséo) sur la *Réalisation d'un démonstrateur de réseau de caméras intelligentes*, dans le cadre d'un projet pour une start-up (50 000 €).

1.5 Formation Doctorale

Au cours des 5 dernières années, mes activités liées à la formation doctorale comprennent :
- les co-directions de thèse de doctorat, 3 doctorants ayant soutenu et 4 doctorants en cours ;
- les encadrements de 8 stages de recherche ;
- deux cours dispensés en Master de Recherche, un à l'UTT et un à l'UTC - Université Libanaise.

1.5.1 Thèses de doctorat soutenues

J'ai co-dirigé les thèses de doctorat suivantes, présentées par ordre chronologique de soutenances de thèse.

avr10–nov12 **Maya Kallas** co-direction avec Clovis Francis (Université Libanaise)

Intitulé : Méthodes à noyaux en reconnaissance de formes, prédiction et classification. Application aux biosignaux

Financement : Région, Université Libanaise, UTT, programme franco-libanais CEDRE

Collaboration : Hassan Amoud (centre AZM pour la recherche en biotechnologie et ses applications

Publications :

2 articles dans des revues internationales (Signal Processing, Pattern Recognition) [Kallas et al., 2013a, Kallas et al., 2013b]

9 conférences internationales (dont 5 IEEE) [Kallas et al., 2012a, Kallas et al., 2012b, Kallas et al., 2012c, Kallas et al., 2011b, Kanaan et al., 2011, Kallas et al., 2011e, Kallas et al., 2011c, Honeine et al., 2011, Kallas et al., 2010b]

2 conférences francophones à comités de lecture et actes [Kallas et al., 2011d, Kallas et al., 2011a]

2 workshop avec actes [Khodor et al., 2011, Kallas et al., 2010a]

Soutenance : le 23 Novembre 2012 à l'UTT (financement de 36 mois, soutenue au 32-ème mois), devant le jury composé de F. D'Alché-Buc (rapporteur, Univ. d'Evry-Val d'Essonne), S. Canu (rapporteur, INSA de Rouen), C. Francis (Université Libanaise, Liban), P. Honeine (UTT), R. Lengellé (président, UTT), N. Nassif (American University of Beirut, Liban) et C. Richard (Université de Nice Sophia-Antipolis)

Devenir de la doctorante : Maître de Conférences à l'Université de Lorraine depuis septembre 2013

dec09–dec12 **Zineb Noumir** co-direction avec Cédric Richard (Université de Nice Sophia-Antipolis)

Intitulé : Surveillance en temps réel de la qualité de l'eau potable d'un réseau de distribution par apprentissage statistique (Vigires'eau)

Financement : ANR, Programme CSOSG (Concepts, Systèmes et Outils pour la Sécurité Globale)

Partenaire : Partenaire industriel : Suez Environnement - Safège

Equipes : UTT : Lionel Fillatre (resp.), Igor Nikiforov, Hichem Snoussi, Nourddine Azzaoui
Suez Environnement (Safège) : Stéphane Deveughele (resp.), Guillaume Gancel, Pierre-Antoine Jarrige, Hao-Nhiên Pham

Publications

1 chapitre de livre international (Springer) [Noumir et al., 2013]

1 article de revue internationale (Signal Processing) [Honeine et al., 2013b]

7 conférences internationales (dont 3 IEEE) [Noumir et al., 2012b, Noumir et al., 2012e, Noumir et al., 2012d, Noumir et al., 2012c, Noumir et al., 2012a, Deveughèle et al., 2012, Noumir et al., 2011b]

1 conférence francophone à comités de lecture et actes [Noumir et al., 2011a]

3 workshops avec actes [Yin et al., 2012, Fillatre et al., 2011, Fillatre et al., 2010]

Soutenance : le 11 décembre 2012 (financement de 36 mois, soutenue au 36-ème mois), devant le jury composé de F. Abdallah (rapporteur, UTC, Compiègne), D. Brie (rapporteur, Université de Lorraine, Nancy), P. Borgnat (ENS Lyon), G. Gelle (président, Université de Reims Champagne-Ardenne), P. Honeine (UTT, Troyes) et C. Richard (Université de Nice Sophia-Antipolis), et comme invité : F. Campan (ONDEO Systems - Le Pecq), A. Dembele (ONDEO Systems - Le Pecq), et P.-A. Jarrige (SAFEGE - Nanterre)

Devenir de la doctorante : ATER à l'Université Paris-Sud depuis septembre 2013

oct09–jan13 **Jie Chen** co-direction avec Cédric Richard (Université de Nice Sophia-Antipolis)

Intitulé : System identification under non-negativity constraints – Applications in adaptive filtering and hyperspectral image analysis

Financement : Allocations du China Scholarship Council (coopération avec les UT et les INSA)

Collaborations : Université de Nice Sophia-Antipolis : Henri Lantéri et Céline Theys

Publications :

4 articles de revues internationales (dont 3 IEEE)
[Chen et al., 2013d, Chen et al., 2013e, Chen et al., 2011b, Honeine et al., 2010]

9 conférences internationales (dont 4 IEEE)
[Chen et al., 2013b, Chen et al., 2012b, Chen et al., 2012a, Chen et al., 2011c, Chen et al., 2011a, Chen et al., 2011f, Chen et al., 2011e, Chen et al., 2010a, Chen et al., 2010b]

4 conférences francophones à comités de lecture et actes
[Chen et al., 2013a, Chen et al., 2011d, Richard et al., 2011, Chen et al., 2010c]

Soutenance : le 28 Janvier 2013 (financement de thèse de 42 mois, soutenue au 40-ème mois), devant le jury composé de C. Jutten (rapporteur, Grenoble INP, GIPSA-Lab), J.-Y. Tourneret (rapporteur, INP Toulouse), J.-C. M. Bermudez (Federal University of Santa Catarina, Brazil), J. Chanussot (président, Grenoble INP, GIPSA-Lab), P. Larzabal (ENS Cachan, SATIE), P. Honeine (UTT, Troyes) et C. Richard (Université de Nice Sophia-Antipolis).

Devenir du doctorant : actuellement post-doctorant à l'Université de Nice Sophia-Antipolis puis, dès début 2014, post-doctorant à Michigan University

1.5.2 Thèses de doctorat en cours

oct12–... **Sandy Mahfouz** co-direction avec Farah Mourad (UTT)

Intitulé : Détection et surveillance de la diffusion d'espèce biochimique nocive à base de réseaux de capteurs

Financement : Projet WiDiD financé par la Région Champagne-Ardenne (Programme Essaimage)

Publications :

2 conférences internationales (dont 1 IEEE)
[Mahfouz et al., 2013c, Mahfouz et al., 2013a]

1 conférence francophone à comités de lecture et actes
[Mahfouz et al., 2013b]

oct12–... **Patric Nader** co-direction avec Pierre Beauseroy (UTT)

Intitulé : Détection de cyber-intrusions par apprentissage statistique

Financement : Projet SCALA « Surveillance Continue d'Activité et Localisation d'Agression » (ANR, programme CSOSG)

Publications :

1 conférence internationale
[Nader et al., 2013]

nov12–... **Nisrine Ghadban** co-direction avec Clovis Francis (Université Libanaise)

Intitulé : Techniques de surveillance de phénomènes physiques par fusion de l'information dans les réseaux de capteurs.

Collaborations : Farah Mourad (UTT) et Joumana Farah (USEK, Liban)

Financement : Programme de cotutelle Université Libanaise et UT/INSA

Publications :

2 conférences francophone à comités de lecture et actes
[Ghadban et al., 2013b, Ghadban et al., 2013a]

oct13–... **Fei ZHU** co-direction avec Régis Lengellé (UTT)

Intitulé : Démélange en imagerie hyperspectrale.

Financement : Allocations du China Scholarship Council, depuis octobre 2013.

1.5.3 Participations à d'autres thèses

J'ai également participé aux travaux de thèse de trois doctorants.

Mehdi Essoloh :

Intitulé : Méthodes d'apprentissage à noyau pour l'estimation distribuée dans les réseaux de capteurs sans fil

Directeurs de thèse : Cédric Richard et Hichem Snoussi

Soutenance : le 3 décembre 2009

J'ai participé à 6 publications parmi les 8 réalisées au cours de cette thèse :
[Essoloh et al., 2009, Honeine et al., 2009a, Essoloh et al., 2008, Honeine et al., 2008a, Honeine et al., 2008c, Richard et al., 2008]

Chafic Saïdé :

Intitulé : Filtrage adaptatif à l'aide de méthodes à noyau. Application au contrôle d'un palier magnétique actif

Directeur de thèse : Régis Lengellé

Soutenance : le 19 septembre 2013

Activité professionnelle : Chafic Saïdé est professor à l'*American University of Sciences and Technology*, Beyrouth

J'ai participé à toutes les publications de cette thèse :
[Saidé et al., 2013b, Saidé et al., 2013a, Saidé et al., 2012]

Nguyen Hoang Nguyen :

Thèse sur le démélange en imagerie hyperspectrale

Directeurs de thèse : Cédric Richard et Céline Theys

Soutenance : prévue le 3 décembre 2013

J'ai participé à toutes les publications de cette thèse :
[Honeine et al., 2013c, Nguyen et al., 2013, Nguyen et al., 2012]

Jurys de thèse :

En plus des jurys de thèse de mes trois doctorants, j'ai participé en tant qu'examinateur aux jurys de thèses des doctorants suivants :

- *Chafic Saïdé*, thèse dirigée par Régis Lengellé (UTT) et soutenue en septembre 2013 à Troyes.
- *Jihan Khoder*, thèse dirigée par Fethi Ben Ouezdou (Université de Versailles) et Rafic Younes (Université Libanaise) et soutenue le 24 octobre 2013 à Versailles.
- *Nguyen Hoang Nguyen*, thèse dirigée par Cédric Richard et Céline Theys (Université de Nice Sophia-Antipolis) et soutenue le 3 décembre 2013 à Nice.

1.5.4 Stages de recherche

A l'UTT, j'ai participé à l'encadrement de plusieurs stages de Master 2 Recherche :

- *Wissam Sammouri* (2010), sur la détection de changement dans les réseaux de capteurs. Il est actuellement en thèse à IFSTTAR.
- *Nisrine Ghadban* (2011), sur le traitement distribué de l'information dans les réseaux de capteurs sans fil. Elle est actuellement en thèse sous la direction.
- *Fei Zhu* (2013), sur le traitement en imagerie hyperspectrale. Elle est actuellement en thèse sous ma direction.

J'ai aussi participé au co-encadrement des étudiants suivants :

- *Lara Kanaan* et *Dalia Merheb* (2010), co-encadrement à l'Université Saint Esprit de Kaslik, avec Clovis Francis sur la classification de signaux ECG.
- *Nadine Khoder* (2011), co-encadrement à l'Université Libanaise avec Hassan Amoud sur le problème de pré-image avec contraintes. Elle est actuellement en thèse sous la direction de Guy Carrault à l'Université de Rennes.
- *Rita Ammanouil* et *Jean Abou Melhem* (2013), co-encadrement à l'Université Saint Esprit de Kaslik, avec Joumana Farah sur la classification en imagerie hyperspectrale. Rita Ammanouil est actuellement en thèse à l'Université de Nice Sophia-Antipolis sous la direction d'André Ferrari et de Cédric Richard.

1.6 Responsabilités collectives et rayonnement

1.6.1 Sociétés

Je suis correspondant du l'Institut Charles Delaunay (UMR CNRS 6279) auprès du groupement de recherche GDR ISIS , depuis 2012. Je veille aux inscriptions des membres du laboratoire et je sers d'intermédiaire entre eux et le GDR en cas de besoin.

Je suis membre des sociétés savantes IEEE (signal processing society), EURASIP, et Club EEA.

1.6.2 Organisation

J'ai participé à l'organisation du Colloque GRETSI 2007, à Troyes. Plus précisément, j'ai été jeune porteur d'un projet oséo Anvar pour l'organisation du colloque. Le projet, de durée 9 mois, avait pour objet la communication sur les thèmes scientifiques et technologiques stratégiques en traitement du signal et des images auprès des conférenciers et des partenaires académiques, groupes industriels, PME et PMI.

J'ai été président de séance dans les conférences suivantes :

- IEEE WHISPERS 2013, session *Classification* ;
- Colloque ROADEF 2013, session *Aide à la décision dans les réseaux collaboratifs* ;
- 19th ICT 2012, session *Compressed Sensing and Signal Processing Techniques* ;
- 18th IEEE SSP workshop 2011, session *Learning Theory and Pattern Recognition*.

J'ai été aussi membre du comité de programme du 5th *International Congress on Image and Signal Processing*, 2012.

1.6.3 Activités d'expertise

Je suis relecteur pour les revues Trans. Signal Processing (IEEE, $\approx$ 50 acticles depuis 2008), Trans. Geoscience and Remote Sensing (IEEE), Trans. Neural Networks and Learning Systems (IEEE), Signal Processing (Elsevier), Neurocomputing (Elsevier) et Engineering Applications of Artificial Intelligence (Elsevier).

Je suis relecteur occasionnel pour les revues International Journal of Communication Systems (Wiley), Statistical Analysis and Data Mining (Wiley), Digital Signal Processing (Elsevier), Journal of Signal Processing Systems (Springer), et Traitement du Signal.

Je suis relecteur pour les conférences IEEE ICASSP (depuis 2012), IEEE EMBC (depuis 2012), GRETSI (depuis 2011), ainsi que les conférences IEEE WHISPERS (2013), IEEE Globecom (2008 et 2012), Eusipco (2012), ICT (2012), ISSPA (2012), CISP (2012), BMEI (2012), IEEE SSP (2011).

J'ai rapporté sur des projets soumis à l'ANR (1 projet en 2013 et 2 projets en 2012) et l'ANRT-CIFRE (1 projet en 2012).

1.7 Enseignements

Parallèlement à mes activités de recherche, j'effectue des activités d'enseignement en tant que Maître de Conférences à l'Université de technologie de Troyes. J'ai ainsi effectué 1310 heures depuis septembre 2008 (jusqu'en été 2013), soit une moyenne de 260 heures d'enseignement par an, dont 21% de cours, 46% de travaux dirigés et 33% de travaux pratiques. A titre indicatif, le volume d'enseignement de l'année 2012–2013 est le suivant : 89 heures de cours, 108 heures de travaux dirigés, et 76 heures de travaux pratiques, soit 32.6% de cours, 39.6% de travaux dirigés et 27.8% de travaux pratiques. Par ailleurs, je participe à l'encadrement de stages d'ingénieur d'une durée de 6 mois, de 12 élèves-ingénieurs par an en moyenne.

La plupart de mes enseignements est dispensée aux élèves-ingénieurs de l'UTT. Je suis responsable de 3 Unités de Valeur (UV) (voir la section 1.7.3 pour plus de détails) :

- **NF05 : « Introduction au langage C »**, niveau 2-ème année, cycle ingénieur ;
- **IF01 : « Théorie et codage de l'information »**, niveau 3/4-ème année, spécialité Systèmes d'Information et Télécommunications ;
- **SY05 : « Outils d'aide à la décision et théorie des jeux »**, niveau 4/5-ème année, spécialité Systèmes Industriels.

A ces 3 cours classiques pour les élèves-ingénieurs, se rajoutent 2 cours en Master 2 Recherche. Il s'agit de cours centrés sur mes activités de recherche et qui permettent de sensibiliser les étudiants aux travaux de recherche que je mène :

- **OS14 : « Reconnaissance des formes et applications en surveillance »**, dispensée en Master 2 Recherche de l'UTT, option Optimisation et Sûreté des Systèmes. Je suis responsable de la première partie de cette UV, soit 47% des heures d'enseignement.
- **TS02 : « Estimation et prédiction »**, dispensée aux étudiants du Master 2 Recherche « Contrôle Industriel », double diplôme délivré par l'Université de technologie de Compiègne (UTC) et l'Université Libanaise. Je suis responsable de ce cours.

1.7.1 Création de contenus pédagogiques

La plupart des UV a été considérablement modifiée, notamment avec la création de contenus pédagogiques :

- NF05 : création de transparents de cours et des énoncés de travaux pratiques.
- SY05 : création de transparents de cours et des énoncés de travaux dirigés , en collaboration avec Nacima Labadie (laboratoire d'optimisation des systèmes industriels, UTT) et Roberto Wolfler Calvo (Université Paris 13).
- TS02 : création de transparents de cours, en collaboration avec Régis Lengellé.

Un effort considérable a été entrepris pour la création de OS14. Je suis responsable de la première partie de cette UV dispensée en Master 2 Recherche. Cette partie, qui correspond à 47% des heures d'enseignements de l'UV, concerne les fondements de l'apprentissage statistique et les méthodes à noyaux. J'ai ainsi développé un support de cours et des transparents de cours.

Les différents contenus pédagogiques sont disponible sur ma page web : http://honeine.fr/paul/.

1.7.2 Enseignements par niveau de formation

Enseignements en formation initiale, premier cycle

Depuis 2011, j'assure des travaux dirigés de « Mesure physique et instrumentation » (MS11), dispensée aux étudiants en première année. Volume annuel : 34 heures de travaux dirigés.

Depuis 2011, je suis responsable de « Introduction au langage C » (NF05), dispensée aux étudiants en deuxième année, cycle ingénieur. Volume annuel : 34 heures de cours et 28 heures de travaux pratiques.

Enseignements en formation initiale, deuxième cycle

Depuis 2008, j'assure des travaux dirigés de « Statistiques pour l'ingénieur » (SY02), pour un volume annuel 34 heures.

De 2008 à 2009, j'ai été en charge des travaux pratiques de « Techniques mathématiques de l'ingénieur ». Volume annuel : 64 heures de travaux pratiques.

Spécialité Systèmes d'Information et Télécommunications

Depuis 2008, je suis responsable de « Théorie et codage de l'information » (IF01), dispensée aux étudiants dans la spécialité Systèmes d'Information et Télécommunications (années 3/4). Volume annuel : 34 heures de cours et 34 heures de travaux dirigés.

J'interviens également en « Traitement du signal » (SY06), dispensée aux étudiants dans la spécialité Systèmes d'Information et Télécommunications (année 4/5) au niveau des travaux dirigés depuis 2009. Volume annuel : 25 heures. Il s'agissait dans un premier temps de travaux pratiques, de 2006 à 2009, avec un volume annuel de 48 h. Depuis 2009, j'assure des travaux dirigés avec un volume annuel de 25 heures par an.

Spécialité Systèmes Industriels

Depuis 2008, je suis responsable de « Outils d'aide à la décision et théorie des jeux » (SY05), dispensée aux étudiants dans la spécialité Systèmes Industriels (année 4/5). Volume annuel : 17 heures de cours et 34 heures de travaux dirigés.

Enseignements en Master 2 Recherche, UTT et UTC/UL

Depuis 2008, je suis responsable de la première partie de l'UV « Reconnaissance des formes et applications en surveillance » (OS14), dispensée en Master 2 Recherche de l'UTT, option Optimisation et Sûreté des Systèmes (OSS). Volume annuel : 6 heures de cours et 10 heures de travaux dirigés.

Depuis 2010, je suis responsable de l'UV « Estimation et prédiction » (TS02) dispensée aux étudiants du Master 2 Recherche « Contrôle Industriel », double diplôme délivré par l'Université de technologie de Compiègne (UTC) et l'Université Libanaise. Le cours théorique a lieu pendant la semaine de la Toussaint à l'Université Libanaise, Liban. Volume annuel : 20 h de cours à l'Université Libanaise, Liban.

1.7.3 Résumé par module d'enseignement dont je suis responsable

2011–... **Introduction au langage C (NF05)** [responsabilité]

- UTT (cycle d'ingénieur, tronc commun, année 2)
- Cours (34 h/an) et travaux pratiques (28 h/an), depuis 2011
- Programme : introduction au langage C, environnement de développement d'applications, structures de données classiques en C (tableaux, fichiers, articles...), de l'algorithme au programme, contrôle de code et qualité du logiciel, éléments de C avancé, introduction aux systèmes d'exploitation, aux fichiers
- Support de cours : transparents de cours, énoncés de travaux pratiques.

2008–... **Théorie et codage de l'information (IF01)** [responsabilité]

- UTT (cycle d'ingénieur, branche Systèmes d'Information et Télécommunications, année 3/4)
- Cours (34 h/an) et travaux dirigés (34 h/an), depuis 2008
- En collaboration avec Cédric Richard (Laboratoire Lagrange, Observatoire de la Côte d'Azur, Nice)
- Programme : mesure quantitative de l'information, caractérisation d'une source et codage, modèles de canal discret, méthodes linéaires de codage canal.
- Support de cours : transparents de cours, énoncés de travaux dirigés.

2008–... **Outils d'aide à la décision et théorie des jeux (SY05)** [responsabilité]

- UTT (cycle d'ingénieur / M1, branche Systèmes Industriels, année 4/5)
- Cours (17 h/an) et travaux dirigés (34 h/an), depuis 2008
- En collaboration avec Nacima Labadie (Equipe LOSI, UTT) et Roberto Wolfler Calvo (Université Paris 13)
- Programme : théorie de la décision, valeur des informations, théorie de l'utilité, jeux à somme nulle et jeux à somme non nulle, jeux répétitifs, jeux coopératifs.
- Support de cours : transparents de cours, énoncés de travaux dirigés.

2008–... **Reconnaissance des formes et applications en surveillance (OS14)** [responsabilité de la première partie]

- UTT (cycle d'ingénieur, Master 2 Recherche, spécialité Optimisation et Sûreté des Systèmes)
- Première partie : cours (6 h/an) et travaux dirigés (10 h/an), depuis 2008
- Première partie – Programme : le problème d'apprentissage, régularisation, espaces de Hilbert à noyau reproduisant, méthodes de moindres carrés, mini-projets en Matlab.
- Première partie – Support de cours : manuscrit, transparents de cours, énoncés de travaux dirigés
- Volume annuel total de l'UV : 14 heures de cours et 20 heures de travaux dirigés.
- Support de cours : transparents de cours, manuscrit.

2010–... **Estimation et prédiction (TS02)** [responsabilité]

- Master 2 Recherche Contrôle Industriel (double diplôme UTC France et Université Libanaise)
- Cours (20 h/an) à l'Université Libanaise, Liban, depuis 2010
- En collaboration avec Régis Lengellé (Equipe LM2S)
- Programme : théorie de l'estimation, théorie de la décision binaire, courbes ROC, généralisation aux mesures multiples, hypothèses composites, éléments de détection séquentielle.
- Support de cours : transparents de cours, manuscrit.

1.8 Liste des publications

http://honeine.fr/paul/

Les publications sont triées par catégories, et dans chaque catégorie, par ordre chronologique décroissant.

1.8.1 Articles de revues à comité de lecture et chapitres de livres

2013

[J1] Z. Noumir, B. K. Guépié, L. Fillatre, P. Honeine, I. Nikiforov, H. Snoussi, C. Richard, P.-A. Jarrige, and F. Campan. Detection of contamination in water distribution network. In *Advances in Hydroinformatics : SIMHYDRO - New Frontiers of Simulation, In Eds. Philippe Gourbesville, Jean Cunge, and Guy Caignaert*, Springer Hydrogeology, chapter 12, pages 1–11. Springer Science, 2013.

[J2] J. Chen, C. Richard, and P. Honeine. Nonlinear estimation of material abundances of hyperspectral images with ℓ_1-norm spatial regularization. *IEEE Transactions on Geoscience and Remote Sensing*, (accepted) 2013.

[J3] M. Kallas, P. Honeine, C. Francis, and H. Amoud. Kernel autoregressive models using yule-walker equations. *Signal Processing*, 93(11) :3053–3061, November 2013.

[J4] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Non-negativity constraints on the pre-image for pattern recognition with kernel machines. *Pattern Recognition*, 46(11) :3066–3080, November 2013.

[J5] F. Mourad-Chehade, P. Honeine, and H. Snoussi. Polar interval-based localization in mobile sensor networks. *IEEE Transactions on Aerospace and Electronic Systems*, (in press) 2013.

[J6] J. Chen, C. Richard, and P. Honeine. Nonlinear unmixing of hyperspectral data based on a linear-mixture/nonlinear-fluctuation model. *IEEE Transactions on Signal Processing*, 61(2) :480–492, January 15 2013.

[J7] C. Saidé, R. Lengellé, P. Honeine, and R. Achkar. Online kernel adaptive algorithms with dictionary adaptation for mimo models. *IEEE Signal Processing Letters*, 20(5) :535–538, May 2013.

[J8] P. Honeine, Z. Noumir, and C. Richard. Multiclass classification machines with the complexity of a single binary classifier. *Signal Processing*, 93(5) :1013–1026, May 2013.

[J9] P. Honeine, C. Richard, and N. H. Nguyen. Approches géométriques pour l'estimation des fractions d'abondance en traitement de données hyperspectales. extensions aux modèles de mélange non-linéaires. *Traitement du signal*, 30(1-2) :61–86, 2013.

[J10] N. H. Nguyen, J. Chen, C. Richard, P. Honeine, and C. Theys. Supervised nonlinear unmixing of hyperspectral images using a pre-image method. In *New Concepts in Imaging : Optical and Statistical Models, In Eds. D. Mary, C. Theys, and C. Aime*, volume 59 of *EAS Publications Series*, pages 417–437. EDP Sciences, 2013.

2012

[J11] P. Honeine. Online kernel principal component analysis : a reduced-order model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9) :1814–1826, September 2012.

[J12] P. Honeine and C. Richard. Geometric unmixing of large hyperspectral images : a barycentric coordinate approach. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6) :2185–2195, June 2012.

[J13] P. Borgnat, P. Flandrin, C. Richard, A. Ferrari, H. Amoud, and P. Honeine. Time-frequency learning machines for nonstationarity detection using surrogates. In *Advances in Machine Learning and Data Mining for Astronomy, In Eds. M. Way, J. Scargle, K. Ali, and A. Srivastava*, Data Mining and Knowledge Discovery series, chapter 22, pages 487–503. Chapman and Hall / CRC Press (Taylor and Francis), April 2012.

2011

[J14] P. Honeine and C. Richard. A closed-form solution for the pre-image problem in kernel-based machines. *Journal of Signal Processing Systems*, 65(3) :289–299, December 2011.

[J15] J. Chen, C. Richard, J. C. M. Bermudez, and P. Honeine. Non-negative least-mean-square algorithm. *IEEE Transactions on Signal Processing*, 59(11) :5225–5235, November 2011.

[J16] P. Honeine and C. Richard. Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine*, 28(2) :77–88, March 2011.

[J17] P. Flandrin, C. Richard, P.-O. Amblard, P. Borgnat, P. Honeine, H. Amoud, A. Ferrari, J. Xiao, A. Moghtaderi, and P. Ramirez-Cobo. Stationnarité relative et approches connexes. *Traitement du signal*, 28(6), 2011.

2010

[J18] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao. Testing stationarity with surrogates : A time-frequency approach. *IEEE Transactions on Signal Processing*, 58(7) :3459–3470, July 2010.

[J19] P. Honeine, C. Richard, H. Snoussi, J. C. M. Bermudez, and J. Chen. A decentralized approach for nonlinear prediction of time series data in sensor networks. *Journal on Wireless Communications and Networking*, Special issue on theoretical and algorithmic foundations of wireless ad hoc and sensor networks :12 :1–12 :12, Jan. 2010.

[J20] P. Honeine, C. Richard, and P. Flandrin. Nonstationary signal analysis with time-frequency kernel machines. Handbook of Research on Machine Learning Applications and Trends : Algorithms, Methods and Techniques, In Eds E. Soria, J.D. MartÌn, R. Magdalena, M. MartÌnez, and A.J. Serrano, chapter 10, pages 223–241. Information Science Reference, IGI Global, 2010.

2009

[J21] C. Richard, J. C. M. Bermudez, and P. Honeine. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3) :1058–1067, March 2009.

2008

[J22] P. Honeine and C. Richard. Distribution temps-fréquence à paramétrisation radialement gaussienne optimisée pour la classification. *Traitement du signal*, 2008. (invited paper).

2007

[J23] P. Honeine, C. Richard, and P. Flandrin. Time-frequency learning machines. *IEEE Transactions on Signal Processing*, 55 :3930–3936, July 2007.

1.8.2 Communications dans des congrès à comité de lecture et actes

2013

[C1] P. Honeine, H. Lantéri, and C. Richard. Constrained kaczmarz's cyclic projections for unmixing hyperspectral data. In *Proc. 21th European Conference on Signal Processing*, Marrakech, Morocco, 9–13 September 2013.

[C2] P. Nader, P. Honeine, and P. Beauseroy. Intrusion detection in scada systems using one-class classification. In *Proc. 21th European Conference on Signal Processing*, Marrakech, Morocco, 9–13 September 2013.

[C3] S. Mahfouz, F. Mourad-Chehade, H. Paul, J. Farah, and H. Snoussi. Decentralized localization using fingerprinting and kernel methods in wireless sensor networks. In *Proc. 21th European Conference on Signal Processing*, Marrakech, Morocco, 9–13 September 2013.

[C4] J. Chen, C. Richard, J. C. M. Bermudez, and H. Paul. Non-stationary analysis of the convergence of the non-negative least-mean-square algorithm. In *Proc. 21th European Conference on Signal Processing*, Marrakech, Morocco, 9–13 September 2013.

[C5] N. Ghadban, P. Honeine, C. Francis, F. Mourad-Chehade, J. Farah, and M. Kallas. Mobilité d'un réseau de capteurs sans fil basée sur les méthodes à noyau. In *Actes du 24-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Brest, France, September 2013.

[C6] C. Saidé, P. Honeine, R. Lengellé, C. Richard, and R. Achkar. Adaptation en ligne d'un dictionnaire pour les méthodes à noyau. In *Actes du 24-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Brest, France, September 2013.

[C7] J. Chen, C. Richard, J. C. M. Bermudez, and P. Honeine. Identification en ligne avec régularisation l1. algorithme et analyse de convergence en environnement non-stationnaire. In *Actes du 24-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Brest, France, September 2013.

[C8] P. Honeine and H. Lantéri. Constrained reflect-then-combine methods for unmixing hyperspectral data. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Gainesville, Florida, USA, 25 - 28 June 2013.

[C9] J. Chen, C. Richard, and P. Honeine. Estimating abundance fractions of materials in hyperspectral images by fitting a post-nonlinear mixing model. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Gainesville, Florida, USA, 25 - 28 June 2013.

[C10] S. Mahfouz, F. Mourad-Chehade, P. Honeine, J. Farah, and H. Snoussi. Kernel-based localization using fingerprinting in wireless sensor networks. In *Proc. 14th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Darmstadt, Germany, 16 - 19 June 2013.

[C11] J. Chen, C. Richard, A. Ferrari, and P. Honeine. Nonlinear unmixing of hyperspectral data with partially linear least-squares support vector regression. In *Proc. 38th IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, May 2013.

[C12] N. Ghadban, P. Honeine, C. Francis, F. Mourad-Chehade, J. Farah, and M. Kallas. Estimation locale d'un champ de diffusion par modèles à noyaux. In *Actes de la 14-ème conférence ROADEF de la Société FranÃaise de Recherche Opérationnele et Aide à la Décision*, Troyes, France, 13–15 Février 2013.

[C13] S. Mahfouz, F. Mourad-Chehade, P. Honeine, J. Farah, and H. Snoussi. Localisation par fingerprinting et méthodes à noyaux dans les réseaux de capteurs sans fil. In *Actes de la 14-ème conférence ROADEF de la Société FranÃaise de Recherche Opérationnelle et Aide à la Décision*, Troyes, France, 13–15 Février 2013.

2012

[C14] Z. Noumir, P. Honeine, and C. Richard. Kernels for time series of exponential decay/growth processes. In *Proc. 22nd IEEE workshop on Machine Learning for Signal Processing*, Santander, Spain, 23–26 Sept. 2012.

[C15] Z. Noumir, P. Honeine, and C. Richard. Online one-class machines based on the coherence criterion. In *Proc. 20th European Conference on Signal Processing*, Bucharest, Romania, 27–31 August 2012.

[C16] Z. Noumir, P. Honeine, and C. Richard. One-class machines based on the coherence criterion. In *Proc. IEEE workshop on Statistical Signal Processing*, Ann Arbor, Michigan, USA, 5–8 August 2012.

[C17] C. Saidé, R. Lengellé, P. Honeine, C. Richard, and R. Achkar. Dictionary adaptation for online prediction of time series data with kernels. In *Proc. IEEE workshop on Statistical Signal Processing*, Ann Arbor, Michigan, USA, 5–8 August 2012.

[C18] Z. Noumir, P. Honeine, and C. Richard. On simple one-class classification methods. In *Proc. IEEE International Symposium on Information Theory*, MIT, Cambridge (MA), USA, 1–6 July 2012.

[C19] Z. Noumir, B. K. Guépié, L. Fillatre, P. Honeine, I. Nikiforov, H. Snoussi, C. Richard, P.-A. Jarrige, and F. Campan. Detection of contamination in water distribution network. In *2nd International Conference SimHydro : New trends in simulation hydroinformatics and 3D modeling*, Nice, France, 12-14 September 2012.

[C20] J. Chen, C. Richard, P. Honeine, and J.-Y. Tourneret. Prediction of rain attenuation series with discretized spectral model. In *Proc. IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany, 22 - 27 July 2012.

[C21] N. H. Nguyen, C. Richard, P. Honeine, and C. Theys. Hyperspectral image unmixing using manifold learning : methods derivations and comparative tests. In *Proc. IEEE International Geoscience and Remote Sensing Symposium*, pages 3086–3089, Munich, Germany, 22 - 27 July 2012. IEEE.

[C22] S. Deveughèle, H. Yin, L. Fillatre, P. Honeine, I. Nikiforov, C. Richard, H. Snoussi, N. Azzaoui, B. K. Guépié, and Z. Noumir. Vigires'eau. In *Proc. 10th International Conference on Hydroinformatics*, Hamburg, Germany, 14-18 July 2012.

[C23] J. Chen, C. Richard, and P. Honeine. Nonlinear unmixing of hyperspectral images based on multi-kernel learning. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Shanghai, China, 4 - 7 June 2012.

[C24] F. Mourad, P. Honeine, and H. Snoussi. Indoor localization using polar intervals in wireless sensor networks. In *Proc. 19th International Conference on Telecommunications*, pages 1–6, Jounieh, Lebanon, 23 - 25 April 2012.

[C25] M. Kallas, C. Francis, P. Honeine, H. Amoud, and C. Richard. Modeling electrocardiogram using yule-walker equations and kernel machines. In *Proc. 19th International Conference on Telecommunications*, Jounieh, Lebanon, 23 - 25 April 2012.

[C26] M. Kallas, C. Francis, L. Kanaan, D. Merheb, P. Honeine, and H. Amoud. Multi-class svm classification combined with kernel pca feature extraction of ecg signals. In *Proc. 19th International Conference on Telecommunications*, Jounieh, Lebanon, 23 - 25 April 2012.

[C27] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Prediction of time series using yule-walker equations with kernels. In *Proc. 37th IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25 - 30 March 2012.

[C28] P.-O. Amblard, O. J. Michel, C. Richard, and P. Honeine. A gaussian process regression approach for testing granger causality between time series data. In *Proc. 37th IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25 - 30 March 2012.

2011

[C29] J. Chen, C. Richard, and P. Honeine. A novel kernel-based nonlinear unmixing scheme of hyperspectral images. In *Proc. 45th Asilomar Conference on Signals, Systems, and Computers*, pages 1898–1902, Pacific Grove (CA), USA, 6–9 November 2011. IEEE.

[C30] J. Chen, C. Richard, J. C. M. Bermudez, and P. Honeine. A modified non-negative LMS algorithm and its stochastic behavior analysis. In *Proc. 45th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove (CA), USA, 6–9 November 2011.

[C31] M. Kallas, P. Honeine, C. Francis, and H. Amoud. A comparative study of pre-image techniques : The kernel autoregressive case. In *Proc. IEEE workshop on Signal Processing Systems*, pages 379–384, Beirut, Lebanon, 4–7 October 2011.

[C32] L. Kanaan, D. Merheb, M. Kallas, C. Francis, H. Amoud, and P. Honeine. Pca and kpca of ecg signals with binary svm classification. In *Proc. IEEE workshop on Signal Processing Systems*, pages 344–348, Beirut, Lebanon, 4–7 October 2011.

[C33] P. Honeine and C. Richard. Approches géométriques pour l'estimation des fractions d'abondance en traitement de données hyperspectales. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, September 2011.

[C34] Z. Noumir, P. Honeine, and C. Richard. Classification multi-classes au prix d'un classifieur binaire. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, September 2011.

[C35] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Modèle autorégressif non-linéaire à noyau. une première approche. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, September 2011.

[C36] M. Kallas, P. Honeine, H. Amoud, and C. Francis. Sur le problème de la pré-image en reconnaissance des formes avec contraintes de non-négativité. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, September 2011.

[C37] J. Chen, C. Richard, and P. Honeine. Un nouveau paradigme pour le démélange non-linéaire des images hyperspectrales. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, September 2011.

[C38] C. Richard, J. Chen, P. Honeine, and J. C. M. Bermudez. Filtrage adaptatif avec contrainte de non-négativité. principes de l'algorithme nn-LMS et modèle de convergence. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France, September 2011.

[C39] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Non-negative pre-image in machine learning for pattern recognition. In *Proc. 19th European Conference on Signal Processing*, Barcelona, Spain, 29 Aug. - 2 Sept. 2011.

[C40] J. Chen, C. Richard, H. Lantéri, C. Theys, and P. Honeine. Online system identification under non-negativity and ℓ_1-norm constraints algorithm and weight behavior analysis. In *Proc. 19th European Conference on Signal Processing*, Barcelona, Spain, 29 Aug. - 2 Sept. 2011.

[C41] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Kernel-based autoregressive modeling with a pre-image technique. In *Proc. IEEE workshop on Statistical Signal Processing*, Nice, France, 28–30 June 2011.

[C42] Z. Noumir, P. Honeine, and C. Richard. Multi-class least squares classification at binary-classification complexity. In *Proc. IEEE workshop on Statistical Signal Processing*, Nice, France, 28–30 June 2011.

[C43] J. Chen, C. Richard, H. Lantéri, C. Theys, and P. Honeine. A gradient based method for fully constrained least-squares unmixing of hyperspectral images. In *Proc. IEEE workshop on Statistical Signal Processing*, Nice, France, 28–30 June 2011.

[C44] P. Honeine, F. Mourad, M. Kallas, H. Snoussi, H. Amoud, and C. Francis. Wireless sensor networks in biomedical : body area networks. In *Proc. 7th International Workshop on Systems, Signal Processing and their*

Applications, Algeria, 09–11 May 2011.

2010

[C45] J. Chen, C. Richard, P. Honeine, and J. C. M. Bermudez. Non-negative distributed regression for data inference in wireless sensor networks. In *Proc. 44th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove (CA), USA, 7–10 November 2010.

[C46] M. Kallas, P. Honeine, C. Richard, H. Amoud, and C. Francis. Nonlinear feature extraction using kernel principal component analysis with non-negative pre-image. In *Proc. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Buenos Aires, Argentina, 31 Aug. - 4 Sept. 2010.

[C47] J. Chen, C. Richard, P. Honeine, H. Lantéri, and C. Theys. System identification under non-negativity constraints. In *Proc. 18th European Conference on Signal Processing*, Aalborg, Denmark, 23 - 27 Aug. 2010.

[C48] P. Honeine and C. Richard. A simple scheme for unmixing hyperspectral data based on the geometry of the n-dimensional simplex. In *Proc. IEEE International Geoscience and Remote Sensing Symposium*, Honolulu (Hawaii), USA, 25 - 30 July 2010.

[C49] C. Richard, P. Honeine, H. Snoussi, A. Ferrari, and C. Theys. Distributed learning with kernels in wireless sensor networks for physical phenomena modeling and tracking. In *Proc. IEEE International Geoscience and Remote Sensing Symposium*, Honolulu (Hawaii), USA, 25 - 30 July 2010.

[C50] P. Honeine and C. Richard. The angular kernel in machine learning for hyperspectral data classification. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Reykjavik, Iceland, 14 - 16 June 2010.

[C51] C. Richard, A. Ferrari, H. Amoud, P. Honeine, P. Flandrin, and P. Borgnat. Statistical hypothesis testing with time-frequency surrogates to check signal stationarity. In *Proc. 35th IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, 14 - 19 March 2010.

[C52] J. Chen, C. Richard, P. Honeine, H. Snoussi, H. Lantéri, and C. Theys. Techniques d'apprentissage non-linéaires en ligne avec contraintes de positivité. In *Actes de la VI-ème Conférence Internationale Francophone d'Automatique*, Nancy, France, 2 - 4 Juin 2010.

2009

[C53] P. Honeine and C. Richard. Solving the pre-image problem in kernel machines : a direct method. In *Proc. 19th IEEE workshop on Machine Learning for Signal Processing*, Grenoble, France, September 2009. - best paper award -.

[C54] H. Amoud, P. Honeine, C. Richard, P. Borgnat, and P. Flandrin. Time-frequency learning machines for nonstationarity detection using surrogates. In *Proc. IEEE workshop on Statistical Signal Processing*, Cardiff (Wales), UK, 31 Aug. - 3 Sept 2009.

[C55] P. Honeine, C. Richard, J. C. M. Bermudez, H. Snoussi, M. Essoloh, and F. Vincent. Functional estimation in hilbert space for distributed learning in wireless sensor networks. In *Proc. 34th IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, April 2009.

[C56] H. Amoud, C. Richard, P. Honeine, P. Flandrin, and P. Borgnat. Sur la caractérisation de non-stationnarités par la méthode des substituts. In *Actes du 22-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Dijon, France, September 2009.

[C57] P. Honeine, C. Richard, and H. Snoussi. Auto-localisation dans les réseaux de capteurs sans fil par régression de matrices de gram. In *Actes du 22-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Dijon, France, September 2009.

[C58] M. Essoloh, P. Honeine, C. Richard, and H. Snoussi. Apprentissage non-linéaire en ligne dans les réseaux de capteurs sans fil. In *Actes du 22-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Dijon, France, September 2009.

2008

[C59] P. Honeine, C. Richard, J. C. M. Bermudez, and H. Snoussi. Distributed prediction of time series data with kernels and adaptive filtering techniques in sensor networks. In *Proc. 42nd Annual ASILOMAR Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, October 2008. invited paper.

[C60] M. Essoloh, C. Richard, H. Snoussi, and P. Honeine. Distributed localization in wireless sensor networks as a pre-image problem in a reproducing kernel hilbert space. In *Proc. 16th European Conference on Signal Processing*, Lausanne, Switzerland, August 2008.

[C61] P. Honeine, M. Essoloh, C. Richard, and H. Snoussi. Distributed regression in sensor networks with a reduced-order kernel model. In *Proc. 51st IEEE GLOBECOM Global Communications Conference*, New Orleans, LA, USA, 2008.

[C62] P. Honeine, C. Richard, M. Essoloh, and H. Snoussi. Localization in sensor networks - a matrix regression approach. In *Proc. 5th IEEE Sensor Array and Multichannel Signal Processing Workshop*, Darmstadt, Germany, July 2008.

2007

[C63] P. Honeine and C. Richard. Signal-dependent time-frequency representations for classification using a radially gaussian kernel and the alignment criterion. In *Proc. IEEE workshop on Statistical Signal Processing*, Madison, WI, USA, August 2007.

[C64] P. Honeine, C. Richard, and J. C. M. Bermudez. On-line nonlinear sparse approximation of functions. In *Proc. IEEE International Symposium on Information Theory*, pages 956–960, Nice, France, June 2007.

[C65] P. Honeine, C. Richard, and J. C. M. Bermudez. Modélisation parcimonieuse non linéaire en ligne par une méthode à noyau reproduisant et un critère de cohérence. In *Actes du XXI-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Troyes, France, September 2007.

[C66] P. Honeine and C. Richard. Distribution temps-fréquence à noyau radialement gaussien : optimisation pour la classification par le critère d'alignement noyau-cible. In *Actes du XXI-ème Colloque GRETSI sur le*

Traitement du Signal et des Images, Troyes, France, September 2007.

2006

[C67] P. Honeine, C. Richard, P. Flandrin, and J.-B. Pothin. Optimal selection of time-frequency representations for signal classification : A kernel-target alignment approach. In *Proc. 31st IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006.

2005

[C68] P. Honeine, C. Richard, and P. Flandrin. Reconnaissance des formes par méthodes à noyau dans le domaine temps-fréquence. In *Actes du XX-ème Colloque GRETSI sur le Traitement du Signal et des Images*, pages 969–972, Louvain-la-Neuve, Belgium, 2005.

1.8.3 Workshop (avec actes) et journées thématiques GdR CNRS

2012

[W1] H. Yin, F. Campan, B. K. Guépié, Z. Noumir, L. Fillatre, P. Honeine, I. Nikiforov, C. Richard, H. Snoussi, P.-A. Jarrige, and C. Morio. Vigires'eau : Surveiller un réseau de distribution d'eau potable. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'12), (ANR - CSOSG)*, pages 1–8, Troyes, France, 2012.

2011

[W2] F. Septier, Y. Delignon, P. Armand, H. Snoussi, and P. Honeine. Malice : Localisation de sources polluantes depuis un réseau de capteurs. In *4-ème Workshop du Groupement d'Intérêt Scientifique : Surveillance, Sûreté, Sécurité des Grands Systèmes (GIS-3SGS'11)*, page 1, Valenciennes, France, 12–13 octobre 2011.

[W3] N. Khodor, H. Amoud, M. Kallas, P. Honeine, and C. Francis. Le problème de pré-image dans la reconnaissance des formes. In *Proc. 1st International Conference on Advances in Biomedical Engineering*, pages 1–2, Tripoli, Lebanon, 6–8 July 2011.

[W4] P. Honeine. Problème de pré-image en apprentissage et reconnaissance des formes. applications en traitement du signal et des images. In *Journée apprentissage et reconnaissance des formes en signal et images, journées thématiques au GdR ISIS*, 7 avril 2011.

[W5] L. Fillatre, P. Honeine, I. Nikiforov, C. Richard, H. Snoussi, N. Azzaoui, B. K. Guépié, Z. Noumir, S. Deveughèle, and H. Yin. Vigires'eau : Surveillance en temps réel de la qualité de l'eau potable d'un réseau de distribution en vue de la détection d'intrusions. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'11), (ANR - CSOSG)*, pages 1–7, Troyes, France, 2011.

2010

[W6] P. Flandrin, P. Borgnat, A. Moghtaderi, C. Richard, P. Honeine, H. Amoud, P.-O. Amblard, and P. Ramirez-Cobo. Starac : Stationnarité relative et approches connexes. In *Grand Colloque STIC 2010*, Paris - Cité des sciences et de l'industrie, France, 5-7 janvier 2010.

[W7] M. Kallas, P. Honeine, H. Amoud, C. Francis, and C. Richard. Constrained pattern recognition with non-linear principal component analysis. In *Journées Scientifiques à l'Ecole Doctorale de Sciences et Technologie*, Liban, 8-9 décembre 2010.

[W8] L. Fillatre, P. Honeine, I. Nikiforov, C. Richard, H. Snoussi, and N. Azzaoui. Vigires'eau : Surveillance en temps réel de la qualité de l'eau potable d'un réseau de distribution en vue de la détection d'intrusions. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'10), (ANR - CSOSG)*, pages 1–7, Troyes, France, 26-27 janvier 2010.

2008

[W9] C. Richard, P. Honeine, H. Snoussi, M. Essoloh, and J. C. M. Bermudez. Distributed learning in wireless sensor networks. In *5th Workshop on Sensor Networks (CNRS RECAP Sensor and Self-Organized Networks)*, 13 - 14 november 2008.

[W10] P. Honeine and C. Richard. Sur l'usage de critères de représentation parcimonieuse pour la rdf par méthodes à noyau. In *Journée représentations parcimonieuses, journées thématiques au GdR ISIS*, 17 avril 2008.

2007

[W11] C. Richard and P. Honeine. Filtrage adaptatif non linéaire par méthode à noyau. In *Journée signal, reconnaissance des formes et machines à noyaux, journées thématiques au GdR ISIS*, 8 juin 2007.

1.8.4 Brevet

2010

[B1] H. Snoussi, C. Richard, and P. Honeine. System and method for locating a target using a transceiver array (fr : Système et procédé de localisation de cible par un réseau d'émetteurs/récepteurs), 2010.

1.8.5 Rapports de recherche

2007

[R1] P. Honeine. *Méthodes à noyau pour l'analyse et la décision en environnement non-stationnaire.* PhD thesis, mémoire de thèse de doctorat en Optimisation et Sûreté des Systèmes, Ecole doctoral SSTO - UTT, Troyes, France, 2007.

2003

[R2] P. Honeine. Théorie de l'information pour l'analyse du typage sonore de véhicules,. Master's thesis, mémoire de DEA, UTT(LM2S) – PSA Peugeot Citroen (centre DRIA/SARA/EMSA/PEFH), Troyes, France, 2003.

Deuxième partie

Synthèse des travaux scientifiques

Table des cadres

Glossaire

Notation	Signification
Espaces, ensembles :	
$\mathbb{N}$	Ensemble des entiers naturels
$\mathbb{R}$	Ensemble des nombres réels
$\mathbb{R}_+$	Ensemble des nombres réels positifs
$\mathbb{R}^d$	Espace vectoriel réel de dimension d
$\mathbb{X}$	Espace des échantillons ou d'observation, $\mathbb{X} \subset \mathbb{R}^d$
$\mathbb{Y}$	Espace des réponses, $\mathbb{Y} \subset \mathbb{R}$ (réponse scalaire) ou $\mathbb{Y} \subset \mathbb{R}^\ell$ (réponse vectorielle)
$\mathbb{H}$	Espace de Hilbert à noyau reproduisant, de fonctions réelles définies sur $\mathbb{X}$
$\lvert\mathcal{V}\rvert$	Taille (i.e., cardinalité) de l'ensemble $\mathcal{V}$
$\dim(\mathbb{X})$	Dimension de l'espace $\mathbb{X}$
Notations :	
$x, y, \alpha, \ldots$	Scalaires en minuscule ou majuscule mais sans gras
$\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\alpha}, \ldots$	Vecteurs (colonnes) en gras et en minuscule
$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{A}, \ldots$	Matrices notées en gras et en majuscule
$[\boldsymbol{w}]_i$	i-ème élément du vecteur $\boldsymbol{w}$
$[\boldsymbol{W}]_{i,j}$	(i,j)-ème élément de la matrice $\boldsymbol{W}$
Statistiques :	
$P(\cdot)$	Distribution de probabilité qui génère les échantillons
$\mathbb{E}[\cdot]$	Espérance mathématique associée à la distribution $P(\cdot)$, avec $\mathbb{E}[\kappa(\boldsymbol{x}, \cdot)] = \int_{\mathbb{X}} \kappa(\boldsymbol{x}, \cdot)\, dP(\boldsymbol{x})$
$\mu_\infty(\cdot)$	Espérance mathématique associée à la distribution des fonctions noyau, *i.e.,* $\mathbb{E}[\kappa(\boldsymbol{x}, \cdot)]$
$\mu_n(\cdot)$	Estimation de $\mu_\infty(\cdot)$ à partir de l'ensemble d'apprentissage : $\frac{1}{n}\sum_{i=1}^n \kappa(\boldsymbol{x}_i, \cdot)$

Algèbre linéaire :

$\boldsymbol{W}^\top$	Matrice transposée de la matrice $\boldsymbol{W}$
$\boldsymbol{W}^{-1}$	Matrice inverse de la matrice $\boldsymbol{W}$
$\langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle$	Produit scalaire, $\langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle = \boldsymbol{w}_i^\top \boldsymbol{w}_j$
$\mathrm{diag}(\boldsymbol{w})$	Matrice diagonale formée par les éléments du vecteur $\boldsymbol{w}$
$\mathrm{diag}(\boldsymbol{W})$	Vecteur formé par les éléments diagonaux de la matrice $\boldsymbol{W}$
$\det \boldsymbol{W}$	Déterminant de la matrice $\boldsymbol{W}$
$\mathrm{tr}(\boldsymbol{W})$	Trace de la matrice $\boldsymbol{W}$
$\|\boldsymbol{w}\|$	Norme euclidienne du vecteur $\boldsymbol{w}$, avec $\|\boldsymbol{w}\|^2 = \sum_i [\boldsymbol{w}]_i^2$
$\|\boldsymbol{w}\|_1$	Norme vectorielle ℓ_1 du vecteur $\boldsymbol{w}$, avec $\|\boldsymbol{w}\|_1 = \sum_i \|[\boldsymbol{w}]_i\|$
$\|\boldsymbol{W}\|_{1,1}$	Somme des normes vectorielles ℓ_1 des colonnes de la matrice $\boldsymbol{W}$
$\|\boldsymbol{W}\|_F$	Norme de Frobenius de $\boldsymbol{W}$, avec $\|\boldsymbol{W}\|_F^2 = \sum_{i,j} [\boldsymbol{W}]_{i,j}^2 = \mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{W})$
$\mathrm{LT}(\boldsymbol{W})$	Matrice triangulaire inférieure formée par les éléments sous la diagonale de $\boldsymbol{W}$
$\mathbf{1}, \mathbf{1}_n$	Vecteur unité de taille appropriée, avec $\mathbf{1}_n$ de taille $(n \times 1)$
$\mathbf{I}, \mathbf{I}_{n,n}$	Matrice identité de taille appropriée, avec $\mathbf{I}_{n,n}$ de taille $(n \times n)$
$\mathbf{0}, \mathbf{0}_n$	Vecteur nul de taille appropriée, avec $\mathbf{0}_n$ de taille $(n \times 1)$
$\mathbf{0}, \mathbf{0}_{n,n}$	Matrice nulle de taille appropriée, avec $\mathbf{0}_{n,n}$ de taille $(n \times n)$

Apprentissage :

n	Nombre de données d'apprentissage
N	Nombre de données test/validation
$\boldsymbol{x}_i$	Observation, $\boldsymbol{x}_i \in \mathbb{X}$
$\boldsymbol{X}$	Matrice $[\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \cdots \ \boldsymbol{x}_n]$
y_i	Sortie désirée (scalaire) associée à l'observation $\boldsymbol{x}_i$, $y_i \in \mathbb{Y} \subset \mathbb{R}$
$\boldsymbol{y}_i$	Sortie désirée (vectorielle) associée à l'observation $\boldsymbol{x}_i$, $\boldsymbol{y}_i \in \mathbb{Y} \subset \mathbb{R}^\ell$
$\boldsymbol{Y}$	Matrice $[\boldsymbol{y}_1 \ \ \boldsymbol{y}_2 \ \ \cdots \ \ \boldsymbol{y}_n]$
$\mathcal{C}^{(k)}$	k-ème classe en classification multi-classes
$\boldsymbol{y}^{(k)}$	Etiquette vectorielle associée au k-ème sous-problème OvA, *i.e.,* $[\boldsymbol{y}^{(k)}]_i = [\boldsymbol{y}_i]_k$

Paramètres :

η, η_1, η_t	Paramètres de régularisation, pas de convergence
ν, ν_0, ν_1	Paramètres de seuil
σ	Paramètre de largeur de bande du noyau Gaussien $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \mathrm{e}^{-\frac{1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}$
$\overline{m}, \overline{\lvert\mathcal{V}\rvert}$	Valeurs moyennes

Méthodes à noyaux :

$\phi(\cdot)$	Transformation de $\mathbb{X}$ dans $\mathbb{H}$
$\psi(\cdot)$	Fonction de décision à valeur réelle définie sur $\mathbb{X}$, $\psi(\cdot) \in \mathbb{H}$
$\boldsymbol{\psi}(\cdot)$	Fonction vectorielle (*i.e.,* uplet de fonctions) d'éléments $\psi^{(1)}(\cdot), \psi^{(2)}(\cdot), \ldots$
$\kappa(\cdot, \cdot)$	Fonction noyau défini positif, de $\mathbb{X} \times \mathbb{X}$ dans $\mathbb{R}$

$\langle \psi_i(\cdot), \psi_j(\cdot) \rangle_{\mathbb{H}}$	Produit scalaire entre les deux fonctions $\psi_i(\cdot)$ et $\psi_j(\cdot)$ de l'espace $\mathbb{H}$
$\|\psi(\cdot)\|_{\mathbb{H}}$	Norme de $\psi(\cdot)$ dans l'espace $\mathbb{H}$
$\mathcal{C}(\cdot, \cdot)$	Fonction coût mesurant l'écart entre ses deux arguments
$\mathcal{R}(\cdot)$	Fonction de régularisation, monotone croissante sur $\mathbb{R}_+$
$\mathcal{J}(\cdot)$	Fonction coût (risque empirique) à minimiser
$\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x})$	Gradient de la fonction $\mathcal{J}(\cdot)$ par rapport à $\boldsymbol{x}$

Dictionnaire :

$\mathcal{D}$	Dictionnaire, *i.e.,* sous-ensemble d'échantillons : $\{\boldsymbol{x}_{\omega_1}, \ldots, \boldsymbol{x}_{\omega_m}\} \subset \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$
m	Taille (i.e., cardinalité) du dictionnaire : $m = \lvert\mathcal{D}\rvert$
$\boldsymbol{X}_{\mathcal{D}}$	Matrice $[\boldsymbol{x}_{\omega_1}\ \boldsymbol{x}_{\omega_2}\ \cdots\ \boldsymbol{x}_{\omega_m}]$
$\boldsymbol{K}_{\mathcal{D}}$	Matrice de noyau associée aux éléments de $\mathcal{D}$: $[\boldsymbol{K}_{\mathcal{D}}]_{i,j} = \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$
$\boldsymbol{\kappa}_{\mathcal{D}}(\psi(\cdot))$	Vecteur de l'évaluation de $\psi(\cdot)$ en chaque élément de $\mathcal{D}$: $[\boldsymbol{\kappa}_{\mathcal{D}}(\psi(\cdot))]_j = \psi(\boldsymbol{x}_{\omega_j})$
$\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_i)$	Vecteur de taille $(m \times 1)$ d'éléments $[\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_i)]_j = \kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_i)$
$\mathrm{vol}(\mathcal{D})$	Volume du simplexe dont les sommets sont les éléments de $\mathcal{D}$
$\mathbf{P}_{\mathcal{D}}$	Opérateur de projection dans l'espace engendré par les fonctions noyau de $\mathcal{D}$

$\max, \min$ **:**

$\max\{x; 0\}$	Partie positive du scalaire x
$\max_i u_i$	Valeur maximale parmi les scalaires u_i
$\operatorname{argmax}_u \mathcal{J}(u)$	Valeur de u qui donne le maximum de $\mathcal{J}(u)$
$\min_i u_i$	Valeur minimale parmi les scalaires u_i
$\operatorname{argmin}_u \mathcal{J}(u)$	Valeur de u qui donne le minimum de $\mathcal{J}(u)$

Abbréviation	Signification / Acronyme de
2D	Bidimensionnel
2D	Tridimensionnel
ACP	Analyse en composante principale
APA	Algorithme de projection affine ; *Affine Projection Algorithm*
AR	Modèle autorégressif
AVIRIS	*Airborne Visible/Infrared Imaging Spectrometer*
DAG	Graphe orienté acyclique ; *Directed Acyclic Graph*
ICE	Extraction de composants purs ; *Iterated Constrained Endmembers*
i.e.	« c'est à dire », acronyme de *id est*
LLE	Approche localement linéaire ; *Locally Linear Embedding*
LMS	Algorithme du gradient stochastique ; *Least Mean Squares*
LS-SVM	Machines à moindres carrés ; *Least-Squares SVM*
LS, LSM	Machines à moindres carrés ; *Least-Squares Machines*
MDS	Positionnement multidimentionnel ; *MultiDimensional Scaling*
N-Findr	Algorithme d'extraction de composants purs
NMF	Factorisation en matrices non négatives ; *Nonnegative Matrix Factorization*
OSP	Extraction de composants purs ; *Orthogonal Subspace Projection*
OvA	Stratégie un-contre-tous ; *One versus All the rest classes*
OvO	Stratégie un-contre-un ; *One versus One*
RKHS	Espace de Hilbert à noyau reproduisant ; *Reproducing Kernel Hilbert Space*
RLS	Algorithme des moindres carrés récursif ; *Recursive Least-Squares*
RLSC	Machines à moindres carrés ; *Regularized Least-Squares Classification*
RSSI	Mesures de portée inter-capteurs ; *Received Signal Strength Indication*
SGA	Extraction de composants purs ; *Simplex Growing Algorithm*
SVM	Machines à vecteurs de support ; *Support Vector Machines*
VCA	Extraction de composants purs ; *Vertex Component Analysis*

Habilité à être Dirigé par la Recherche.
[anonyme]

2

Résumé des activités de recherche

Sommaire

Le présent chapitre expose d'une manière synthétique l'activité de recherche que j'ai menée après ma thèse de doctorat. Mes travaux se placent principalement dans le cadre de l'apprentissage statistique en traitement du signal. Je me suis intéressé essentiellement aux méthodes à noyaux en reconnaissance des formes, régression, classification, et détection. Derrière la diversité des problèmes traités et des solutions proposées, une vision cohérente du traitement du signal se dessine, généralisant le principe de « l'apprentissage en pénurie d'information statistique *a priori* » que j'ai tant défendu pendant ma thèse de doctorat.

Contexte

Mes activités de recherche se sont déroulées en totalité à l'Université de technologie de Troyes, en thèse de doctorat, puis en post-doctorat, et actuellement en Maître de Conférences. Bien qu'une telle immobilité puisse être un handicap pour la recherche, les projets de recherche et les collaborations que j'ai pu mener montrent le contraire. L'hétérogénéité des projets de recherche auquel j'ai participé m'a permis de développer et de consolider divers axes de recherche. Il s'agit aussi bien de projets de recherche fondamentale que de projets avec des partenaires industriels. Nous retrouvons ainsi des projets selon plusieurs niveaux.

– Au niveau UTT-UTC, j'ai été coordinateur d'un projet d'Abondement Carnot (2009–2012), sur le traitement de l'information dans les réseaux de capteurs sans fil. Ce projet a permis

de développer et de consolider la collaboration entre trois équipes, deux de l'UTT, LM2S et ERA, et une équipe du laboratoire Heudiasyc de l'UTC. D'autres projets ont poursuivi cette collaboration, dont le projet « Mobiloc ».
- Au niveau régional, je suis le coordinateur du projet WiDiD « Wireless Diffusion Detection » (2012–2015) dans le cadre de la plateforme CAPSEC sur les réseaux de capteurs.
- Au niveau national, j'ai participé très activement à 5 projets ANR au cours de ces 5 dernières années. Le dernier en date, HYPANEMA (2012–2015), associe 4 laboratoires de recherche. Je suis coordinateur de l'équipe UTT de ce projet.
- Dans le cadre international, j'ai été coordinateur d'un projet (2011–2012) du programme franco-libanais « CEDRE ». Dans ce cadre, j'ai co-dirigé la thèse de Mlle Maya Kallas, qui est actuellement Maître de Conférences à l'Université de Lorraine.

2.1 Préambule général

L'apprentissage statistique consiste à déterminer une fonction qui traduit le mieux possible une relation entre des échantillons successifs recueillies sur un système, ou encore entre ses entrées et sorties, à partir d'un ensemble d'apprentissage [Vapnik, 1995]. Ce problème est mal-posé, puisqu'il existe une infinité de fonctions continues qui vérifient les conditions discrètes induites par les échantillons d'apprentissage. Ceci est par exemple le cas des problèmes de régression, où l'on cherche une fonction passant par certains points tandis qu'il en existe une infinité. Il est alors nécessaire de réduire l'espace fonctionnel dans lequel la solution est recherchée, en y incluant une certaine connaissance *a priori* du comportement du système. Il peut s'agir de contraintes physiques ou encore d'une régularisation. La théorie de la régularisation introduite par Tikhonov et Arsenin dans [Tikhonov and Arsenin, 1977] propose une solution élégante à ce problème : restreindre la recherche à un espace de fonctions régulières.

Un type d'espace fonctionnel particulier est l'espace de Hilbert à noyau reproduisant, un concept introduit par Aronszajn dans [Aronszajn, 1950]. Ses propriétés sont exploitées par le Théorème de Représentation, initialement proposé pour les problèmes de régression par Kimeldorf et Wahba dans [Kimeldorf and Wahba, 1971, Wahba, 1990] et récemment généralisé à d'autres problèmes d'apprentissage par Schölkopf *et coll.* dans [Schölkopf et al., 2000]. La simplicité des méthodes dites à noyaux est principalement due au coup du noyau, plus communément désigné par *kernel trick* en anglais, qui permet de transformer des algorithmes linéaires en des méthodes non-linéaires sans surcoût calculatoire considérable. Cette notion de non-linéarité par usage de noyau a été proposée par Aizerman *et coll.* dans [Aizerman et al., 1964] et renforcée par Vapnik dans [Vapnik, 1995] avec la théorie de l'apprentissage statistique.

2.1.1 Apprentissage statistique

Dans le cadre de l'apprentissage supervisé, on cherche la fonction $\psi(\cdot)$ qui détermine la relation entre un compact $\mathbb{X}$ de $\mathbb{R}^d$ dit espace des échantillons ou d'observation, et un compact $\mathbb{Y}$ de $\mathbb{R}$ dit espace des réponses. La distribution de probabilité, définie pour tout couple $(\boldsymbol{x}, y) \in \mathbb{X} \times \mathbb{Y}$, est en effet inconnue. Celle-ci n'est connue qu'à partir d'un ensemble fini de réalisations, appelé ensemble d'apprentissage, que l'on note $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ avec $(\boldsymbol{x}_k, y_k) \in \mathbb{X} \times \mathbb{Y}$. En considérant une fonction coût $\mathcal{C}(\cdot, \cdot)$ mesurant l'erreur commise entre la sortie désirée y et la sortie estimée $\psi(\boldsymbol{x})$, le problème d'optimisation se traduit par la minimisation du risque empirique régularisé, selon

$$\psi(\cdot) = \underset{\psi(\cdot) \in \mathbb{H}}{\operatorname{argmin}} \sum_{i=1}^{n} \mathcal{C}(\psi(\boldsymbol{x}_i), y_i) + \eta\, \mathcal{R}(\|\psi(\cdot)\|_{\mathbb{H}}^2),$$

où $\mathbb{H}$ est l'espace de fonctions candidates, et η contrôle le compromis entre le risque empirique qui mesure l'adéquation entre les sorties estimées et les sorties désirées (premier terme), et la pénalisation (second terme) permettant d'obtenir des solutions plus régulières. Sans celle-ci, le problème serait mal-posé puisqu'il existerait une infinité de fonctions qui minimisent le premier terme. Un choix particulier de l'espace fonctionnel est l'espace de Hilbert à noyau reproduisant. Dans ce qui suit, on rappelle le concept de noyaux reproduisants et des espaces de Hilbert associés, avant de les exploiter au travers du Théorème de Représentation.

2.1.2 Noyau reproduisant et espace de Hilbert associé

Un noyau désigne une fonction $\kappa(\cdot,\cdot)$ de $\mathbb{X}\times\mathbb{X}$ dans $\mathbb{R}$, à symétrie Hermitienne, c'est à dire telle que $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\kappa(\boldsymbol{x}_j,\boldsymbol{x}_i)$ pour tout $\boldsymbol{x}_i,\boldsymbol{x}_j\in\mathbb{X}$. Un noyau $\kappa(\cdot,\cdot)$ est dit défini positif sur $\mathbb{X}$ s'il vérifie $\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\,a_j\,\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)\geq 0$, pour tout $n\in\mathbb{N}$, $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\in\mathbb{X}$ et $a_1,\ldots,a_n\in\mathbb{R}$. Un noyau $\kappa(\cdot,\cdot)$ est dit noyau reproduisant d'un espace de Hilbert $\mathbb{H}$ (RKHS, acronyme de Reproducing Kernel Hilbert Space), sous réserve que celui-ci en admette un, si et seulement si la fonction $\kappa(\boldsymbol{x}_i,\cdot)$ appartient à $\mathbb{H}$, quel que soit $\boldsymbol{x}_i\in\mathbb{X}$; et on a $\psi(\boldsymbol{x}_j)=\langle\psi(\cdot),\kappa(\boldsymbol{x}_j,\cdot)\rangle_{\mathbb{H}}$ pour tout $\boldsymbol{x}_j\in\mathbb{X}$ et $\psi(\cdot)\in\mathbb{H}$. Dans cette expression, $\langle\cdot\ ,\cdot\rangle_{\mathbb{H}}$ désigne le produit scalaire associé à l'espace en question, et on désigne par $\|\cdot\|_{\mathbb{H}}$ la norme correspondante. En combinant ces deux propriétés, on retrouve le coup du noyau, à savoir $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\kappa(\boldsymbol{x}_i,\cdot),\kappa(\boldsymbol{x}_j,\cdot)\rangle_{\mathbb{H}}$. Cette propriété fondamentale stipule que tout noyau reproduisant $\kappa(\cdot,\cdot)$ d'un espace de Hilbert $\mathbb{H}$ est un produit scalaire dans cet espace. Le théorème de Moore-Aronszajn [Aronszajn, 1950] établit qu'à tout noyau défini positif, il correspond un RKHS unique, et réciproquement. Par abus de langage, on remplace dans la suite la dénomination de noyau défini positif par noyau reproduisant.

Les méthodes à noyaux sont pour la plupart issues d'algorithmes linéaires auxquels on a pu appliquer le résultat clé qu'est le coup du noyau [Aizerman et al., 1964]. Le coup du noyau permet de transformer des algorithmes linéaires en des méthodes non-linéaires sans surcoût calculatoire considérable, sous réserve que ceux-ci puissent s'exprimer uniquement par des produits scalaires entre les données. Cette notion de non-linéarité par usage de noyau a été proposée par Aizerman *et coll.* dans [Aizerman et al., 1964] dans le cadre d'un problème de classification, et renforcée par Vapnik dans [Vapnik, 1995] avec le théorie de l'apprentissage statistique dans un contexte plus général de classification et régression.

2.1.3 Théorème de Représentation

Le coup du noyau offre une interprétation du noyau reproduisant en tant que produit scalaire et permet d'élaborer des méthodes non linéaires à partir d'algorithmes linéaires. Pour que ce principe soit opérationnel, il nécessite souvent d'être associé au Théorème de Représentation. Ce dernier, à usage multidisciplinaire aujourd'hui, est issu des travaux précurseurs de Kimeldorf et Wahba dans le domaine de la théorie de l'approximation [Kimeldorf and Wahba, 1971, Wahba, 1990]. Plus récemment, il a été repris dans le cadre de la résolution de problèmes inverses [Kurková, 2004], ainsi qu'en théorie de l'apprentissage [Cucker and Smale, 2002]. La formulation suivante du Théorème de Représentation est principalement due à Schölkopf *et coll.* dans [Schölkopf et al., 2000].

Théorème de Représentation. *Soient $\mathbb{X}$ un compact, $\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_n,y_n)\}$ un ensemble d'apprentissage donné avec $\boldsymbol{x}_i\in\mathbb{X}$ l'ensemble des échantillons et $y_i\in\mathbb{R}$ l'ensemble des sorties désirées, $\mathcal{C}$ une fonction coût arbitraire et $\mathcal{R}(\cdot)$ une fonction monotone croissante sur $\mathbb{R}+$. Soit $\mathbb{H}$ un espace de Hilbert induit par le noyau $\kappa(\cdot,\cdot)$ défini positif sur $\mathbb{X}$. Toute fonction $\psi(\cdot)\in\mathbb{H}$ minimisant la fonctionnelle de risque régularisée*

$$\sum_{i=1}^{n}\mathcal{C}(\psi(\boldsymbol{x}_i),y_i)+\eta\,\mathcal{R}(\|\psi(\cdot)\|_{\mathbb{H}}^2),$$

peut s'écrire sous la forme

$$\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \, \kappa(\boldsymbol{x}_j, \cdot). \qquad \texttt{(Th.Rep)}$$

Démonstration. Ce résultat est démontré en notant que toute fonction $\psi(\cdot)$ de $\mathbb{H}$ se décompose selon $\psi(\cdot) = \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \cdot) + \psi^{\perp}(\cdot)$, avec $\langle \psi^{\perp}(\cdot), \kappa(\boldsymbol{x}_i, \cdot) \rangle_{\mathbb{H}} = 0$ pour tout $i = 1, \ldots, n$. D'une part, la valeur de $\psi(\boldsymbol{x}_j)$ n'est pas affectée par $\psi^{\perp}(\cdot)$, pour $j = 1, \ldots, n$, puisque $\psi(\boldsymbol{x}_j) = \langle \psi(\cdot), \kappa(\boldsymbol{x}_j, \cdot) \rangle_{\mathbb{H}}$. D'autre part, comme $\mathcal{R}(\cdot)$ est une fonction monotone croissante, alors $\mathcal{R}(\|\psi(\cdot)\|_{\mathbb{H}}^2) = \mathcal{R}(\| \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \cdot) + \psi^{\perp}(\cdot) \|_{\mathbb{H}}^2) \geq \mathcal{R}(\| \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \cdot) \|_{\mathbb{H}}^2)$ par le théorème de Pythagore. Par conséquent, une valeur nulle de $\psi^{\perp}(\cdot)$ minimise le terme de régularisation, sans affecter la fidélité du modèle mesurée par le risque empirique. ■

L'importance de ce théorème réside dans l'existence d'une solution unique à une fonctionnelle de risque régularisée, celle-ci pouvant s'exprimer comme un développement en série fini de fonctions noyau. La minimisation de la fonction coût ci-dessus se ramène à un problème d'optimisation à n dimensions, celui de la détermination des coefficients optimaux $\alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R}$. Pour fixer les idées, on rappelle le problème de moindres carrés correspondant à la minimisation du risque empirique $\sum_{i=1}^{n} |y_i - \psi(\boldsymbol{x}_i)|^2 + \eta \, \|\psi(\cdot)\|_{\mathbb{H}}^2$, où une pénalisation quadratique est utilisée. En reprenant la forme (`Th.Rep`) de la solution, on obtient le problème d'optimisation $\min_{\boldsymbol{\alpha}} \|\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha}\|^2 + \eta \, \boldsymbol{\alpha}^{\top} \boldsymbol{K} \boldsymbol{\alpha}$, où $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ \cdots \ \alpha_n]^{\top}$, $\boldsymbol{y} = [y_1 \ y_2 \ \cdots \ y_n]^{\top}$ et $\boldsymbol{K}$ est la matrice de Gram d'éléments $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ pour $i, j = 1, 2, \ldots, n$. La solution optimale est alors obtenue par la résolution d'un système linéaire de n équations à n inconnues, avec $(\boldsymbol{K} + \eta \, \mathbf{I}) \, \boldsymbol{\alpha} = \boldsymbol{y}$, où $\mathbf{I}$ est la matrice identité de taille $(n \times n)$. Ceci conduit à $\psi(\boldsymbol{x}) = \boldsymbol{y}^{\top} (\boldsymbol{K} + \eta \mathbf{I})^{-1} \boldsymbol{\kappa}(\boldsymbol{x})$, où $\boldsymbol{\kappa}(\boldsymbol{x})$ désigne le vecteur d'éléments $\kappa(\boldsymbol{x}_i, \boldsymbol{x})$, pour $i = 1, 2, \ldots, n$.

Dans le cadre de l'apprentissage non-supervisé, nous retrouvons l'analyse en composantes principale (ACP) [Jolliffe, 1986]. Elle consiste à déterminer les axes (*i.e.*, directions) principaux, en terme de la plus grande variance des échantillons projetés. Il s'agit d'estimer les vecteurs propres associés aux plus grandes valeurs propres de la matrice de covariance, selon $\boldsymbol{C}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$. Selon le formalisme des espaces de Hilbert à noyau reproduisant, l'homologue de l'ACP est désigné par ACP-à-noyaux [Schölkopf et al., 1998]. Dans ce cadre, les directions principales deviennent des fonctions principales de la forme (`Th.Rep`). Le problème se ramène à l'estimation des n coefficients de pondération α_j pour chaque fonction principale.

2.2 Défis et motivations

Le Théorème de Représentation est un théorème d'existence. Il montre que la solution optimale s'exprime par un développement en série fini de fonctions noyau, selon la forme (`Th.Rep`). Malgré son importance dans la communauté, le Théorème de Représentation soulève plusieurs problèmes à traiter, comme résumé par les points suivants et détaillé dans la section suivante :

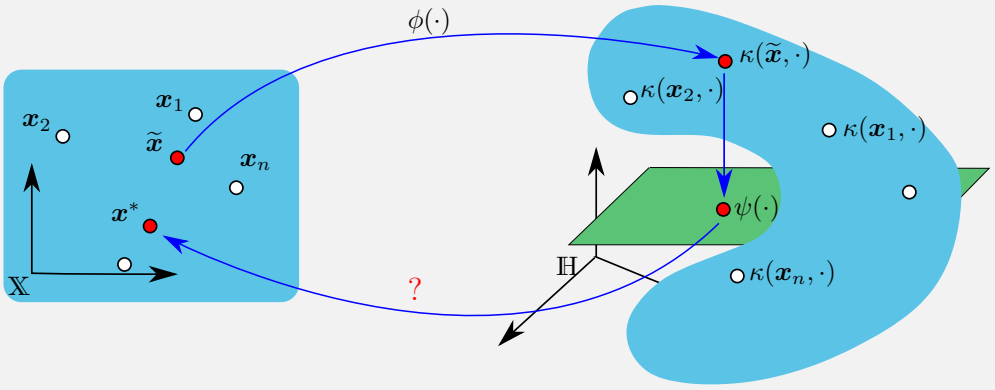
- La caractéristique ainsi obtenue $\psi(\cdot)$ appartient à l'espace de Hilbert à noyau reproduisant $\mathbb{H}$, un espace accessible dans un seul sens, par la transformation non linéaire implicite au noyau utilisé $\kappa(\cdot, \cdot)$. En d'autres termes, cette transformation permet d'injecter de nouveaux échantillons dans l'espace RKHS. Le retour inverse, du RKHS à l'espace des observations, est souvent crucial dans divers problèmes. Il s'agit alors de trouver l'élément de l'espace des observations dont l'image, par la fonction noyau correspondante, soit la plus proche de la caractéristique en question. Il s'agit d'un problème mal-posé, dit de pré-image. Alors que l'on craint souvent de traiter ce problème, dit souvent malédiction de la pré-image, apprendre à le résoudre ouvre la voie à de nouvelles classes de méthodes non linéaires.
  → Le chapitre 3 est dédié à l'étude du problème de pré-image. Il présente des éléments de solution et décrit de nouvelles méthodes à noyaux dans des domaines d'application aussi multiples que variés.

- En apprentissage en ligne, la difficulté est liée à la solution optimale dont la taille du modèle, selon le Théorème de Représentation, est égale au nombre de couples entrée/sortie utilisés. Pour esquiver cette difficulté, il est nécessaire de considérer une solution sous-optimale en contrôlant l'ordre du modèle sans en compromettre les performances. Il est alors indispensable de définir, d'une part un sous-ensemble d'échantillons dans le développement (`Th.Rep`), et d'autre part une règle de mise à jour récursive des coefficients du modèle.
  ⇝ Le chapitre 4 traite le problème de l'apprentissage en ligne, avec une extension à l'apprentissage collaboratif dans les réseaux.
- Le Théorème de Représentation ne montre pas comment déterminer cette solution optimale. Le défi est de plus en plus grand à mesure que l'on s'intéresse à une solution qui représente une interprétation physique. Il est clair que le choix de la fonction noyau joue un rôle primordial (voir [Noumir et al., 2012b, Chen et al., 2012a, Honeine and Richard, 2010a] pour des noyaux adaptés aux processus étudiés). Des contraintes sont souvent imposées sur les coefficients et sur les échantillons nécessaires à la définition du modèle, comme c'est le cas du problème de démélange de données hyperspectrales.
  ⇝ Le chapitre 5 étudie le problème de démélange, avec divers modèles sous contraintes inspirées par la physique, dont la combinaison de noyaux et le modèle de mélange post-non-linéaire. L'utilisation de l'information spatio-spectrale dans la résolution du problème est aussi traitée.
- L'essence de ce résultat réside d'avoir autant de paramètres que d'observations, et pas plus. Malheureusement, ce principe n'est pas souvent respecté dans l'ingénierie de méthodes à noyaux. C'est le cas en particulier du problème de classification multi-classes avec les stratégies un-contre-tous et un-contre-un. Pour cette dernière, il s'agit de la fusion des solutions de ℓ sous-problèmes, ℓ étant le nombre de classes en compétition. Chaque solution est de la forme (`Th.Rep`). Il est ainsi nécessaire d'estimer $n \times \ell$ paramètres.
  ⇝ Le défi réside à repousser les limites du Théorème de Représentation, afin de se ramener à n paramètres à optimiser, tout en simplifiant les algorithmes d'optimisation développés. Voir le chapitre 6.

Je me suis intéressé à tous ces défis majeurs dans le cadre des méthodes à noyaux, comme résumé dans ce document. Bien évidemment, les méthodes proposées sont loin d'être en mesure de résoudre tous ces problèmes posés. Je n'ai pas non plus pour ambition de prétendre combler tous les vides existants.

Au delà de ces défis en méthodes d'apprentissage statistique, je suis aussi motivé à montrer la pertinence de ces méthodes dans de nouveaux domaines d'application qui ne se sont pas encore appropriés ni les méthodes à noyaux, ni la théorie qui les accompagne. Je suis intéressé aux domaines de recherche suivants :

- Le traitement collaboratif et distribué de l'information. Il s'agit de la thématique des réseaux de capteurs sans fil, avec deux problématiques qui sont l'auto-localisation des capteurs et d'estimation d'un champ physique. Ces travaux de recherches ont été soutenues dans le cadre du projet ANR KernSig (2006–2009), la Plateforme CAPSEC du CPER (2010–...), deux projets « Abondement Carnot » (2009–2014) et un projet Région « WiDiD » (2012–2015).
- Le traitement de signaux non-stationnaires et le test de non-stationnarité. Le cadre proposé permet de profiter des avancées récentes en apprentissage statistique dans ce domaine. Cette thématique a été traitée au cours de mes études de doctorat, puis en tant que Maître de Conférences dans le cadre du projet ANR StaRAC (2007–2010).
- La détection d'attaques dans les réseaux physiques ou cyber-physiques. C'est le cas dans le cadre du projet ANR Vigires'eau (2009–2012) qui porte sur la détection de contamination dans les réseaux de distribution d'eau potable, et le projet ANR SCALA (2012–2016) qui traite la détection d'intrusion cyber-physique dans les systèmes SCADA.



[CADRE 1] **Illustration du problème de pré-image.**

2.3 Synthèse des travaux

La présente section est dédiée à résumer les récentes contributions qui fournissent des éléments de réponse aux divers défis soulevés.

2.3.1 Le problème de pré-image en méthodes à noyaux

Les méthodes dites à noyaux puisent leur force dans le coup du noyau : un noyau correspond à un produit scalaire dans un espace transformé. Heureusement, nul n'a besoin d'expliciter la transformation induite par le noyau. Bien que cette transformation soit centrale, le retour inverse peut être d'une grande utilité en ouvrant la voie à de nouveaux domaines d'application des méthodes à noyaux. C'est le cas par exemple du débruitage de signaux/images. Bien que le débruitage se fasse dans l'espace fonctionnel, en projetant sur un espace pertinent identifié par l'ACP-à-noyaux par exemple, le résultat final devra apparaître dans le même espace des observations.

Malheureusement, il s'avère que le retour inverse n'existe pas en général, et seuls peu d'éléments dans l'espace caractéristique ont une pré-image valide dans l'espace des observations. Il s'agit essentiellement des images, par la fonction noyau utilisée, des échantillons d'apprentissage. Ces derniers n'ont donc aucune importance réelle. Le problème, dit de pré-image, est illustré dans le CADRE 1. Il consiste à déterminer l'élément de l'espace des observations dont l'image, par la fonction noyau utilisée, est une bonne approximation de la caractéristique étudiée, c'est à dire

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}\in\mathbb{X}}{\operatorname{argmin}} \|\kappa(\boldsymbol{x}, \cdot) - \psi(\cdot)\|^2_{\mathbb{H}}, \tag{2.1}$$

où $\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \, \kappa(\boldsymbol{x}_j, \cdot)$ est obtenue par une méthode à noyaux.

L'objectif des travaux réalisés dans cette thématique est double : d'abord, développer une étude théorique pour mieux comprendre le problème de pré-image et y proposer des éléments de solution ; ensuite, montrer que sa résolution ouvre la voie à de nouvelles classes de méthodes non linéaires. Nos contributions portant sur ces deux axes clés sont multiples, comme résumé dans la suite et décrit dans le chapitre 3.

Principales contributions sur la résolution du problème de pré-image

L'évolution historique de la résolution du problème de pré-image montre une diversité des méthodes proposées dans la littérature [Mika et al., 1999, Bakir, 2005, Kwok and Tsang, 2003]. Ce problème est étroitement lié au problème de réduction de dimension, la méthode *multidimensional scaling* (MDS) en est le fer de lance. Cette méthode réside dans la conservation des distances dans les deux espaces, une hypothèse toutefois pas naturelle pour les méthodes à noyaux puisqu'elle est en contradiction avec le principe de transformation non linéaire par noyau [Kwok and Tsang, 2003]. Pour combler ce défaut, nous avons proposé une nouvelle méthode, en considérant la conservation des produits scalaires. La méthode proposée, dite de transformation conforme, opère en deux étapes.

La première étape consiste à construire un repère dans l'espace RKHS qui soit en isométrie avec celui de l'espace des observations. Soient $\psi_\ell(\cdot) = \sum_{i=1}^n \theta_{i,\ell}\, \kappa(\boldsymbol{x}_i, \cdot)$ les fonctions qui définissent ce repère, où les coefficients $\theta_{i,\ell}$ sont à déterminer. Les coordonnées de tout $\kappa(\boldsymbol{x}_i, \cdot)$ dans ce repère sont de la forme $\langle \psi_\ell(\cdot), \kappa(\boldsymbol{x}_i, \cdot)\rangle_{\mathbb{H}} = \psi_\ell(\boldsymbol{x}_i)$. En regroupant ces coordonnées dans un vecteur $\boldsymbol{\psi}_{\boldsymbol{x}_i}$, les paramètres sont obtenus par l'isométrie avec l'espace des observations, en minimisant l'erreur quadratique moyenne entre $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ et $\boldsymbol{\psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\psi}_{\boldsymbol{x}_j}$, pour tout $i, j = 1, 2, \ldots, n$.

Dans la seconde étape, la caractéristique $\psi(\cdot)$ est représentée dans le repère ainsi construit. Dans ce cas, sa ℓ-ème coordonnée se ramène à $\langle \psi_\ell(\cdot), \psi(\cdot)\rangle_{\mathbb{H}} = \sum_{i,j=1}^n \theta_{i,\ell}\, \alpha_j\, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. En utilisant le principe d'isométrie, son homologue dans l'espace des observations est la pré-image. Nous avons montré dans [Honeine and Richard, 2011b, Honeine and Richard, 2009] que la pré-image est donnée par

$$\boldsymbol{x}^* = (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} - \eta_0\, \boldsymbol{K}^{-1})\ \boldsymbol{\alpha}, \tag{2.2}$$

où $\boldsymbol{X}^\top\boldsymbol{X}$ désigne la matrice du produit scalaire des observations dans l'espace $\mathbb{X}$, $\boldsymbol{K}$ désigne les produits scalaires dans l'espace RKHS et η_0 est un terme de régularisation.

Cette approche résulte en une expression analytique de la solution, qui est plus simple à implémenter et plus performante que les méthodes étudiées dans la littérature. Elle est décrite en détail dans la section 3.2.2 et dans [Honeine and Richard, 2011b, Honeine and Richard, 2009]. Nous avons montré son intérêt au travers de deux domaines d'application inédits :

- le problème d'auto-localisation des capteurs dans les réseaux sans fil, en utilisant le principe de régression matricielle (voir la section 3.4.2) ;
- le démélange spectral en imagerie hyperspectrale, avec l'intégration d'une régularisation spatiale dans l'image (voir la section 5.5.2)

Pré-image avec contraintes de non-négativité

En forçant des contraintes, un problème mal-posé peut devenir bien-posé. Nous avons considéré d'imposer des contraintes dans le problème de pré-image. Nous nous sommes intéressés en particulier à des contraintes de type non-négativité sur la pré-image résultante ou sur les coefficients qui définissent son modèle. Nous avons proposé un cadre unique pour la résolution de ses deux problèmes sous contraintes, avec une approche itérative selon une mise à jour de la forme

$$\boldsymbol{x}_{t+1}^* = \boldsymbol{x}_t^* + \eta_t\, \mathrm{diag}(\boldsymbol{x}_t^*)\nabla_{\boldsymbol{x}_t^*}\mathcal{J}(\boldsymbol{x}_t^*),$$

où $\mathrm{diag}(\boldsymbol{x}_t)$ désigne la matrice diagonale formée par les éléments de $\boldsymbol{x}_t$, $\nabla_{\boldsymbol{x}}\mathcal{J}(\boldsymbol{x})$ est le gradient de la fonction coût dans (2.1) par rapport à $\boldsymbol{x}$ et η_t est le pas d'adaptation. Une telle formulation permet de simplifier la contrainte imposée sur le pas pour imposer la non-négativité, qui ne doit pas dépasser une certaine valeur $\eta_t \leq \min_i 1/[\nabla_{\boldsymbol{x}_t^*}\mathcal{J}(\boldsymbol{x}_t^*)]_i$. Cette approche est étudiée en détail dans [Kallas et al., 2013b, Kallas et al., 2011e, Kallas et al., 2010b] et résumée dans la section 3.2.3. Voir aussi Cadre 5 (page 58) pour une illustration.

Domaines d'application : au delà du débruitage

Le débruitage d'images par l'ACP-à-noyaux, motivation initiale de la résolution du problème de pré-image, a toujours été au centre de la plupart des travaux dans la littérature. La résolution de

ce problème a énormément évolué pendant la dernière décennie, et les méthodes à noyaux se sont de plus en plus diversifiées. Toutefois, peu de travaux ont été menés sur des applications autres que le débruitage en imagerie. Nous avons montré avec succès que de nouvelles méthodes peuvent profiter de ses avancées. C'est le cas du modèle autorégressif-à-noyaux pour la modélisation en séries temporelles, selon la forme

$$\psi_i(\cdot) = \sum_{j=1}^{p} \alpha_{p-j+1} \kappa(\boldsymbol{x}_{i-j}, \cdot),$$

où p désigne l'ordre du modèle. Nous avons montré que les coefficients α_{p-j+1} peuvent être facilement estimés, au sens des moindres carrés ou encore par des équations *à la Yule-Walker*. Dans tous les cas, la résolution de la pré-image est nécessaire pour faire la prédiction dans l'espace des observations, en déterminant l'échantillon $\boldsymbol{x}_i^*$ dont l'image est la plus proche de $\psi_i(\cdot)$. L'approche proposée est résumée dans la section 3.3.2, elle est décrite en détail dans [Kallas et al., 2013a, Kallas et al., 2012c, Kallas et al., 2011b].

Au delà de l'analyse de séries temporelles, nous avons proposé de profiter de la résolution du problème de pré-image dans d'autres domaines. Dans un premier temps, il est question du problème d'auto-localisation dans les réseaux de capteurs sans fil. Deux méthodes d'estimation ont été proposées, comme présenté dans la section 3.4 :

- la première méthode résout explicitement le problème de pré-image (voir la section 3.4.1 et [Essoloh et al., 2008]) ;
- la seconde s'inspire du problème de la régression matricielle afin de résoudre implicitement le problème de pré-image (voir la section 3.4.2 et [Honeine et al., 2008c, Honeine et al., 2009b].

L'intérêt de la pré-image a aussi été démontré dans la résolution du problème de démélange spectral en imagerie hyperspectrale. Pour cela, deux étapes sont considérées : la première utilise l'estimation de la pré-image selon la transformation conforme sus-mentionnée ; la seconde étape considère des contraintes physiques ainsi qu'une régularisation spatiale de la solution. Voir la section 5.5.2 (page 116) et [Nguyen et al., 2013] pour plus de détails.

Pour compléter ces diverses méthodes que nous avons traitées, nous présentons [1] une étude de cas dans la section 3.5, avec le problème de la factorisation en matrices non négatives (NMF pour *nonnegative matrix factorization*) à noyaux. Nous revisitons le problème de la NMF linéaire au jour des méthodes à noyaux. Nous montrons explicitement la malédiction de la pré-image et nous proposons une solution directe. Les algorithmes de mise à jour additive et multiplicative sont décrits dans le cadre des méthodes à noyaux, montrant ainsi la simplicité de l'approche.

2.3.2 Avancées récentes en apprentissage en ligne par méthodes à noyaux

Confronté à un environnement non-stationnaire et dynamique, un apprentissage en ligne peut s'avérer incontournable. Les méthodes à noyaux n'apportent hélas pas de réponse directe et satisfaisante à cette question, la taille des modèles qu'elles engendrent étant égale au nombre de couples entrée/sortie utilisés. Il est alors crucial de contrôler l'ordre du modèle, selon le modèle d'ordre m à l'instant ℓ suivant :

$$\psi_\ell(\cdot) = \sum_{j=1}^{m} \alpha_{\ell,j}\, \kappa(\boldsymbol{x}_{\omega_j}, \cdot).$$

Ici, le « dictionnaire » à l'instant ℓ est l'ensemble $\{\kappa(\boldsymbol{x}_{\omega_2}, \cdot), \kappa(\boldsymbol{x}_{\omega_2}, \cdot), \ldots, \kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}$ formé par les fonctions sélectionnées, $\omega_j \in \{1, 2, \ldots, \ell\}$.

Au cours de mes études de doctorat, j'ai abordé cette question par le biais de critères classiquement utilisés par la communauté ayant trait aux représentations parcimonieuses, que j'ai étudié au jour des méthodes à noyaux [Honeine, 2007]. L'attention s'est portée en particulier sur le critère de cohérence du dictionnaire, qui permet un contrôle de la taille des modèles avec une complexité calculatoire linéaire. Le critère de cohérence est défini comme suit, pour un seuil de cohérence $\nu_0 \in [0, 1]$

1. Travail qui n'a pas été publié.

qui détermine le niveau de parcimonie : à l'instant ℓ, la fonction $\kappa(\boldsymbol{x}_\ell, \cdot)$ est introduite dans le dictionnaire si

$$\max_{j=1,\ldots,m} \mathrm{coh}(\kappa(\boldsymbol{x}_\ell, \cdot), \kappa(\boldsymbol{x}_{\omega_j}, \cdot)) > \nu_0, \tag{2.3}$$

avec la cohérence entre deux fonctions noyau

$$\mathrm{coh}(\kappa(\boldsymbol{x}_i, \cdot), \kappa(\boldsymbol{x}_j, \cdot)) = \frac{|\langle \kappa(\boldsymbol{x}_i, \cdot), \kappa(\boldsymbol{x}_j, \cdot)\rangle_{\mathbb{H}}|}{\|\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}} \|\kappa(\boldsymbol{x}_j, \cdot)\|_{\mathbb{H}}};$$

dans le cas contraire, le dictionnaire reste inchangé. Cette expression se ramène à $\max_i |\kappa(\boldsymbol{x}_\ell, \boldsymbol{x}_{\omega_i})| > \nu_0$ pour les noyaux de norme unité (*i.e.,* $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) = 1$ pour tout $\boldsymbol{x}_i \in \mathbb{X}$). J'ai démontré plusieurs résultats théoriques, dont la taille finie du dictionnaire résultant. J'ai aussi établi des liens avec divers critères plus gourmands en coût calculatoire, dont la procédure d'ACP. Des exemples de mise en œuvre ont été présentés par le développement d'algorithmes d'identification de systèmes non linéaires [Honeine, 2007].

Il convient à présent de décrire succinctement les travaux élaborés plus récemment.

Critère de parcimonie *a posteriori*

Les critères de parcimonie souvent utilisés sont de nature *a priori* : ils s'appliquent sur les $\kappa(\boldsymbol{x}_{\omega_j}, \cdot)$ et sont indépendants de la fonction $\psi_\ell(\cdot)$ résultante et de ses paramètres. C'est le cas du critère de cohérence d'un dictionnaire ou encore celui de l'approximation linéaire [Csató and Opper, 2001, Engel et al., 2004] et de l'entropie [Girolami, 2002, Suykens et al., 2002]. Nous avons proposé un critère de parcimonie *a posteriori*, en appliquant la cohérence à la fonctionnelle déterminant le modèle. Ainsi, la cohérence fonctionnelle est définie entre le modèle adapté et le modèle précédent. Soit la fonction $\psi_\ell(\cdot)$ déterminée en ajoutant $\kappa(\boldsymbol{x}_\ell, \cdot)$ dans le développement. Le critère consiste à déterminer si $\psi_\ell(\cdot)$ n'est pas assez pertinente, par rapport aux précédentes fonctions calculées, avec $\max_{k=1,\ldots,\ell-1} \mathrm{coh}(\psi_\ell(\cdot), \psi_k(\cdot)) > \nu_0'$, où $\psi_k(\cdot)$ désigne la fonction optimale obtenue à l'instant k. Cette expression correspond à une extension de (2.3), en considérant une limite supérieure du cosinus des angles entre les fonction. Bien que cela nécessite garder toutes les fonctions précédentes, nous avons dérivé une condition suffisante, avec le critère

$$\mathrm{coh}(\psi_\ell(\cdot), \psi_\ell^\perp(\cdot)) > \nu_1',$$

pour un seuil donné $\nu_1' \in [0, 1]$, où $\psi_\ell^\perp(\cdot)$ est la projection de $\psi_\ell(\cdot)$ sur l'espace engendré par les $\ell - 1$ fonctions. Par ailleurs, nous n'avons pas besoin de garder toutes ces fonctions en mémoire, puisque cet espace est aussi engendré par les fonctions noyau retenues. Voir la section 4.2.1 et [Honeine et al., 2009a, Honeine et al., 2008b] pour plus de détails, ainsi que [Honeine et al., 2009a] où la pertinence de ce critère de parcimonie est montrée en apprentissage distribué dans les réseaux de capteurs sans fil.

Adaptation des éléments du dictionnaire

Les critères de parcimonie abordés jusqu'à présent reposent uniquement sur la topologie des éléments du dictionnaire. Ils garantissent que les fonctions noyau sélectionnées reflètent une couverture raisonnable de l'espace de Hilbert correspondant. Toutefois, la constitution de tels dictionnaires n'est jamais remise en question, étant donné l'unique dépendance du critère de parcimonie vis-à-vis des entrées du système. Pour y remédier, nous avons proposé l'adaptation des éléments du dictionnaire afin de suivre l'évolution du système et donc réduire l'erreur de modélisation commise. Ceci consiste à optimiser le modèle conjointement selon les paramètres $\alpha_{\ell,i}$ du modèle et les éléments $\boldsymbol{x}_{\omega_i}$ du dictionnaire. L'adaptation de ces derniers est réalisée par la minimisation de l'erreur quadratique instantanée e_ℓ^2, avec $e_\ell = y_\ell - \sum_{i=1}^m \alpha_{\ell,i}\, \kappa(\boldsymbol{x}_\ell, \boldsymbol{x}_{\omega_i})$, et ceci sans enfreindre la règle de cohérence. En considérant une approche de gradient stochastique, la règle d'adaptation consiste à remplacer chaque élément $\boldsymbol{x}_{\omega_i}$ du dictionnaire par l'élément amélioré

$$\boldsymbol{x}_{\omega_i} - \eta_\ell\, \boldsymbol{g}_{\boldsymbol{x}_{\omega_i}},$$

où $\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}}$ désigne le gradient de e_ℓ^2 par rapport à $\boldsymbol{x}_{\omega_i}$, donné par l'expression (4.9). La valeur du pas de l'ajustement η_ℓ est déterminée sous la contrainte de cohérence du dictionnaire résultant. Une description détaillée de cette méthode est présentée dans la section 4.2.2. Voir [Saidé et al., 2013b, Saidé et al., 2013a, Saidé et al., 2012] pour plus de détails.

Nouvelle classe de méthodes à noyaux en ligne

Les travaux entamés jusqu'à présent étudient le problème d'identification en ligne de systèmes non linéaires, à sortie unique. Récemment, nous nous sommes intéressés à d'autres problèmes d'application très importants dans plusieurs domaines. Nos travaux se sont articulés autour des axes suivants :

- identification en ligne de systèmes non linéaires à sorties multiples, résumée dans [Saidé et al., 2013b] avec une application à l'analyse de signaux EMG ;
- analyse en composantes principales à noyaux avec un algorithme en ligne, décrit en détail dans la section 4.3 et [Honeine, 2012] ;
- détection séquentielle par approche mono-classe en ligne, présentée dans la section 6.2 et [Noumir et al., 2012e, Noumir et al., 2012d, Noumir et al., 2012c] ;
- traitement collaboratif de l'information dans les réseaux de capteurs. Voir la section 2.3.4 pour un résumé, et la section 4.4 pour plus de détails [Honeine et al., 2010, Honeine et al., 2009a, Honeine et al., 2008a, Honeine et al., 2008b].

A titre illustratif, nous présentons dans la suite de cette partie le problème de l'estimation de vecteurs propres en apprentissage statistique et reconnaissance des formes. L'analyse en composantes principales (ACP) et l'ACP-à-noyaux en sont le fer de lance. La règle d'Oja [Oja, 1982] propose une mise à jour adaptative du premier vecteur propre. En l'adaptant au formalisme fonctionnel basé sur les RKHS, la mise à jour à l'instant ℓ s'écrit selon

$$\psi_{\ell+1}(\cdot) = \psi_\ell(\cdot) + \eta_\ell \left(y_\ell \kappa(\boldsymbol{x}_\ell, \cdot) - y_\ell^2 \, \psi_\ell(\cdot)\right),$$

où $y_\ell = \psi_\ell(\boldsymbol{x}_\ell)$ est la valeur obtenue par la projection de $\kappa(\boldsymbol{x}_\ell, \cdot)$ sur $\psi_\ell(\cdot)$, avec $\psi_\ell(\boldsymbol{x}) = \langle \psi_\ell(\cdot), \kappa(\boldsymbol{x}, \cdot) \rangle_{\mathbb{H}} = \sum_k \alpha_{k,\ell} \, \kappa(\boldsymbol{x}_k, \boldsymbol{x})$. Il est clair qu'à chaque instant, la fonction principale $\psi_{\ell+1}(\cdot)$ s'enrichit d'un nouveau terme en $\kappa(\boldsymbol{x}_\ell, \cdot)$. Il s'avère nécessaire de contrôler l'ordre du modèle. Nous avons proposé la mise en œuvre de divers critères de parcimonie, avec l'adaptation récursive des paramètres du modèle d'ordre réduit. Nous avons étendu cette approche à l'extraction de multiples fonctions principales, et étudié des problèmes inhérents dont le centrage dans l'espace RKHS et le débruitage par pré-image. Ces travaux sont décrits en détail dans la section 4.3 et [Honeine, 2012].

2.3.3 Démélange linéaire et non linéaire de données hyperspectrales

Au delà du domaine de l'astrophysique ou encore de la cartographie terrestre, l'imagerie hyperspectrale se démocratise dans divers domaines industriels et médicaux, permettant l'identification de matière à l'aide de sa signature spectrale. Le problème de démélange spectral constitue un problème fondamental en traitement d'images hyperspectrales. Il vise à décomposer chaque pixel (*i.e.,* vecteur spectral) d'une image hyperspectrale sur une collection de signatures spectrales de composants purs, et à estimer la proportion de ces derniers dans le mélange [Bioucas-Dias et al., 2012, Keshava and Mustard, 2002]. Les défis du démélange en imagerie hyperspectrale sont multiples : (*a*) l'interprétation physique des résultats obtenus, ce qui nécessite la satisfaction de certaines contraintes telles que la complétude de la décomposition et l'additivité des contributions ; (*b*) la grande taille de ces cubes de données, avec des centaines de bandes spectrales et une résolution spatiale de plus en plus fine, nécessitant des algorithmes à faible complexité de calcul ; (*c*) la non-linéarité des mélanges spectraux, comme préconisé dans de récentes études ; et (*d*) l'intégration de l'information spatiale dans la résolution du problème de démélange spectral.

En désignant par $\{\boldsymbol{x}_{\omega_1}, \ldots, \boldsymbol{x}_{\omega_m}\}$ l'ensemble des signatures spectrales des m composants purs, le modèle de mélange linéaire est donné pour chaque pixel $\boldsymbol{x}$ par

$$\boldsymbol{x} = \sum_{j=1}^{m} \alpha_j \, \boldsymbol{x}_{\omega_j} + \boldsymbol{\epsilon}, \qquad \text{sous contraintes :} \quad \sum_{j=1}^{m} \alpha_j = 1, \quad \text{et} \quad \alpha_1, \ldots, \alpha_m \geq 0,$$

où $\boldsymbol{\epsilon}$ intègre la présence de bruit et autres composantes non-définies par le modèle. Les contraintes de somme unité et de non-négativité permettent une interprétation physique des fractions d'abondance α_j. Ce modèle linéaire s'écrit également selon chaque bande spectrale, avec $[\boldsymbol{x}]_i = \boldsymbol{z}_i^\top \boldsymbol{\alpha}$, où $\boldsymbol{z}_i$ désigne le vecteur des signatures spectrales des composants purs à la i-ème bande spectrale et $\boldsymbol{\alpha} = [\alpha_1 \;\; \alpha_2 \;\; \cdots \;\; \alpha_m]^\top$.

Diverses techniques d'identification des composants purs sont souvent préconisées dans la littérature, dont N-Findr [Winter, 1999, Plaza and Chang, 2005], SGA [Chang et al., 2006], VCA [Nascimento and Dias, 2004], OSP [Harsanyi and Chang, 1994, Chang, 2005], ICE [Berman et al., 2004]. L'estimation des fractions d'abondance ouvre la voie à de nouveaux développements, aussi bien en démélange linéaire qu'en démélange non linéaire. Les méthodes de démélange hyperspectral existantes sont essentiellement de deux natures : géométriques ou statistiques, avec diverses extensions pour le démélange non linéaire.

Les développements que nous avons proposés pour résoudre le problème de démélange sont essentiellement structurés autour de ces axes historiques, avec une extension pour intégrer l'information spatiale dans le problème d'optimisation. Les travaux entamés sont présentés dans le chapitre 5 et résumés dans la suite de cette partie.

En démélange par la géométrie, la plupart des méthodes de démélange traite le problème de démélange en deux étapes : d'abord, l'identification des composants purs, avec N-Findr, SGA, VCA, OSP, etc. ; ensuite, l'estimation des fractions d'abondance qui est souvent réalisée par la résolution d'un problème inverse secondaire. Nous avons montré que ces techniques géométriques d'extraction des composants purs permettent d'estimer conjointement les fractions d'abondance, pour un coût calculatoire supplémentaire négligeable. Voir la section 5.2.1 et [Honeine and Richard, 2012, Honeine and Richard, 2011a, Honeine and Richard, 2010b]. Nous avons proposé d'étendre ce socle commun pour le démélange non linéaire, en s'appropriant des techniques non linéaires de réduction de dimension, dont une approche de type géodésique (ISOMAP) et une approche localement linéaire (LLE). Voir la section 5.2.2 et [Honeine et al., 2013c, Nguyen et al., 2012].

En démélange linéaire par méthodes statistiques, notre avons montré que les contraintes de non-négativité et de somme unité peuvent être facilement imposées en modifiant certaines techniques classiques. D'une part, nous avons exploré une méthode de descente de gradient qui vérifie ces contraintes. Voir la section 5.3.1 et [Chen et al., 2011e]. D'autre part, nous avons proposé de forcer ces contraintes avec des techniques de projections multiples, dont les célèbres techniques de Cimmino et de Kaczmarz. Voir la section 5.3.2 et [Honeine and Lantéri, 2013, Honeine et al., 2013a].

En démélange non linéaire, la mise en œuvre de modèles non linéaires de la forme $[\boldsymbol{x}]_i = \psi(\boldsymbol{z}_i)$, où $\psi(\cdot)$ est un élément d'un espace fonctionnel arbitraire, se heurte à une difficulté majeure. Il est crucial d'exhiber les abondances $\boldsymbol{\alpha}$ dans le modèle afin de pouvoir imposer les contraintes totales, c'est à dire la non-négativité et la somme unité des α_j. Pour cela, nous avons exposé un nouveau paradigme sur l'hypothèse que le mécanisme de mélange peut être décrit par un mélange linéaire avec une fluctuation additive non linéaire [Chen et al., 2013e, Chen et al., 2011d]. L'interaction entre les bandes spectrales est donnée selon l'expression $[\boldsymbol{x}]_i \approx \boldsymbol{z}_i^\top \boldsymbol{\alpha} + \psi_{\text{nlin}}(\cdot)$. En décrivant cette fluctuation $\psi_{\text{nlin}}(\cdot)$ dans un espace de Hilbert à noyau reproduisant, nous avons exploité divers modèles de non-linéarité :

- non-linéarité indépendante des abondances, avec $\psi_{\text{nlin}}(\boldsymbol{z}_i)$ [Chen et al., 2011c] ;
- non-linéarité selon une forme convexe entre la composante linéaire et la composante non linéaire [Chen et al., 2013b, Chen et al., 2012a] ;
- non-linéarité par mélange post-non-linéaire, selon $\psi_{\text{nlin}}(\boldsymbol{z}_i^\top \boldsymbol{\alpha})$ [Chen et al., 2013c].

Les problèmes d'optimisation correspondants sont de difficulté croissante. Nous avons proposé des techniques appropriées pour l'estimation des différents paramètres de ces modèles. Voir la section 5.4, ainsi que la section 5.4.2 pour une étude sur le choix de la non-linéarité et l'interprétation physique de certains noyaux.

Nous avons complété ces méthodes par l'intégration de l'information spatiale, et ainsi profiter de la dualité spatiale-spectrale pour la résolution du problème de démélange. Pour l'estimation des fractions d'abondance $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N$ d'une image de N pixels, nous avons proposé la minimisation d'une fonction coût de la forme

$$\mathcal{J}_{\text{err}}(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N) + \eta\, \mathcal{J}_{\text{sp}}(\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_N),$$

où le premier terme représente l'erreur de modélisation et le second est un terme de régularisation pour favoriser la similarité des abondances de pixels voisins. Privilégiant l'homogénéité spatiale, ce dernier utilise la norme ℓ_1 entre chaque $\boldsymbol{\alpha}_i$ et ceux des pixels voisins, selon l'expression (5.23) (page 114). Dans ce cadre, nous avons étudié deux problèmes d'optimisation, selon le modèle investi :

- le modèle sus-mentionné combinant un mélange linéaire et une fluctuation non linéaire [Chen et al., 2013d] ;
- la solution du problème de pré-image décrite par l'expression (2.2). Voir [Nguyen et al., 2013] pour plus de détails.

La section 5.5 est dédiée au problème de la régularisation spatiale.

Au delà de la résolution du problème de démélange, nous nous sommes intéressés aussi au problème de classification multi-classes de vecteurs hyperspectraux, permettant ainsi une segmentation automatique de l'image hyperspectrale. Voir [Noumir et al., 2011b, Honeine and Richard, 2010a] pour un résumé de nos récents travaux sur la classification en imagerie hyperspectrale, avec un noyau adapté à ce type de données. Le chapitre 6 est dédié au problème de classification multi-classes, avec la description d'algorithmes à faible coût calculatoire.

2.3.4 Traitement collaboratif de l'information dans les réseaux de capteurs

Le domaine des réseaux de capteurs sans fil fait actuellement l'objet d'un intérêt considérable de la part des communautés académique et industrielle. La dispersion d'une multitude de capteurs bon marché dans une région donnée, l'élaboration d'un protocole de routage adéquat, et une implémentation algorithmique efficace, ouvrent en effet de nombreuses perspectives d'applications civiles et militaires telles que la surveillance et la sécurité. Chaque nœud du réseau est constitué d'un système miniaturisé, énergétiquement autonome, doté de capacités d'acquisition et de traitement de données. Une technologie sans fil leur permet de communiquer de proche en proche, sans hiérarchie centrale, de façon dynamique et instantanée, reconfigurable en fonction de l'évolution de la population de capteurs. Ce mode distribué présente l'avantage d'être particulièrement robuste aux attaques extérieures et à la défaillance de nœuds puisqu'il est prévu que la perte de composants ne compromette pas l'efficacité du réseau dans son ensemble. Deux problèmes fondamentaux ont fait l'objet de mes travaux de recherche dans ce domaine : l'auto-localisation des capteurs et l'apprentissage distribué et collaboratif.

Auto-localisation des capteurs

En l'absence d'information sur la position des éléments d'un réseau de capteurs sans fil, au sein de l'environnement où ils sont déployés, les mesures récoltées peuvent s'avérer d'une utilité limitée. Une étape préalable à tout traitement consiste donc à estimer la position de ces nœuds, à partir de mesures de portée inter-capteurs telles que les RSSI (*Received Signal Strength Indication*), et de la position supposée connue d'une fraction de capteurs appelés ancres. Ce problème d'auto-localisation a fait l'objet de nombreux travaux de recherche, les solutions proposées varient selon le type de mesures de portée inter-capteurs considérées, la nature des hypothèses relatives à la propagation des signaux correspondants, etc. Ainsi des méthodes classiques d'estimation statistique côtoient-elles des techniques d'analyse de données de type MDS, ou encore de programmation semi-définie. Les

méthodes non-paramétriques font l'objet d'une attention particulière, compte tenu de leur flexibilité particulièrement appréciable dans le cadre des réseaux de capteurs.

Dans ce contexte, nous avons proposé des techniques adaptées à la résolution de ce problème, notamment par des méthodes non paramétriques en apprentissage statistique. En adoptant les méthodes à noyaux, nous considérons qu'aucune hypothèse n'est faite sur les lois de probabilité régissant le réseau. Les positions des capteurs dans le réseaux sans fil sont ainsi estimées par des algorithmes de traitement distribué. Les développements que nous avons proposés pour résoudre le problème d'auto-localisation sont structurés autour des axes suivants :

- résolution du problème de pré-image dans les méthodes à noyaux, avec des techniques explicites [Honeine and Richard, 2011c, Essoloh et al., 2008] ou implicites par régression matricielle [Honeine et al., 2009b, Honeine et al., 2008c] ;
- méthodes à noyaux pour l'auto-localisation à partir de cartes radio par *fingerprints* [Mahfouz et al., 2013c, Mahfouz et al., 2013a, Mahfouz et al., 2013b] ;
- méthode d'auto-localisation par la définition d'intervalles polaires en théorie des intervalles [Mourad-Chehade et al., 2013, Mourad et al., 2012] ;
- utilisation de modèles de mobilité pour corriger l'auto-localisation dans les réseaux de capteurs mobiles [Mourad-Chehade et al., 2013, Ghadban et al., 2013b].

Par souci de clarté, seul le premier axe est décrit en détail dans ce document, dans la section 3.4 ; le chapitre 3 étant dédié au problème de pré-image.

Apprentissage collaboratif et algorithmes distribués

Nous nous sommes intéressés au problème d'identification distribuée d'un champ, par exemple de concentration de gaz ou d'une espèce chimique, et le suivi de son évolution au cours du temps. Le mode de calcul distribué est inhérent au caractère réparti des nœuds du réseau, dont la tâche est d'acquérir et de traiter localement les mesures. L'efficacité de la procédure d'identification est conditionnée par les interactions des nœuds, dictées par la topologie du réseau.

Dans ce contexte, chaque nœud ℓ, de coordonnées identifiées par sa position $\boldsymbol{x}_\ell \in \mathbb{X}$ supposée connue, accède à des réalisations temporelles, désignées par $y_{\ell,t}$ pour l'instant t. Le problème consiste à estimer, au sens des moindres carrés, une fonction $\psi(\cdot)$ d'un espace de Hilbert $\mathbb{H}$ à noyau reproduisant $\kappa(\cdot,\cdot)$ en minimisant un risque empirique régularisé. La solution d'un tel problème d'apprentissage est donnée par le Théorème de Représentation, avec

$$\psi_t(\cdot) = \sum_{j=1}^{m} \alpha_{j,t}\, \kappa(\boldsymbol{x}_{\omega_j}, \cdot).$$

Il s'agit alors, d'une part de l'estimation distribuée des coefficients de pondération avec une mise à jour adaptative, et d'autre part de l'identification du dictionnaire $\{\boldsymbol{x}_{\omega_1}, \ldots, \boldsymbol{x}_{\omega_m}\}$ des capteurs qui définissent le modèle. Nos contributions dans ce cadre sont multiples, comme résumé dans la suite :

- algorithmes décentralisés de type filtrage adaptatif, avec un dictionnaire reposant uniquement sur la topologie du réseau [Honeine et al., 2010, Honeine et al., 2008a] ;
- adaptation du dictionnaire pour le suivi de l'évolution du champ estimé [Honeine et al., 2009a, Honeine et al., 2008b] ;
- algorithmes de coopération pour les méthodes à noyaux, avec le mode incrémental et le mode de diffusion [Richard et al., 2010b, Essoloh et al., 2009] ;
- estimation distribuée sous contrainte, avec contraintes de non-négativité [Chen et al., 2010a] ou de voisinage [Ghadban et al., 2013a].

Dans la section 4.4, nous décrivons en détail une de ces méthodes, avec les deux principes de coopération, modes incrémental et de diffusion.

Au delà des réseaux sans fil

A ces développements théoriques, nous avons traité le problème de détection d'intrusion dans les réseaux. Il s'agit en particulier d'un réseau de distribution d'eau potable, où la mesure de Chlore en

certains nœuds permet de contrôler la qualité de l'eau tout en détectant des injections de matières nocives. Nos contributions dans ce domaine se résument par la proposition de techniques de détection en ligne par approche mono-classe à noyaux [Noumir et al., 2012e, Noumir et al., 2012d], le choix de noyaux optimisés aux mesures de Chlore traitées [Noumir et al., 2012b], et la détection localisée dans un réseau de capteurs [Noumir et al., 2012a, Deveughèle et al., 2012]. Voir aussi [Noumir et al., 2013, Yin et al., 2012, Fillatre et al., 2011, Fillatre et al., 2010] pour une revue des contributions apportées dans le cadre du projet ANR Vigires'eau, en partenariat avec Suez Environnement. Ces travaux continuent dans le cadre du projet ANR SCALA, sur la détection d'intrusion cyber-physique dans les réseaux SCADA (*supervisory control and data acquisition*) [Nader et al., 2013].

2.3.5 Contributions à la classification multi-classes

Les réseaux de neurones artificiels [Rosenblatt, 1958] ont permis des avancées considérables en classification automatique. La force de ces méthodes dites connexionnistes réside dans un grand nombre de paramètres qui permet une remarquable capacité d'apprentissage, au prix d'un problème d'optimisation non-convexe et du surapprentissage. Ces inconvénients ont été surmontés par les méthodes à noyaux avec la résolution d'un problème d'optimisation régularisé. Comme démontré par le Théorème de Représentation (`Th.Rep`), le nombre de paramètres est égal au nombre d'échantillons disponibles pour l'apprentissage.

Le problème de classification binaire (*i.e.,* à deux classes) a été largement étudié et les performances de diverses techniques de classification ont été bien établies. C'est le cas en particulier des SVM [Vapnik, 1998] et des machines à moindres carrés (RLSC ou LS-SVM) [Rifkin et al., 2003]. Une généralisation de ces résultats est souhaitée pour des tâches de discrimination multi-classes [Guermeur, 2008]. L'approche « diviser-pour-conquérir » permet d'étendre les méthodes de classification binaire pour des tâches à plusieurs classes. Elle consiste à décomposer le problème multi-classes en un ensemble de sous-problèmes binaires. La solution multi-classes finale est obtenue en fusionnant les résultats de ces sous-problèmes. Les stratégies de décomposition les plus courantes sont : un-contre-tous (OvA pour *one versus all the rest classes*) [Rifkin and Klautau, 2004] et un-contre-un [Fürnkranz, 2002].

La stratégie de décomposition un-contre-tous a montré son efficacité pour la mise en œuvre des SVM pour la classification multi-classes [Weston and Watkins, 1999, Crammer and Singer, 2002]. Les capacités de cette stratégie ont été démontrées dans plusieurs publications [Rifkin et al., 2003, Suykens et al., 2002, Rifkin and Klautau, 2004]. Le prix à payer réside dans un nombre accru de paramètres. En effet, pour un problème de classification à ℓ classes, chacun des sous-problèmes définit une fonction de décision de la forme (`Th.Rep`) avec n inconnues. Il s'agit donc d'un problème d'estimation de $n \times \ell$ paramètres inconnus.

Nos principales contributions en classification multi-classes montrent que le nombre d'inconnues peut être réduit de manière significative, sans sacrifier les performances. En effet, nous avons proposé de considérer autant de paramètres que d'échantillons d'apprentissage (c'est à dire n), et ainsi indépendamment du nombre ℓ de classes. Ces considérations peuvent sembler étonnantes, puisqu'elles sont en opposition aux méthodes couramment utilisées dans la littérature. Pourtant, nous rappelons que le principe sous-jacent du Théorème de Représentation stipule qu'il y ait autant de paramètres α_i que d'échantillons $\boldsymbol{x}_i$, et cela indépendamment du type (binaire, réel ou vectoriel) des étiquettes $\boldsymbol{y}_i$.

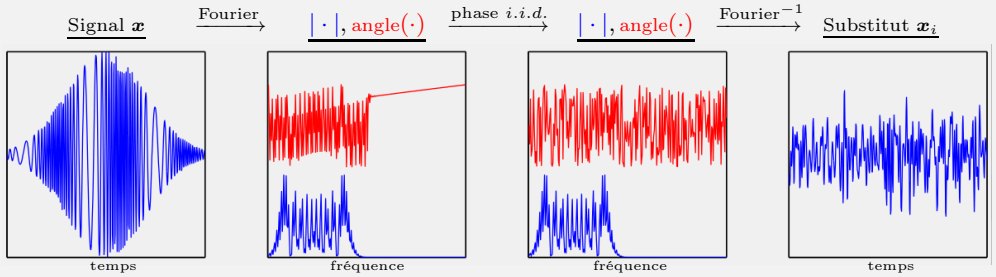
Nous avons développé des méthodes de classification multi-classes qui ont essentiellement la même complexité de calcul qu'un seul classifieur binaire [Honeine et al., 2013b, Noumir et al., 2011a, Noumir et al., 2011b]. Pour ce faire, deux approches ont été investies :

- D'une part, nous avons proposé un nouveau cadre de classifieurs multi-classes en étudiant des machines à sortie vectorielle. Pour cela, nous avons revisité des méthodes de classification binaire classiques afin d'en décrire des versions à sortie vectorielle, sur la base des mêmes routines d'optimisation et sans en augmenter véritablement le coût calculatoire. Plus précisément, nous avons étudié en détail trois méthodes largement reconnues dans la littérature :

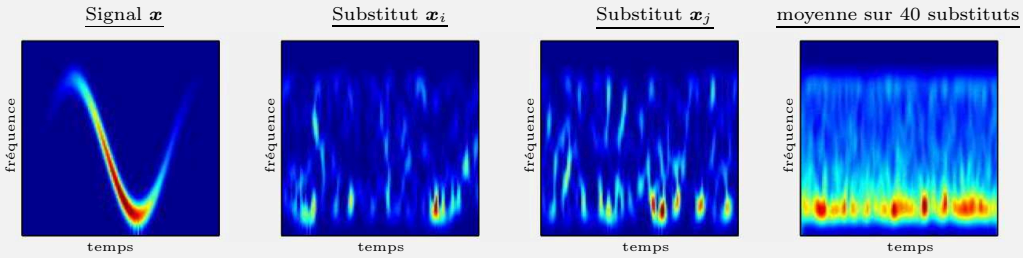
[Cadre 2] Substituts temps-fréquence pour l'analyse de non-stationnarité.

Signaux substituts (en anglais *surrogates*) : cette méthode, initialement proposée dans [Theiler et al., 1992] pour tester la non-linéarité, définit une hypothèse nulle.

Synthèse de substituts : passer dans le domaine fréquentiel et brouiller la phase.



Substituts temps-fréquence – Nouvelle interprétation : cette procédure stationnarise le signal dans le domaine temps-fréquence. Elle permet de caractériser une hypothèse nulle de stationnarité.



SVM [Vapnik, 1998], LSM (ou LS-SVM) [Suykens and Vandewalle, 1999b] et RLSC (pour *regularized least-squares classification*) [Rifkin, 2002]. Voir la section 6.4.1 pour plus de détails.

– D'autre part, nous avons étudié le couplage entre la stratégie de décomposition un-contre-tous et la résolution de sous-problèmes avec optimalité au sens des moindres carrés. Nous avons déterminé des relations inhérentes entre les inconnues des différents sous-problèmes. Nous avons ainsi montré qu'il s'agit bien d'un problème d'estimation de n paramètres inconnus, ceux-ci peuvent être estimés par l'inversion d'une seule matrice de taille $(n \times n)$. La description de la méthode et les résultats théoriques sont donnés dans la section 6.4.2

Voir le Cadre 26 et le Cadre 28 pour une analyse comparative des différentes méthodes, en terme de taux d'erreur de classification et en temps de calcul.

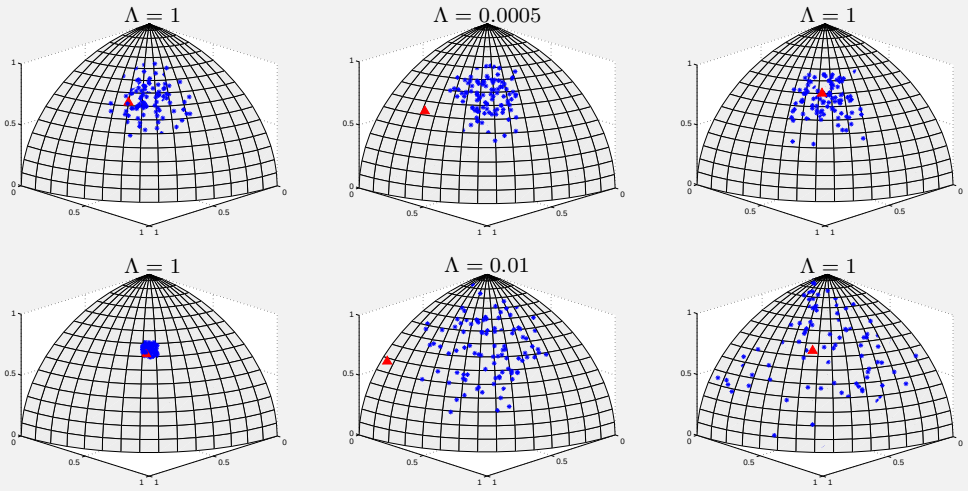
En classification multi-classes, le problème du choix du codage des étiquettes s'impose. Il existe divers codages d'étiquettes dans la littérature, dont le codage ± 1 [Dietterich and Bakiri, 1995, Allwein et al., 2001], le codage standard par les fonctions indicatrices [Bishop, 1995, Huang et al., 2012], le codage par alignement noyau-cible [Guermeur, 2008], le codage par principe inductif [Lee et al., 2004] et le codage par corrélation minimale [Szedmak et al., 2006]. Avec le foisonnement des différents codages, nos contributions dans cette thématique n'ont pas été de proposer « encore » un autre codage d'étiquettes. Paradoxalement, nous avons démontré que ces différentes étiquettes sont équivalentes dans le cadre d'un problème de moindres carrés. Ces résultats théoriques sur le problème de moindres carrés ont été complétés par des tests statistiques sur les autres méthodes multi-classes sus-mentionnées. Nous avons montré qu'il n'existe pas de différences significatives entre les codages d'étiquettes. Voir la section 6.4.3 pour l'étude du codage des étiquettes, ainsi que le Cadre 27.

[Cadre 3] Détection de non-stationnarité par approche mono-classe.

Nous avons proposé la mise en œuvre de méthodes à noyaux pour l'analyse et le test de non-stationnarité, dont la détection par approche mono-classe. Voir [Amoud et al., 2009b, Amoud et al., 2009a, Richard et al., 2010a, Borgnat et al., 2010, Borgnat et al., 2012]. Les expérimentations suivantes sont étudiées en détail dans [Amoud et al., 2009a].

Soient les signaux de modulation d'amplitude avec $x(t) = (1 + m\sin(2\pi t/T_0))\, e(t)$, où $m \leq 1$, $e(t)$ est un bruit blanc gaussien, et T_0 est la période de la modulation. Soient les signaux de modulation de fréquence avec $x(t) = \sin 2\pi(f_0 t + m\sin(2\pi t/T_0))\, e(t)$, où f_0 est la fréquence de la porteuse. Selon les valeurs de T_0 par rapport à la durée T du signal, trois cas peuvent être soulignés pour chaque type de modulation :
– $T \gg T_0$: la périodicité due au nombre important d'oscillations indique un régime stationnaire.
– $T \approx T_0$: une seule oscillation est visible dans le signal. Il est considéré comme non stationnaire.
– $T \ll T_0$: il n'y a pas de variation de l'amplitude ou de la fréquence. Il s'agit d'un régime stationnaire.

Les figures ci-dessous illustrent une représentation sphérique 3D des substituts (∗) et du signal test (▲), pour la modulation d'amplitude (première ligne) et la modulation de fréquence (seconde ligne), avec $T \gg T_0$ (gauche), $T \approx T_0$ (centre) et $T \ll T_0$ (droite). La valeur de Λ désigne l'indice de stationnarité obtenu par l'approche mono-classe proposée dans [Amoud et al., 2009a].



2.3.6 Analyse de signaux non-stationnaires et test de non-stationnarité

Le concept de stationnarité est omniprésent en traitement du signal. En accepter l'hypothèse représente un prérequis au bon usage de méthodes standards dévolues à l'étude de phénomènes en régime établi ; s'en écarter peut constituer en soi une information importante dans un environnement évolutif. La notion de stationnarité est parfaitement définie comme une invariance statistique des variables d'intérêt par rapport à toute translation temporelle. Elle est souvent assortie de considérations pratiques ayant le mérite d'en étendre la portée à des réalisations simples à horizon fini, éventuellement ponctuées de changements brusques et selon une échelle d'observation. Ainsi donne-t-on concrètement une dimension relative à ce concept, par rapport à une référence traduisant la notion de permanence dans le cadre expérimental défini par le praticien.

Abordant la notion de stationnarité dans une telle perspective relative, Patrick Flandrin et Pierre Borgnat ont initialement pris le parti de développer une méthode de simulation générant des références stationnarisées du signal étudié [Flandrin and Borgnat, 2008, Xiao et al., 2007]. Ces substituts (an anglais *surrogates*) permettent, le cas échéant, de rejeter l'hypothèse nulle de stationnarité

au terme d'un test statistique de type paramétrique, offrant alors la possibilité de quantifier un degré de non-stationnarité et d'en identifier une échelle caractéristique.

Nous avons contribué à ces travaux dans le cadre du projet StaRAC « Stationnarité relative et approches connexes » (ANR, programme blanc, 2007–2010), en proposant la mise en œuvre des méthodes à noyaux pour l'analyse et le test de non-stationnarité. Les différentes avancées dans cette thématique sont résumées dans la suite de cette section. Le Cadre 2 présente le problème de détection de non-stationnarité par méthode mono-classe. Voir aussi [Borgnat et al., 2010, Borgnat et al., 2012] pour une synthèse des travaux réalisés dans le cadre de ce projet.

Synthèse de nos contributions

Soit $\boldsymbol{x}$ le signal étudié. Une famille de substituts est générée, comme décrit dans le Cadre 2. Ces signaux, désignés par $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, forment l'ensemble d'apprentissage. Cette base d'apprentissage permet d'élaborer un détecteur non-paramétrique $\psi(\cdot)$ de type mono-classe (la première partie du chapitre 6 est dédiée à la classification mono-classe). Ainsi les coefficients de pondération α_j dans l'expression (`Th.Rep`) du Théorème de Représentation sont-ils estimés. Le test de non-stationnarité est donné par la règle de décision

$$\psi(\boldsymbol{x}) \underset{\text{stat.}}{\overset{\text{nonstat.}}{\gtrless}} \nu_0,$$

pour un seuil ν_0 prédéfini, c'est à dire

$$\sum_{j=1}^{n} \alpha_j \, \kappa(\boldsymbol{x}_j, \boldsymbol{x}) \underset{\text{stat.}}{\overset{\text{nonstat.}}{\gtrless}} \nu_0.$$

Le choix du noyau est étudié dans [Honeine and Richard, 2007, Honeine et al., 2006, Honeine et al., 2005], où nous avons proposé divers noyaux à interprétation temps-fréquence.

Tout en se situant dans ce contexte, nos récentes contributions sont les suivantes :

- Nous avons poursuivi cette étude en examinant dans quelle mesure le test permettrait d'aller au-delà de la détection, en proposant une caractérisation du type de non-stationnarité dont le signal testé ferait l'objet. Puisqu'aucun cadre ne permet de couvrir toutes les classes de non-stationnarités, nous nous sommes penchés à une classe de signaux mêlant modulations d'amplitude et de fréquence à des degrés respectifs choisis [Amoud et al., 2009b].
- L'originalité des travaux précédents est d'extraire des caractéristiques temps-fréquence à partir de l'ensemble des références stationnarisées du signal étudié, et de les utiliser pour définir l'hypothèse nulle de stationnarité. Nous avons proposé un cadre général combinant les techniques de l'apprentissage statistique à l'analyse temps-fréquence. Sur la base d'un apprentissage de machine mono-classe, notre approche utilise la totalité des représentations temps-fréquence et ne nécessite pas l'extraction de caractéristiques [Amoud et al., 2009a].
- Dans la littérature, la stationnarité des références stationnarisées du signal étudié a été montrée au sens large, c'est à dire que leur moments de première et second ordre sont invariants par translation. En démontrant la stationnarité au sens strict de ses substituts, nous avons exploité cette propriété pour estimer les distributions asymptotiques du spectrogramme correspondant, ainsi que la densité spectrale de puissance. Un test d'hypothèse statistique a été proposé pour vérifier la stationnarité du signal [Richard et al., 2010a].

2.4 Une vue d'ensemble

Il n'est pas évident de synthétiser mes travaux de recherche, puisqu'il s'agit bien d'une activité « non linéaire ». Le tableau suivant présente une vue synthétique de la situation, montrant ainsi la transversalité récurrente entre les principales méthodologies que j'ai développées et les domaines d'application traités. Pour rendre la présentation plus claire, j'ai choisi de décrire certains de mes travaux de recherche selon quatre chapitres :

– le chapitre 3 étudie le problème de pré-image dans les méthodes à noyaux, avec divers domaines applicatifs dont le traitement d'images, la modélisation de séries temporelles, l'auto-localisation dans les réseaux de capteurs sans fil et la factorisation en matrices non négatives ;
– le chapitre 4 traite la parcimonie et en particulier le problème de l'apprentissage en ligne avec deux études de cas : la première concerne un algorithme en ligne pour l'ACP-à-noyaux et la deuxième étudie le traitement collaboratif de l'information dans les réseaux de capteurs ;
– le chapitre 5 présente nos contributions sur le problème de démélange en imagerie hyperspectrale, en décrivant plusieurs méthodes de traitement linéaire et non linéaire. Ce chapitre est complété par une méthode de démélange qui utilise la résolution du problème de pré-image ;
– le chapitre 6 étudie le problème de classification, mono-classe et multi-classes. Nos contributions selon ces deux axes se résument par une simplification des techniques souvent utilisées dans la littérature.

Méthodologie

Domaines d'application		Pré-image	Parcimonie et dictionnaire	Apprentissage en ligne et collaboratif	Mono-classe et multi-classes	Noyau et représentation optimaux	Contraintes sur le Théorème de Représentation $\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \kappa(\boldsymbol{x}_j, \cdot) \Downarrow \psi_\ell(\cdot) = \sum_{j=1}^{m} \alpha_{\ell,j} \kappa(\boldsymbol{x}_{\omega_j}, \cdot)$: Choix des éléments du dictionnaire : $\mathcal{D} = \{\boldsymbol{x}_{\omega_1}, \ldots, \boldsymbol{x}_{\omega_m}\}$	Contraintes sur les coefficients : $\alpha_{\ell,1}, \ldots, \alpha_{\ell,m}$
Séries temporelles et signaux	Modélisation/prédiction AR	✓					*first in / first out*	
	Identification en ligne		✓	✓			critère de parcimonie	
	Détection en ligne		✓	✓	✓		critère de parcimonie	
	Extraction de caractéristiques	✓	✓	✓			critère de parcimonie	
	Débruitage signal/image	✓						$\psi_\ell(\cdot) \geq 0$ (facultatif)
	Temps-fréquence	✓			✓	✓	*surrogates*	
Réseaux de capteurs	Auto-localisation	✓	✓	✓	✓	✓	voisinage	
	Estimation distribuée		✓	✓			voisinage	$\psi_\ell(\cdot) \geq 0$ (facultatif)
Imagerie hyperspectrale	Démélange linéaire		✓				composants purs	$\alpha_{\ell,j} \geq 0$ et $\sum_j \alpha_{\ell,j} = 1$
	Démélange non linéaire	✓	✓			✓	composants purs	$\alpha_{\ell,j} \geq 0$ et $\sum_j \alpha_{\ell,j} = 1$
	Classification / segmentation				✓	✓	vérité terrain	

Quelle rage a-t-on d'apprendre ce qu'on craint toujours de savoir.
[Pierre-Augustin C. Beaumarchais]

3

Le problème de pré-image en méthodes à noyaux : changer la malédiction en bénédiction

Sommaire

Le problème de pré-image désigne le retour inverse de l'espace caractéristique à l'espace des observations. Il s'agit d'un problème mal-posé difficile à résoudre, à l'opposé du coup du noyau dont l'application directe et intuitive a permis le foisonnement des méthodes à noyaux. Alors que l'on craint souvent de traiter ce problème, dit souvent malédiction de la pré-image, apprendre à le résoudre ouvre la voie à de nouvelles classes de méthodes non linéaires dans des domaines applicatifs variés.

Depuis l'obtention de mon doctorat, je me suis activement engagé dans cette thématique de recherche. Bien qu'il s'agissait au début d'une *recherche en amont*, mes travaux ont suscité un intérêt croissant pour plusieurs projets financés. Depuis, les doctorants Mehdi Essoloh, Maya Kallas et Nguyen Hoang Nguyen y ont participé, chacun dans un champ d'application totalement différent. Ces travaux ont fait l'objet de quatre publications de revues [1], un chapitre de livre [2], et une dizaine

1. [Kallas et al., 2013a, Kallas et al., 2013b, Honeine and Richard, 2011c, Honeine and Richard, 2011b]
2. [Nguyen et al., 2013]

d'articles publiés dans des actes de conférences [3], ainsi que le papier [Honeine and Richard, 2009] lauréat en 2009 du *best paper award* au *19th IEEE international workshop on Machine Learning for Signal Processing*.

Dans la suite, je vais résumer très brièvement les travaux réalisés dans cette thématique. L'objectif poursuivi est double. Il est question de développer, dans un premier temps, une étude théorique pour mieux comprendre le problème de pré-image et y proposer des éléments de solution. C'est le cas en particulier avec la méthode de résolution par transformation conforme qui permet de résoudre des problèmes difficiles sous contraintes, comme illustré dans la section 5.5.2 du chapitre 5. Dans un second temps, nous cherchons à montrer que sa résolution ouvre la voie à de nouvelles classes de méthodes non linéaires. Ainsi l'intérêt d'étudier le problème de pré-image est-il montré au travers de champs d'application aussi multiples que variés :

- analyse de séries temporelles par modèle autorégressif à noyaux (voir la section 3.3.2) ;
- auto-localisation de capteurs dans les réseaux sans fil (voir la section 3.4) ;
- extraction de caractéristiques en signal et image (voir le chapitre 4) ;
- démélange spectral en imagerie hyperspectrale (voir la section 5.5.2 du chapitre 5, page 116).

Dans la section 3.5, nous étudions en détail le problème de la factorisation en matrices non négatives, sa mise en œuvre dans le cadre des méthodes à noyaux, la malédiction de la pré-image, et la résolution de ce problème.

3.1 Problématique et état de l'art

Les méthodes à noyaux sont pour la plupart issues d'algorithmes linéaires auxquels on a pu appliquer les deux résultats clés que sont le Théorème de Représentation [Wahba, 1990, Schölkopf et al., 2000] et le coup du noyau [Aizerman et al., 1964]. Le premier démontre que la solution d'un problème d'apprentissage régularisé est de la forme

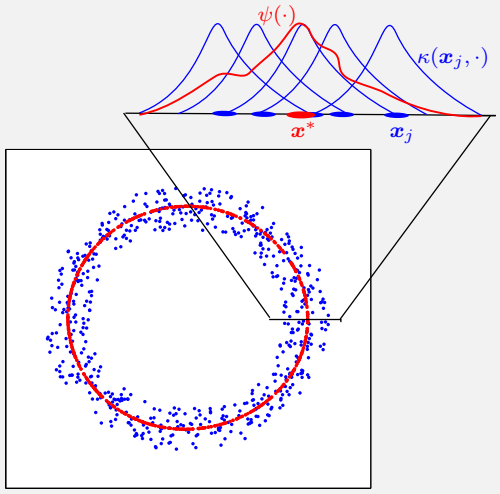
$$\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \, \kappa(\boldsymbol{x}_j, \cdot),$$

alors que le second stipule qu'on n'a pas besoin d'exhiber la transformation non linéaire $\phi(\cdot)$ induite par le noyau utilisé, puisque seuls les $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ sont nécessaires pour estimer les coefficients de pondération α_j. Comme illustré au travers des principales motivations des méthodes à noyaux en apprentissage statistique, à savoir les problèmes de classification et de régression, seule l'évaluation de la fonction $\psi(\cdot)$ est nécessaire. C'est le cas par exemple en détection et classification, où la discrimination d'une observation $\boldsymbol{x}$ est donnée par une règle de décision qui consiste à comparer $\psi(\boldsymbol{x})$ à un seuil.

Toutefois, rien n'empêche de s'intéresser à $\psi(\cdot)$, c'est à dire la caractéristique dans l'espace des RKHS $\mathbb{H}$. Bien qu'une caractéristique fonctionnelle ne peut être exploitable qu'en l'évaluant, nous nous intéressons à son équivalent dans l'espace observable $\mathbb{X}$. Le retour inverse de l'espace fonctionnel à l'espace des observations est alors primordial. A titre illustratif, c'est le cas du débruitage de signaux/images. Bien que le débruitage se fasse dans l'espace fonctionnel, par exemple en projetant sur un espace pertinent identifié par l'ACP-à-noyaux, le résultat final devra apparaître dans l'espace des observations. Un signal devra être débruité en un signal, une image en une image de même taille. Au delà du débruitage, le retour inverse ouvre la voie à de nouveaux domaines d'application des méthodes à noyau.

Malheureusement, il s'avère que le retour inverse n'existe pas en général, et seuls peu d'éléments de l'espace caractéristique ont une pré-image valide dans l'espace des observations. Il s'agit essentiellement des images, par la fonction noyau, des échantillons d'apprentissage. Ces derniers n'ont

3. [Kallas et al., 2012a, Kallas et al., 2012c, Kallas et al., 2011b, Kallas et al., 2011d, Kallas et al., 2011a, Kallas et al., 2011e, Kallas et al., 2011c, Kallas et al., 2010b, Honeine et al., 2009b, Essoloh et al., 2008, Honeine et al., 2008c]



[Cadre 4] Schéma illustrant le problème de pré-image dans le cas du noyau gaussien.

Considérons la caractéristique obtenue en appliquant l'ACP-à-noyaux sur des échantillons (points bleus) générés par une distribution en forme d'anneau. Représentée dans une coupe radiale, la pré-image (point rouge) correspond au centre de la gaussienne *la plus proche* de la somme (pondérée) de gaussiennes centrées sur les échantillons d'apprentissages (courbe rouge).

La malédiction de la pré-image illustrée ici ne concerne pas uniquement l'ACP-à-noyaux, ni le noyau gaussien en particulier.

donc pas d'importance réelle. Le problème dit de pré-image consiste à déterminer l'élément $\boldsymbol{x}^*$ de l'espace des observations $\mathbb{X}$ dont l'image $\kappa(\boldsymbol{x}^*, \cdot)$ est une bonne approximation de la caractéristique étudiée $\psi(\cdot)$. Ce problème peut être abordé à l'instar du problème de réduction de dimension, dit aussi *out-of-sample*, de $\mathbb{H}$ dans $\mathbb{X}$; ces deux problèmes sont intimement liés dans leur évolution historique [Honeine and Richard, 2011c]. De plus, le point de vue « fonctionnel » sort de la perspective de réduction de dimension, tout en enrichissant le problème de pré-image, comme illustré par le noyau gaussien.

Le point de vue gaussien

La forme induite par le noyau gaussien a été naturellement étudiée dans plusieurs domaines, comme c'est le cas en interpolation depuis les années 1970 [Schagen, 1979]. Nous illustrons dans la suite le problème de pré-image dans le cadre du noyau gaussien en méthodes à noyaux. Il s'agit de transformer chaque échantillon en une fonction gaussienne (« cloche ») centrée sur lui, selon

$$\begin{aligned} \phi\colon \mathbb{X} &\to \mathbb{H} \\ \boldsymbol{x}_j &\mapsto \exp(-\|\boldsymbol{x}_j - \cdot\,\|^2/2\sigma^2), \end{aligned}$$

où σ désigne le paramètre de largeur de bande. Le Théorème de Representation définit une combinaison linéaire de gaussiennes centrées sur les échantillons d'apprentissage. Cependant, il est bien connu qu'une somme de gaussiennes, centrées en différentes positions, ne peut pas être écrite comme une unique gaussienne. Ainsi, dans le cas général, toute fonction $\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \exp(-\|\boldsymbol{x}_j - \cdot\|^2/2\sigma^2)$ ne peut-elle être une gaussienne ; en d'autres termes, il ne s'agit pas d'une image, selon la fonction $\phi(\cdot) = \exp(-\|\boldsymbol{x} - \cdot\,\|^2/2\sigma^2)$ ci-avant, d'un certain $\boldsymbol{x}$ de $\mathbb{X}$. Trouver un élément $\boldsymbol{x}^*$ de $\mathbb{X}$, dont l'image est assez proche de la fonction $\psi(\cdot)$, est le problème de pré-image. Ce principe est illustré dans le Cadre 4.

Ses résultats ne concernent pas uniquement les méthodes à noyaux, ni le noyau gaussien en particulier. En effet, les modèles définis par une somme (pondérée ou pas) de fonctions de base ont été largement étudiés dans la littérature. Nous citons entre autres les problèmes d'interpolation [Schagen, 1979, Micchelli, 1986] et plus récemment le cadre de l'apprentissage statistique [Girosi et al., 1993].

Etat de l'art

Les différents éléments de réponse se regroupent essentiellement en deux catégories : d'une part, les techniques d'optimisation classique dont la descente de gradient et la technique du point-fixe,

et d'autre part, les méthodes se basant sur les échantillons d'apprentissage dont une variante de la technique MDS et l'apprentissage de la fonction inverse.

Soit une caractéristique $\psi(\cdot)$ de $\mathbb{H}$, obtenue par une méthode à noyaux, à savoir $\psi(\cdot) = \sum_{j=1}^{n} \alpha_j \, \kappa(\boldsymbol{x}_j, \cdot)$. Le problème de pré-image est donné par le problème d'optimisation suivant :

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \in \mathbb{X}}{\operatorname{argmin}} \left\| \kappa(\boldsymbol{x}, \cdot) - \sum_{j=1}^{n} \alpha_j \, \kappa(\boldsymbol{x}_j, \cdot) \right\|_{\mathbb{H}}^2, \tag{3.1}$$

qui, selon le coup du noyau, consiste à déterminer $\boldsymbol{x}^*$ qui minimise la fonction coût

$$\mathcal{J}(\boldsymbol{x}) = \kappa(\boldsymbol{x}, \boldsymbol{x}) - 2 \sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}, \boldsymbol{x}_i), \tag{3.2}$$

où le terme indépendant de $\boldsymbol{x}$ a été ignoré. Ce problème d'optimisation est non convexe et non linéaire, par la nature même de la fonction noyau utilisée.

La technique d'optimisation la plus simple est donnée par une descente de gradient, selon l'expression récursive

$$\boldsymbol{x}_{t+1}^* = \boldsymbol{x}_t^* - \eta_t \, \nabla_{\boldsymbol{x}_t^*} \mathcal{J}(\boldsymbol{x}_t^*),$$

où η_t désigne le pas à l'itération t et $\nabla_{\boldsymbol{x}_t^*} \mathcal{J}(\boldsymbol{x}_t^*)$ désigne le gradient de la fonction coût (3.2) en $\boldsymbol{x}_t^*$. Pour décrire une technique du point fixe, il suffit d'annuler le gradient à l'optimum, c'est à dire $\nabla_{\boldsymbol{x}^*} \mathcal{J}(\boldsymbol{x}^*) = 0$. Par conséquent, nous retrouvons l'expression associée au noyau gaussien, présentée dans [Mika et al., 1999], avec

$$\boldsymbol{x}_{t+1}^* = \frac{\sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_t^*, \boldsymbol{x}_i) \; \boldsymbol{x}_i}{\sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_t^*, \boldsymbol{x}_i)}, \tag{3.3}$$

ou encore pour le noyau polynomial $\kappa_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c)^p$ [Kwok and Tsang, 2003] :

$$\boldsymbol{x}_{t+1}^* = \sum_{i=1}^{n} \alpha_i \left(\frac{\langle \boldsymbol{x}_t^*, \boldsymbol{x}_i \rangle + c}{\langle \boldsymbol{x}_t^*, \boldsymbol{x}_t^* \rangle + c} \right)^{p-1} \boldsymbol{x}_i. \tag{3.4}$$

Malheureusement, les travaux menés dans cette direction ont montré que seul le noyau gaussien est susceptible de donner des résultats acceptables [Kwok and Tsang, 2003]. Toutefois, ces schémas d'optimisation classiques restent inadaptés à une telle fonction coût non linéaire et non convexe. Ceci se manifeste en particulier par des minima locaux, ou encore une instabilité due à un dénominateur qui tend vers zéro. Une première tentative pour remédier à cette dernière situation réside dans la formulation d'une solution régularisée, avec l'adjonction d'un terme additif dans les dénominateurs des expressions (3.3) et (3.4) [Abrahamsen and Hansen, 2009].

Les méthodes précédentes n'exploitent pas le lien qui existe déjà entre les deux espaces, à savoir les échantillons d'apprentissage et leurs images par la fonction noyau. Les méthodes ci-après mettent en évidence ce lien.

Le problème de pré-image peut s'écrire sous la forme d'un problème d'apprentissage, en construisant une machine pour apprendre la fonction de retour de $\mathbb{H}$ dans $\mathbb{X}$, comme suit. Le problème consiste à déterminer une fonction $\Gamma^*(\cdot)$ telle que $\Gamma^*(\kappa(\boldsymbol{x}_i, \cdot)) \approx \boldsymbol{x}_i$, pour $i = 1, 2, \ldots, n$. La pré-image de $\psi(\cdot)$ est alors définie par $\Gamma^*(\psi(\cdot))$. Afin de dériver une solution à ce problème, les deux considérations suivantes ont été apportées dans [Bakir et al., 2004, Bakir, 2005]. Tout d'abord, la fonction $\Gamma^*(\cdot)$ est définie sur un espace vectoriel, en décomposant tout élément de l'espace $\mathbb{H}$ sur une base orthonormée, cette dernière étant définie par les r fonctions principales de l'ACP-à-noyaux. Ensuite, la fonction à définir est décomposée en plusieurs fonctions, une par dimension de l'espace $\mathbb{X}$. A partir de ces considérations, chacune des fonctions $\Gamma_1^*(\cdot), \Gamma_2^*(\cdot), \ldots, \Gamma_{\dim(\mathbb{X})}^*(\cdot)$, avec $\Gamma_m^* \colon \mathbb{R}^r \to \mathbb{R}$, est obtenue par la résolution du problème d'optimisation suivant :

$$\Gamma_m^*(\cdot) = \underset{\Gamma(\cdot)}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \left| [\boldsymbol{x}_i]_m - \Gamma(\psi) \right|^2 + \eta \, \|\Gamma(\cdot)\|^2,$$

où un terme de régularisation a été introduit, et $[\cdot]_m$ désigne la m-ème composante. Il s'agit alors d'un problème de régression par moindres carrés. Cette approche d'apprentissage est étudiée davantage dans la littérature, en intégrant des informations de voisinage [Zheng and Lai, 2006] ou encore par une pénalisation [Zheng et al., 2010b], toujours en combinant d'une part une ACP-à-noyaux et d'autre part une résolution d'un ensemble de problèmes de régression. Il est à noter que cette dernière peut se réaliser en une seule étape, comme nous démontrons dans le chapitre 6, dans un cadre plus général avec l'apprentissage multi-tâches.

Le problème de pré-image peut être davantage simplifié, grâce à la technique MDS [Cox and Cox, 2000]. Voir la section 5.2.2 pour la formulation de la technique MDS, ainsi que [Williams, 2002] pour une connexion entre les techniques MDS et ACP-à-noyaux. La méthode MDS exploite le lien entre les deux espaces en utilisant les distances deux-à-deux dans les deux espaces, comme montré dans [Kwok and Tsang, 2003] et présenté dans la suite. En considérant les distances dans l'espace caractéristique $\delta_i = \|\psi(\cdot) - \kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}}$, pour $i = 1, 2, \ldots, n$, et celles dans l'espace des observations, $\|\boldsymbol{x}^* - \boldsymbol{x}_i\|$, le problème d'optimisation consiste à trouver un $\boldsymbol{x}^*$ tel que

$$\|\boldsymbol{x}^* - \boldsymbol{x}_i\|^2 \approx \|\psi(\cdot) - \kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}}^2, \tag{3.5}$$

pour tout $i = 1, 2, \ldots, n$. Ceci se traduit par les n équations $2\langle \boldsymbol{x}^*, \boldsymbol{x}_i \rangle = \langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle + \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle - \delta_i^2$. La quantité $\langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle$ est facilement estimée pour des échantillons centrés, puisque la moyenne de ces équations permet d'écrire :

$$\langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle = \frac{1}{n} \sum_{i=1}^{n} \left(\delta_i^2 - \langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle \right).$$

Soit $\boldsymbol{\epsilon}$ le vecteur qui contient ces quantités, alors le problème s'écrit sous la forme matricielle

$$2\boldsymbol{X}^\top \boldsymbol{x}^* = \operatorname{diag}\left(\boldsymbol{X}^\top \boldsymbol{X}\right) - \left[\delta_1^2 \ \delta_2^2 \ \cdots \ \delta_n^2\right]^\top + \boldsymbol{\epsilon},$$

où $\boldsymbol{X} = [\boldsymbol{x}_1 \ \boldsymbol{x}_2 \ \cdots \ \boldsymbol{x}_n]$ et $\operatorname{diag}(\cdot)$ désigne l'opérateur diagonal, avec $\operatorname{diag}(\boldsymbol{X}^\top \boldsymbol{X})$ étant le vecteur colonne d'éléments $\langle \boldsymbol{x}_i, \boldsymbol{x}_i \rangle$. La pré-image est alors obtenue par la solution aux moindres carrés, avec

$$\boldsymbol{x}^* = \tfrac{1}{2}(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\left(\operatorname{diag}(\boldsymbol{X}^\top \boldsymbol{X}) - [\delta_1^2 \ \delta_2^2 \ \cdots \ \delta_n^2]^\top\right),$$

où $(\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}\boldsymbol{\epsilon}$ vaut zéro grâce à l'hypothèse des échantillons centrés.

Malheureusement, toutes ces hypothèses ne sont pas satisfaites dans la majorité des cas, et en particulier l'hypothèse d'échantillons de moyenne nulle. C'est le cas par exemple en traitement des images où chaque image est constituée d'un ensemble de pixels non négatifs.

3.2 Synthèse des contributions sur la résolution du problème de pré-image

Nos multiples contributions sur la résolution du problème de pré-image sont résumées dans la suite autour de trois parties : dans la première partie, des avancées théoriques sur le problème sont présentées, autour de deux théorèmes. Une méthode de résolution très simple et performante est décrite dans la deuxième partie. La troisième partie traite la résolution du problème avec contraintes de non-négativité. Nous complétons ces contributions méthodologiques en montant dans les sections suivantes la pertinence de la résolution du problème de pré-image, dans des domaines d'application divers tels que la modélisation de séries temporelles, l'auto-localisation de capteurs dans les réseaux sans fil, et la factorisation en matrices non négatives.

3.2.1 Résultats théoriques

Dans cette partie, nous montrons que la pré-image s'écrit sous la forme $\boldsymbol{x}^* = \sum_{i=1}^{n} \beta_i^* \boldsymbol{x}_i$, pour des coefficients optimaux $\beta_1^*, \beta_2^*, \ldots, \beta_n^*$ à estimer. Ainsi l'espace d'hypothèse est-il contrôlé, par opposition aux techniques de descente de gradient qui explorent l'espace entier. Ce résultat a

été conjecturé par certaines techniques de pré-image dans la littérature [Kwok and Tsang, 2003, Honeine and Richard, 2009, Zheng et al., 2010a], toutefois sans aucune explication théorique. Le théorème suivant valide cette thèse, en l'étendant aux vastes classes des noyaux radiaux et projectifs.

Théorème 1 (Expression de la pré-image)**.**
Pour un problème d'apprentissage sur un ensemble d'échantillons $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, toute pré-image $\boldsymbol{x}^*$ s'écrit sous la forme

$$\boldsymbol{x}^* = \sum_{i=1}^{n} \beta_i^* \, \boldsymbol{x}_i. \tag{3.6}$$

Démonstration. En reprenant l'expression du gradient de la fonction coût (3.2), et en l'annulant à l'optimum, nous avons

$$\sum_{i=1}^{n} \alpha_i \nabla_{\boldsymbol{x}^*} \kappa(\boldsymbol{x}^*, \boldsymbol{x}_i) = \frac{1}{2} \nabla_{\boldsymbol{x}^*} \kappa(\boldsymbol{x}^*, \boldsymbol{x}^*),$$

où $\nabla_{\boldsymbol{x}^*}$ désigne le gradient par rapport à $\boldsymbol{x}^*$. Il suffit alors de considérer les noyaux radiaux de la forme

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = g(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2), \tag{3.7}$$

pour obtenir

$$\boldsymbol{x}^* = \sum_{i=1}^{n} \alpha_i \, \frac{g'(\|\boldsymbol{x}_i - \boldsymbol{x}^*\|^2)}{\sum_{j=1}^{n} \alpha_j \, g'(\|\boldsymbol{x}_j - \boldsymbol{x}^*\|^2)} \, \boldsymbol{x}_i,$$

ou encore les noyaux projectifs de la forme

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = f(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle), \tag{3.8}$$

pour avoir

$$\boldsymbol{x}^* = \sum_{i=1}^{n} \alpha_i \, \frac{f'(\langle \boldsymbol{x}_i, \boldsymbol{x}^* \rangle)}{f'(\langle \boldsymbol{x}^*, \boldsymbol{x}^* \rangle)} \, \boldsymbol{x}_i.$$

Ici, $g'(\cdot)$ et $f'(\cdot)$ désignent les premières dérivées des fonctions par rapport à leurs arguments. ■

Notons toutefois qu'il ne s'agit pas d'une simple combinaison linéaire, puisque les coefficients de pondération dépendent des $\boldsymbol{x}_i$. Nous retrouvons les deux cas particuliers les plus connus, le noyau gaussien $g(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) = \exp(\frac{-1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$ et le noyau polynomial $\kappa_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + c)^p$, avec les expressions respectives :

$$\boldsymbol{x}^* = \frac{\sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}^*) \, \boldsymbol{x}_i}{\sum_{i=1}^{n} \alpha_i \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}^*)}, \qquad \text{et} \qquad \boldsymbol{x}^* = \frac{\sum_{i=1}^{n} \alpha_i \, \kappa_{p-1}(\boldsymbol{x}_i, \boldsymbol{x}^*) \, \boldsymbol{x}_i}{\kappa_{p-1}(\boldsymbol{x}^*, \boldsymbol{x}^*)}.$$

Ce théorème exploite la première dérivée de la fonction coût (3.2). Sa dérivée seconde donne une idée plus précise sur sa convexité, tel qu'il découle du théorème suivant.

Théorème 2 (Sur la convexité du problème de pré-image)**.**
Pour la classe des noyaux radiaux, une condition suffisante pour la convexité de la fonction coût (3.2) *est donnée par la non-négativité des coefficients $\alpha_1, \alpha_2, \ldots, \alpha_n$.*

Démonstration. D'une part, nous rappelons que les noyaux radiaux partagent la propriété de monotonicité complète [Cucker and Smale, 2002] (voir aussi [Burges, 1999, Proposition 7.2]), c'est à dire que la k-ème dérivée de $g(\cdot)$ par rapport à son argument vérifie

$$(-1)^k g^{(k)}(\zeta) \geq 0,$$

pour tout $\zeta > 0$ et $k \geq 0$, et en particulier $g(\zeta) \geq 0$, $g'(\zeta) \leq 0$, et $g''(\zeta) \geq 0$. D'autre part, nous considérons la seconde dérivée de la fonction coût (3.2) par rapport à $\boldsymbol{x}$, avec

$$\begin{aligned}\nabla^2_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}) &= \nabla_{\boldsymbol{x}} \Big(2 \sum_{i=1}^{n} \alpha_i \, (\boldsymbol{x}_i - \boldsymbol{x}) \, g'(\|\boldsymbol{x}_i - \boldsymbol{x}\|^2) \Big) \\ &= 2 \sum_{i=1}^{n} \alpha_i \Big(- g'(\|\boldsymbol{x}_i - \boldsymbol{x}\|^2) + 2 (\boldsymbol{x}_i - \boldsymbol{x})^2 \, g''(\|\boldsymbol{x}_i - \boldsymbol{x}\|^2) \Big).\end{aligned}$$

En combinant ces deux résultats, il est clair que le terme entre parenthèses est positif. Donc, une condition suffisante pour que la fonction coût soit convexe, réside dans la non-négativité des α_i. ■

3.2.2 Solution par transformation conforme

Comme l'illustre la section 3.1, l'évolution historique de la résolution du problème de pré-image montre une diversité des techniques proposées dans la littérature. Il est clair que la méthode MDS ouvre la voie à une nouvelle classe de méthodes de réduction de dimension, y compris par exemple l'approche localement linéaire (LLE pour *locally linear embedding*) [Roweis and Saul, 2000] et l'approche de type géodésique (ISOMAP) [Tenenbaum et al., 2000]. Voir la section 5.2.2 pour plus de détails sur les méthodes non linéaires de réduction de dimension. La simplicité de la méthode MDS montre que seules sont nécessaires les distances deux-à-deux entre les échantillons représentés dans chacun des deux espaces.

La méthode MDS réside dans la conservation des distances dans les deux espaces. Comme illustré avec l'identité (3.5), cette hypothèse est toutefois non naturelle pour les méthodes à noyaux, puisqu'elle peut être en parfaite contradiction avec le principe de transformation non linéaire par noyau. Une première tentative pour remédier à ce problème est proposée dans [Etyngier et al., 2007], avec une conservation locale des distances à l'instar de l'approche LLE. Notre méthode présentée dans la suite est complètement différente, en proposant d'étudier la conservation des produits scalaires [Honeine and Richard, 2011b, Honeine and Richard, 2009]. Pour cela, nous commençons par construire un repère dans l'espace caractéristique $\mathbb{H}$ qui est en isométrie avec celui de l'espace des observations $\mathbb{X}$. Nous précisons qu'aucune contrainte n'est imposée sur ce repère, à l'opposé de l'ACP-à-noyaux où une base orthonormée en résulte. Ensuite, en représentant la caractéristique en question dans ce repère, on obtient par isométrie les coordonnées de sa pré-image dans l'espace des observations. Ces deux étapes sont détaillées dans la suite.

Pour un ensemble de n échantillons d'apprentissage, soient $\psi_1(\cdot), \psi_2(\cdot), \ldots, \psi_m(\cdot)$, pour $m \leq n$, les m fonctions qui définissent le repère en question. En vertu du Théorème de Représentation, chacune de ces fonctions peut s'écrire comme une combinaison linéaire des images disponibles, selon

$$\psi_\ell(\cdot) = \sum_{i=1}^{n} \theta_{i,\ell} \, \kappa(\boldsymbol{x}_i, \cdot).$$

Soit $\boldsymbol{\Theta}$ la matrice de taille $(n \times m)$ des coefficients $\theta_{i,\ell}$ à déterminer. Tout élément de l'espace $\mathbb{H}$ est ainsi décrit dans ce repère par sa projection sur les fonctions sus-mentionnées. En d'autres termes, chaque fonction $\kappa(\boldsymbol{x}_i, \cdot)$ est représentée par ses m coordonnées dans

$$\begin{aligned}\boldsymbol{\psi}_{\boldsymbol{x}_i} &= \begin{bmatrix} \langle \psi_1(\cdot), \kappa(\boldsymbol{x}_i, \cdot) \rangle_{\mathbb{H}} & \langle \psi_2(\cdot), \kappa(\boldsymbol{x}_i, \cdot) \rangle_{\mathbb{H}} & \cdots & \langle \psi_m(\cdot), \kappa(\boldsymbol{x}_i, \cdot) \rangle_{\mathbb{H}} \end{bmatrix}^\top \\ &= \begin{bmatrix} \psi_1(\boldsymbol{x}_i) & \psi_2(\boldsymbol{x}_i) & \cdots & \psi_m(\boldsymbol{x}_i) \end{bmatrix}^\top.\end{aligned}$$

En d'autres termes, $\psi_{\boldsymbol{x}_i} = \boldsymbol{\Theta}^\top \boldsymbol{\kappa}(\boldsymbol{x}_i)$, où $\boldsymbol{\kappa}(\boldsymbol{x}_i) = [\kappa(\boldsymbol{x}_1, \boldsymbol{x}_i) \quad \kappa(\boldsymbol{x}_2, \boldsymbol{x}_i) \quad \cdots \quad \kappa(\boldsymbol{x}_n, \boldsymbol{x}_i)]^\top$. Par construction, nous souhaitons que les produits scalaires soient conservés dans les deux systèmes de coordonnées, celui de l'espace (euclidien) des observations $\mathbb{X}$ et celui défini par ce repère dans $\mathbb{H}$. En conséquence, le modèle considéré est

$$\boldsymbol{x}_i^\top \boldsymbol{x}_j = \boldsymbol{\psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\psi}_{\boldsymbol{x}_j} + \epsilon_{i,j}, \tag{3.9}$$

pour tout $i, j = 1, 2, \ldots, n$, où $\epsilon_{i,j}$ intègre la présence de bruit et autres composantes non-définies par le modèle. Nous considérons le problème d'optimisation par la minimisation de la variance empirique (sur tous les couples) des $\epsilon_{i,j}$, à savoir

$$\min_{\psi_1,\ldots,\psi_\ell} \sum_{i,j=1}^{n} \left| \boldsymbol{x}_i^\top \boldsymbol{x}_j - \boldsymbol{\psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\psi}_{\boldsymbol{x}_j} \right|^2 + \eta \sum_{\ell=1}^{m} \|\psi_\ell(\cdot)\|_{\mathbb{H}}^2, \tag{3.10}$$

où une régularisation de type ℓ_2 a été introduite, avec η un paramètre qui contrôle le compromis entre la fidélité du modèle et sa régularité. Sous forme matricielle, nous écrivons le problème d'optimisation selon

$$\boldsymbol{\Theta}^* = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{K}\,\boldsymbol{\Theta}^\top \boldsymbol{\Theta}\,\boldsymbol{K}\|_F^2 + \eta\, \mathrm{tr}(\boldsymbol{\Theta}^\top \boldsymbol{\Theta} \boldsymbol{K}), \tag{3.11}$$

où $\mathrm{tr}(\cdot)$ désigne la trace de la matrice et $\|\cdot\|_F$ la norme de Frobenius, c'est à dire $\|\boldsymbol{M}\|_F^2 = \mathrm{tr}(\boldsymbol{M}^\top \boldsymbol{M})$. En annulant la première dérivée de cette fonction coût par rapport à $\boldsymbol{\Theta}^\top \boldsymbol{\Theta}$ directement, plutôt que $\boldsymbol{\Theta}$, on obtient

$$\boldsymbol{\Theta}^{*\top} \boldsymbol{\Theta}^* = \boldsymbol{K}^{-1} \left(\boldsymbol{X}^\top \boldsymbol{X} - \eta\, \boldsymbol{K}^{-1} \right) \boldsymbol{K}^{-1}. \tag{3.12}$$

Dans la suite, nous verrons que $\boldsymbol{\Theta}^{*\top} \boldsymbol{\Theta}^*$ suffit pour l'estimation de la pré-image.

Une fois que ce repère est défini à partir de cette expression, nous sommes en mesure de déterminer la pré-image d'une caractéristique arbitraire $\psi(\cdot) = \sum_{i=1}^{n} \alpha_i\, \kappa(\boldsymbol{x}_i, \cdot)$. Ses coordonnées dans le repère ainsi construit sont données par

$$\langle \psi_\ell(\cdot), \psi(\cdot) \rangle_{\mathbb{H}} = \sum_{i,j=1}^{n} \theta_{i,\ell}\, \alpha_j\, \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j),$$

pour $\ell = 1, 2, \ldots, m$. La conservation des produits scalaires dans chaque espace conduit à l'application du modèle (3.9) à $\psi(\cdot)$. Par analogie avec (3.11), le problème d'optimisation est

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{X}^\top \boldsymbol{x} - \boldsymbol{K}\, \boldsymbol{\Theta}^{*\top} \boldsymbol{\Theta}^*\, \boldsymbol{K} \boldsymbol{\alpha}\|^2,$$

c'est à dire, en utilisant la solution optimale (3.12) :

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{X}^\top \boldsymbol{x} - \left(\boldsymbol{X}^\top \boldsymbol{X} - \eta\, \boldsymbol{K}^{-1} \right) \boldsymbol{\alpha}\|^2. \tag{3.13}$$

La solution de ce problème est donnée par la résolution du système linéaire suivant :

$$\boldsymbol{X}^\top \boldsymbol{x}^* = \boldsymbol{K} \boldsymbol{\Theta}^{*\top} \boldsymbol{\Theta}^*\, \boldsymbol{K} \boldsymbol{\alpha},$$

ou encore $\boldsymbol{X}^\top \boldsymbol{x}^* = (\boldsymbol{X}^\top \boldsymbol{X} - \eta\, \boldsymbol{K}^{-1})\, \boldsymbol{\alpha}$ où l'expression (3.12) est utilisée. La solution optimale, au sens des moindres carrés, est donc donnée par le pseudo-inverse selon :

$$\boldsymbol{x}^* = \left((\boldsymbol{X} \boldsymbol{X}^\top)^{-1} \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X} - \eta\, \boldsymbol{K}^{-1}) \right) \boldsymbol{\alpha}.$$

Il est clair que la méthode proposée est indépendante du type du noyau utilisé. De plus, cette technique peut être facilement étendue pour identifier les pré-images d'un ensemble d'éléments dans l'espace caractéristique, puisque le terme entre parenthèses n'est calculé qu'une seule fois. En effet, il s'agit d'un problème de complétion de matrice. Ce problème est un problème de régression matricielle, comme celui introduit en bioinfomatique dans [Yamanishi and Vert, 2007]. Dans notre cas, la matrice des produits scalaires est complétée à partir d'une matrice de noyau. Le lien entre le problème de pré-image et la régression matricielle est explicité en détail dans la section 3.4.2, au travers du problème d'auto-localisation des capteurs.

La méthode présentée ici est investie dans la section 5.5.2, pour la résolution du problème de démélange spectral en imagerie hyperspectrale. Dans ce cadre, deux modifications majeures sont intégrées dans la résolution du problème (3.13). D'une part, afin d'imposer une interprétation physique du résultat, deux contraintes sont imposées dans le problème d'optimisation : la somme unité et la non-négativité de la pré-image. D'autre part, ce problème sous contraintes est étudié dans la section 5.5.2 avec l'intégration d'une régularisation spatiale dans l'image.

Remarque 1 : pré-image d'une fonction noyau

Un cas particulier réside dans la pré-image d'une fonction noyau $\kappa(\boldsymbol{x}_0, \cdot)$, avec $\boldsymbol{x}_0$ n'appartenant pas à l'ensemble d'apprentissage. Dans ce cas, cette fonction noyau est représentée par sa projection dans le sous-espace défini par les fonctions noyau de l'ensemble d'apprentissage, avec $\sum_{i=1}^{n} \beta_i^* \, \kappa(\boldsymbol{x}_i, \cdot)$. Les coefficients β_i^* sont obtenus en minimisant l'erreur de reconstruction, selon

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \, \|\kappa(\boldsymbol{x}_0, \cdot) - \sum_{i=1}^{n} \beta_i \, \kappa(\boldsymbol{x}_i, \cdot)\|_{\mathrm{H}}^2.$$

La solution optimale est donnée par la résolution du système linéaire $\boldsymbol{K}\boldsymbol{\beta}^* = \boldsymbol{\kappa}(\boldsymbol{x}_0)$, où $\boldsymbol{\kappa}(\boldsymbol{x}_0)$ est le vecteur de taille $(n \times 1)$ d'éléments $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_0)$, pour $i = 1, 2, \ldots, n$. En reprenant le problème d'optimisation (3.13), on obtient

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \, \|\boldsymbol{X}^\top \boldsymbol{x} - \big(\boldsymbol{X}^\top \boldsymbol{X} - \eta \, \boldsymbol{K}^{-1}\big)\boldsymbol{K}^{-1}\boldsymbol{\kappa}(\boldsymbol{x}_0)\|^2,$$

et ainsi la solution

$$\boldsymbol{x}^* = \Big((\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} - \eta \, \boldsymbol{K}^{-1})\Big) \, \boldsymbol{K}^{-1}\boldsymbol{\kappa}(\boldsymbol{x}_0).$$

Remarque 2 : régularisation avec la norme de Frobenius

Dans le problème d'optimisation (3.10), nous avons considéré une régularisation avec la norme ℓ_2 sur les fonctions, ce qui a induit un terme de régularisation en $\operatorname{tr}(\boldsymbol{\Theta}^\top \boldsymbol{\Theta} \, \boldsymbol{K})$ dans (3.11). Puisque $\operatorname{tr}(\cdot) \leq \|\cdot\|_F^2$, rien n'empêche d'utiliser le terme de régularisation suivant : $\|\boldsymbol{\Theta}^\top \boldsymbol{\Theta} \, \boldsymbol{K}\|_F^2$. Dans ce cas, l'expression de la solution (3.12) devient

$$\boldsymbol{\Theta}^{*\top}\boldsymbol{\Theta}^* = \boldsymbol{K}^{-1}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{K}\big(\boldsymbol{K}^\top \boldsymbol{K} + \eta\mathbf{I}\big)^{-1},$$

et le problème de pré-image dans (3.13) devient

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \, \frac{1}{2}\|\boldsymbol{X}^\top \boldsymbol{x} - \boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{K}\big(\boldsymbol{K}^\top \boldsymbol{K} + \eta\mathbf{I}\big)^{-1}\boldsymbol{K}\boldsymbol{\alpha}\|^2.$$

3.2.3 Pré-image avec contraintes de non-négativité

La conception de nouveaux outils de résolution du problème de pré-image est plus ambitieuse quand on exige un sens physique à la solution. L'intégration de contraintes, dans un problème mal-posé par nature, permet souvent de le rendre bien-posé. Le chapitre 5 est dédié à la mise en œuvre de méthodes de démélange de données hyperspectrales, avec des contraintes issues d'une interprétation physique. Dans cette section, nous étudions des contraintes dans le problème de pré-image, en s'intéressant en particulier à deux types de contraintes de non-négativité, sur la pré-image et sur les coefficients dans le développement défini par le théorème 1 (page 54).

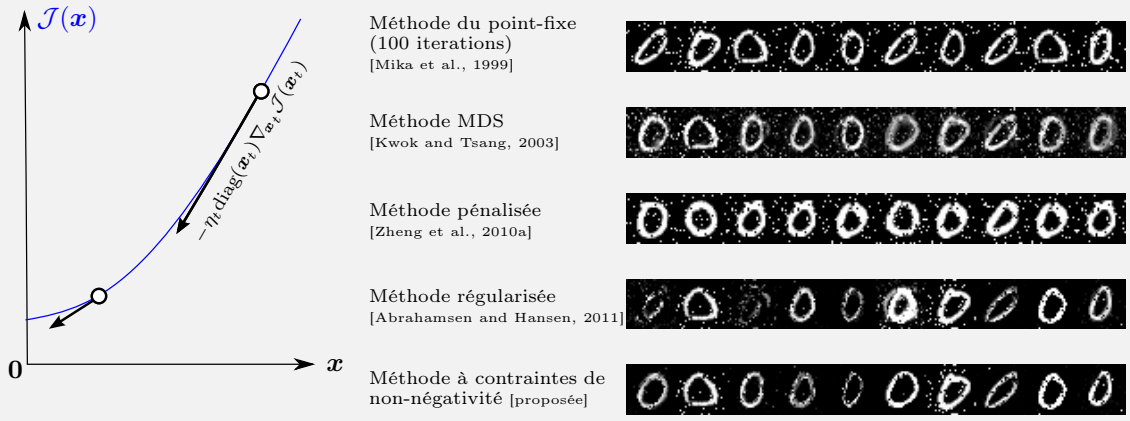
Nous considérons dans un premier temps les contraintes de non-négativité sur la pré-image, avec $\boldsymbol{x}^* \geq \mathbf{0}$. Le problème d'optimisation correspondant est défini par

$$\boldsymbol{x}^* = \underset{\boldsymbol{x} \geq \mathbf{0}}{\operatorname{argmin}} \, \mathcal{J}(\boldsymbol{x}),$$

où la fonction coût est donnée par (3.2). Nous proposons une résolution itérative selon

$$\boldsymbol{x}_{t+1}^* = \boldsymbol{x}_t^* - \eta_t(\boldsymbol{x}_t^*) \, \nabla_{\boldsymbol{x}_t^*} \mathcal{J}(\boldsymbol{x}_t^*), \tag{3.14}$$

où le pas $\eta_t(\boldsymbol{x}_t^*)$, une fonction de $\boldsymbol{x}_t^*$, permet de contrôler la convergence sans sortir du domaine admissible de la non-négativité. Comme préconisé dans la section 5.3.1 (bien loin du problème de pré-image, voir aussi [Chen et al., 2011b]), nous écrivons $\eta_t(\boldsymbol{x}_t^*) = \eta_t \operatorname{diag}(\boldsymbol{x}_t^*)$, où $\operatorname{diag}(\boldsymbol{x}_t^*)$ est la matrice diagonale formée par les éléments de $\boldsymbol{x}_t^*$. Une telle formulation du pas favorise la parcimonie, puisqu'elle permet une convergence rapide vers les valeurs nulles avec des pas plus importants pour



[CADRE 5] **Problème de pré-image avec contraintes de non-négativité.**
Schéma de gauche : illustration de la pondération du pas η_t par les valeurs dans $\boldsymbol{x}$. La descente de gradient est plus prononcée dans les directions des valeurs élevées dans $\boldsymbol{x}$.
Figures de droite : un ensemble de dix images de « 0 » corrompues par un bruit sel-et-poivre de densité 0,1 (première ligne). Une collection de 500 images est exploitée pour l'apprentissage de l'ACP-à-noyaux, avec un noyau gaussien de largeur de bande $\sigma = 500$ et 50 vecteurs propres sont retenus. Les résultats des différentes méthodes montrent la pertinence d'imposer la contrainte de non-négativité.

les grandes valeurs dans le vecteur $\boldsymbol{x}_t^*$, comme illustré dans le schéma du CADRE 5. En outre, une telle pondération du pas η_t par $\operatorname{diag}(\boldsymbol{x}_t^*)$ permet de simplifier la contrainte imposée sur le pas, puisqu'en conséquence il doit satisfaire l'inégalité suivante :

$$\eta_t \leq \min_i \frac{1}{[\nabla_{\boldsymbol{x}_t^*} \mathcal{J}(\boldsymbol{x}_t^*)]_i}.$$

En utilisant le modèle $\boldsymbol{x}^* = \sum_{i=1}^n \beta_i^* \, \boldsymbol{x}_i$ décrit dans le théorème 1, nous pouvons étendre notre approche à une méthode qui impose la non-négativité des coefficients β_i^*. Soit $\boldsymbol{\beta}^* = [\beta_1^* \;\; \beta_2^* \;\; \cdots \;\; \beta_n^*]^\top$, c'est à dire $\boldsymbol{x}^* = \boldsymbol{\beta}^{*\top} \boldsymbol{X}$. Dans ce cas, le problème d'optimisation avec contraintes devient

$$\boldsymbol{\beta}^* = \operatorname*{argmin}_{\boldsymbol{\beta} \geq \mathbf{0}} \mathcal{J}(\boldsymbol{\beta}^\top \boldsymbol{X}).$$

En tenant compte de la linéarité entre $\boldsymbol{x}^*$ et $\boldsymbol{\beta}^*$, nous avons une relation entre les gradients associés, où le gradient de $\mathcal{J}(\boldsymbol{\beta}^\top \boldsymbol{X})$ par rapport à $\boldsymbol{\beta}$ est égal à $\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}) \, \boldsymbol{X}^\top$. Ainsi une modeste modification du précédent algorithme (3.14) permet-elle de décrire la nouvelle règle itérative, selon

$$\boldsymbol{\beta}_{t+1}^* = \boldsymbol{\beta}_t^* - \eta_t \operatorname{diag}(\boldsymbol{\beta}_t^*) \nabla_{\boldsymbol{x}_t^*} \mathcal{J}(\boldsymbol{x}_t^*) \, \boldsymbol{X}^\top.$$

Une illustration de la pertinence de cette méthode est donnée dans le CADRE 5. Voir [Kallas et al., 2011a, Kallas et al., 2011e, Kallas et al., 2010b] pour plus de détails, avec des applications sur des enregistrements de l'activité cérébrale (EEG, potentiel évoqué).

3.3 Domaines d'application : au delà du débruitage

Le débruitage d'images par l'ACP-à-noyaux a été la motivation initiale de la résolution du problème de pré-image et reste toujours au centre de la plupart des études dans la littérature. La résolution de ce problème a énormément évolué pendant cette dernière décennie, et les méthodes à noyaux se sont de plus en plus diversifiées. Toutefois, peu de travaux ont été menés sur des applications autres

que le débruitage d'images, à l'exception de l'étude des espaces structurés dont l'analyse de séquences biologiques [Sonnenburg et al., 2008] et l'analyse de chaines de caractères [Cortes et al., 2005]. La suite de ce chapitre est dédiée à montrer l'importance de la résolution du problème de pré-image au travers de nouvelles méthodes à noyaux.

3.3.1 Nouvelles classes de méthodes à noyaux

Pour cela, nous commençons par un problème bien connu en traitement du signal, qui est la modélisation d'un système autorégressif (AR). En revisitant les équations de Yule-Walker dans le cadre des méthodes à noyaux, nous décrivons le modèle AR-à-noyaux qui, en vue d'une prédiction, nécessite la résolution du problème de pré-image. Un schéma illustratif de cette approche est donné dans le Cadre 6, avec une analyse comparative des performances. Voir la section 3.3.2 ci-après pour le détail du calcul.

Un autre problème bien connu dans la littérature est l'analyse en composantes indépendantes, notamment en séparation aveugle de sources [Hyvärinen et al., 2001]. L'algorithme FastICA en est le fer de lance. Il est facile de décrire, dans l'espace RKHS, chaque étape de l'algorithme FastICA classique. L'algorithme résultant, dit FastICA-à-noyaux, permet de définir des composantes indépendantes dans cet espace. Toutefois, une étape supplémentaire est nécessaire afin de résoudre le problème de pré-image, et ainsi revenir à l'espace des échantillons. Le Cadre 9 illustre l'analogie entre le fameux algorithme FastICA, et sa version non linéaire FastICA-à-noyaux.

A ces nouvelles méthodes à noyaux, s'ajoutent deux autres problèmes que nous étudions en détail dans la suite de ce chapitre. Le premier concerne l'auto-localisation de capteurs dans les réseaux sans fil, en proposant deux méthodes par résolution du problème de pré-image. Ces deux méthodes sont décrites dans la section 3.4. Le second étudie le problème de factorisation en matrices non négatives à noyaux, en illustrant la malédiction de la pré-image et sa résolution. Voir la section 3.5.

L'intérêt de la pré-image est aussi illustré dans le chapitre 5 pour la résolution du problème de démélange spectral en imagerie hyperspectrale. Voir la section 5.5.2 (page 116) où la transformation conforme décrite dans la section 3.2.2 est considérée sous contraintes.

3.3.2 Modèle AR-à-noyaux en traitement de séries temporelles

Dans cette partie, nous montrons que le traitement de séries temporelles peut profiter de ces avancées, en développant un modèle autorégressif à noyaux (AR-à-noyaux) pour l'analyse et la prédiction.

Soit une série temporelle définie par les échantillons $x_1, x_2, \ldots, x_n \in \mathbb{X} \subset \mathbb{R}$. Nous considérons un modèle AR d'ordre m sur les images par la fonction noyau, $\kappa(x_1, \cdot), \ldots, \kappa(x_n, \cdot)$, c'est à dire dans l'espace caractéristique $\mathbb{H}$, selon

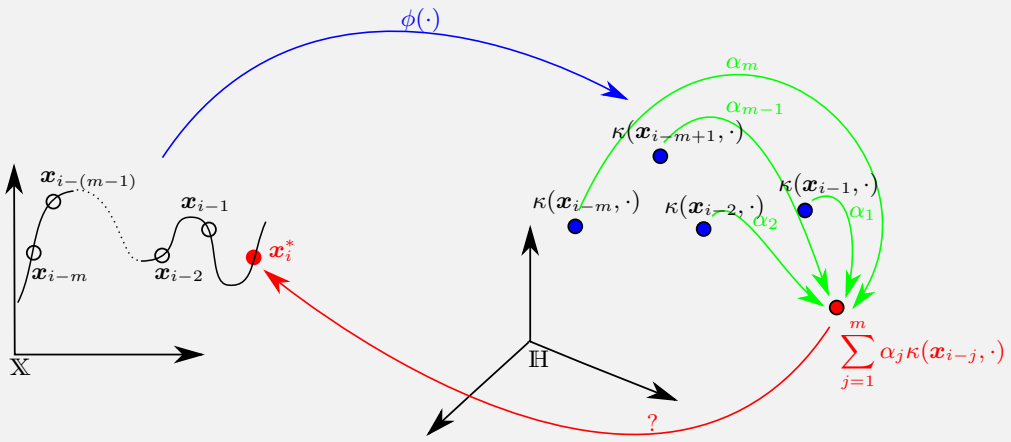
$$\kappa(x_i, \cdot) = \sum_{j=1}^{m} \alpha_j \, \kappa(x_{i-j}, \cdot) + \varepsilon_i^{\phi}(\cdot), \tag{3.15}$$

où $\varepsilon_i^{\phi}(\cdot) \in \mathbb{H}$ désigne un terme d'erreur, supposé décorrélé de tout $\kappa(x_{i-\tau}, \cdot)$ pour un retard τ fixé. En considérant les espérances des deux membres de ce modèle, nous avons $\mathbb{E}[\varepsilon_i^{\phi}(\cdot)] = (1-\sum_{j=1}^{m} \alpha_j)\mu_{\infty}(\cdot)$ sous l'hypothèse de stationnarité. Dans cette expression, $\mu_{\infty}(\cdot) = \mathbb{E}[\kappa(x, \cdot)]$ désigne l'espérance mathématique avec $\mu_{\infty}(\cdot) = \int_{\mathbb{X}} \kappa(x, \cdot) \, dP(x)$, où $P(x)$ est la distribution de probabilité inconnue qui génère les échantillons de la série temporelle. Nous obtenons par conséquent :

$$\kappa(x_i, \cdot) - \mu_{\infty}(\cdot) = \sum_{j=1}^{m} \alpha_j \Big(\kappa(x_{i-j}, \cdot) - \mu_{\infty}(\cdot) \Big) + \varepsilon_i^{\phi}(\cdot) - \Big(1 - \sum_{j=1}^{m} \alpha_j \Big) \mu_{\infty}(\cdot).$$

En considérant le produit scalaire de chaque membre de cette équation avec $(\kappa(x_{i-\tau}, \cdot) - \mu_{\infty}(\cdot))$, pour un retard τ fixé, les espérances des différents termes permettent d'écrire

$$\mathbb{E}[\kappa_c(x_i, x_{i-\tau})] = \sum_{j=1}^{m} \alpha_j \, \mathbb{E}[\kappa_c(x_{i-j}, x_{i-\tau})],$$



	Laser	MG_{30}	*Ikeda*	*Lorenz*
Perceptron multicouche	1,4326	0,0461	0,00071	0,2837
Régression à vecteurs de support	0,2595	0,0313	0,00081	0,1811
Filtre de Kalman non linéaire	0,2325	0,0307	0,00077	0,3133
AR-à-noyaux avec pré-image	0,0702	0,0008	0,00088	0,1792

[CADRE 6] **Méthode AR-à-noyaux avec pré-image : illustration et étude comparative.**
Nous comparons la méthode AR-à-noyaux aux méthodes suivantes : perception multicouches [Haykin, 1999], SVM pour la régression [Vapnik, 1998], et filtre de Kalman non linéaire [Ralaivola and d'Alché-Buc, 2005]. Le modèle AR linéaire n'est malheureusement pas adapté aux séries temporelles utilisées, puisqu'elles exhibent un comportement non linéaire.
Nous considérons la même configuration étudiée dans [Ralaivola and d'Alché-Buc, 2005], qui est rappelée dans la suite. L'apprentissage est réalisé sur les $n = 300$ premiers échantillons, ainsi que l'estimation des différents paramètres (ordre du modèle, largeur de bande du noyau gaussien, ...). Les performances sont mesurées sur les 300 suivants échantillons, avec l'erreur quadratique moyenne $\frac{1}{n}\sum_{i=n+1}^{2n} \|x_i^* - x_i\|^2$, où x_i^* désigne la valeur prédite à l'instant i et x_i la valeur réelle de la série temporelle au même instant. Il est à noter que l'approche proposée est de loin plus simple à mettre en œuvre, avec moins de paramètres (avec $m \leq 6$ en pratique), et une complexité calculatoire nettement plus faible que toutes les autres méthodes. Se référer à [Kallas et al., 2012c, Kallas et al., 2012a, Kallas et al., 2011b, Kallas et al., 2011d] pour de plus amples explications et une étude comparative à d'autres méthodes sur les performances et le coût calculatoire.

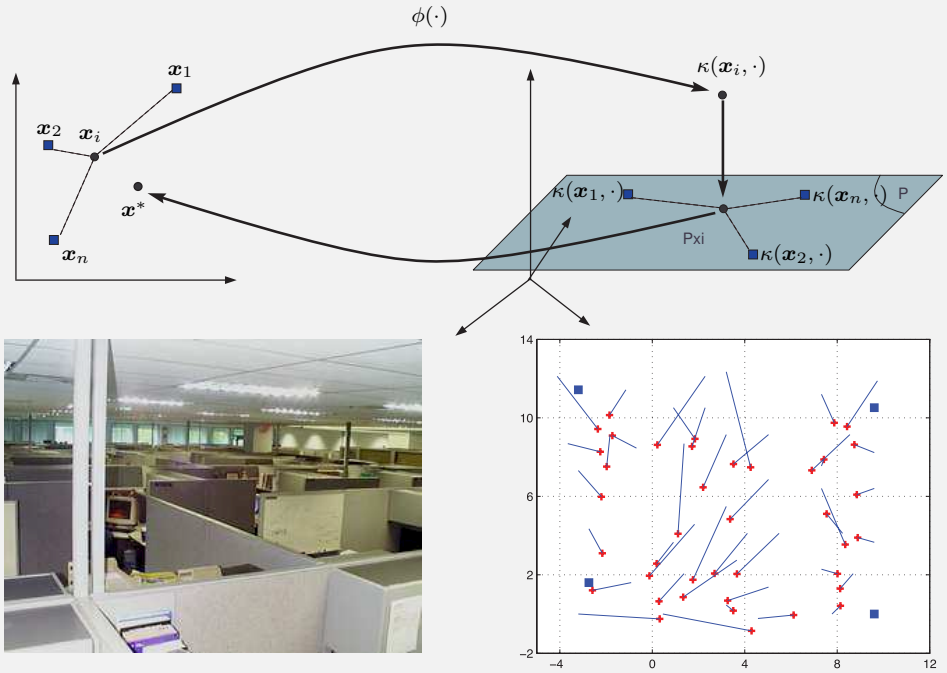
où $\kappa_c(\cdot,\cdot)$ est la version « centrée » du noyau $\kappa(\cdot,\cdot)$, avec

$$\begin{aligned}\kappa_c(x_i,x_j) &= \langle \kappa(x_i,\cdot) - \mu_\infty(\cdot), \kappa(x_j,\cdot) - \mu_\infty(\cdot)\rangle_{\mathbb{H}} \\ &= \kappa(x_i,x_j) - \mu_\infty(x_i) - \mu_\infty(x_j) + \|\mu_\infty(\cdot)\|^2_{\mathbb{H}}.\end{aligned}$$

La forme matricielle est obtenue en regroupant toutes les valeurs de retards, selon

$$\begin{bmatrix} \mathbb{E}[\kappa_c(x_i,x_{i-1})] \\ \mathbb{E}[\kappa_c(x_i,x_{i-2})] \\ \vdots \\ \mathbb{E}[\kappa_c(x_i,x_{i-m})] \end{bmatrix} = \begin{bmatrix} \mathbb{E}[\kappa_c(x_i,x_i)] & \mathbb{E}[\kappa_c(x_i,x_{i-1})] & \cdots & \mathbb{E}[\kappa_c(x_i,x_{i-m+1})] \\ \mathbb{E}[\kappa_c(x_i,x_{i-1})] & \mathbb{E}[\kappa_c(x_i,x_i)] & \cdots & \mathbb{E}[\kappa_c(x_i,x_{i-m+2})] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[\kappa_c(x_i,x_{i-m+1})] & \mathbb{E}[\kappa_c(x_i,x_{i-m+2})] & \cdots & \mathbb{E}[\kappa_c(x_i,x_i)] \end{bmatrix} \boldsymbol{\alpha}.$$

En conséquence, le vecteur des coefficients est obtenu par l'inversion d'une matrice de taille $(m \times m)$. Une fois ces coefficients estimés, nous pouvons prédire un échantillon inédit en utilisant le même modèle $\sum_{j=1}^{m} \alpha_j \, \kappa(x_{i-j},\cdot)$. Bien que cette dernière se trouve dans l'espace des caractéristiques $\mathbb{H}$,



[Cadre 7] **Auto-localisation de capteurs par la résolution du problème de pré-image.** Nous considérons les mesures RSSI acquises au cours d'une expérience réalisée dans les locaux du centre Motorola à Plantation en Floride (Etats-Unis) [Patwari et al., 2003]. Il s'agit d'un local de bureaux de 14×13 mètres, avec des partitions de hauteur 1.8 mètres (voir figure de gauche). Le réseau est constitué de 40 capteurs à positions inconnues, et 4 ancres déployées près des coins. Il faut noter la difficulté de la résolution d'une telle tâche, compte tenu de l'environnement, en présence des parois de cabines, des supports métalliques et en béton. Les locaux sont occupés par des bureaux, bibliothèques, classeurs métalliques et en bois, et des ordinateurs.
Le choix du noyau gaussien est corroboré par les expérimentations sur les mesures réelles [Patwari et al., 2003], sa largeur de bande étant estimée par le critère d'alignement noyau-cible. La figure de droite illustre les ancres (■), les positions estimées (✚), ainsi que les erreurs (lignes −). La méthode proposée produit une erreur quadratique moyenne de localisation de 4.53 m^2, un résultat à comparer avec l'estimateur du maximum de vraisemblance (qui s'est avéré biaisé) avec 4.75 m^2. Il faut noter la difficulté d'une telle tâche puisque certains capteurs ne sont pas situés dans le domaine de convexité des ancres. Voir [Essoloh et al., 2008] pour plus de détails et d'autres expérimentations.

il est nécessaire de retourner à l'espace des échantillons afin de déterminer l'échantillon prédit. Il s'agit alors du problème de pré-image qui peut être abordé par les différentes méthodes décrites dans le présent chapitre. Le Cadre 6 illustre l'approche AR-à-noyaux par la résolution du problème de pré-image, et présente une étude comparative sur les performances.

3.4 Auto-localisation de capteurs dans les réseaux sans fil

Les récentes avancées en micro-électronique et en communication numérique ont favorisé l'émergence des réseaux de capteurs sans fil, ouvrant la voie à de nouvelles applications dans des domaines tels que la sûreté, la surveillance et la sécurité. Chaque nœud du réseau est constitué d'un système miniaturisé, énergétiquement autonome, doté de capacités d'acquisition de données et de traitement. Une technologie sans fil leur permet de communiquer de proche en proche, sans hiérarchie centrale, de façon dynamique et instantanée, reconfigurable en fonction de l'évolution de la population de capteurs. Ce mode distribué présente l'avantage d'être particulièrement robuste aux attaques exté-

rieures et à la défaillance de nœuds puisqu'il est prévu que la perte de composants ne compromette pas l'efficacité du réseau dans son ensemble.

En l'absence d'information sur la position des nœuds d'un réseau de capteurs sans fil, au sein de l'environnement où ils sont déployés, les mesures récoltées peuvent s'avérer d'une utilité limitée. Le problème traité concerne l'auto-localisation de chacun de ces nœuds à partir de quelques capteurs dits ancres dont la position est connue, et de mesures de portée inter-capteurs. Afin de clarifier les propos, sans que cela nuise toutefois au caractère générique de l'approche, nous supposons qu'il s'agit de RSSI (acronyme de *Received Signal Strength Indication*), qui est une mesure de la puissance du signal reçu.

Le problème d'auto-localisation a fait l'objet de nombreux travaux de recherche, les solutions proposées variant selon le type de mesures de portée inter-capteurs considérées, la nature des hypothèses relatives à la propagation des signaux correspondants, etc. Ainsi des méthodes classiques d'estimation statistique [Patwari et al., 2003] côtoient-elles des techniques d'analyse de données de type MDS [Costa et al., 2006, Shang et al., 2003], ou encore de programmation semi-définie [Bachrach and Taylor, 2005, Doherty et al., 2001]. Les méthodes non-paramétriques ont récemment fait l'objet d'une attention particulière, compte tenu de leur flexibilité particulièrement appréciable dans le cadre des réseaux de capteurs. Dans [Patwari and Hero, 2004] par exemple, les auteurs proposent une technique de réduction de dimension basée sur l'apprentissage de la variété. L'auto-localisation est étudiée comme un problème de classification dans [Nguyen et al., 2005a], les données relatives aux ancres constituant l'ensemble d'apprentissage.

Reposant sur le concept de noyau reproduisant, nous proposons dans la suite deux méthodes non-paramétriques pour l'auto-localisation de capteurs. La première décrit l'auto-localisation sous la forme d'un problème de réduction de dimension dans l'espace RKHS. Afin de revenir à l'espace des positions des capteurs, il est nécessaire de résoudre explicitement le problème de pré-image. La seconde méthode exploite une technique de régression matricielle introduite dans le domaine de la bioinformatique [Yamanishi and Vert, 2007], où le problème résultant est similaire à celui de l'approche par transformation conforme décrit dans la section 3.2.2.

Considérons un réseau de $N+n$ capteurs constitué de n ancres dont les positions sont connues, et de N capteurs de coordonnées inconnues, avec $N \gg n$. Soit $\boldsymbol{x}_i$ le vecteur des coordonnées du i-ème capteur, avec $i \in \{1, 2, \ldots, n\}$ lorsqu'il s'agit d'une ancre. Soient $\boldsymbol{K}$ la matrice des mesures de RSSI entre chaque couple de capteurs, et $\boldsymbol{K}_{aa}$ la sous-matrice de $\boldsymbol{K}$ des informations ancre-ancre.

En s'appuyant sur le théorème de Moore-Aronszajn [Aronszajn, 1950], toute matrice symétrique définie positive $\boldsymbol{K}$ permet de définir un noyau reproduisant $\kappa(\cdot\,,\cdot)$ et son espace associé RKHS $\mathbb{H}$. L'espace $\mathbb{H}$ est engendré par les fonctions $\kappa(\boldsymbol{x}_i, \cdot)$, et complété de sorte que toute suite de Cauchy y converge. Le (i,j)-ème élément de $\boldsymbol{K}$ est alors $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \kappa(\boldsymbol{x}_i, \cdot), \kappa(\boldsymbol{x}_j, \cdot) \rangle_{\mathbb{H}}$. Il convient de préciser que les mesures de RSSI ne conduisent pas nécessairement à une matrice $\boldsymbol{K}$ qui soit symétrique définie positive. Il suffira dans ce cas de lui appliquer au préalable l'une des nombreuses transformations possibles afin de lui conférer cette propriété. Voir [Muñoz and de Diego, 2006] par exemple.

3.4.1 Auto-localisation par résolution explicite du problème de pré-image

La première méthode d'auto-localisation se décompose en trois phases, chacune de ces étapes peut être exécutée de manière distribuée à travers le réseau :

1. la première phase consiste à définir un noyau reproduisant approximant au mieux les mesures de portée inter-ancres, et ainsi construire un espace RKHS $\mathbb{H}$ approprié ;
2. la deuxième phase est destinée à construire un sous-espace pertinent de $\mathbb{H}$, à partir d'une technique de réduction de dimension appliquée aux informations liées aux éléments ancres ;
3. la troisième phase a pour but d'estimer la position d'un capteur (initialisée d'une manière aléatoire), par projection sur le sous-espace pertinent, puis par retour inverse à l'espace euclidien grâce à la résolution du problème de pré-image.

Ces étapes sont illustrées dans le schéma du Cadre 7, et résumées dans la suite.

Une étape préliminaire à tout traitement consiste à déterminer le noyau reproduisant qui permet au mieux de représenter les mesures de portée disponibles, et à estimer ses paramètres optimaux.

Ainsi cette étape induit-elle un sens physique aux résultats obtenus dans les étapes suivantes. Le critère d'alignement noyau-cible [Cristianini et al., 2002] permet d'optimiser, pour des mesures de similarité particulières, les paramètres d'un noyau donné, en maximisant « le cosinus de l'angle entre les matrices correspondantes ». D'une part, nous avons les mesures RSSI entre toute couple d'ancres (i, j), et d'autre part le noyau reproduisant appliqué à leurs coordonnées $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Le problème d'optimisation se réduit alors à l'estimation des paramètres optimaux de ce dernier. En considérant le noyau gaussien, avec $\kappa_\sigma(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2)$, sa largeur de bande σ est optimisée par la maximisation de l'expression suivante

$$\frac{\langle \boldsymbol{K}_\sigma, \boldsymbol{K}_{aa} \rangle_F}{\sqrt{\langle \boldsymbol{K}_\sigma, \boldsymbol{K}_\sigma \rangle_F \langle \boldsymbol{K}_{aa}, \boldsymbol{K}_{aa} \rangle_F}},$$

où le produit scalaire de Frobenius est utilisé. La méthode du Lagrangien permet de résoudre facilement ce problème. Voir [Essoloh et al., 2008] pour la résolution du problème d'optimisation.

L'identification du noyau reproduisant adapté aux mesures permet de définir un espace RKHS approprié. Dans cet espace, un sous-espace pertinent est déterminé à partir des mesures de portée ancres-ancres. Un tel espace est identifié à partir de techniques de type ACP-à-noyaux. Dans ce cas, les vecteurs propres $\boldsymbol{w}_k$ et les valeurs propres correspondantes λ_k sont obtenus par la résolution de l'équation suivante :

$$n\lambda_k \boldsymbol{w}_k = \boldsymbol{K}_{aa} \boldsymbol{w}_k.$$

Dans le cadre des réseaux de capteurs, la résolution de ce problème peut se faire d'une manière distribuée et/ou en ligne. C'est le cas de l'algorithme ACP-à-noyaux en ligne étudié au paragraphe 4.3 du chapitre suivant. Voir aussi [Honeine, 2012].

Ainsi, chaque capteur est représenté par son image dans l'espace caractéristique $\mathbb{H}$ déjà défini. La projection de cette image sur le sous-espace pertinent (défini par l'ACP-à-noyaux sur les ancres) permet de décrire le capteur dans la variété des ancres. Une initialisation aléatoire se retrouve alors dans le sous-espace de $\mathbb{H}$ défini par les ancres. Il reste toutefois le problème d'estimation des coordonnées dans l'espace initial, c'est à dire l'espace des positions. Ceci n'est autre que le problème de pré-image. A titre d'exemple, le CADRE 7 illustre la tâche d'auto-localisation de capteurs dans un local de bureaux.

3.4.2 Auto-localisation par régression de matrices de Gram

La méthode proposée ici exploite une technique de régression matricielle, afin de résoudre le problème de pré-image d'une manière implicite. La régression matricielle consiste en la détermination d'une fonction de régression liant deux matrices, symétriques définies positives dans notre contexte de matrices de Gram. La première matrice, totalement connue, rassemble ici les mesures de portée inter-capteurs, c'est à dire les RSSI. La seconde matrice regroupe les produits scalaires des vecteurs de coordonnées des capteurs, et n'est donc que partiellement connue au travers des couples ancre–ancre. La méthode proposée repose sur les informations ancre–ancre de chaque matrice, que nous utilisons pour compléter les entrées ancre–capteur de la matrice des produits scalaires.

Soit $\boldsymbol{P}$ la matrice de Gram $[\boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \cdots \;\; \boldsymbol{x}_{N+n}]^\top [\boldsymbol{x}_1 \;\; \boldsymbol{x}_2 \;\; \cdots \;\; \boldsymbol{x}_{N+n}]$ constituée à partir des coordonnées des capteurs. Celle-ci peut être décomposée en quatre sous-matrices $\boldsymbol{P}_{aa}, \boldsymbol{P}_{ac}, \boldsymbol{P}_{ca}$ et $\boldsymbol{P}_{cc}$ selon qu'elles impliquent les ancres $(_a)$ et/ou capteurs $(_c)$. Etant donnée la sous-matrice $\boldsymbol{P}_{aa}$, connue, nous souhaitons déterminer les autres sous-matrices. Nous disposons pour cela de la matrice $\boldsymbol{K}$ des mesures RSSI pour chaque couple de capteurs. Nous décomposons de même cette matrice en quatre sous-matrices $\boldsymbol{K}_{aa}, \boldsymbol{K}_{ac}, \boldsymbol{K}_{ca}$ et $\boldsymbol{K}_{cc}$. A partir des deux sous-matrices $\boldsymbol{P}_{aa}$ et $\boldsymbol{K}_{aa}$ associées aux couples ancre–ancre, le problème consiste à déterminer une application associant les deux matrices $\boldsymbol{P}$ et $\boldsymbol{K}$, comme illustré dans le schéma suivant où les matrices en gris sont inconnues :

$$\left[\begin{array}{c|c} \boldsymbol{K}_{aa} & \boldsymbol{K}_{ac} \\ \hline \boldsymbol{K}_{ca} & \boldsymbol{K}_{cc} \end{array}\right] \longrightarrow \left[\begin{array}{c|c} \boldsymbol{P}_{aa} & P_{ac} \\ \hline P_{ca} & P_{cc} \end{array}\right].$$

En appliquant le théorème de Moore-Aronszajn à la sous-matrice $\boldsymbol{K}_{aa}$, elle-même définie positive si $\boldsymbol{K}$ l'est, l'espace RKHS qui lui est associé est donc engendré par les n fonctions noyau $\kappa(\boldsymbol{x}_1,\cdot),\kappa(\boldsymbol{x}_2,\cdot),\ldots,\kappa(\boldsymbol{x}_n,\cdot)$. Nous proposons d'extraire m caractéristiques de cet espace, avec $m \leq n$, qui reflètent implicitement la topologie des nœuds du réseau. Il s'agit ici d'un ensemble de fonctions $\psi_1(\cdot),\psi_2(\cdot),\ldots,\psi_m(\cdot)$ qui, du fait de leur appartenance à l'espace fonctionnel susmentionné, peuvent se décomposer ainsi :

$$\psi_\ell(\cdot) = \sum_{i=1}^{n} \theta_{i,\ell}\, \kappa(\boldsymbol{x}_i,\cdot),$$

pour $\ell = 1,2,\ldots,m$. Soient $\boldsymbol{\psi}_{\boldsymbol{x}_i} = [\psi_1(\boldsymbol{x}_i) \quad \psi_2(\boldsymbol{x}_i) \quad \cdots \quad \psi_m(\boldsymbol{x}_i)]^\top$, c'est à dire le vecteur des évaluations des fonctions extraites $\psi_\ell(\cdot)$ en un $\boldsymbol{x}_i$ donné, et $\boldsymbol{\Theta}$ la matrice de taille $(n \times m)$ des coefficients de pondération $\theta_{i,\ell}$. Cette matrice est déterminée dans la suite en optimisant un critère liant les matrices $\boldsymbol{P}_{aa}$ et $\boldsymbol{K}_{aa}$. Pour cela, nous proposons le modèle (3.9) de conservation des produits scalaires, selon

$$\boldsymbol{x}_i^\top \boldsymbol{x}_j = \boldsymbol{\psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\psi}_{\boldsymbol{x}_j} + \epsilon_{i,j}$$

pour $i,j = 1,2,\ldots,n$, où $\epsilon_{i,j}$ est l'erreur de modélisation. En considérant la minimisation de la variance de cette dernière, nous retrouvons le même problème d'optimisation que (3.10)-(3.11), avec

$$\boldsymbol{\Theta}^* = \arg\min_{\boldsymbol{\Theta}} \|\boldsymbol{P}_{aa} - \boldsymbol{K}_{aa}\,\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\,\boldsymbol{K}_{aa}\|_F^2 + \eta\,\mathrm{tr}(\boldsymbol{\Theta}^\top\boldsymbol{\Theta}\,\boldsymbol{K}_{aa}).$$

La solution est donc donnée par (3.12), soit ${\boldsymbol{\Theta}^*}^\top\boldsymbol{\Theta}^* = \boldsymbol{K}_{aa}^{-1}\left(\boldsymbol{P}_{aa} - \eta\,\boldsymbol{K}_{aa}^{-1}\right)\boldsymbol{K}_{aa}^{-1}$.

Nous pouvons utiliser cette application associant les deux matrices $\boldsymbol{K}$ et $\boldsymbol{P}$ pour compléter cette dernière. Nous sommes à présent en mesure de déterminer les positions des capteurs non-ancres, en procédant en deux étapes. A partir du modèle considéré, nous notons tout d'abord que

$$\boldsymbol{P}_{ac}^* = \boldsymbol{K}_{aa}{\boldsymbol{\Theta}^*}^\top\boldsymbol{\Theta}^*\,\boldsymbol{K}_{ac}. \tag{3.16}$$

Nous remarquons ensuite que $\boldsymbol{P}_{aa} = \boldsymbol{X}_a^\top\boldsymbol{X}_a$ est une matrice de rang 2, puisque les capteurs sont dans un espace de dimension 2. Ceci permet de décomposer cette matrice sous la forme valeurs propres / vecteurs propres suivante $\boldsymbol{P}_{aa} = \boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}^\top$, les 2 valeurs propres non-nulles étant données par les éléments de la matrice diagonale $\boldsymbol{\Lambda}$, et les vecteurs propres correspondants constituant les deux colonnes de $\boldsymbol{W}$. Nous pouvons alors écrire $\boldsymbol{X}_a^\top = \boldsymbol{W}\boldsymbol{\Lambda}^{1/2}$ et, puisque $\boldsymbol{P}_{ac} = \boldsymbol{X}_a^\top\boldsymbol{X}_c$, on obtient

$$\boldsymbol{X}_c^* = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{W}^\top\boldsymbol{P}_{ac}^*.$$

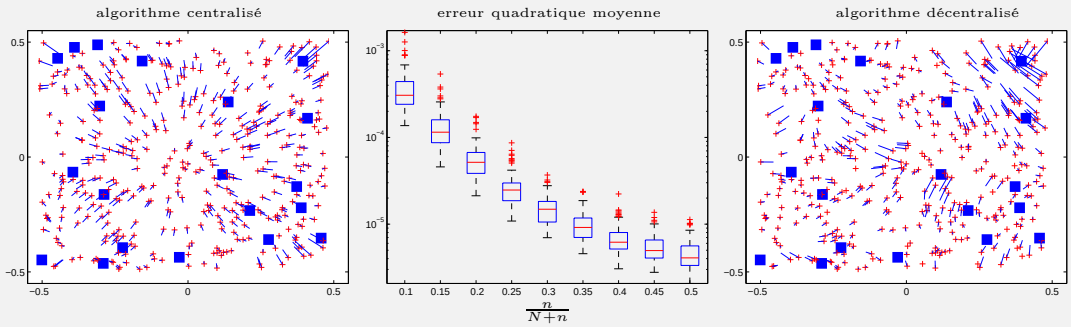
Précisons que ces coordonnées sont obtenues dans le repère défini par les vecteurs propres, ce qui nécessite un post-traitement par transformation affine pour les aligner sur le repère des ancres.

Le CADRE 8 illustre la pertinence de cette approche. Cette méthode est étudiée en détail dans [Honeine et al., 2008c]. Une approche distribuée, inhérente au caractère réparti des capteurs sans fil, est décrite dans [Honeine et al., 2009b].

3.5 Etude de cas : le problème NMF-à-noyaux [4]

Tout au long de ce chapitre, l'importance de la résolution du problème de pré-image est montrée. Chaque méthode à noyaux considérée se décompose en deux étapes : d'abord, un algorithme appliqué dans l'espace RKHS (souvent en s'inspirant d'un algorithme linéaire classique), ensuite une technique de résolution du problème de pré-image. C'est le cas en particulier avec la méthode AR-à-noyaux dans 3.3.2, l'auto-localisation dans les réseaux de capteurs sans fil dans 3.4.1, ou encore l'algorithme FastICA-à-noyaux présenté dans le CADRE 9. Jusqu'à présent, les techniques de pré-image s'opèrent en dernière étape de l'algorithme. Dans cette section, nous étudions la possibilité d'intégrer le problème de pré-image dans l'algorithme initial, afin de proposer un problème d'optimisation global, avec une résolution en une seule étape. Dans la suite, nous démontrons cette approche dans le cadre de la factorisation en matrices non négatives (NMF pour *nonnegative matrix factorization*) à noyaux.

4. C'est la seule partie du manuscrit qui n'a pas été publiée.



[Cadre 8] **Auto-localisation de capteurs par régression de matrices de Gram.**
Afin d'illustrer les performances de la méthode proposée, une expérience semblable à [Patwari and Hero, 2004] est considérée. L'atténuation des signaux y croît avec la distance inter-capteurs selon la loi $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma^2\right)$, où $\sigma = 0.75$ traduit la portée des capteurs. Dans notre étude, nous considérons un réseau de capteurs comprenant 420 nœuds uniformément distribués sur une région carrée de surface unité. Dans un premier temps, nous étudions l'algorithme centralisé, avec 20 ancres (■). La figure de gauche représente les positions réelles et estimées pour chacun des capteurs, ainsi que l'erreur d'estimation (lignes –). L'erreur quadratique moyenne sur l'estimation est de 0.0287. L'influence de la fraction d'ancres parmi les capteurs sur cette dernière est illustrée dans la figure au milieu, le nombre total des capteurs étant fixé à $N + n = 420$. La figure de droite concerne la mise en œuvre de l'algorithme sous forme distribuée, appliqué au même réseau. Chaque capteur y estime sa position à partir des informations provenant des 5 ancres les plus proches. Dans le cas de l'implémentation distribuée, l'erreur quadratique moyenne est de 0.0326.

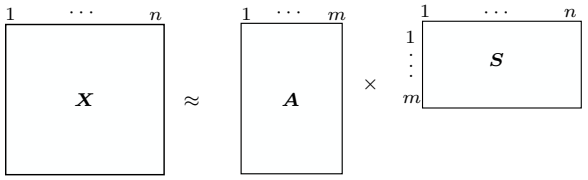
3.5.1 Introduction à la NMF linéaire

La factorisation en matrices non négatives (NMF) est très connue en traitement du signal et des images [Paatero and Tapper, 1994, Lee and Seung, 1999, Comon and Jutten, 2010]. Elle a été appliquée avec succès pour la classification d'images [Buchsbaum and Bloch, 2002], la reconnaissance de l'expression du visage [Buciu and Pitas, 2004], la reconnaissance d'objets [Liu and Zheng, 2004, Wild et al., 2004], ou encore en bioinformatique [Devarajan, 2008, Kim and Tidor, 2003, Li and Ding, 2006].

Elle consiste à approximer une matrice non négative par deux matrices non négatives de faible rang [Paatero and Tapper, 1994, Lee and Seung, 1999]. Pour une matrice $\boldsymbol{X}$ non négative, le problème consiste à estimer les deux matrices non négatives $\boldsymbol{A}$ et $\boldsymbol{S}$, telles que

$$\boldsymbol{X} \approx \boldsymbol{A}\boldsymbol{S}, \tag{3.17}$$

sous contraintes $\boldsymbol{A} \geq 0$ et $\boldsymbol{S} \geq 0$. Le schéma suivant illustre les matrices en question :



Soient $\boldsymbol{X} = [\boldsymbol{x}_1 \ \ \boldsymbol{x}_2 \ \cdots \ \boldsymbol{x}_n]$, $\boldsymbol{A} = [\boldsymbol{a}_1 \ \ \boldsymbol{a}_2 \ \cdots \ \boldsymbol{a}_m]$, et $s_{j,i}$ la (j,i)-ème composante de la matrice $\boldsymbol{S}$. La NMF consiste à estimer $\boldsymbol{a}_j \geq 0$ et $s_{j,i} \geq 0$, pour tout $j = 1, \ldots, m$ et $i = 1, \ldots, n$, tels que

$$\boldsymbol{x}_i \approx \sum_{j=1}^{m} s_{j,i}\, \boldsymbol{a}_j. \tag{3.18}$$

FastICA	FastICA-à-noyaux
1.Centrage et blanchiment des échantillons $\boldsymbol{x}_i \in \mathbb{X}$	1.Centrage et blanchiment dans l'espace $\mathbb{H}$
2.Initialisation aléatoire $\boldsymbol{W} = [\boldsymbol{w}_1 \ \cdots \ \boldsymbol{w}_m]^\top$ avec normalisation $\boldsymbol{w}_k \leftarrow \boldsymbol{w}_k / (\boldsymbol{w}_k^\top \boldsymbol{w}_k)^{\frac{1}{2}}$	2.Initialisation aléatoire $\boldsymbol{A} = [\boldsymbol{\alpha}_1 \ \cdots \ \boldsymbol{\alpha}_m]^\top$ avec normalisation $\boldsymbol{\alpha}_k \leftarrow \boldsymbol{\alpha}_k / (\boldsymbol{\alpha}_k^\top \boldsymbol{K} \boldsymbol{\alpha}_k)^{\frac{1}{2}}$
3.Orthogonalisation, avec $\boldsymbol{W} \leftarrow (\boldsymbol{W}\boldsymbol{W}^\top)^{-\frac{1}{2}} \boldsymbol{W}$	3.Orthogonalisation, avec $\boldsymbol{A} \leftarrow \boldsymbol{A} (\boldsymbol{A}^\top \boldsymbol{K} \boldsymbol{A})^{-\frac{1}{2}}$
4.Pour chaque ligne $\boldsymbol{w}_k^\top$ de $\boldsymbol{W}$, $\boldsymbol{w}_k \leftarrow \frac{1}{n}\big([g(\boldsymbol{w}_k^\top \boldsymbol{x}_1) \cdots g(\boldsymbol{w}_k^\top \boldsymbol{x}_n)]\boldsymbol{X}\big)^\top - \frac{1}{n}\sum_i g'(\boldsymbol{w}_k^\top \boldsymbol{x}_i)\, \boldsymbol{w}_k$	4.Pour chaque ligne $\boldsymbol{\alpha}_k^\top$ de $\boldsymbol{A}$, $\boldsymbol{\alpha}_k \leftarrow \frac{1}{n}[g(\psi_k(\boldsymbol{x}_1)) \cdots g(\psi_k(\boldsymbol{x}_n))]^\top - \frac{1}{n}\sum_i g'(\psi_k(\boldsymbol{x}_i))\boldsymbol{\alpha}_k$
5.Orthogonalisation, avec $\boldsymbol{W} \leftarrow (\boldsymbol{W}\boldsymbol{W}^\top)^{-\frac{1}{2}} \boldsymbol{W}$	5.Orthogonalisation, avec $\boldsymbol{A} \leftarrow \boldsymbol{A} (\boldsymbol{A}^\top \boldsymbol{K} \boldsymbol{A})^{-\frac{1}{2}}$
6.Retour à l'étape 4, jusqu'à convergence de $\boldsymbol{W}$.	6.Retour à l'étape 4, jusqu'à convergence de $\boldsymbol{A}$.
7.Composantes indépendantes : $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m \in \mathbb{X}$	7.Composantes indépendantes : $\psi_1(\cdot), \ldots, \psi_m(\cdot) \in \mathbb{H}$
	8.Pré-image pour trouver les composantes dans $\mathbb{X}$

[CADRE 9] **Analogie entre les algorithmes FastICA et FastICA-à-noyaux par pré-image.** Gauche : l'algorithme classique de FastICA permet d'estimer les composantes indépendantes $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_m$, en maximisant une mesure de non gaussianité [Hyvärinen et al., 2001]. Ici, $g(\cdot)$ est la dérivée première d'une fonction non quadratique et non linéaire, et $g'(\cdot)$ sa dérivée seconde. Les exemples les plus connus sont $g(\zeta) = \zeta \exp(-\zeta^2/2)$ et $g(\zeta) = \tanh(\zeta)$.
Droite : la version non linéaire, dite FastICA-à-noyaux, détermine des fonctions $\psi_1(\cdot), \ldots, \psi_m(\cdot) \in \mathbb{H}$ selon la forme donnée par le Théorème de Représentation (`Th.Rep`), avec $\psi_k(\cdot) = \sum_{i=1}^n \alpha_{k,i} \kappa(\boldsymbol{x}_i, \cdot)$. Soient les vecteurs des coefficients $\boldsymbol{\alpha}_1 \ldots, \boldsymbol{\alpha}_m$. Nous retrouvons une analogie avec l'algorithme classique de FastICA, où le produit scalaire $\boldsymbol{w}_k^\top \boldsymbol{x}_1$ est désormais $\psi_k(\boldsymbol{x}_1)$, puisque ce dernier correspond à un produit scalaire dans $\mathbb{H}$ avec $\psi_k(\boldsymbol{x}_1) = \langle \psi_k(\cdot), \kappa(\boldsymbol{x}_1, \cdot) \rangle_{\mathbb{H}}$. De même, la norme (quadratique) $\|\boldsymbol{w}_k\|^2 = \boldsymbol{w}_k^\top \boldsymbol{w}_k$ devient $\|\psi_k(\cdot)\|_H^2 = \sum_{i,j=1}^n \alpha_{k,i} \alpha_{k,j} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\alpha}_k^\top \boldsymbol{K} \boldsymbol{\alpha}_k$. La différence majeure dans cette analogie réside dans la dernière étape, avec la nécessité de résolution du problème de pré-image afin d'estimer les m composantes dans l'espace des observations $\mathbb{X}$.

Par souci de simplification, nous utilisons par convention $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, avec $m < n$. Avec la notation proposée ici, nous avons un ensemble d'apprentissage $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$, avec $\boldsymbol{x}_i \in \mathbb{X}$ où $\mathbb{X}$ désigne l'espace des observations.

3.5.2 La malédiction de la pré-image

Les algorithmes classiques de NMF, ainsi que la plupart de ses variantes, sont basés sur une hypothèse de mélange linéaire. Fournir des modèles non linéaires pour la NMF est un problème difficile. Les méthodes à noyaux fournissent un cadre élégant pour dériver des techniques non linéaires.

Récemment, plusieurs tentatives ont été faites pour développer des méthodes NMF non linéaires, dans le cadre des méthodes à noyaux [Zhang et al., 2006, Ding et al., 2010, Li and Ngom, 2012]. L'approche qui a été considérée consiste à appliquer une fonction non linéaire $\phi(\cdot)$ aux colonnes de $\boldsymbol{X}$, transformant ainsi chaque $\boldsymbol{x}_i$ en $\phi(\boldsymbol{x}_i)$. Soit $\kappa(\cdot, \cdot)$ le noyau reproduisant associé à cette transformation non linéaire, et $\mathbb{H}$ l'espace associé. Écrit dans cet espace, le modèle NMF devient

$$\phi(\boldsymbol{x}_i) \approx \sum_{j=1}^{m} s_{j,i}\, \boldsymbol{a}_j^\phi. \tag{3.19}$$

Ici, les éléments $\boldsymbol{a}_j^\phi$ appartiennent à l'espace $\mathbb{H}$. Soit $\boldsymbol{X}^\phi = \big[\phi(\boldsymbol{x}_1) \quad \phi(\boldsymbol{x}_2) \quad \cdots \quad \phi(\boldsymbol{x}_n)\big]$, le modèle (3.19) devient

$$\boldsymbol{X}^\phi \approx \big[\boldsymbol{a}_1^\phi \quad \boldsymbol{a}_2^\phi \quad \cdots \quad \boldsymbol{a}_m^\phi\big]\, \boldsymbol{S}.$$

Le modèle présenté ici a été investi dans [Zhang et al., 2006, Ding et al., 2010, Li and Ngom, 2012]. Malheureusement, ce modèle souffre d'une faiblesse majeure, héritée des méthodes à noyaux : nous

n'avons pas accès aux éléments du RKHS. Cette malédiction de la pré-image est illustrée ici avec les éléments $\boldsymbol{a}_j^\phi$ qui appartiennent à $\mathbb{H}$, ce qui nécessite de résoudre le problème de pré-image.

Il existe un autre inconvénient majeur du modèle (3.19) : il n'est pas direct d'imposer la non-négativité des éléments dans l'espace fonctionnel $\mathbb{H}$, et en particulier $\boldsymbol{a}_j^\phi$. Par conséquent, la contrainte $\boldsymbol{a}_j^\phi \geq 0$ devrait être abandonnée. Seuls les coefficients $s_{j,i}$ peuvent être non négatifs. Dans ce cas, il ne s'agit plus du problème NMF, mais plutôt du problème semi-NMF, où seule la contrainte $\boldsymbol{S} \geq 0$ est imposée, comme présenté dans [Li and Ngom, 2012]. Pour surmonter cette difficulté, les auteurs de [Pan et al., 2011] proposent d'approximer le noyau par une fonction associée à une transformation non négative, ce qui nécessite de résoudre un autre problème d'optimisation en pré-traitement, avant l'application de la NMF. En outre, le problème de pré-image doit être résolu par la suite.

Pour toutes ces raisons, l'application de la factorisation en matrices non négatives dans l'espace RKHS a été limitée jusqu'ici à des problèmes de classification, ou encore au noyau homogène [Buciu et al., 2008]. Dans la suite, nous démontrons que nous pouvons estimer les matrices $\boldsymbol{A}$ et $\boldsymbol{S}$ dans l'espace des observations, sans souffrir de la malédiction du problème de pré-image.

3.5.3 Méthode NMF-à-noyaux

Nous proposons une nouvelle méthode NMF-à-noyaux, où les matrices résultantes sont définies dans l'espace d'entrée, et donc sans la nécessité de résoudre le problème de pré-image. A cette fin, nous considérons le modèle de factorisation suivant :

$$\boldsymbol{X}^\phi \approx \boldsymbol{A}^\phi \boldsymbol{S},$$

où $\boldsymbol{A}^\phi = [\phi(\boldsymbol{a}_1) \quad \phi(\boldsymbol{a}_2) \quad \cdots \quad \phi(\boldsymbol{a}_m)]$. Par conséquent, nous avons le modèle suivant :

$$\phi(\boldsymbol{x}_i) \approx \sum_{j=1}^{m} s_{j,i}\, \phi(\boldsymbol{a}_j).$$

Cela signifie que nous estimons les éléments $\boldsymbol{a}_j$ directement dans l'espace des observations $\mathbb{X}$, par opposition au modèle (3.19) où les éléments $\boldsymbol{a}_j^\phi$ appartiennent à $\mathbb{H}$. La contrainte de non-négativité est imposée à la matrice $\boldsymbol{S}$ et aux vecteurs $\boldsymbol{a}_j$, pour tout $j = 1, 2, \ldots, m$.

Afin d'estimer tous les $\boldsymbol{a}_j$ et $s_{j,i}$, nous considérons une technique de moindres carrés alternés pour résoudre le problème d'optimisation suivant :

$$\min_{s_{j,i}, \boldsymbol{a}_j} \frac{1}{2} \sum_{i=1}^{n} \|\phi(\boldsymbol{x}_i) - \sum_{j=1}^{m} s_{j,i}\, \phi(\boldsymbol{a}_j)\|_{\mathbb{H}}^2. \tag{3.20}$$

En développant l'expression ci-dessus, le problème d'optimisation devient :

$$\min_{s_{j,i}, \boldsymbol{a}_j} \sum_{i=1}^{n} \Big(- \sum_{j=1}^{m} s_{j,i}\, \kappa(\boldsymbol{a}_j, \boldsymbol{x}_i) + \frac{1}{2} \sum_{j=1}^{m} \sum_{j'=1}^{m} s_{j,i} s_{j',i}\, \kappa(\boldsymbol{a}_j, \boldsymbol{a}_{j'}) \Big).$$

Soit $\mathcal{J}$ la fonction coût dans cette expression. En considérant sa dérivée par rapport à $s_{j,i}$, on obtient l'expression suivante :

$$\nabla_{s_{j,i}} \mathcal{J} = -\kappa(\boldsymbol{a}_j, \boldsymbol{x}_i) + \sum_{j'=1}^{m} s_{j',i}\, \kappa(\boldsymbol{a}_j, \boldsymbol{a}_{j'}).$$

En prenant le gradient de $\mathcal{J}$ par rapport au vecteur $\boldsymbol{a}_j$, nous avons :

$$\nabla_{\boldsymbol{a}_j} \mathcal{J} = \sum_{i=1}^{n} s_{j,i} \Big(- \nabla_{\boldsymbol{a}_j} \kappa(\boldsymbol{a}_j, \boldsymbol{x}_i) + \sum_{j'=1}^{m} s_{j',i}\, \nabla_{\boldsymbol{a}_j} \kappa(\boldsymbol{a}_j, \boldsymbol{a}_{j'}) \Big). \tag{3.21}$$

Ici, $\nabla_{\boldsymbol{a}_j} \kappa(\boldsymbol{a}_j, \cdot)$ désigne le gradient du noyau par rapport à son argument $\boldsymbol{a}_j$. Les expressions peuvent être facilement obtenues pour les différents noyaux, comme considéré dans la section 3.2.1. Voir la section 3.5.3.c pour le cas des noyaux linéaire et gaussien. Mais avant, nous décrivons deux algorithmes itératifs pour résoudre le problème NMF-à-noyaux, en alternant l'estimation de $s_{j,i}$ et $\boldsymbol{a}_j$. Dans la suite, nous désignons par $s_{j,i,t+1}$ et $\boldsymbol{a}_{j,t+1}$ leurs estimations à l'itération t.

3.5.3.a Algorithme avec une mise à jour additive

Dans le premier algorithme, nous proposons une règle de mise à jour additive pour résoudre le problème d'optimisation. Chaque itération est basée, d'abord sur une descente de gradient, en alternant sur $s_{j,i}$ et $\boldsymbol{a}_j$, et ensuite sur une fonction de redressement pour imposer la non-négativité.

En utilisant une descente de gradient, le coefficient $s_{j,i,t}$ est actualisé à l'itération t selon

$$s_{j,i,t+1} = s_{j,i,t} - \eta_{j,i,t}\, \nabla_{s_{j,i,t}} \mathcal{J},$$

où le pas $\eta_{j,i,t}$ peut prendre des valeurs différentes pour chaque couple (j,i). En remplaçant $\nabla_{s_{j,i,t}} \mathcal{J}$ par son expression, on obtient la règle de mise à jour suivante :

$$s_{j,i,t+1} = s_{j,i,t} - \eta_{j,i,t}\Big(\sum_{j'=1}^{m} s_{j',i,t}\, \kappa(\boldsymbol{a}_{j,t}, \boldsymbol{a}_{j',t}) - \kappa(\boldsymbol{a}_{j,t}, \boldsymbol{x}_{i,t})\Big). \tag{3.22}$$

Une procédure similaire est appliquée pour estimer les vecteurs $\boldsymbol{a}_j$. La règle de mise à jour obtenue à l'itération t est alors

$$\boldsymbol{a}_{j,t+1} = \boldsymbol{a}_{j,t} - \eta_{j,t} \nabla_{\boldsymbol{a}_{j,t}} \mathcal{J}, \tag{3.23}$$

où $\eta_{j,t}$ désigne le pas selon la j-ème direction et $\nabla_{\boldsymbol{a}_{j,t}} \mathcal{J}$ est le gradient donné par l'expression (3.21).

Afin d'imposer la non-négativité des matrices, les valeurs négatives obtenues par les mises à jour ci-dessus sont mises à zéro. Pour cette rectification, il suffit de remplacer chaque élément $s_{j,i,t+1}$ par $\max\{s_{j,i+1}; 0\}$. De même pour les éléments des vecteurs $\boldsymbol{a}_{j,t+1}$.

3.5.3.b Algorithme avec une mise à jour multiplicative

La règle de mise à jour additive est une procédure simple. Toutefois, la convergence est généralement lente, et dépend directement de la valeur du pas utilisée. Afin de surmonter ces problèmes, nous proposons une règle de mise à jour multiplicative, dans le même esprit que celle proposée dans [Lee and Seung, 1999] pour la méthode NMF conventionnelle.

Afin de proposer une règle de mise à jour multiplicative de $s_{j,i,t}$ à l'itération t, le pas $\eta_{j,i,t}$ est choisi de telle sorte que le premier et le troisième termes du membre de droite de l'équation (3.22) s'annulent, c'est à dire $\eta_{j,i,t} = s_{j,i,t} / \sum_{j'=1}^{m} s_{j',i,t}\, \kappa(\boldsymbol{a}_{j,t}, \boldsymbol{a}_{j',t})$. Par conséquent, en substituant cette expression du pas dans (3.22), on obtient la règle de mise à jour suivante :

$$s_{j,i,t+1} = s_{j,i,t} \times \frac{\kappa(\boldsymbol{a}_{j,t}, \boldsymbol{x}_i)}{\displaystyle\sum_{j'=1}^{m} s_{j',i,t}\, \kappa(\boldsymbol{a}_{j,t}, \boldsymbol{a}_{j',t})}. \tag{3.24}$$

Une procédure similaire est appliquée pour estimer les vecteurs $\boldsymbol{a}_{j,t}$, pour $j = 1, 2, \ldots, m$. L'astuce est que l'expression du gradient (3.21) peut toujours être décomposée selon $\nabla_{\boldsymbol{a}_j,t} \mathcal{J} = P - Q$, où P et Q sont deux termes non négatifs. Cette décomposition est connue dans la littérature par la méthode *split gradient* [Lantéri et al., 2011]. Il est évident que cette décomposition n'est pas unique. Pourtant, nous pouvons décrire une mise à jour multiplicative pour une fonction noyau donnée, comme indiqué dans la suite.

3.5.3.c Algorithmes pour différents noyaux

De retour à la NMF classique

Une propriété fondamentale de la méthode NMF-à-noyaux proposée est que la NMF classique en est un cas particulier, comme indiqué dans la suite. Pour cela, il suffit de considérer le noyau linéaire $\kappa(\boldsymbol{a}_j, \cdot) = \cdot^{\top} \boldsymbol{a}_j$, et dont le gradient est donné par $\nabla_{\boldsymbol{a}_j} \kappa(\boldsymbol{a}_j, \cdot) = \cdot$. En substituant ce résultat dans

les expressions ci-avant, on obtient les règles de mise à jour additive

$$\begin{aligned} s_{j,i,t+1} &= s_{j,i,t} - \eta_{j,i,t}\Big(\sum_{j'=1}^{m} s_{j',i,t}\,\boldsymbol{a}_{j,t}^{\top}\boldsymbol{a}_{j',t} - \boldsymbol{x}_i^{\top}\boldsymbol{a}_{j,t}\Big), \\ \boldsymbol{a}_{j,t+1} &= \boldsymbol{a}_{j,t} - \eta_{j,t}\sum_{i=1}^{n} s_{j,i,t}\Big(\sum_{j'=1}^{m} s_{j',i,t}\,\boldsymbol{a}_{j',t} - \boldsymbol{x}_i\Big), \end{aligned}$$

ainsi que les règles de mise à jour multiplicative

$$s_{j,i,t+1} = s_{j,i,t} \times \frac{\boldsymbol{x}_i^{\top}\boldsymbol{a}_{j,t}}{\sum_{j'=1}^{m} s_{j',i,t}\,\boldsymbol{a}_{j,t}^{\top}\boldsymbol{a}_{j',t}}, \qquad \boldsymbol{a}_{j,t+1} = \boldsymbol{a}_{j,t} \times \frac{\sum_{i=1}^{n} s_{j,i,t}\,\boldsymbol{x}_i}{\sum_{i=1}^{n} s_{j,i,t}\sum_{j'=1}^{m} s_{j',i,t}\,\boldsymbol{a}_{j',t}}.$$

Ces expressions sont celles de la bien connue NMF classique. Il est à noter que dans le cas du noyau linéaire, à savoir lorsque la transformation $\phi(\cdot)$ est l'opérateur identité, le problème d'optimisation (3.20) est équivalent à la minimisation de la norme de Frobenius entre les deux matrices $\boldsymbol{X}$ et $\boldsymbol{AS}$.

Cas du noyau gaussien

Le noyau gaussien $\kappa(\boldsymbol{a}_j, \cdot) = \exp(\frac{-1}{2\sigma^2}\|\boldsymbol{a}_j - \cdot\ \|^2)$ admet un gradient selon $\nabla_{\boldsymbol{a}_j}\kappa(\boldsymbol{a}_j, \cdot) = -\frac{1}{\sigma^2}\kappa(\boldsymbol{a}_j, \cdot)(\boldsymbol{a}_j - \cdot)$. Il est facile de déterminer la règle de mise à jour de $s_{j,i}$, aussi bien selon la règle additive que multiplicative. Pour la mise à jour de $\boldsymbol{a}_j$, on obtient la règle additive suivante :

$$\boldsymbol{a}_{j,t+1} = \boldsymbol{a}_{j,t} - \eta_{j,i,t}\Big(+\frac{1}{\sigma^2}\sum_{i=1}^{n} s_{j,i,t}\,\kappa(\boldsymbol{a}_{j,t}, \boldsymbol{x}_i)(\boldsymbol{a}_{j,t} - \boldsymbol{x}_i) - \frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j'=1}^{m} s_{j,i,t}s_{j',i,t}\,\kappa(\boldsymbol{a}_{j,t}, \boldsymbol{a}_{j',t})(\boldsymbol{a}_{j,t} - \boldsymbol{a}_{j',t})\Big).$$

Pour l'algorithme multiplicatif, nous écrivons le gradient selon une soustraction de deux termes non négatifs. Ceci est possible puisque toutes les matrices sont non négatives, ainsi que les valeurs du noyau. Nous obtenons la règle de mise à jour suivante :

$$\boldsymbol{a}_{j,t+1} = \boldsymbol{a}_{j,t} \otimes \frac{\sum_{i=1}^{n} s_{j,i,t}\Big(\boldsymbol{x}_i\,\kappa(\boldsymbol{a}_{j,t}, \boldsymbol{x}_i) + \sum_{j'=1}^{m} s_{j',i,t}\,\boldsymbol{a}_{j,t}\,\kappa(\boldsymbol{a}_{j,t}, \boldsymbol{a}_{j',t})\Big)}{\sum_{i=1}^{n} s_{j,i,t}\Big(\boldsymbol{a}_{j,t}\,\kappa(\boldsymbol{a}_{j,t}, \boldsymbol{x}_i) + \sum_{j'=1}^{m} s_{j',i,t}\,\boldsymbol{a}_{j',t}\,\kappa(\boldsymbol{a}_{j,t}, \boldsymbol{a}_{j',t})\Big)},$$

où $\otimes$ désigne le produit matriciel de Hadamard, et la division utilisée ici est aussi une opération matricielle terme à terme.

3.6 Conclusion et perspectives

Dans ce chapitre, j'ai décrit, d'une manière synthétique, l'un des travaux entamés après mes études doctorales, sur la définition et la résolution du problème de pré-image. Ce travail présente une nouvelle méthodologie pour résoudre ce problème. Son intérêt est décrit au travers de divers champs d'application en reconnaissance des formes et traitement du signal et des images. Les applications décrites sont celles du débruitage d'images sous contraintes, de la modélisation AR-à-noyaux de séries temporelles et de l'auto-localisation dans les réseaux de capteurs sans fil. Un effort particulier est porté sur le problème de la factorisation en matrices non négatives à noyaux, illustrant la malédiction de la pré-image. La solution proposée ouvre un nouvel espoir sur l'intégration de la résolution du problème de pré-image dans les méthodes à noyaux en vue d'une optimisation globale. Ces diverses

contributions sont complétées dans la section 5.5.2 (page 116), où nous proposons de traiter le problème de démélange en imagerie hyperspectrale par la résolution du problème de pré-image avec contraintes.

Le problème de pré-image reste un problème ouvert. Nous sommes persuadés que divers domaines devront profiter des avancées récentes sur le problème de pré-image et sa résolution. Ceci a déjà été démontré récemment pour l'analyse de séquences biologiques [Sonnenburg et al., 2008] et l'analyse de chaines de caractèrs [Cortes et al., 2005]. Nous sommes aussi convaincus que ces avancées devront permettre de développer des méthodes à noyaux dans des terrains souvent délaissés, dont l'automatique. Un thème majeur en automatique, ainsi qu'en traitement du signal, est le filtre de Kalman. Dans ce cas, il est naturel de définir deux espaces, celui des observations et celui des caractéristiques (ou espace d'état). En utilisant le cadre non linéaire des méthodes à noyaux, le coup du noyau permet d'estimer les états, et la résolution de la pré-image détermine la solution dans l'espace des observations.

La nature n'est qu'un dictionnaire.
[Charles Baudelaire]

Tout est prédit par le dictionnaire.
[Paul Valéry]

4

Apprentissage en ligne et apprentissage collaboratif : apprendre avec parcimonie

Sommaire

Confronté à un environnement non-stationnaire et dynamique, un apprentissage en ligne peut s'avérer incontournable. Les méthodes à noyaux n'apportent hélas pas de réponse directe et satisfaisante à cette question, la taille des modèles qu'elles engendrent étant égale au nombre de couples entrée/sortie utilisés. Il est alors primordial de contrôler l'ordre du modèle. Pour cela, un critère de parcimonie est nécessaire pour restreindre le modèle à un sous-espace pertinent, ce dernier étant défini par une sélection appropriée des éléments d'apprentissage. Ce problème est inhérent en traitement de l'information dans les réseaux de capteurs sans fil, où un apprentissage parcimonieux et distribué est naturel au caractère réparti des nœuds du réseau.

J'ai commencé cette activité de recherche au cours de mes études de doctorat, dans le cadre de l'identification non linéaire de systèmes non stationnaires à sortie unique. Depuis, je l'ai poursuivi activement pour développer davantage l'apprentissage en ligne, avec une méthodologie originale pour répondre à de récents défis dans des domaines aussi multiples que variés. Dans le cadre du projet KernSig (ANR, programme blanc), mes travaux en post-doctorat m'ont permis d'étendre cette activité au problème de l'apprentissage collaboratif. Depuis, elle a été financée par plusieurs projets sur le traitement de l'information dans les réseaux de capteurs sans fil, dont la plateforme

CAPSEC du Contrat de Projets État-Région. Les doctorants Mehdi Essoloh (sur les réseaux de capteurs), Jie Chen (sur l'identification en ligne et distribuée), Zineb Noumir (sur l'apprentissage mono-classe en ligne) et Chafic Saidé (sur l'adaptation du dictionnaire) ont contribué à cette activité de recherche. Ces travaux ont donné lieu à quatre publications de revues [1], un chapitre de livre [2], et une vingtaine d'articles publiés dans des actes de conférences [3]. Cette activité a permis d'initier et de développer une collaboration internationale avec José C. M. Bermudez, de Federal University of Santa Catarina (Brazil). La poursuite de ces travaux est évidement d'actualité, notamment dans le cadre du projet récemment accepté ODISSEE (ANR, programme ASTRID, 2014) auquel je participe activement.

Ce chapitre s'efforce de décrire, d'une manière synthétique, les deux objectifs visés : d'abord, au niveau méthodologique avec l'élaboration de critères de parcimonie efficaces, d'une part en exploitant la sortie désirée et l'erreur de modélisation, et d'autre part en adaptant les éléments du dictionnaire pour une meilleure couverture ; ensuite, au niveau applicatif avec le développement de nouvelles classes d'algorithmes d'apprentissage en ligne, selon les axes suivants :

- identification en ligne de systèmes non linéaires à sorties multiples (voir [Saidé et al., 2013b]) ;
- analyse en composantes principales à noyaux avec un algorithme en ligne (voir la section 4.3) ;
- traitement collaboratif de l'information dans les réseaux de capteurs (voir la section 4.4).

Les travaux résumés dans ce chapitre se poursuivent dans le chapitre 6, avec la détection séquentielle par approche mono-classe en ligne. Voir la section 6.2 à la page 123.

4.1 Problématique et état de l'art

Dans la cadre d'un apprentissage en ligne, une nouvelle observation $(\boldsymbol{x}_\ell, y_\ell)$ est disponible à l'instant ℓ et le modèle est mis à jour de manière récursive basée sur cette nouvelle information. Toutefois, compte tenu du Théorème de Représentation avec un modèle de la forme (`Th.Rep`), l'ordre du modèle est alors en perpétuelle augmentation. Afin de contourner cet inconvénient, il faut le contrôler, en ne conservant qu'une petite fraction de fonctions noyau dans le développement (`Th.Rep`). En désignant par m le nombre de ces éléments, nous écrivons

$$\psi_\ell(\cdot) = \sum_{j=1}^{m} \alpha_{\ell,j}\, \kappa(\boldsymbol{x}_{\omega_j}, \cdot), \tag{4.1}$$

où $\mathcal{D} = \{\boldsymbol{x}_{\omega_1}, \boldsymbol{x}_{\omega_2}, \ldots, \boldsymbol{x}_{\omega_m}\}$ désigne un ensemble, dit dictionnaire, des m échantillons sélectionnés parmi ceux disponibles jusqu'à l'instant ℓ et $\alpha_{\ell,j}$ désignent les coefficients optimaux à cet instant. La constitution du dictionnaire dépend de l'instant ℓ (à lire $\mathcal{D}_\ell$), ainsi que l'ordre du modèle m (à lire m_ℓ) et les coefficients de pondération (écrits aussi avec un indice $_\ell$).

Le critère de sélection (dit de parcimonie) détermine s'il est nécessaire de conserver la fonction noyau candidate $\kappa(\boldsymbol{x}_\ell, \cdot)$ dans le modèle, pour l'enrichir d'un nouveau terme. Le dictionnaire est alors augmenté de l'élément courant. Ainsi la représentation parcimonieuse (compacte) est-elle définie par le critère de parcimonie utilisé. Plusieurs critères sont étudiés dans la littérature des méthodes à noyaux, comme rappelé dans la suite.

Notation

Dans la suite, nous désignons par $\boldsymbol{K}_{\mathcal{D}}$ la matrice de Gram de taille $(m \times m)$ associée aux éléments du dictionnaire $\mathcal{D}$, c'est à dire d'éléments $\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$, pour tout $(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j}) \in \mathcal{D} \times \mathcal{D}$. Le vecteur $\boldsymbol{\kappa}_{\mathcal{D}}(\cdot)$

1. [Saidé et al., 2013b, Honeine, 2012, Honeine et al., 2010, Richard et al., 2009]
2. [Noumir et al., 2013]
3. [Mahfouz et al., 2013c, Ghadban et al., 2013b, Saidé et al., 2013a, Mahfouz et al., 2013a, Noumir et al., 2012e, Saidé et al., 2012, Noumir et al., 2012d, Chen et al., 2012b, Noumir et al., 2012c, Richard et al., 2011, Chen et al., 2011f, Chen et al., 2010a, Chen et al., 2010b, Richard et al., 2010b, Chen et al., 2010c, Essoloh et al., 2009, Honeine et al., 2009a, Honeine et al., 2008b, Honeine et al., 2008a, Honeine et al., 2007a, Honeine et al., 2007b]

désigne le vecteur de taille $(m \times 1)$ représentant l'évaluation de son argument en chaque élément du dictionnaire $\mathcal{D}$. En d'autres termes, pour toute fonction $\psi(\cdot)$ de $\mathbb{H}$, le j-ème élément de $\boldsymbol{\kappa}_{\mathcal{D}}(\psi(\cdot))$ n'est autre que l'évaluation de $\psi(\cdot)$ en $\boldsymbol{x}_{\omega_j}$, puisque

$$[\boldsymbol{\kappa}_{\mathcal{D}}(\psi(\cdot))]_j = \langle \kappa(\boldsymbol{x}_{\omega_j}, \cdot), \psi(\cdot) \rangle_{\mathbb{H}} = \psi(\boldsymbol{x}_{\omega_j}).$$

En particulier, nous avons $[\boldsymbol{\kappa}_{\mathcal{D}}(\kappa(\boldsymbol{x}_i, \cdot))]_j = \langle \kappa(\boldsymbol{x}_{\omega_j}, \cdot), \kappa(\boldsymbol{x}_i, \cdot) \rangle_{\mathbb{H}} = \kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_i)$. Par abus de notation, nous noterons par $\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_i)$ le vecteur $\boldsymbol{\kappa}_{\mathcal{D}}(\kappa(\boldsymbol{x}_i, \cdot))$. Pour terminer, les coefficients $\alpha_{\ell,j}$ sont regroupés dans le vecteur $\boldsymbol{\alpha}_\ell$ de taille $(m \times 1)$.

4.1.1 Etat de l'art

Une première approche concerne l'approximation linéaire, présentée initialement dans [Baudat and Anouar, 2001] pour la classification et la régression et dans [Csató and Opper, 2001] pour les processus gaussiens, puis étendue aux méthodes à noyaux dans [Engel et al., 2004]. La fonction noyau candidate est alors écartée du modèle si elle peut être suffisamment représentée par les éléments de celui-ci. Ceci est possible en comparant la fonction noyau candidate à sa projection sur l'espace engendré par les m éléments du dictionnaire. L'élément $\kappa(\boldsymbol{x}_\ell, \cdot)$ est alors ajouté au dictionnaire si la règle d'approximation linéaire suivante est satisfaite :

$$\min_{\boldsymbol{\gamma}} \|\kappa(\boldsymbol{x}_\ell, \cdot) - \sum_{j=1}^{m} \beta_j \, \kappa(\boldsymbol{x}_{\omega_j}, \cdot)\|_{\mathbb{H}}^2 > \epsilon_0^2, \tag{4.2}$$

où ϵ_0 est un seuil prédéfini qui détermine le niveau de parcimonie du modèle. De l'algèbre linéaire, on obtient $\kappa(\boldsymbol{x}_\ell, \boldsymbol{x}_\ell) - \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell)^\top \boldsymbol{K}_{\mathcal{D}}^{-1} \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell) > \epsilon_0^2$, ce qui nécessite l'inversion de la matrice de Gram avec une complexité calculatoire en $\mathcal{O}(m^3)$.

Une approche moins gourmande en coût calculatoire est le critère de l'entropie quadratique de Rényi. En effet, elle permet de mesurer la quantité de désordre dans un dictionnaire. Son approximation selon l'estimateur de Parzen avec une fenêtre gaussienne est donnée par

$$-\log\left(\frac{1}{m^2}\sum_{i,j=1}^{m} \frac{\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})}{(2\pi\sigma^2)^{\dim(\mathbb{X})/2}}\right). \tag{4.3}$$

Cette expression montre que la somme des termes de la matrice de Gram caractérise la diversité des éléments du dictionnaire [Girolami, 2002]. Ainsi cette propriété est-elle investie dans [Suykens et al., 2002] pour proposer une technique d'élagage baptisée *fixed-size least-squares support vector machines*.

4.1.2 Travaux réalisés en doctorat

Nous avons montré dans [Honeine et al., 2007b, Richard et al., 2009] qu'il existe un critère plus efficace que le critère de l'approximation linéaire (4.2), par une simple analyse du dictionnaire ainsi construit. Il s'agit du critère de cohérence, ou encore de la famille de critères qu'il engendre dont la fonction de Babel. Le critère de cohérence pour définir un dictionnaire est décrit comme suit : la fonction noyau candidate $\kappa(\boldsymbol{x}_\ell, \cdot)$ n'est pas ajoutée dans le dictionnaire si la cohérence de ce dernier dépasse un seuil prédéfini, c'est à dire si

$$\max_{j=1,\ldots,m} \operatorname{coh}(\kappa(\boldsymbol{x}_\ell, \cdot), \kappa(\boldsymbol{x}_{\omega_j}, \cdot)) > \nu_0, \tag{4.4}$$

où la cohérence entre deux fonctions noyau est

$$\operatorname{coh}(\kappa(\boldsymbol{x}_i, \cdot), \kappa(\boldsymbol{x}_j, \cdot)) = \frac{|\langle \kappa(\boldsymbol{x}_i, \cdot), \kappa(\boldsymbol{x}_j, \cdot) \rangle_{\mathbb{H}}|}{\|\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}} \|\kappa(\boldsymbol{x}_j, \cdot)\|_{\mathbb{H}}}. \tag{4.5}$$

Il s'agit alors d'imposer une borne supérieure au cosinus de l'angle entre chaque couple de fonctions noyau, c'est à dire

$$\max_{\substack{i,j=1,2,\ldots,m \\ i\neq j}} \operatorname{coh}(\kappa(\boldsymbol{x}_{\omega_i},\cdot),\kappa(\boldsymbol{x}_{\omega_j},\cdot)) \leq \nu_0. \tag{4.6}$$

Le dictionnaire ainsi construit est dit ν_0-cohérent. Le seuil ν_0 détermine le niveau de parcimonie, où une valeur nulle produit un dictionnaire de bases orthogonales. Ce critère est très efficace d'un point de vue calculatoire, puisque le numérateur de l'expression (4.5) est donné par $|\kappa(\boldsymbol{x}_\ell,\boldsymbol{x}_{\omega_j})|$ et son dénominateur par $(\kappa(\boldsymbol{x}_\ell,\boldsymbol{x}_\ell)\,\kappa(\boldsymbol{x}_{\omega_j},\boldsymbol{x}_{\omega_j}))^{\frac{1}{2}}$, c'est à dire

$$\operatorname{coh}(\kappa(\boldsymbol{x}_i,\cdot),\kappa(\boldsymbol{x}_j,\cdot)) = \frac{|\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)|}{\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}}.$$

La valeur du dénominateur vaut 1 pour les noyaux de norme unité, comme c'est le cas du noyau gaussien.

Nous avons démontré plusieurs propriétés d'un dictionnaire issu du critère de cohérence, et a fortiori du modèle résultant, comme résumé succinctement dans la suite avec quelques éléments de preuves. Sans perte de généralité, nous présentons ici le cadre des noyaux de norme unité. Voir [Honeine, 2007] pour plus de détails, ainsi que les démonstrations.

- La taille du dictionnaire ainsi obtenu est finie et *a fortiori* l'ordre du modèle $\psi_\ell(\cdot)$. Pour démontrer cela, nous rappelons que la propriété de compacité de l'espace d'observation $\mathbb{X}$ est généralisée à l'ensemble $\{\kappa(\boldsymbol{x},\cdot)\}_{\boldsymbol{x}\in\mathbb{X}}$, grâce à la continuité de $\kappa(\boldsymbol{x},\cdot)$. Il existe alors un ensemble de boules de rayons non-nuls, définies selon la norme ℓ_2, pouvant couvrir ces fonctions noyau. Or un dictionnaire ν_0-cohérent vérifie $\|\kappa(\boldsymbol{x}_{\omega_i},\cdot)-\kappa(\boldsymbol{x}_{\omega_j},\cdot)\|^2_{\mathrm{H}} = 2-2\kappa(\boldsymbol{x}_{\omega_i},\boldsymbol{x}_{\omega_j}) \geq 2-2\nu_0$. Cette borne impose alors un nombre fini de ces boules.
- Les m fonctions noyau d'un dictionnaire ν_0-cohérent sont linéairement indépendantes si $(m-1)\,\nu_0 < 1$. Cette condition suffisante est démontrée en rappelant un résultat très connu en algèbre linéaire : une matrice à diagonale dominante est non singulière. Appliqué à la matrice de Gram associée au dictionnaire, nous avons $|\kappa(\boldsymbol{x}_{\omega_j},\boldsymbol{x}_{\omega_j})| > \sum_{i\neq j}^{m} |\kappa(\boldsymbol{x}_{\omega_i},\boldsymbol{x}_{\omega_j})|$, pour tout $j=1,2,\ldots,m$, c'est à dire la condition suivante $(m-1)\,\nu_0 < 1$ pour un dictionnaire ν_0-cohérent.
- Tout dictionnaire déterminé à l'aide de la règle de cohérence vérifie le critère d'approximation linéaire (4.2), avec $\epsilon_0^2 > 1-(m-1)\nu_0^2/(1-(m-1)\nu_0)$. Ainsi garantissons-nous un résultat équivalent avec un coût calculatoire moindre. Voir [Honeine, 2007, Proposition 7.4 et Proposition 7.6] pour la démonstration.
- Il existe un lien direct avec l'entropie quadratique de Rényi, puisqu'un dictionnaire issu du critère de cohérence admet une entropie quadratique de Rényi bornée inférieurement par une fonction croissante de ν_0 et décroissante de m. Pour un noyau de norme unité, cette borne vaut $C_0 - \log((1+(m-1)\nu_0)/m)$, où C_0 désigne une constante dépendante du noyau.
- Le lien avec l'ACP-à-noyaux est moins évident. Dans [Honeine, 2007, Proposition 7.7], nous avons établi une borne supérieure sur l'erreur quadratique d'approximer une fonction principale par les éléments d'un dictionnaire ν_0-cohérent obtenu par le critère de cohérence. Cette borne, inversement proportionnelle à la valeur propre correspondante, montre que les fonctions propres associées aux plus grandes valeurs propres appartiennent, à une faible erreur près, à l'espace engendré par les éléments du dictionnaire.

4.2 Synthèse des contributions récentes

Les critères de parcimonie les plus étudiés dans la littérature sont le critère d'approximation linéaire défini par l'expression (4.2), l'entropie de Rényi avec l'estimateur (4.3), ainsi que le critère de cohérence donné par l'expression (4.4). Ces critères examinent les propriétés du dictionnaire afin que ses éléments puissent, à un niveau donné, couvrir le domaine. Ils permettent d'exploiter les travaux réalisés sur les dictionnaires en théorie de l'approximation [Gilbert et al., 2003, Tropp, 2004], ou encore des résultats bien

connus dans le cadre de dictionnaires temps-fréquence [Mallat and Zhang, 1993]. Ces critères de parcimonie sont souvent appliqués pour proposer des techniques d'identification en ligne [Engel et al., 2004, Honeine et al., 2007b, Richard et al., 2009], de classification et de régression [Baudat and Anouar, 2001, Csató and Opper, 2001, Suykens et al., 2002].

Les critères étudiés jusqu'à présent sont de nature *a priori*, c'est à dire qu'ils sont appliqués en pré-traitement, indépendamment du système étudié et du modèle résultant. En d'autres termes, ils ne tiennent pas compte que de l'observation $\boldsymbol{x}_\ell$. Pourtant, le domaine d'application a toujours privilégié l'apprentissage supervisé, et en particulier l'identification en ligne, en présence d'un couple entrée-sortie $(\boldsymbol{x}_\ell, y_\ell)$ disponible à chaque instant ℓ. Paradoxalement, la sortie désirée n'est pas prise en compte dans la règle de construction du dictionnaire, bien qu'il soit facile de déterminer l'erreur entre la sortie désirée y_ℓ et celle estimée $\psi_\ell(\boldsymbol{x}_\ell)$. Ces critiques que nous avons faites nous ont ouvert la voie à de développements originaux.

Nos récentes contributions sont multiples, aussi bien au niveau de la construction du dictionnaire qu'au niveau de nouveaux algorithmes en ligne. Au niveau de la construction du dictionnaire, deux axes sont considérés. Dans le premier axe, nous revisitons le critère de cohérence afin de proposer un critère de parcimonie dit *a posteriori*, le dictionnaire étant construit en examinant la fonction candidate $\psi_\ell(\cdot)$ et non seulement d'échantillon courant $\boldsymbol{x}_\ell$. Ainsi, une information sur le modèle résultant est-elle retenue pour la construction du dictionnaire, à l'opposé des approches précédentes où seul le niveau de couverture du dictionnaire est considéré. Dans le second axe, nous proposons d'adapter les éléments du dictionnaire, afin de les améliorer en minimisant l'erreur entre la sortie désirée y_ℓ et celle estimée $\psi_\ell(\boldsymbol{x}_\ell)$. Ainsi les éléments du dictionnaire sont-ils remis en question à chaque instant ℓ. L'approche proposée est alors complémentaire à tous les critères de parcimonie utilisés jusqu'à présent où les éléments du dictionnaire sont fixes vis-à-vis des échantillons en entrée du système étudié, et demeurent inchangés.

Nous proposons de diversifier les problèmes traités, afin d'ouvrir la voie à de champs d'application aussi multiples que variés. Pour cela, nous montrons l'intérêt de la parcimonie au travers des développements suivants :

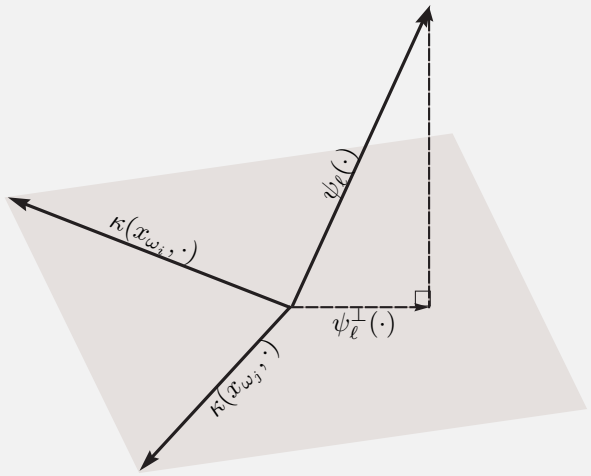
- identification en ligne de systèmes non linéaires à sorties multiples (voir [Saidé et al., 2013b] pour un résumé avec une application sur l'analyse de signaux EMG) ;
- ACP-à-noyaux avec un algorithme en ligne (voir la section 4.2.3.a pour un résumé et section 4.3 pour plus de détails) ;
- traitement collaboratif de l'information dans les réseaux de capteurs sans fil (voir la section 4.4) ;
- détection séquentielle par approche mono-classe en ligne (voir la section 4.2.3.b et section 6.2 à la page 123 pour plus de détails).

Il convient à présent de décrire, d'une manière synthétique, ces récentes contributions.

4.2.1 Parcimonie *a posteriori* : le critère de cohérence, revisité

Les critères présentés précédemment sont de nature *a priori*, reposant uniquement sur la topologie des échantillons construisant le dictionnaire, indépendamment de la pertinence du modèle résultant. C'est le cas des critères de l'approximation linéaire, de l'entropie quadratique de Rényi, ou encore de la cohérence du dictionnaire. Aujourd'hui encore, peu de travaux considèrent des critères *a posteriori* dans la construction du dictionnaire, c'est à dire en analysant le modèle résultant $\psi_\ell(\cdot)$. Citons toutefois [Dodd and Harris, 2002, Dodd et al., 2003] où les auteurs opèrent par projections séquentielles où le critère de parcimonie est donné par la valeur de l'erreur de projection.

Nous proposons un critère de parcimonie *a posteriori*, qui prend en compte la pertinence du modèle résultant, et donc sa mise à jour à chaque instant. Pour cela, nous revisitons la mesure de cohérence en l'appliquant à la fonctionnelle $\psi_\ell(\cdot)$ qu'est le modèle, au lieu de $\kappa(\boldsymbol{x}_\ell, \cdot)$, ce qui permet de définir un critère de cohérence fonctionnelle comme suit. Soit la fonction $\psi_\ell(\cdot)$ déterminée avec $\kappa(\boldsymbol{x}_\ell, \cdot)$ dans le développement, c'est à dire le modèle augmenté à l'ordre $m+1$. Nous proposons de réduire l'ordre du modèle, en éliminant cette contribution, si $\psi_\ell(\cdot)$ n'est pas assez pertinente, par



[CADRE 10] **Illustration de la cohérence fonctionnnelle, comme critère de parcimonie *a posteriori*.**

rapport aux fonctions précédemment estimées, avec

$$\max_{k=1,\ldots,\ell-1} \mathrm{coh}(\psi_\ell(\cdot), \psi_k(\cdot)) > \nu_0$$

où $\psi_k(\cdot)$ désigne la fonction optimale obtenue à l'instant k. Cette expression de cohérence que nous proposons permet de la définir sur tout $\mathbb{H} \times \mathbb{H}$, et non pas seulement sur les fonctions noyau comme c'était le cas dans la définition (4.5). En conséquence, cette borne sur la cohérence fonctionnelle correspond à une extension de (4.6), en considérant la limite supérieure du cosinus des angles entre les fonctions, afin « d'étendre au plus large » le sous-espace correspondant.

Bien que cela nécessite de garder toutes les fonctions précédentes, nous proposons une condition suffisante avec le critère

$$\mathrm{coh}(\psi_\ell(\cdot), \psi_\ell^\perp(\cdot)) > \nu_1, \tag{4.7}$$

pour un seuil donné $\nu_1 \in [0, 1]$ et où $\psi_\ell^\perp(\cdot)$ désigne la projection orthogonale de $\psi_\ell(\cdot)$ sur le sous-espace engendré par les $\ell - 1$ fonctions. Voir l'illustration dans le CADRE 10. Par ailleurs, on n'a pas besoin de garder toutes ces fonctions en mémoire, puisque le sous-espace correspondant est aussi engendré par les fonctions noyau retenues. Il convient d'évaluer $\psi_\ell^\perp(\cdot) = \sum_{j=1}^m \beta_j \, \kappa(\boldsymbol{x}_{\omega_j}, \cdot)$, en minimisant $\|\psi_\ell(\cdot) - \psi_\ell^\perp(\cdot)\|_{\mathbb{H}}^2$, soit

$$\min_{\boldsymbol{\beta}} \left\| \alpha_{m+1} \, \kappa(\boldsymbol{x}_\ell, \cdot) - \sum_{j=1}^m (\beta_j - \alpha_j) \, \kappa(\boldsymbol{x}_{\omega_j}, \cdot) \right\|_{\mathbb{H}}^2. \tag{4.8}$$

En annulant la dérivée de cette fonction coût quadratique par rapport aux variables β_j, l'optimum est donné par

$$\boldsymbol{\beta} = \boldsymbol{\alpha} + \alpha_\ell \, \boldsymbol{K}_{\mathcal{D}}^{-1} \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell),$$

avec $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, et $\kappa(\boldsymbol{x}_\ell)$ les vecteurs de taille $(m \times 1)$ d'éléments α_j, β_j, et $\kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_\ell)$, respectivement. La matrice $\boldsymbol{K}_{\mathcal{D}}$ est de terme général $\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$, pour $i = 1, 2, \ldots, m$.

Dans la section 4.4 à la page 85, nous montrons la pertinence de ce critère de parcimonie *a posteriori*, dans le cadre de l'apprentissage collaboratif dans les réseaux de capteurs sans fil.

4.2.2 Adaptation des éléments du dictionnaire

A priori ou *a posteriori*, tous les critères de parcimonie abordés jusqu'à présent reposent sur un principe intuitif très attractif : les fonctions noyau du dictionnaire reflètent une couverture raisonnable de l'espace en question. Nous déplorons toutefois que les éléments du dictionnaire ne soient jamais remis en question, étant donné qu'ils sont fixes vis-à-vis des entrées $\boldsymbol{x}_{\omega_j}$. Cette dernière demeure inchangée même si la non-stationnarité du système étudié rend sa contribution faible dans l'estimation de la sortie courante.

Il apparaît alors opportun d'adapter les éléments du dictionnaire $\mathcal{D}$ pour obtenir un dictionnaire amélioré $\mathcal{D}^{\mathrm{a}}$. Afin de suivre au mieux l'évolution du système en considérant toujours le modèle (4.1), nous proposons d'optimiser conjointement les coefficients $\alpha_{\ell,j}$ et les éléments $\boldsymbol{x}_{\omega_j}$ du dictionnaire. Pour cela, le problème étudié est la minimisation de l'erreur quadratique instantanée e_ℓ^2, avec

$$e_\ell = y_\ell - \sum_{j=1}^{m} \alpha_{\ell,j}\, \kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_\ell).$$

Il est clair qu'il s'agit d'un problème d'optimisation non linéaire et non convexe en $\boldsymbol{x}_{\omega_j}$. Pour y remédier, la résolution de ce problème est effectuée en deux phases. La première consiste à trouver le vecteur des coefficients optimaux $\alpha_{\ell,j}$ du modèle (4.1), en appliquant un algorithme d'identification en ligne. Soit $\boldsymbol{\alpha}_\ell$ le vecteur regroupant ces coefficients optimaux. La seconde phase a pour but d'adapter les éléments du dictionnaire sans enfreindre le critère de parcimonie considéré, en particulier la règle de cohérence fixée par (4.4).

Nous proposons d'adapter le i-ème élément du dictionnaire par une perturbation dans le sens opposé au gradient de l'erreur quadratique instantanée par rapport à $\boldsymbol{x}_{\omega_i}$. L'expression du gradient, désigné par $\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}}$, est donnée par

$$\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} = \nabla_{\boldsymbol{x}_{\omega_i}} e_\ell^2 = \nabla_{\boldsymbol{x}_{\omega_i}} \Big(y_\ell - \sum_{j=1}^{m} \alpha_{\ell,j}\, \kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_\ell)\Big)^2 = -2\, e_\ell\, \alpha_{\ell,i}\, \nabla_{\boldsymbol{x}_{\omega_i}} \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_\ell). \tag{4.9}$$

La règle d'adaptation consiste alors à remplacer chaque élément $\boldsymbol{x}_{\omega_i}$, pour tout $i = 1, 2, \ldots, m$, par l'élément amélioré

$$\boldsymbol{x}_{\omega_i}^{\mathrm{a}} = \boldsymbol{x}_{\omega_i} - \eta_\ell\, \boldsymbol{g}_{\boldsymbol{x}_{\omega_i}}. \tag{4.10}$$

La valeur du pas de l'ajustement η_ℓ est déterminée sous la contrainte de cohérence du dictionnaire résultant. Dans la suite, nous décrivons les développements dans le cas des noyaux radiaux. Le lecteur pourra se référer à nos récentes publications [Saidé et al., 2013a, Saidé et al., 2012] pour plus de détails avec une extension au noyau polynomial, ainsi qu'à [Saidé et al., 2013b] pour une généralisation à l'identification en ligne de systèmes à sorties multiples.

De la forme (3.7), les noyaux radiaux permettent d'écrire $\nabla_{\boldsymbol{x}_{\omega_i}} \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_\ell) = 2(\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_\ell)\, g'(\|\boldsymbol{x}_\ell - \boldsymbol{x}_{\omega_i}\|^2)$, où nous retrouvons en particulier le noyau gaussien $g(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) = \exp(\frac{-1}{2\sigma^2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$, avec $\nabla_{\boldsymbol{x}_{\omega_i}} \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_\ell) = -\frac{1}{\sigma^2}(\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_\ell)\, \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_\ell)$. L'adaptation de tous les éléments du dictionnaire est donnée par la mise à jour (4.10) où, évidemment, le choix de la valeur du pas η_ℓ n'est par arbitraire mais il est soumis à des contraintes strictes. En effet, le dictionnaire doit toujours vérifier le critère de cohérence (4.6), ce qui se traduit par

$$g(\|\boldsymbol{x}_{\omega_i}^{\mathrm{a}} - \boldsymbol{x}_{\omega_j}^{\mathrm{a}}\|^2) \leq \nu_0,$$

pour tout $i, j = 1, 2, \ldots, m$, c'est à dire $g(\|(\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j}) - \eta_\ell\, (\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} - \boldsymbol{g}_{\boldsymbol{x}_{\omega_j}})\|^2) \leq \nu_0$. Le développement en série de Taylor du terme de gauche de cette inégalité autour de $\eta_\ell \approx 0$ permet d'écrire :

$$\begin{aligned} g(\|(\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j}) - \eta_\ell\, (\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} - \boldsymbol{g}_{\boldsymbol{x}_{\omega_j}})\|^2) \;=\; & g(\|(\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j})\|^2) - 2\eta_\ell \Big((\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j})^\top (\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} - \boldsymbol{g}_{\boldsymbol{x}_{\omega_j}}) \\ & - \eta_\ell \|\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} - \boldsymbol{g}_{\boldsymbol{x}_{\omega_j}}\|^2\Big) g'(\|\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j}\|^2) + \ldots \end{aligned}$$

Avec cette approximation, la condition de ν_0-cohérence du dictionnaire adapté devient

$$-2\Big(\|\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} - \boldsymbol{g}_{\boldsymbol{x}_{\omega_j}}\|^2 \eta_\ell^2 - (\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j})^\top (\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} - \boldsymbol{g}_{\boldsymbol{x}_{\omega_j}}) \eta_\ell\Big) g'(\|\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j}\|^2) + \nu_0 - g(\|\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j})\|^2) \geq 0.$$

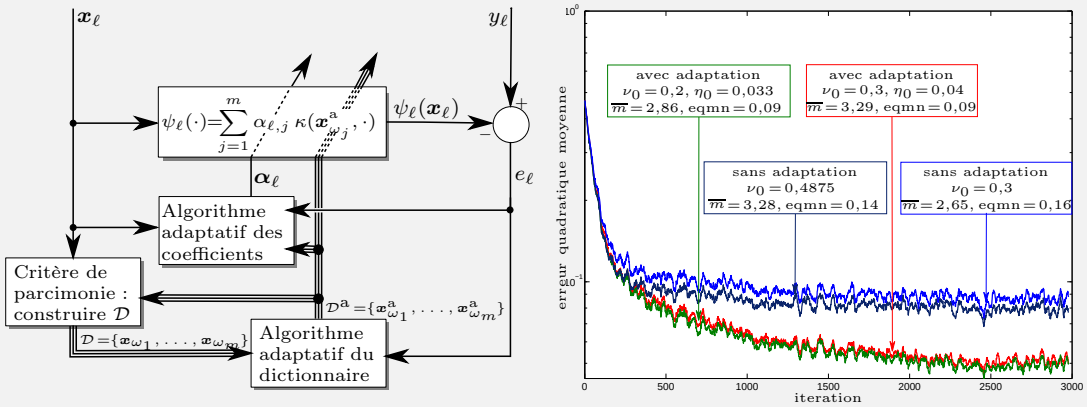


schéma illustratif résultats de convergence

[CADRE 11] **Adaptation des éléments du dictionnaire : schéma illustratif et résultats.**
Gauche : schéma illustratif de l'adaptation conjointe des coefficients du modèle (les $\alpha_{\ell,j}$) et des éléments du dictionnaire (les $\boldsymbol{x}^{\mathrm{a}}_{\omega_j}$).
Droite : le système étudié, emprunté de [Richard et al., 2009], est défini par la série temporelle : $y_t = v_t^2$, avec $v_t = 1{,}1 \exp(-|v_{t-1}|) + x_t$, où x_t et y_t sont respectivement l'entrée et la sortie désirée du système à l'instant t. La condition initiale est $v_0 = 0{,}5$ et les entrées x_t sont générées selon une loi normale de moyenne nulle et d'écart type 0,25. La sortie y_t est noyée dans un bruit blanc gaussien de variance unité. Le modèle étudié est $\psi_t(x_t)$.
Pour l'adaptation des coefficients du modèle, tout algorithme adaptatif peut être utilisé, dont les algorithmes à noyaux des moindres carrés récursif (RLS) [Honeine et al., 2007a], du gradient stochastique (LMS) [Honeine et al., 2007b] ou encore de projection affine (APA) [Richard et al., 2009]. Sans perte de généralité, ce dernier est utilisé dans la suite avec les paramètres préconisés dans [Richard et al., 2009] : taille de la fenêtre court-terme fixée à 3, paramètre du contrôle du pas de l'algorithme adaptatif fixé à 0,01, et facteur de régularisation fixé à 0,07. De même, le noyau gaussien est utilisé avec une largeur de bande $\sigma = 0{,}42$.
Un ensemble de 200 séries temporelles de 3000 échantillons chacune est utilisé pour comparer les résultats de simulation, sans et avec adaptation. Les mesures de performances sont la taille moyenne (sur les 200 séries) $\overline{m}$ du dictionnaire final, et l'erreur quadratique moyenne normalisée (eqmn), estimée sur les derniers 500 échantillons selon l'expression : eqmn $= \mathbb{E}\big(\frac{\sum_{t=2501}^{3000} |y_t - \psi_t(x_t)|^2}{\sum_{t=2501}^{3000} y_t^2}\big)$. Les courbes de convergence (figure de droite) montrent plusieurs résultats. D'une part, pour le même seuil de cohérence $\nu_0 = 0{,}3$, et après adaptation, une augmentation de $\overline{m}$ de 23% mène à une diminution de 41% du l'eqmn. D'autre part, pour une taille de dictionnaire comparable $\overline{m}$, obtenue selon des valeurs différentes de ν_0, l'adaptation permet une réduction de l'eqmn de 38%.

La valeur du discriminant Δ de ce polynôme de second degré en η_ℓ permet de définir les contraintes sur celle-ci, tout en rappelant la monotonicité du noyau. Si le discriminant est strictement négatif, il n'y a pas de racines réelles et, en conséquence, il n'y a aucune contrainte sur le choix du pas. En revanche, s'il est positif, on obtient deux racines $\eta_{i,j-}$ et $\eta_{i,j+}$ qui sont de même signe, car le terme constant du polynôme en question est positif vu que le dictionnaire est toujours ν_0-cohérent. Ces deux racines sont données par :

$$\eta_{i,j\pm} = \frac{-(\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j})^\top (\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} - \boldsymbol{g}_{\boldsymbol{x}_{\omega_j}})\, g'(\|(\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j})\|^2) \pm \sqrt{\Delta}}{-\|\boldsymbol{g}_{\boldsymbol{x}_{\omega_i}} - \boldsymbol{g}_{\boldsymbol{x}_{\omega_j}}\|^2 g'(\|\boldsymbol{x}_{\omega_i} - \boldsymbol{x}_{\omega_j}\|^2)}$$

Le domaine des valeurs admissibles de η_ℓ est $]-\infty, \eta_{i,j-}] \cup [\eta_{i,j+}, +\infty[$ et la valeur $\eta_\ell = 0$ appartient toujours à cet intervalle puisque, dans ce cas, il n'y a pas adaptation du dictionnaire et qu'il est ν_0-cohérent par construction.

4.2.3 Nouvelle classe de méthodes à noyaux en ligne

Les travaux entamés jusqu'à présent étudient le problème d'identification en ligne de systèmes non linéaires. Récemment, nous nous sommes intéressés à d'autres problèmes très importants dans plusieurs domaines. Nos travaux s'articulent autour de deux axes de recherche : le premier concerne l'estimation de vecteurs propres avec un algorithme en ligne d'ACP-à-noyaux, et le second traite la détection séquentielle avec la mise en œuvre d'une approche mono-classe en ligne. Ces deux axes sont introduit brièvement dans la suite, avant d'être décrits en détail aux sections 4.3 et 6.2, respectivement.

4.2.3.a ACP-à-noyaux en ligne

Bien que l'ACP et l'ACP-à-noyaux soient le fer de lance de l'estimation des vecteurs propres, l'intérêt d'algorithmes en ligne a souvent été soulevé. Les règles d'Oja [Oja, 1982] et de Sanger [Sanger, 1989a] ont permis le développement d'algorithmes itératifs pour l'ACP linéaire. En un mot (voir la section suivante pour une description détaillée), un vecteur principal est donné par la règle récursive d'Oja :

$$\boldsymbol{w}_{\ell+1} = \boldsymbol{w}_\ell + \eta_\ell(\boldsymbol{x}_\ell y_\ell - y_\ell^2 \boldsymbol{w}_\ell),$$

où $y_\ell = \boldsymbol{w}_\ell^\top \boldsymbol{x}_\ell$, sous l'hypothèse d'échantillons de moyenne nulle. La mise en œuvre d'une telle règle dans le cadre des méthodes à noyaux, afin d'estimer une fonction principale, est donnée par

$$\psi_{\ell+1}(\cdot) = \psi_\ell(\cdot) + \eta_\ell \left(y_\ell \kappa(\boldsymbol{x}_\ell, \cdot) - y_\ell^2 \, \psi_\ell(\cdot)\right),$$

où $y_\ell = \psi_\ell(\boldsymbol{x}_\ell)$. Il est clair qu'à chaque instant, la fonction principale $\psi_{\ell+1}(\cdot)$ s'enrichit d'un nouveau terme en $\kappa(\boldsymbol{x}_\ell, \cdot)$. Il s'avère nécessaire de contrôler l'ordre du modèle. Plusieurs défis naissent, dont la mise en œuvre d'un critère de parcimonie approprié, l'extraction de multiples fonctions principales, ainsi que le choix du pas, et le problème de centrage des échantillons dans le sous-espace. Ces divers sujets sont traités en détail dans la section 4.3.

4.2.3.b Détection mono-classe en ligne

Le problème d'apprentissage d'une seule classe, dit mono-classe (en anglais *one-class*) est un problème élémentaire en méthodes à noyaux. Il s'agit d'identifier la sphère qui englobe la totalité des échantillons dans l'espace caractéristique. Il consiste principalement à estimer le centre de la sphère, puisque le rayon est donné par la distance des échantillons à celui-ci. Une formulation parcimonieuse consiste à représenter le barycentre de l'ensemble $\{\kappa(\boldsymbol{x}_1, \cdot), \kappa(\boldsymbol{x}_2, \cdot), \ldots, \kappa(\boldsymbol{x}_n, \cdot)\}$, c'est à dire $\mu_n(\cdot) = \frac{1}{n}\sum_{i=1}^n \kappa(\boldsymbol{x}_i, \cdot)$ par une expression de la forme (4.1), c'est à dire $\sum_{j=1}^m \alpha_j \, \kappa(\boldsymbol{x}_{\omega_j}, \cdot)$. Le dictionnaire des $\boldsymbol{x}_{\omega_j}$ peut être constitué par n'importe quel critère de parcimonie présenté dans le présent chapitre. L'estimation des coefficients optimaux, au sens des moindres carrés, est donnée par

$$\min_{\alpha_1,\ldots,\alpha_m} \left\| \frac{1}{n}\sum_{i=1}^n \kappa(\boldsymbol{x}_i, \cdot) - \sum_{j=1}^m \alpha_j \, \kappa(\boldsymbol{x}_{\omega_j}, \cdot) \right\|_{\mathbb{H}}^2.$$

La solution finale, hors ligne, est alors $\boldsymbol{\alpha} = \boldsymbol{K}_{\mathcal{D}}^{-1} \boldsymbol{\kappa}_{\mathcal{D}}(\mu_\ell(\cdot))$, où $\boldsymbol{K}_{\mathcal{D}}$ et $\boldsymbol{\kappa}_{\mathcal{D}}(\mu_n(\cdot))$ ont les termes généraux $\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$ et $\mu_n(\kappa(\boldsymbol{x}_{\omega_i}, \cdot)) = \frac{1}{n}\sum_{i=1}^n \kappa(\boldsymbol{x}_i, \boldsymbol{x}_{\omega_i})$ respectivement.

La section 6.2 du chapitre 6 décrit la résolution en ligne de la détection mono-classe. L'objectif poursuivi est double. Du point de vue algorithmique, nous étudions des critères de parcimonie adaptés, et proposons la mise en œuvre d'un algorithme adaptatif. Du point de vue théorique, nous établissons des bornes sur l'erreur d'approximation et le risque de première espèce. Ces travaux ont pour vocation de faire un pont entre, d'une part l'approche de classification mono-classe par méthodes à noyaux, et d'autre part la détection séquentielle de rupture et la théorie qui l'accompagne [Basseville and Nikiforov, 1993].

4.3 Mise en œuvre algorithmique : ACP-à-noyaux en ligne

Le concept de vecteurs propres / valeurs propres est essentiel dans de nombreux domaines en mathématiques pures et appliquées, de la factorisation matricielle à la mécanique quantique, en passant par de nombreux champs applicatifs. L'estimation de vecteurs propres / valeurs propres est souvent liée à divers outils en analyse des données et réduction de dimension, dont l'ACP est le fer de lance. Ceci s'est traduit naturellement par l'ACP-à-noyaux dans le formalisme fonctionnel dans un espace RKHS en apprentissage statistique. Dans cette section, nous développons l'algorithme en ligne de l'ACP-à-noyaux (pour plus de détails, voir [Honeine, 2012]). Afin de préparer le terrain, le tableau suivant retrace succinctement l'évolution des algorithmes d'ACP, en terme de linéarité du modèle et du mode de traitement :

	Modèle	Mode
Analyse en composantes principales [Jolliffe, 1986]	linéaire	batch
ACP-à-noyaux [Schölkopf et al., 1998]	non linéaire	batch
Règles d'Oja [Oja, 1982] et de Sanger [Sanger, 1989a]	linéaire	en ligne
ACP-à-noyaux itératif [Kim et al., 2005]	non linéaire	itératif
ACP-à-noyaux en ligne (proposée ici, [Honeine, 2012])	non linéaire	en ligne

L'ACP classique détermine les axes (*i.e.*, les directions) de plus grande variance des échantillons. Il est donné par $\boldsymbol{w}$, la solution du problème $\boldsymbol{C}\boldsymbol{w} = \lambda\,\boldsymbol{w}$, où $\boldsymbol{C} = \frac{1}{\ell}\sum_{i=1}^{\ell} \boldsymbol{x}_i\boldsymbol{x}_i^\top$ désigne une estimation de la matrice de covariance des échantillons (supposés d'espérance nulle). La direction principale correspond au vecteur propre associé à la plus grande valeur propre de cette matrice. Bien que l'algorithme de l'ACP classique ne soit pas adapté au traitement en ligne de l'information, Oja suggère dans [Oja, 1982] une règle itérative d'estimation du vecteur propre associé à la plus grande valeur propre. Le premier axe principal est estimé selon la règle d'Oja :

$$\boldsymbol{w}_{\ell+1} = \boldsymbol{w}_\ell + \eta_\ell\,(\boldsymbol{x}_\ell\, y_\ell - y_\ell^2\,\boldsymbol{w}_\ell),$$

où η_ℓ désigne le paramètre du pas et $y_\ell = \boldsymbol{w}_\ell^\top\,\boldsymbol{x}_\ell = \boldsymbol{x}_\ell^\top\,\boldsymbol{w}_\ell$. Cette règle itérative converge bien vers le premier axe principal, puisqu'à l'état d'équilibre $\boldsymbol{w}$ on a $\boldsymbol{x}_\ell y_\ell = y_\ell^2\boldsymbol{w}$, une identité qui s'exprime également selon $\boldsymbol{x}_\ell\boldsymbol{x}_\ell^\top\boldsymbol{w} = \boldsymbol{w}^\top\boldsymbol{x}_\ell\boldsymbol{x}_\ell^\top\boldsymbol{w}\boldsymbol{w}$. Il suffit simplement de faire la moyenne sur tous les échantillons pour obtenir $C\boldsymbol{w} = \boldsymbol{w}^\top C\boldsymbol{w}\,\boldsymbol{w}$, avec la valeur propre $\boldsymbol{w}^\top C\boldsymbol{w}$ correspondant à y^2 que l'on souhaite maximiser. Cette règle est étendue par Sanger dans [Sanger, 1989a] pour l'estimation de multiples vecteurs propres.

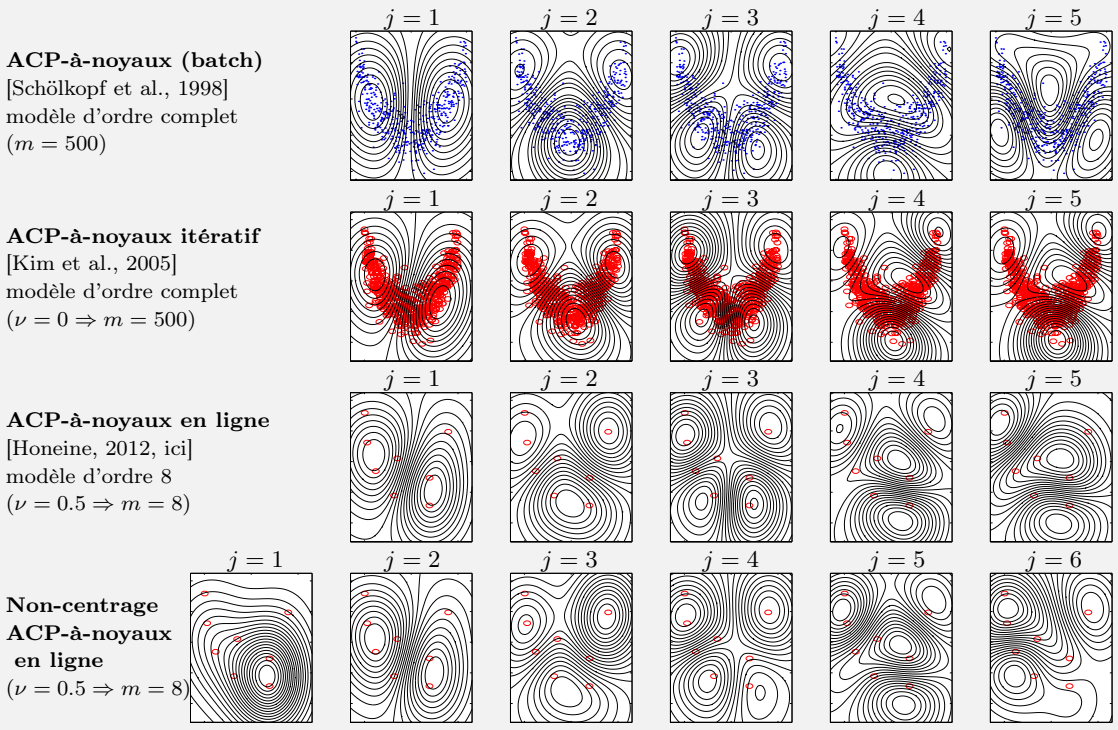
Dans un formalisme fonctionnel basé sur les espaces de Hilbert à noyau reproduisant, il s'agit d'estimer une fonction principale, appartenant à l'espace engendré par les fonctions noyau $\kappa(\boldsymbol{x}_1,\cdot), \kappa(\boldsymbol{x}_2,\cdot), \ldots, \kappa(\boldsymbol{x}_\ell,\cdot)$. Nous proposons alors une règle fonctionnelle d'Oja, dans $\mathbb{H}$, où la fonction principale à l'instant ℓ est donnée par

$$\psi_{\ell+1}(\cdot) = \psi_\ell(\cdot) + \eta_\ell\,(y_\ell\kappa(\boldsymbol{x}_\ell,\cdot) - y_\ell^2\,\psi_\ell(\cdot)), \tag{4.11}$$

où $y_\ell = \psi_\ell(\boldsymbol{x}_\ell)$ désigne la valeur associée à la projection de $\kappa(\boldsymbol{x}_\ell,\cdot)$ sur $\psi_\ell(\cdot)$, avec $\psi_\ell(\boldsymbol{x}) = \langle\psi_\ell(\cdot), \kappa(\boldsymbol{x},\cdot)\rangle_{\mathbb{H}}$. Il est clair que le modèle s'enrichit d'un terme nouveau à la présence d'un échantillon inédit, $\kappa(\boldsymbol{x}_\ell,\cdot)$ à l'instant ℓ, avec

$$\psi_\ell(\cdot) = \sum_{k=1}^{\ell} \alpha_{\ell,k}\,\kappa(\boldsymbol{x}_k,\cdot). \tag{4.12}$$

Nous préconisons l'utilisation d'un critère de parcimonie pour contrôler en ligne l'ordre du modèle, permettant ainsi d'élaborer un modèle d'ordre réduit de la forme (4.1). Tous les critères de parcimonie présentés dans la section 4.1 peuvent être appliqués pour contrôler l'ordre du modèle. Indépendamment de ce choix, on présente dans la suite la méthode d'extraction de la première fonction principale, avant de l'étendre au cas de plusieurs fonctions principales dans la section 4.3.2. Des discussions supplémentaires sont considérées dans la section 4.3.3. Voir le Cadre 12 pour une illustration, et [Honeine, 2012] pour une description détaillée.



[Cadre 12] **Illustration de l'algorithme en ligne de l'ACP-à-noyaux.**
Les 500 échantillons, *i.i.d.* régies par une distribution en forme de « banane », sont représentées dans la première rangée par les points « · ». Ces figures illustrent les contours associés aux premières fonctions principales, dans le cas du noyau gaussien avec une largeur de bande fixée à $\sigma = 0.5$, et un pas défini par les deux paramètres $(\eta_0, \tau) = (0.5, 10^5)$. Les trois premières rangées de figures présentent les résultats de l'ACP-à-noyaux [Schölkopf et al., 1998], l'approche itérative avec un modèle à ordre complet [Kim et al., 2005] et l'approche en ligne avec un contrôle de l'ordre du modèle [Honeine, 2012, dans ce chapitre]. La dernière rangée de figures illustre le cas où les fonctions ne sont pas centrées, avec des résultats assez similaires aux versions centrées, à part la première fonction principale qui saisit le centre des échantillons. Les éléments du dictionnaire qui participent dans l'expression du modèle sont illustrés par des anneaux « ○ ».

4.3.1 Extraction de la première fonction principale

En présence d'un nouvel échantillon $\boldsymbol{x}_\ell$ à l'instant ℓ, le critère de parcimonie détermine s'il est nécessaire d'ajouter sa fonction noyau $\kappa(\boldsymbol{x}_\ell, \cdot)$ dans le modèle ou, dans le cas contraire, de garder l'ordre du modèle inchangé. Dans chacun des deux cas, les coefficients du modèle sont adaptés correctement dans un schéma récursif.

– Cas 1. Dictionnaire inchangé
Dans ce cas, la fonction noyau est approximée par sa projection, $\kappa_p(\cdot)$. Cette dernière est de la forme $\sum_{k=1}^{m} \beta_k \, \kappa(\boldsymbol{x}_{\omega_k}, \cdot)$, soit $\kappa_p(\cdot) = \boldsymbol{\beta}_\ell^\top \boldsymbol{\kappa}_{\mathcal{D}}(\cdot)$ où $\boldsymbol{\beta}_\ell = \operatorname{argmin}_{\boldsymbol{\beta}} \|\kappa(\boldsymbol{x}_\ell, \cdot) - \boldsymbol{\beta}^\top \boldsymbol{\kappa}_{\mathcal{D}}(\cdot)\|_{\mathrm{H}}^2$ avec $\boldsymbol{\kappa}_{\mathcal{D}}(\cdot)$ le vecteur d'éléments $\kappa(\boldsymbol{x}_{\omega_i}, \cdot)$, pour $\boldsymbol{x}_{\omega_i} \in \mathcal{D}$. La solution est donnée par $\boldsymbol{\beta}_\ell = \boldsymbol{K}_{\mathcal{D}}^{-1} \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell)$ avec $\boldsymbol{K}_{\mathcal{D}}$ la matrice de Gram associée au dictionnaire. En remplaçant $\kappa(\boldsymbol{x}_\ell, \cdot)$ par sa projection dans la règle fonctionnelle d'Oja (4.11), on obtient la règle récursive suivante :

$$\boldsymbol{\alpha}_{\ell+1} = \boldsymbol{\alpha}_\ell + \eta_\ell \, (y_\ell \boldsymbol{\beta}_\ell - y_\ell^2 \, \boldsymbol{\alpha}_\ell). \tag{4.13}$$

Dans cette expression, $y_\ell = \psi_\ell(\boldsymbol{x}_\ell) = \sum_{k=1}^{m} \alpha_{\ell,k}\, \kappa(\boldsymbol{x}_{\omega_k}, \boldsymbol{x}_\ell)$, qui correspond aussi à la sortie associée à l'élément projeté [4].

– Cas 2. Dictionnaire augmenté
Dans ce cas, l'échantillon $\boldsymbol{x}_\ell$ est ajouté au dictionnaire, avec l'indice $\omega_{m+1} = \ell$, ainsi qu'un nouveau coefficient dans le modèle. La mise à jour du vecteur des coefficients $\boldsymbol{\alpha}_{\ell+1} = [\alpha_{\ell+1,1} \;\; \cdots \;\; \alpha_{\ell+1,m} \;\; \alpha_{\ell+1,m+1}]^\top$ est alors

$$\boldsymbol{\alpha}_{\ell+1} = \begin{bmatrix} \boldsymbol{\alpha}_\ell \\ 0 \end{bmatrix} + \eta_\ell\, y_\ell \left(\boldsymbol{\beta}_\ell - y_\ell \begin{bmatrix} \boldsymbol{\alpha}_\ell \\ 0 \end{bmatrix} \right), \tag{4.14}$$

où $\boldsymbol{\beta}_\ell = [0\; 0\; 0\; \cdots\; 0\; 0\; 1]^\top$, qui s'exprime également selon $\boldsymbol{\beta}_\ell = [\mathbf{0}_m^\top \; 1]^\top$. Dans cette expression, la sortie du modèle est définie par

$$y_\ell = \psi_\ell(\boldsymbol{x}_\ell) = \sum_{k=1}^{m+1} \alpha_{\ell,k}\, \kappa(\boldsymbol{x}_{\omega_k}, \boldsymbol{x}_\ell) = \boldsymbol{\alpha}_\ell^\top \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell), \tag{4.15}$$

qui est équivalente au cas précédent. Il est à noter la modification de la matrice de Gram, en remplaçant $\boldsymbol{K}_{\mathcal{D}}$ par

$$\begin{bmatrix} \boldsymbol{K}_{\mathcal{D}} & \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell) \\ \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell)^\top & \kappa(\boldsymbol{x}_\ell, \boldsymbol{x}_\ell) \end{bmatrix}. \tag{4.16}$$

Son inverse, nécessaire pour l'étape de projection dans le Cas 1., est obtenu à partir du lemme d'inversion matricielle, selon

$$\begin{bmatrix} \boldsymbol{K}_{\mathcal{D}}^{-1} & \mathbf{0}_m \\ \mathbf{0}_m^\top & 0 \end{bmatrix} + \frac{1}{\kappa(\boldsymbol{x}_\ell, \boldsymbol{x}_\ell) - \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell)^\top \boldsymbol{\beta}_\ell} \begin{bmatrix} -\boldsymbol{\beta}_\ell \\ 1 \end{bmatrix} \begin{bmatrix} -\boldsymbol{\beta}_\ell^\top & 1 \end{bmatrix}. \tag{4.17}$$

Le dénominateur dans cette expression, qui n'est autre que le complément de Schur de l'ancienne matrice $\boldsymbol{K}_{\mathcal{D}}$ dans la matrice augmentée (4.16), correspond à l'erreur quadratique de projection utilisée dans le critère d'approximation linéaire. Par conséquent, on obtient la matrice de Gram et son inverse sans la nécessité de recalculer à nouveau des matrices entières à chaque incrémentation de l'ordre du modèle.

4.3.2 Extraction de multiples fonctions principales

L'extension de la précédente méthode à de multiples fonctions principales est directe, en utilisant l'algorithme Hebbien proposé par Sanger pour l'algorithme linéaire de l'analyse en composantes principales [Sanger, 1989a]. Soit $\{\psi_{\ell,1}(\cdot), \psi_{\ell,2}(\cdot), \ldots, \psi_{\ell,r}(\cdot)\}$ la collection des r fonctions principales à déterminer, triée dans l'ordre décroissant des valeurs propres. En opérant une orthogonalisation de type Gram-Schmidt, la j-ème fonction principale est donnée par la règle itérative suivante :

$$\psi_{\ell,j+1}(\cdot) = \psi_{\ell,j}(\cdot) + \eta_\ell \Big(y_{\ell,j}\, \kappa(\boldsymbol{x}_\ell, \cdot) - y_{\ell,j} \sum_{i=1}^{j} y_{\ell,i}\, \psi_{\ell,i}(\cdot) \Big),$$

où $y_{\ell,j} = \psi_{\ell,j}(\boldsymbol{x}_\ell)$, pour $j = 1, 2, \ldots, r$. En effet, cette expression s'exprime également sous la forme

$$\psi_{\ell,j+1}(\cdot) = \psi_{\ell,j}(\cdot) + \eta_\ell \Big(y_{\ell,j} \big(\kappa(\boldsymbol{x}_\ell, \cdot) - \sum_{i=1}^{j-1} y_{\ell,i}\, \psi_{\ell,i}(\cdot) \big) - y_{\ell,j}^2\, \psi_{\ell,j}(\cdot) \Big),$$

4. La sortie du modèle (*i.e.*, l'évaluation de $\psi_\ell(\cdot)$) en tout élément correspond à celle obtenue en sa projection. Cela peut être démontré en décomposant $\kappa(\boldsymbol{x}_\ell, \cdot)$ en deux composantes, $\kappa_\perp(\cdot)$ qui est orthogonale au sous-espace des m fonctions noyau retenues, et $\kappa_p(\cdot)$ qui est la projection sur ce sous-espace. La sortie du modèle est alors donnée par

$$y_\ell = \langle \psi_\ell(\cdot), \kappa(\boldsymbol{x}_\ell, \cdot) \rangle_{\mathrm{H}} = \langle \psi_\ell(\cdot), \kappa_p(\cdot) \rangle_{\mathrm{H}} + \langle \psi_\ell(\cdot), \kappa_\perp(\cdot) \rangle_{\mathrm{H}}.$$

Puisque $\kappa_\perp(\cdot)$ est orthogonale au sous-espace des $\psi_\ell(\cdot)$, c'est à dire $\langle \psi_\ell(\cdot), \kappa_\perp(\cdot) \rangle_{\mathrm{H}} = 0$, nous avons alors $y_\ell = \langle \psi_\ell(\cdot), \kappa_p(\cdot) \rangle_{\mathrm{H}}$.

Nous obtenons également ce résultat en examinant la sortie associée à la composante projetée, à savoir $\langle \psi_\ell(\cdot), \kappa_p(\cdot) \rangle_{\mathrm{H}} = \boldsymbol{\alpha}_\ell^\top \boldsymbol{K}_{\mathcal{D}} \boldsymbol{\beta}_\ell$, où $\boldsymbol{\beta}_\ell = \boldsymbol{K}_{\mathcal{D}}^{-1} \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell)$, et par suite $\langle \psi_\ell(\cdot), \kappa_p(\cdot) \rangle_{\mathrm{H}} = \boldsymbol{\alpha}_\ell^\top \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_\ell)$.

qui n'est autre que la règle d'Oja pour $\psi_{\ell,j}(\cdot)$ avec l'entrée modifiée $\kappa(\boldsymbol{x}_\ell,\cdot)-\sum_{i=1}^{j-1} y_{\ell,i}\,\psi_{\ell,i}(\cdot)$, c'est à dire la composante à l'extérieur de l'espace engendré par les fonctions principales précédentes. Cette condition impose l'orthogonalité des fonctions principales les unes aux autres.

En désignant par $\boldsymbol{\psi}_\ell(\cdot)$ le « vecteur » de ces fonctions et $\boldsymbol{y}_\ell = [y_{\ell,1}\ y_{\ell,2}\ \cdots\ y_{\ell,r}]^\top$, nous avons alors

$$\boldsymbol{\psi}_{\ell+1}(\cdot) = \boldsymbol{\psi}_\ell(\cdot) + \eta_\ell\Big(\boldsymbol{y}_\ell\,\kappa(\boldsymbol{x}_\ell,\cdot) - \mathrm{LT}(\boldsymbol{y}_\ell\boldsymbol{y}_\ell^\top)\,\boldsymbol{\psi}_\ell(\cdot)\Big),$$

où $\mathrm{LT}(\cdot)$ rend son argument triangulaire inférieure en annulant les éléments au-dessus de sa diagonale. L'algorithme résultant est similaire à celui décrit ci-haut, avec les deux règles itératives (4.13) et (4.14) remplacées respectivement par

$$\boldsymbol{A}_{\ell+1} = \boldsymbol{A}_\ell + \eta_\ell\Big(\boldsymbol{y}_\ell\,\boldsymbol{\beta}_\ell - \mathrm{LT}(\boldsymbol{y}_\ell\boldsymbol{y}_\ell^\top)\boldsymbol{A}_\ell\Big) \qquad \text{et} \qquad \boldsymbol{A}_{\ell+1} = \begin{bmatrix}\boldsymbol{A}_\ell\\ \boldsymbol{0}_r^\top\end{bmatrix} + \eta_\ell\left(\boldsymbol{y}_\ell\boldsymbol{\beta}_\ell - \mathrm{LT}(\boldsymbol{y}_\ell\boldsymbol{y}_\ell^\top)\begin{bmatrix}\boldsymbol{A}_\ell\\ \boldsymbol{0}_r^\top\end{bmatrix}\right),$$

où $\boldsymbol{A}_\ell = [\boldsymbol{\alpha}_{\ell,1}\ \boldsymbol{\alpha}_{\ell,2}\ \cdots\ \boldsymbol{\alpha}_{\ell,r}]$ est la matrice de taille $(m\times r)$ dont la j-ème colonne correspond aux coefficients de la j-ème fonction principale avec $\alpha_{\ell,j,1}, \alpha_{\ell,j,2},\ \ldots\ , \alpha_{\ell,j,m}$.

4.3.3 Discussions

4.3.3.a Pas et vitesse de convergence

Il est évident que les fonctions principales associées aux petites valeurs propres ne puissent mûrir correctement qu'après celles associées aux plus grandes valeurs propres. En outre, l'effet de soustraire les plus grandes variances, associées aux fonctions principales déjà extraites, va ralentir l'apprentissage des fonctions principales suivantes. A l'instar des travaux sur l'ACP itérative linéaire [Chen and Chang, 1995] (voir aussi [Sanger, 1989b, page 52] pour la convergence de l'algorithme Hebbien linéaire), il existe des conditions sur la valeur du pas qui permettent de garantir la convergence. Il suffit d'imposer au pas une valeur inférieure à l'inverse de la plus grande valeur propre. Malheureusement, les valeurs propres sont inconnues en général, et l'effort de les estimer est souvent cher et inadapté pour l'apprentissage en ligne [Schraudolph et al., 2007].

La convergence repose sur un ensemble d'hypothèses mathématiques. Essentiellement, le pas ne peut pas être constant, mais doit décroître avec le temps, comme par exemple $\eta_\ell = \eta_0/\ell$ où η_0 est un paramètre positif constant. Ce choix est fréquent dans la littérature de l'approximation stochastique, avec une convergence lente lorsque η_0 est petit et une divergence pour les grandes valeurs. Afin de remédier à ces inconvénients et de parvenir à une convergence, une approche de type « recherche puis convergence » est préconisée dans [Darken et al., 1992], selon l'expression

$$\eta_\ell = \frac{\eta_0}{1+\ell/\tau}. \tag{4.18}$$

Le paramètre de réglage τ détermine la durée de l'étape de recherche, avec $\eta_\ell \approx \eta_0$ (lorsque $\ell \ll \tau$), avant l'étape de convergence où η_ℓ diminue selon η_0/ℓ (lorsque $\ell \gg \tau$). Nous préconisons donc l'adaptation du pas selon l'expression (4.18). Le choix du paramètre de réglage τ dépend de l'application.

4.3.3.b Sur le centrage dans le sous-espace

Indépendamment de la stratégie d'estimation en ACP linéaire (inversion matricielle, règle d'Oja, etc.), les échantillons sont supposés centrés, c'est à dire de moyenne nulle. La transformation non linéaire induite par le noyau utilisé conduit, presque toujours, à un décentrage dans l'espace caractéristique $\mathbb{H}$. Le re-centrage dans cet espace est réalisé directement par une modification de la matrice de Gram. C'est le cas de l'algorithme classique d'ACP-à-noyaux où, appliqué à un ensemble de n échantillons, la matrice de Gram correspondante $\boldsymbol{K}$ est remplacée par

$$\boldsymbol{K} - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\,\boldsymbol{K} - \tfrac{1}{n}\boldsymbol{K}\,\mathbf{1}_n\mathbf{1}_n^\top + \tfrac{1}{n^2}\mathbf{1}_n\mathbf{1}_n^\top\boldsymbol{K}\,\mathbf{1}_n\mathbf{1}_n^\top, \tag{4.19}$$

où $\mathbf{1}_n$ est le vecteur unité de taille $(n\times 1)$, c'est à dire

$$\big(\mathbf{I} - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big)\,\boldsymbol{K}\,\big(\mathbf{I} - \tfrac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\big), \tag{4.20}$$

où $\mathbf{I}$ est la matrice identité de taille $(n \times n)$.

Concernant l'algorithme proposé ici pour l'ACP-à-noyaux en ligne, et où tout est représenté dans le sous-espace engendré par les éléments du dictionnaire, il en est de même pour le re-centrage. Pour cela, chaque fonction noyau $\kappa(\boldsymbol{x}_i, \cdot)$ est remplacée par

$$\kappa(\boldsymbol{x}_i, \cdot) - \sum_{i'=1}^{n} \Big(\sum_{j=1}^{m} \beta_{n,j}\, \kappa(\boldsymbol{x}_{\omega_j}, \cdot) \Big),$$

où le terme entre parenthèses correspond, à un facteur multiplicatif $1/n$ près, à la projection de $\kappa(\boldsymbol{x}_{i'}, \cdot)$ sur le sous-espace engendré par les m fonctions noyau induites par le dictionnaire $\mathcal{D}$. Le terme à droite est de la forme $\overline{\boldsymbol{\beta}}_n^\top \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x})$, où le vecteur $\overline{\boldsymbol{\beta}}_n$ peut être mis à jour de manière récursive, comme préconisé dans [Honeine, 2012, Annexe]. Finalement, cela conduit à une expression semblable à (4.19), avec

$$\boldsymbol{K}_{\mathcal{D}} - \mathbf{1}_m \overline{\boldsymbol{\beta}}_n^\top \boldsymbol{K}_{\mathcal{D}} - \boldsymbol{K}_{\mathcal{D}} \overline{\boldsymbol{\beta}}_n \mathbf{1}_m^\top + \mathbf{1}_m \overline{\boldsymbol{\beta}}_n^\top \boldsymbol{K}_{\mathcal{D}} \overline{\boldsymbol{\beta}}_n \mathbf{1}_m^\top.$$

Le vecteur $\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_n)$ est aussi corrigé, en le remplaçant par sa version centrée

$$\overline{\boldsymbol{\kappa}_{\mathcal{D}}}(\boldsymbol{x}_n) = \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_n) - \mathbf{1}_m \overline{\boldsymbol{\beta}}_n^\top \boldsymbol{\kappa}(\boldsymbol{x}_n) - \boldsymbol{K}_{\mathcal{D}} \overline{\boldsymbol{\beta}}_n + \mathbf{1}_m \overline{\boldsymbol{\beta}}_n^\top \boldsymbol{K}_{\mathcal{D}} \overline{\boldsymbol{\beta}}_n.$$

La sortie du modèle résultant est alors donnée par $y_\ell = \boldsymbol{\alpha}_n^\top \overline{\boldsymbol{\kappa}_{\mathcal{D}}}(\boldsymbol{x}_n)$ pour une seule fonction principale, et $\boldsymbol{y}_n = \boldsymbol{A}_n^\top \overline{\boldsymbol{\kappa}_{\mathcal{D}}}(\boldsymbol{x}_n)$ dans le cas de multiples fonctions principales.

Pour résumer, l'approche proposée s'effectue selon l'une des deux variantes : une associée au centrage, et une sans centrage. Avant d'aller plus loin, une courte discussion s'impose. Dans le cas de la version non-centrée, les fonctions principales résultantes sont des combinaisons linéaires des fonctions noyau non-centrées. L'ACP correspond alors à une décomposition en vecteurs propres / valeurs propres de la matrice du moment non-centré d'ordre 2 des échantillons (dans l'espace caractéristique). Ceci est différent de la version centrée, où un re-centrage des échantillons est opéré. Cette dernière version, qui considère les moments centrés, est plus appropriée d'un point de vue purement statistique. Mais rien n'empêche d'utiliser l'une ou l'autre version. Ainsi demeure-t-elle ouverte la question d'extraction de caractéristiques à partir d'échantillons non-centrés ou centrés, même dans le cas conventionnel linéaire. Dans le cas d'une ACP avec centrage, la variabilité par rapport au centre des échantillons est concernée, par opposition à la variabilité par rapport à l'origine dans le cas non-centré. Pourtant, les liens entre les deux variantes sont solides, comme étudié dans [Cadima and Jolliffe, 2009] et résumés comme suit : les valeurs propres dans le problème non-centré sont entrelacées avec celles obtenues dans le problème centré, et de nombreux vecteurs propres sont généralement semblables. De plus, le premier vecteur propre dans le cas non-centré est souvent très proche de la direction qui relie l'origine au centre des échantillons. Tous ces résultats, issus de l'ACP linéaire, peuvent être facilement étendus à l'ACP-à-noyaux, et en particulier à l'approche en ligne que nous proposons. Les résultats expérimentaux mettent en évidence ces résultats théoriques avec succès, comme illustré dans le Cadre 12.

4.3.3.c Débruitage dans le sous-espace

En vue du débruitage d'une observation $\boldsymbol{x}$, sa fonction noyau $\kappa(\boldsymbol{x}, \cdot)$ est projetée sur le sous-espace engendré par les r plus pertinentes fonctions principales. Comme ces fonctions définissent une base de cet espace, cette projection s'écrit alors selon

$$\mathcal{P}_{\boldsymbol{\psi}}(\kappa(\boldsymbol{x}, \cdot)) = \sum_{j=1}^{r} \langle \psi_j(\cdot), \kappa(\boldsymbol{x}, \cdot) \rangle\, \psi_j(\cdot) = \sum_{k=1}^{m} \sum_{j=1}^{r} \psi_j(\boldsymbol{x})\, \alpha_{k,j}\, \kappa(\boldsymbol{x}_{\omega_k}, \cdot) = \sum_{k=1}^{m} [\boldsymbol{A}_\ell \boldsymbol{y}]_k\, \kappa(\boldsymbol{x}_{\omega_k}, \cdot),$$

où $\boldsymbol{y} = \boldsymbol{A}_\ell^\top \overline{\boldsymbol{\kappa}_{\mathcal{D}}}(\boldsymbol{x})$. Bien que la projection résultante soit toujours *centrée*, le décentrage est obtenu en ajoutant la moyenne à $\boldsymbol{P}_{\boldsymbol{\psi}}(\kappa(\boldsymbol{x}, \cdot))$, conduisant à réécrire cette projection selon $(\boldsymbol{A}_\ell \boldsymbol{y} + \overline{\boldsymbol{\beta}}_\ell)^\top \boldsymbol{\kappa}(\cdot)$.

Bien que ce résultat soit toujours dans le RKHS, la résolution du problème de pré-image est alors nécessaire pour retrouver le résultat dans l'espace des observations. Les diverses méthodes de résolution peuvent être appliquées, dont celles passées en revue dans le chapitre 3. Précisons toutefois

que le contrôle de l'ordre du modèle, comme c'est le cas avec les différents critères de parcimonie présentés dans ce chapitre, permet de rendre le problème de pré-image plus abordable, voire bien-posé. En considérant par exemple le problème de pré-image comme défini par (3.1)-(3.2), la fonction coût s'écrit

$$\mathcal{J}(\boldsymbol{x}) = \|(\boldsymbol{A}_\ell \boldsymbol{y} + \overline{\boldsymbol{\beta}}_\ell)^\top \boldsymbol{\kappa}(\cdot) - \kappa(\boldsymbol{x}, \cdot)\|^2_{\mathrm{H}},$$

et son gradient est donné pour le noyau gaussien par

$$\nabla_{\boldsymbol{x}} \mathcal{J}(\boldsymbol{x}) = \frac{1}{\sigma^2} \sum_{k=1}^{m} [\boldsymbol{A}_\ell \boldsymbol{y}]_k \, (\boldsymbol{x}_{\omega_k} - \boldsymbol{x}) \exp(\tfrac{-1}{2\sigma^2} \|\boldsymbol{x}_{\omega_k} - \boldsymbol{x}\|^2).$$

Il est évident que le modèle d'ordre réduit est plus facile à manipuler, avec une sommation sur m termes, par opposition à n termes dans le cas du modèle à ordre complet. Un autre exemple concerne la solution par transformation conforme présentée dans la section 3.2.2. Dans ce cas, il s'agit de la résolution du système

$$\boldsymbol{X}_{\mathcal{D}}^\top \boldsymbol{x}^* = \big(\boldsymbol{X}_{\mathcal{D}}^\top \boldsymbol{X}_{\mathcal{D}} - \eta \boldsymbol{K}_{\mathcal{D}}^{-1}\big)(\boldsymbol{A}_\ell \boldsymbol{y} + \overline{\boldsymbol{\beta}}_\ell)$$

où $\boldsymbol{X}_{\mathcal{D}} = [\boldsymbol{x}_{\omega_1} \; \boldsymbol{x}_{\omega_2} \; \cdots \; \boldsymbol{x}_{\omega_m}]$. Précisons que la matrice $\boldsymbol{K}_{\mathcal{D}}$ est inversible pour certains critères de parcimonie, comme nous l'avons démontré pour le critère de cohérence dans [Honeine et al., 2007b, Richard et al., 2009] et plus récemment dans [Honeine, 2012] pour le critère d'approximation linéaire.

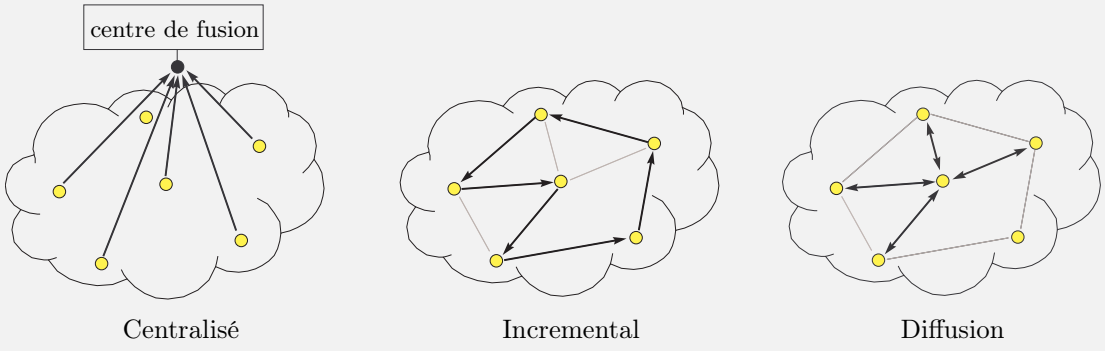
4.4 Traitement de l'information dans les réseaux de capteurs

Le domaine des réseaux de capteurs sans fil fait actuellement l'objet d'un intérêt considérable de la part des communautés académique et industrielle. La dispersion d'une multitude de capteurs bon marché dans une région donnée, l'élaboration d'un protocole de routage adéquat, et une implémentation algorithmique efficace, ouvrent en effet de nombreuses perspectives d'applications civiles et militaires.

Le problème traité ici est l'identification distribuée d'un champ, par exemple de gaz ou de concentration d'une espèce biochimique, et le suivi de son évolution au cours du temps. Le mode de calcul distribué est inhérent au caractère réparti des nœuds du réseau sur la région surveillée, dont la tâche est d'acquérir et de traiter localement les mesures. L'efficacité de la procédure d'identification est conditionnée par les interactions des nœuds, dictées par la topologie du réseau. Deux principes de coopération sont principalement traités dans la littérature, et illustrés par le Cadre 13. En mode incrémental, l'information transite de façon séquentielle et cyclique d'un nœud voisin à l'autre. Le coût énergétique, largement dicté par le volume des communications, tend à être minimum [Rabbat and Nowak, 2004, Lopes and Sayed, 2007]. En mode de diffusion, chaque nœud coopère avec l'ensemble de ses voisins pour une meilleure qualité d'estimation [Lopes and Sayed, 2008, Cattivelli et al., 2008].

La complexité des applications considérées ici, telle que la diffusion d'une grandeur physique qui illustre la suite de ce chapitre, nécessite de recourir à des méthodes d'identification adéquates. Si plusieurs approches existent dans la littérature (voir [Tu and Sayed, 2021, Abdolee et al., 2012] et les citations incluses), notre travail tire toute son originalité de l'usage des méthodes d'apprentissage statistique du formalisme des espaces de Hilbert à noyau reproduisant. Il en résulte des modèles non linéaires et des algorithmes fonctionnels adaptés à l'analyse de phénomènes complexes. Les deux stratégies d'apprentissage en ligne considérées, reposant sur un mode coopératif de type incrémental ou par diffusion, démontrent d'excellentes capacités de suivi des évolutions du système tout en affichant un coût calculatoire réduit.

Dans le contexte de l'apprentissage distribué dans un réseau de n capteurs sans fil, le problème traité ici est la modélisation d'un phénomène physique tel que la diffusion d'un champ de gaz en fonction des coordonnées du plan défini par $\mathbb{X}$. Chaque nœud ℓ, de coordonnées identifiées par sa position $\boldsymbol{x}_\ell \in \mathbb{X}$ supposée connue et fixe, accède à des réalisations temporelles, désignées par (la variable aléatoire) $y_{\ell,t}$ pour l'instant t. Le problème consiste à estimer, au sens des moindres carrés,



[CADRE 13] **Schémas des modes de traitement de l'information dans les réseaux.**
Schémas illustrant les trois modes de traitement de l'information dans les réseaux de capteurs sans fil. Le mode centralisé est basé sur la présence d'un centre de fusion des données, sans traitement local de l'information [Nguyen et al., 2005b, Nguyen et al., 2005a]. Les modes de traitement « en-réseau », dont les modes coopératifs de type incrémental [Rabbat and Nowak, 2004, Lopes and Sayed, 2007] et par diffusion [Predd et al., 2006, Lopes and Sayed, 2008], veillent au respect des contraintes d'énergie et de puissance, inhérentes aux réseaux sans fil.
A titre illustratif, considérons le problème d'estimation de la moyenne de toutes les mesures. Soit un réseau de n capteurs distribués uniformément sur une région de superficie unitaire, avec T mesures issues de chaque capteur. Le mode centralisé nécessite un coût de transmission en $\mathcal{O}(nT)$ sur une distance en $\mathcal{O}(1)$ en moyenne. Les modes de coopération inter-capteurs permettent de réduire cette complexité. C'est le cas en particulier de l'approche incrémentale qui a un coût de transmission en $\mathcal{O}(n)$ sur une distance en $\mathcal{O}(1/\sqrt{n})$.

une fonction $\psi(\cdot)$ d'un RKHS $\mathbb{H}$ en minimisant la fonction coût quadratique

$$\mathcal{J}(\psi) = \sum_{k=1}^{n} \mathbb{E}[|\psi(\boldsymbol{x}_k) - y_{k,t}|^2], \tag{4.21}$$

où $\mathbb{E}[\cdot]$ désigne l'espérance mathématique des variables aléatoires en fonction du temps t. Le CADRE 14 est consacré aux expérimentations, où nous comparons succinctement les modes de coopération incrémental et par diffusion, qui font l'objet de la suite de cette section.

4.4.1 Mode de coopération incrémental

Pour proposer le mode de coopération incrémental, nous décomposons la fonction coût (4.21) en une somme de n critères individuels, selon $\mathcal{J}(\psi(\cdot)) = \sum_{k=1}^{n} \mathcal{J}_k(\psi(\cdot))$, où $\mathcal{J}_k(\psi(\cdot)) = \mathbb{E}[|\psi(\boldsymbol{x}_k) - y_{k,t}|^2]$. Pour déterminer la solution optimale, une méthode simple de descente de gradient est mise en œuvre, selon

$$\psi_t(\cdot) = \psi_{t-1}(\cdot) - \frac{\eta}{2} \sum_{k=1}^{n} \nabla_{\psi_{t-1}(\cdot)} \mathcal{J}_k(\psi_{t-1}(\cdot)),$$

où l'expression du gradient est donnée par $\nabla_{\psi_{t-1}(\cdot)} \mathcal{J}_k(\psi_{t-1}(\cdot)) = 2\,\mathbb{E}[\psi_{t-1}(\boldsymbol{x}_k) - y_{k,t}]\,\kappa(\boldsymbol{x}_k, \cdot)$ puisque $\psi_{t-1}(\boldsymbol{x}_k) = \langle \psi_{t-1}(\cdot), \kappa(\boldsymbol{x}_k, \cdot) \rangle_{\mathbb{H}}$. En remplaçant l'espérance mathématique par la grandeur instantanée correspondante, il en résulte l'algorithme LMS fonctionnel suivant :

– Stratégie incrémentale : à chaque itération t, répéter

1. $\psi_{0,t}(\cdot) = \psi_{t-1}(\cdot)$
2. $\psi_{\ell,t}(\cdot) = \psi_{\ell-1,t}(\cdot) - \eta\,(\psi_{t-1}(\boldsymbol{x}_\ell) - y_{\ell,t})\,\kappa(\boldsymbol{x}_\ell, \cdot)$, pour $\ell = 1, 2, \ldots, n$
3. $\psi_t(\cdot) = \psi_{n,t}(\cdot)$.

Le réseau est parcouru de manière répétée afin de suivre l'évolution du phénomène physique au cours du temps. Cette approche nécessite, par l'instruction 2., que chaque nœud ait accès à l'information globale $\psi_{t-1}(\cdot)$. Afin d'éviter cet inconvénient, il est possible d'évaluer le gradient localement, en $\psi_{\ell-1,t}(\cdot)$ reçu du précédent nœud. Ceci conduit à la mise à jour

$$\psi_{\ell,t}(\cdot) = \psi_{\ell-1,t}(\cdot) - \eta\,(\psi_{t-1}(\boldsymbol{x}_\ell) - y_{\ell,t})\,\kappa(\boldsymbol{x}_\ell, \cdot).$$

Le modèle est malheureusement enrichi d'un terme à chaque rencontre d'un nouveau capteur, limitant en l'état son application à des réseaux peu denses. Il est toutefois possible d'élaborer des modèles d'ordre réduit en considérant un critère de parcimonie, comme présenté tout au long de ce chapitre. Ainsi préconisons-nous, lors de la visite du capteur ℓ à l'instant t, d'élaborer un modèle d'ordre réduit de la forme

$$\psi_{\ell,t}(\cdot) = \sum_{j=1}^{m} \alpha_j\,\kappa(\boldsymbol{x}_{\omega_j}, \cdot) + \alpha_\ell\,\kappa(\boldsymbol{x}_\ell, \cdot),$$

où $\mathcal{D} = \{\boldsymbol{x}_{\omega_1}, \ldots, \boldsymbol{x}_{\omega_m}\} \subset \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ désigne des nœuds sélectionnés.

Il convient à présent de préciser le critère de parcimonie utilisé, à-même de limiter l'ordre du modèle sans en compromettre les performances. Dans un premier temps, nous proposons l'usage du critère de cohérence des fonctions noyau de (4.4), selon la mesure $\max_{j=1,\ldots,m} \mathrm{coh}(\kappa(\boldsymbol{x}_\ell, \cdot), \kappa(\boldsymbol{x}_{\omega_j}, \cdot))$, qui repose uniquement sur la topologie du réseau. Ce critère de parcimonie garantit que les capteurs ω_j sélectionnés reflètent une couverture raisonnable de la région scrutée. Voir [Honeine et al., 2010] pour une étude de ce critère pour la prédiction de séries temporelles dans les réseaux de capteurs. Nous déplorons toutefois que la constitution de ce dictionnaire ne soit jamais remise en question, quelles que soient les évolutions du système et l'erreur de modélisation commise, étant donné l'unique dépendance du critère vis-à-vis des positions $\boldsymbol{x}_i$ des capteurs. Pour y remédier, nous proposons dans un second temps le critère de cohérence entre les fonctionnelles (4.7), selon $\mathrm{coh}(\psi_{\ell,t}(\cdot), \psi^\perp_{\ell,t}(\cdot))$, où $\psi^\perp_{\ell,t}(\cdot)$ désigne la projection orthogonale de $\psi_{\ell,t}$ sur le sous-espace engendré par les fonctions noyau désignées par $\mathcal{D}$. L'algorithme proposé s'exprime finalement ainsi :

– Stratégie incrémentale avec contrôle de l'ordre du modèle : à chaque instant t, répéter
  1. $\psi_{0,t}(\cdot) = \psi_{t-1}(\cdot)$
  2. $\psi_{\ell,t}(\cdot) = \psi_{\ell-1,t}(\cdot) - \eta\,(\psi_{\ell-1,t}(\boldsymbol{x}_\ell) - y_{\ell,t})\,\kappa(\boldsymbol{x}_\ell, \cdot)$, pour $\ell = 1, 2, \ldots, n$
     Si $\mathrm{coh}(\psi_{\ell,t}(\cdot), \psi^\perp_{\ell,t}(\cdot)) > \nu$, remplacer $\psi_{\ell-1,t}(\cdot)$ par $\psi^\perp_{\ell,t}(\cdot)$
  3. $\psi_t(\cdot) = \psi_{\ell,t}(\cdot)$.

Le test dans 2., lorsqu'il est activé, a pour effet de réduire l'ordre du modèle en retirant $\kappa(\boldsymbol{x}_\ell, \cdot)$ du développement.

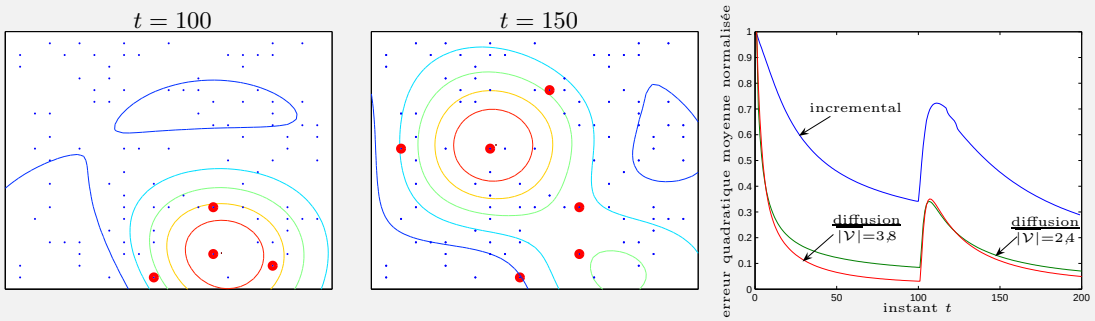
4.4.2 Mode de coopération par diffusion

Lorsque davantage de ressources en communication sont disponibles, il peut être intéressant d'exploiter un schéma de coopération plus sophistiqué. En mode de diffusion, chaque nœud interagit avec l'ensemble de ses voisins pour une meilleure qualité d'estimation [Lopes and Sayed, 2008, Cattivelli et al., 2008]. Plus précisément, à l'itération t, chaque nœud ℓ accomplit trois tâches. Il utilise les mesures acquises localement afin de procéder à une mise à jour provisoire. Il consulte également les nœuds voisins avec lesquels il collabore, le voisinage étant désigné par $\mathcal{V}_\ell$, pour acquérir leurs propres estimées et/ou mesures brutes. Finalement, il agrège l'ensemble de ces informations afin de mettre à jour son estimée.

Afin de mettre cette stratégie en équation, nous considérons comme précédemment la fonction coût quadratique (4.21), en la décomposant en chaque nœud ℓ du réseau par

$$\mathcal{J}(\psi(\cdot)) = \mathcal{J}_\ell(\psi(\cdot)) + \sum_{\substack{k=1 \\ k\neq\ell}}^{n} \mathcal{J}_k(\psi(\cdot)),$$

où le premier terme est $\mathcal{J}_\ell(\psi(\cdot)) = \sum_{k\in\mathcal{V}_\ell} c_{k,\ell}\,\mathbb{E}[|\psi(\boldsymbol{x}_k) - y_{k,t}|^2]$. Ici, les paramètres $c_{k,\ell}$ sont librement choisis sous les contraintes $c_{k,\ell} = 0$ si $k \notin \mathcal{V}_\ell$, et $\sum_k c_{k,\ell} = \sum_\ell c_{k,\ell} = 1$, sous peine de modifier



[CADRE 14] **Estimation d'un champ de diffusion dans les réseaux de capteurs sans fil.** Nous nous intéressons à l'estimation d'un champ de diffusion physique régi par l'équation différentielle suivante :

$$\frac{\partial \Theta(\boldsymbol{x},t)}{\partial t} - c\,\nabla^2_{\boldsymbol{x}}\Theta(\boldsymbol{x},t) = Q(\boldsymbol{x},t),$$

où $\Theta(\boldsymbol{x},t)$ est la quantité physique dépendante de la position et du temps, $\nabla^2_{\boldsymbol{x}}$ est l'opérateur spatial de Laplace et $Q(\boldsymbol{x},t)$ correspond à la quantité injectée dans la région étudiée, la conductivité thermique de cette dernière étant fixée à $c = 0{,}1$. Soit une collection de $n = 100$ capteurs uniformément distribués sur une grille de taille 21×21. Deux sources de chaleur de 200 W sont activées successivement, la première des instants $t = 1$ à $t = 100$ (située dans le quart-plan inférieur-droit), la seconde de $t = 101$ à $t = 200$ (située dans le quart-plan supérieur-gauche). L'objectif est d'estimer $\Theta(\boldsymbol{x},t)$ via $\psi_k(\boldsymbol{x})$, étant donné les mesures $y_{k,t} = \Theta(\boldsymbol{x}_k,t) + \epsilon_{k,t}$, avec $\epsilon_{k,t}$ un bruit blanc gaussien centré de variance 10^{-4}. Les simulations ont été réalisées avec un noyau gaussien de largeur de bande $\sigma = 0{,}5$ et un pas de gradient $\eta = 0{,}5$. Pour la méthode incrémentale, le seuil de cohérence est fixé à 0,985. Pour la méthode de diffusion, la stratégie « adaptation-puis-agrégation » est retenue ici en accordant un même poids à l'ensemble des voisins de chaque nœud, c'est à dire $b_{k,\ell} = c_{k,\ell} = 1/|\mathcal{V}_\ell|$, où $|\mathcal{V}_\ell|$ désigne la taille (c'est à dire cardinalité) de l'ensemble $\mathcal{V}_\ell$. La notion de voisinage est définie par seuillage de la distance séparant les nœuds. Comme l'indiquent les deux figures de contours, l'algorithme incrémental suit parfaitement les évolutions de la distribution de chaleur par un choix approprié des capteurs représentés dans le modèle et cerclés ici de rouge •. Rappelons que leur nombre correspond à l'ordre du modèle. La convergence de l'erreur quadratique moyenne normalisée de prédiction est illustrée par la figure à droite. Il est à noter la croissance de celle-ci à l'instant $t = 101$, provoquée par l'extinction de la première source et l'allumage de la seconde. Elle est comparée à celle obtenue avec la méthode procédant par diffusion d'information. Cette dernière s'avère plus performante, au prix d'un volume de communications plus élevé même si la taille moyenne des voisinages reste faible.

la fonction coût (4.21) à minimiser. Le second terme permet de pénaliser les écarts entre les solutions de chaque nœud. Nous proposons [5] de relaxer celle-ci afin de limiter le volume des communications, avec

$$\mathcal{J}^{\text{rel}}_\ell(\psi(\cdot)) = \sum_{k\in\mathcal{V}_\ell} c_{k,\ell}\,\mathbb{E}[|\psi(\boldsymbol{x}_k) - y_{k,t}|^2] + \sum_{k\in\mathcal{V}_\ell/\{\ell\}} b_{k,\ell}\,\|\psi(\cdot) - \psi^*_k(\cdot)\|^2_{\mathbb{H}},$$

où $\psi^*_k(\cdot)$ désigne à présent la solution disponible au nœud k et à l'instant courant. Une méthode de

5. En effet, soient $\psi^*_k(\cdot) = \arg\min_{\psi(\cdot)\in\mathbb{H}} \mathcal{J}_k(\psi(\cdot))$ et $\mathbb{H}_k$ le sous-espace de $\mathbb{H}$ engendré par le voisinage du k-ème nœud. Nous démontrons que $\mathcal{J}_\ell(\psi(\cdot))$ est de la forme $\sum_{k\in\mathcal{V}_\ell} \|\psi(\cdot) - \psi^*_k(\cdot)\|^2_{\mathbb{H}_k} + \zeta$ avec ζ une constante indépendante de $\psi(\cdot)$, après avoir complété un carré de sorte à aboutir à une identité remarquable, et au prix d'une modification locale de la métrique repérée ici par $\|\cdot\|_{\mathbb{H}_k}$. Dans ces conditions, la fonction coût s'exprime également sous la forme

$$\sum_{k\in\mathcal{V}_\ell} c_{k,\ell}\,\mathbb{E}[|\psi(\boldsymbol{x}_k) - y_{k,t}|^2] + \sum_{\substack{k=1\\k\neq\ell}}^{n} \|\psi(\cdot) - \psi^*_k(\cdot)\|^2_{\mathbb{H}_k},$$

dont la minimisation nécessite que le nœud ℓ ait accès à l'ensemble des optima locaux $\psi^*_k(\cdot)$ et des métriques $\|\cdot\|_{\mathbb{H}_k}$.

descente de gradient peut être mise en œuvre, selon

$$\psi_{\ell,t}(\cdot) = \psi_{\ell,t-1}(\cdot) - \frac{\eta}{2}\nabla_{\psi_{\ell,t-1}(\cdot)}\mathcal{J}_\ell^{\text{rel}}(\psi_{\ell,t-1}(\cdot)),$$

avec

$$\nabla_{\psi_{\ell,t-1}(\cdot)}\mathcal{J}_\ell^{\text{rel}}(\psi_{\ell,t-1}(\cdot)) = 2\sum_{k\in\mathcal{V}_\ell} c_{k,\ell}\,\mathbb{E}[\psi_{\ell,t-1}(\boldsymbol{x}_k) - y_{k,t}]\,\kappa(\boldsymbol{x}_k,\cdot) + 2\sum_{k\in\mathcal{V}_\ell/\{\ell\}} b_{k,\ell}\,(\psi_{\ell,t-1}(\cdot) - \psi_k^*(\cdot)).$$

Les méthodes d'optimisation itératives s'avèrent utiles pour minimiser une somme de fonctions coût convexes. Elles consistent à itérer selon chaque sous-gradient, dans un ordre prédéfini. Le gradient sus-mentionné est constitué d'une première composante visant à mettre le modèle local à jour à partir des mesures, et d'une seconde composante l'agrégeant aux modèles locaux voisins. Dans les travaux dirigés par A. Sayed [Lopes and Sayed, 2008, Cattivelli et al., 2008], deux stratégies distinctes sont suggérées suivant l'ordre adopté pour l'adaptation et l'agrégation. Nous proposons d'adopter ces stratégies, en présentant leurs homologues selon le formalisme des espaces de Hilbert à noyau reproduisant :

- Stratégie de diffusion « adaptation-puis-agrégation » : à chaque itération t, répéter
  1. $\psi_{\ell,t}(\cdot) = \psi_{\ell,t-1}(\cdot) - \eta\sum_{k\in\mathcal{V}_\ell} c_{k,\ell}\,(\psi_{\ell,t-1}(\boldsymbol{x}_k) - y_{k,t})\,\kappa(\boldsymbol{x}_k,\cdot)$
  2. $\psi_{\ell,t}(\cdot) = \sum_{k\in\mathcal{V}_\ell} b_{k,\ell}\,\psi_{k,t}(\cdot)$.
- Stratégie de diffusion « agrégation-puis-adaptation » : à chaque itération t, répéter
  1. $\psi_{\ell,t-1}(\cdot) = \sum_{k\in\mathcal{V}_\ell} b_{k,\ell}\,\psi_{k,t-1}(\cdot)$
  2. $\psi_{\ell,t}(\cdot) = \psi_{\ell,t-1}(\cdot) - \eta\sum_{k\in\mathcal{V}_\ell} c_{k,\ell}\,(\psi_{\ell,t-1}(\boldsymbol{x}_k) - y_{k,t})\,\kappa(\boldsymbol{x}_k,\cdot)$.

Notons que l'ordre de chacun des modèles locaux est dicté par la taille des voisinages considérés pour chacun des nœuds. Aussi la nécessité d'un contrôle de celui-ci par un test de parcimonie n'est-il pas aussi pressant que dans le cas d'une stratégie incrémentale.

4.5 Conclusion et perspectives

Historiquement, les critères de parcimonie en méthodes à noyaux ont joué un rôle crucial pour développer des techniques d'identification en ligne. Ils permettent de limiter l'ordre du modèle sans en compromettre les performances. Les éléments du dictionnaire ainsi construit possèdent une couverture admissible du domaine, à un niveau de parcimonie donné.

Nos récentes contributions sont décrites dans ce chapitre. Une grande part de ces avancées repose sur la constitution du dictionnaire en considérant la pertinence du modèle résultant. Un critère de parcimonie *a posteriori* est présenté et une adaptation des éléments du dictionnaire est décrite. Ces travaux sont à l'opposé des techniques classiques où les éléments du dictionnaire ne sont jamais remis en question, étant donné qu'ils sont fixes vis-à-vis des entrées du système étudié.

Dans le cadre des méthodes à noyaux en ligne, le problème le plus étudié est l'identification de systèmes. Ce chapitre s'efforce d'élargir le champ d'application, en se focalisant sur deux développements clés : d'une part, l'algorithme d'ACP-à-noyaux en ligne, et d'autre part, le traitement de l'information dans les réseaux. Ces travaux se poursuivent avec l'algorithme mono-classe en ligne présenté dans la section 6.2 à la page 123. En conséquence, le socle commun des méthodes à noyaux en ligne s'élargit pour inclure des méthodes non-supervisées, mais aussi des méthodes d'identification à sorties multiples [Saidé et al., 2013b].

L'apprentissage en ligne avec les méthodes à noyaux suscite un intérêt sans cesse grandissant et permet des avancées dans des champs d'application aussi multiples que variés. L'adaptation du dictionnaire peut être appliquée dans les réseaux de capteurs sans fil où la mobilité de certains capteurs permet d'optimiser le modèle. Nous avons récemment proposé une approche originale dans [Ghadban et al., 2013b], qu'il est nécessaire de compléter par une étude théorique. D'autres algorithmes en ligne peuvent être facilement développés dans le cadre décrit dans ce chapitre. C'est le cas en particulier d'un algorithme d'analyse en composantes indépendantes étudié

dans [Hyvärinen and Oja, 2000], avec une règle d'Oja qui ressemble à celle décrite dans ce chapitre pour l'ACP.

La force des représentations parcimonieuses est claire avec l'augmentation considérable du nombre d'échantillons acquis. Cet intérêt est illustré dans le cadre des activités récentes sur l'acquisition compressée (en anglais *compressed sensing*), depuis les travaux de Candés *et coll.* [Candès et al., 2006, Candès and Wakin, 2008]. En outre, le phénomène de données à grands volumes, dit *Big Data*, est considéré comme l'un des grands défis de cette décennie.

La nature déteste le vide. La nature aime les mélanges.

[anonyme]

Tout est bruit pour qui a peur.

[Sophocle]

5

Démélange linéaire et non linéaire en imagerie hyperspectrale

Sommaire

L'imagerie hyperspectrale vise à acquérir des images dans des centaines de bandes spectrales contigües, avec une grande résolution spatiale. Cette technologie connaît actuellement un développement impressionnant dans de nombreux domaines des sciences de l'observation, de la résolution microscopique aux échelles astronomiques, dans des contextes aussi bien civils que militaires. Le problème de démélange spectral constitue un problème fondamental en traitement d'images hyperspectrales. Il vise à décomposer chaque vecteur spectral sur une collection de composants purs et à estimer la proportion de ces derniers dans le mélange.

Cette thématique de recherche suit naturellement celle développée dans le chapitre précédent, puisqu'il s'agit essentiellement d'un problème d'approximation parcimonieuse où le dictionnaire est formé par les signatures spectrales des composants purs. Le problème de démélange ouvre la voie à de nouveaux défis : (*a*) interprétation physique des résultats obtenus, ce qui nécessite la satisfaction de certaines contraintes telles que la complétude de la décomposition et l'additivité des contributions ; (*b*) grande taille de ces cubes de données, avec des centaines de bandes spectrales et une résolution spatiale de plus en plus fine (jusqu'à 20 cm pour des satellites militaires), nécessitant des

algorithmes à faible complexité de calcul ; (*c*) non-linéarité des mélanges spectraux, comme préconisé dans de récentes études ; et (*d*) intégration de l'information spatiale dans la résolution du problème de démélange spectral.

Cette activité de recherche a démarré par une collaboration avec Cédric Richard de l'Observatoire de la Côte d'Azur. Les doctorants Jie Chen et Nguyen Hoang Nguyen ont contribué à ces travaux, qui ont donné lieu à quatre publications dans des revues [1], un chapitre de livre [2]et une douzaine d'articles publiés dans des actes de conférences [3]. Cette activité se poursuit naturellement dans le cadre du projet HYPANEMA (ANR, programme blanc) qui a démarré en 2013, ainsi que d'autres projets en cours de préparation.

Les développements proposés pour résoudre le problème de démélange sont structurés autour de plusieurs axes pour estimer les fractions d'abondance, comme résumé dans la suite :

- développement d'un cadre géométrique pour l'estimation des abondances sans coût calculatoire supplémentaire, avec une extension au démélange non-linéaire ;
- mise en œuvre de méthodes statistiques itératives pour l'estimation des fractions d'abondance sous contraintes, d'une part avec une approche de type descente de gradient et d'autre part avec des méthodes de projections multiples, de type Cimmino et Kaczmarz ;
- définition d'un nouveau paradigme de démélange non linéaire en combinant un modèle linéaire avec une fluctuation non linéaire, avec divers modèles de non-linéarité et méthodes de résolution appropriées ;
- développement de méthodes de démélange non linéaire en intégrant l'information spatiale, et ainsi profiter de la dualité spatiale-spectrale en imagerie hyperspectrale.

Le présent chapitre s'efforce de résumer ces axes de recherche. Au delà de la résolution du problème de démélange, nous nous sommes intéressés au problème de classification multi-classes de vecteurs hyperspectraux, permettant ainsi une segmentation automatique de l'image hyperspectrale. Les travaux correspondants ne sont pas traités dans ce chapitre. Nous renvoyons le lecteur à [Honeine and Richard, 2010a, Noumir et al., 2011b] pour plus d'informations, ainsi qu'au chapitre suivant qui est dédié à la résolution de problèmes de classification à faible coût calculatoire.

5.1 Problématique et état de l'art

Le problème de démélange en imagerie hyperspectrale consiste à décomposer chaque pixel (c'est à dire vecteur spectral) en une collection de signatures spectrales de composants purs, et à estimer la proportion de ces derniers dans le mélange [Keshava and Mustard, 2002, Bioucas-Dias and Plaza, 2010, Bioucas-Dias et al., 2012]. Selon le modèle de mélange linéaire, chaque vecteur spectral $\boldsymbol{x}$ de l'image vérifie

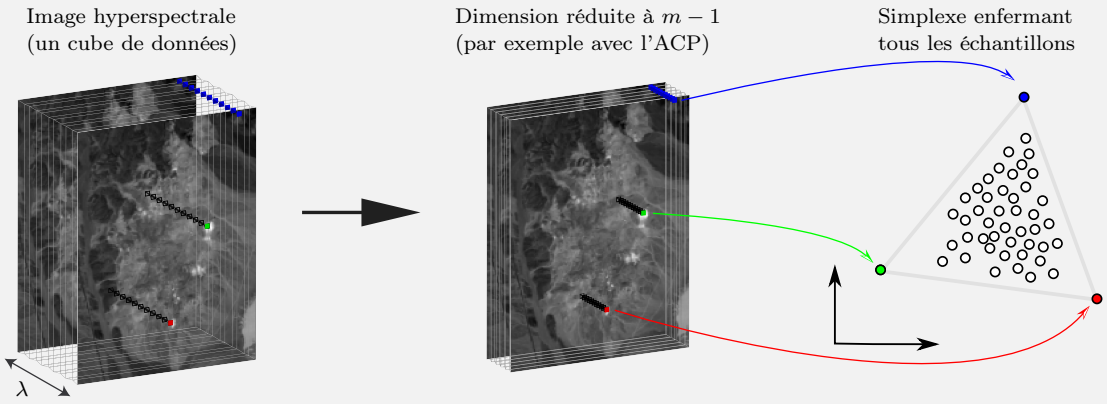
$$\boldsymbol{x} = \sum_{i=1}^{m} \alpha_i \, \boldsymbol{x}_{\omega_i} + \boldsymbol{\epsilon}, \tag{5.1}$$

où $\mathcal{D} = \{\boldsymbol{x}_{\omega_1}, \boldsymbol{x}_{\omega_2}, \ldots, \boldsymbol{x}_{\omega_m}\}$ désigne la collection fixe de signatures spectrales des éléments purs, α_i représente la contribution de $\boldsymbol{x}_{\omega_i}$ dans le mélange $\boldsymbol{x}$, et $\boldsymbol{\epsilon}$ l'erreur de modèle. Le modèle de démélange linéaire s'écrit aussi sous forme matricielle, selon

$$\boldsymbol{x} = \boldsymbol{X}_{\mathcal{D}} \, \boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

où $\boldsymbol{X}_{\mathcal{D}} = [\boldsymbol{x}_{\omega_1} \; \boldsymbol{x}_{\omega_2} \; \cdots \; \boldsymbol{x}_{\omega_m}]$ désigne la matrice des signatures spectrales des composants purs, et $\boldsymbol{\alpha} = [\alpha_1 \; \alpha_2 \; \cdots \; \alpha_m]^\top$ le vecteur de taille $(m \times 1)$ des fractions d'abondance associées au pixel $\boldsymbol{x}$.

1. [Chen et al., 2013d, Chen et al., 2013e, Honeine et al., 2013c, Honeine and Richard, 2012]
2. [Nguyen et al., 2013]
3. [Honeine et al., 2013a, Honeine and Lantéri, 2013, Chen et al., 2013c, Chen et al., 2013b, Nguyen et al., 2012, Chen et al., 2012a, Chen et al., 2011c, Honeine and Richard, 2011a, Chen et al., 2011d, Chen et al., 2011e, Honeine and Richard, 2010b, Honeine and Richard, 2010a]



[Cadre 15] **Schéma illustrant le problème de démélange en imagerie hyperspectrale.** Illustration du problème de démélange géométrique des données hyperspectrales, le nombre de composants purs étant fixé *a priori* (m, ici $m = 3$). Suite à une réduction de dimension à $m-1$, les techniques d'identification de signatures pures (N-Findr, SGA, VCA, ...) permettent de retrouver les composants purs comme étant les sommets du simplexe enfermant tous les échantillons.

Notons que la matrice $\boldsymbol{X}_{\mathcal{D}}$ est de plein rang colonne si les signatures spectrales des composants purs sont linéairement indépendantes. Dans la suite, $\mathbf{1}$ désigne le vecteur unité de taille $(m \times 1)$ et $\mathbf{0}$ désigne le vecteur nul de même taille.

Les méthodes de démélange hyperspectral existantes sont essentiellement de deux natures : géométriques ou statistiques, avec diverses extensions pour le démélange non linéaire [Keshava and Mustard, 2002, Bioucas-Dias and Plaza, 2010, Bioucas-Dias et al., 2012]. La plupart des méthodes de démélange traite le problème en deux étapes, d'abord l'identification des composants purs, ensuite l'estimation des fractions d'abondance. Le nombre m de composants purs est supposé connu *a priori*, ou encore estimé selon [Chang and Du, 2004] par exemple.

Diverses techniques d'identification des composants purs sont préconisées dans la littérature de traitement d'images hyperspectrales. Les plus connues sont N-Findr [Winter, 1999, Plaza and Chang, 2005], SGA [Chang et al., 2006], VCA [Nascimento and Dias, 2004], OSP [Harsanyi and Chang, 1994, Chang, 2005], ICE [Berman et al., 2004], sur lesquelles nous reviendrons par la suite. Afin d'extraire les composants purs, rien n'empêche d'utiliser le principe de dictionnaire parcimonieux comme présenté dans le chapitre précédent. Au-delà de ces stratégies variées tournées vers l'identification des composants purs, l'estimation des fractions d'abondance est souvent réalisée par résolution d'un problème inverse secondaire. La prise en compte de certaines contraintes physiques nécessite des techniques d'optimisation avancées, au prix d'un coût calculatoire supplémentaire important.

Afin que le résultat se prête à une interprétation physique, la complétude de la décomposition et l'additivité des contributions doivent être satisfaites. Ces contraintes, dites totales, se traduisent par les deux contraintes de convexité sur les fractions d'abondance :

- la somme unité, c'est à dire $\mathbf{1}^\top \boldsymbol{\alpha} = 1$. Cette contrainte de complétude permet d'avoir des coefficients de décomposition invariants par erreur spectrale additive. En effet, pour une translation $\boldsymbol{x}_0$ de tous les vecteurs spectraux, l'équivalence suivante est vraie : $\mathbf{1}^\top \boldsymbol{\alpha} = 1$ si et seulement si

$$\boldsymbol{x} + \boldsymbol{x}_0 = \left([\boldsymbol{x}_{\omega_1}\ \boldsymbol{x}_{\omega_2}\ \cdots\ \boldsymbol{x}_{\omega_m}] + \boldsymbol{x}_0 \mathbf{1}^\top \right) \boldsymbol{\alpha} \Leftrightarrow \boldsymbol{x} = [\boldsymbol{x}_{\omega_1}\ \boldsymbol{x}_{\omega_2}\ \cdots\ \boldsymbol{x}_{\omega_m}]\, \boldsymbol{\alpha}\ ;$$

- la non-négativité des coefficients d'abondance, notée avec $\boldsymbol{\alpha} \geq \mathbf{0}$. Cette contrainte implique l'additivité des contributions. Elle est illustrée dans le paragraphe 5.2.1 avec une interprétation géométrique.

Le mélange spectral est alors une combinaison convexe des signatures spectrales des composants purs.

Formulation alternative

La matrice $\boldsymbol{X}_{\mathcal{D}}$, de taille $(d \times m)$, regroupe l'ensemble des signatures spectrales des composants purs. Sa i-ème ligne, associée à une bande spectrale donnée, est désignée par $\boldsymbol{z}_i^\top$, pour $i = 1, 2, \ldots, d$, d étant le nombre total de bandes spectrales. En d'autres termes, $[\boldsymbol{x}_{\omega_1} \;\; \boldsymbol{x}_{\omega_2} \;\cdots\; \boldsymbol{x}_{\omega_m}] = [\boldsymbol{z}_1 \;\; \boldsymbol{z}_2 \;\cdots\; \boldsymbol{z}_d]^\top$. En conséquence, le système $\boldsymbol{x} = \boldsymbol{X}_{\mathcal{D}}\,\boldsymbol{\alpha}$ n'est autre qu'un ensemble de d équations linéaires,

$$x_i = \boldsymbol{z}_i^\top \boldsymbol{\alpha}, \tag{5.2}$$

c'est à dire une équation par bande spectrale, où $x_i = [\boldsymbol{x}]_i$. L'intérêt de cette formulation est montré dans la section 5.3.2 dans le cadre du modèle linéaire, et dans la section 5.4 pour le démélange non linéaire où une interaction entre les bandes spectrales est naturelle.

5.1.1 Démélange par la géométrie

Grâce à ses performances soutenues par des interprétations physiques et sa faible complexité de calcul, l'approche géométrique est fréquemment traitée dans la littérature. Elle se base essentiellement sur trois étapes, comme illustré dans le Cadre 15 : réduction de la dimension qui est portée à $m-1$; extraction de m composants purs par le principe d'un simplexe renfermant tous les échantillons ; et estimation de leur fraction d'abondance par la résolution d'un problème inverse. La suite de cette partie est consacrée à la présentation de ces trois étapes.

5.1.1.a Réduction de dimension par l'ACP

La dimension spectrale est ramenée à $m-1$, où m est le nombre d'éléments purs, en transformant chaque signature spectrale $\boldsymbol{x}$ en $\widetilde{\boldsymbol{x}} \in \mathbb{R}^{m-1}$. L'ACP [Jolliffe, 1986] vise à représenter les spectres dans un espace de faible dimension, avec le moins de perte possible en terme de variance des échantillons. L'ACP détermine la base orthonormée $\{\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{m-1}\}$, où chaque vecteur $\boldsymbol{w}_i$ est donné par maximisation de la variance projetée.

En désignant par $\boldsymbol{C}$ la matrice de covariance des vecteurs spectraux de l'image hyperspectrale étudiée, chaque vecteur $\boldsymbol{w}_i$ est donné par $\boldsymbol{w}_i = \operatorname{argmax}_{\boldsymbol{w}} \boldsymbol{w}^\top \boldsymbol{C}\,\boldsymbol{w}$ sous la contrainte de norme unité. La méthode des multiplicateurs de Lagrange permet de retrouver le problème de décomposition en vecteurs propres / valeurs propres $\boldsymbol{C}\boldsymbol{w}_i = \lambda_i \boldsymbol{w}_i$. Il convient de sélectionner les vecteurs propres associés aux plus grandes valeurs propres, qui mesurent la variance projetée. Ils constituent naturellement une base orthonormée, et la réduction de la dimension spectrale s'effectue finalement selon l'expression $\widetilde{\boldsymbol{x}} = [\boldsymbol{w}_1 \;\; \boldsymbol{w}_2 \;\cdots\; \boldsymbol{w}_{m-1}]^\top \boldsymbol{x}$, ou encore sous forme matricielle

$$\widetilde{\boldsymbol{X}} = [\boldsymbol{w}_1 \;\; \boldsymbol{w}_2 \;\cdots\; \boldsymbol{w}_{m-1}]^\top \boldsymbol{X}.$$

Le modèle de démélange s'écrit alors

$$\widetilde{\boldsymbol{x}} = \sum_{i=1}^{m} \alpha_i\, \widetilde{\boldsymbol{x}}_{\omega_i} + \widetilde{\boldsymbol{\epsilon}}, \tag{5.3}$$

où encore sous la forme matricielle : $\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{X}}_{\mathcal{D}}\,\boldsymbol{\alpha} + \widetilde{\boldsymbol{\epsilon}}$, avec $\widetilde{\boldsymbol{X}}_{\mathcal{D}} = [\widetilde{\boldsymbol{x}}_{\omega_1} \;\; \widetilde{\boldsymbol{x}}_{\omega_2} \;\cdots\; \widetilde{\boldsymbol{x}}_{\omega_m}]$.

5.1.1.b Identification des composants purs

Le modèle de mélange linéaire (5.3) suppose que les fractions d'abondance sont positives et qu'elles satisfont une contrainte de somme unité. Il en résulte que les échantillons sont contenus dans un simplexe dont les sommets sont les composants purs, comme illustré dans le Cadre 15. Pour extraire ces signatures, il suffit d'identifier le simplexe circonscrivant tous les échantillons. Les stratégies les plus utilisées dans la littérature sont présentées ci-après.

Maximisation du volume du simplexe

Cette stratégie, partagée par les algorithmes N-Findr [Winter, 1999] et SGA (pour *simplex growing algorithm*) [Chang et al., 2006], consiste à identifier le simplexe de plus grand volume en visitant chaque échantillon de sorte à faire croitre ce volume. Pour cela, notons que l'expression du volume d'un simplexe, dont les sommets sont données par $\widetilde{\boldsymbol{X}}_{\mathcal{D}} = [\widetilde{\boldsymbol{x}}_{\omega_1}\ \widetilde{\boldsymbol{x}}_{\omega_2}\ \cdots\ \widetilde{\boldsymbol{x}}_{\omega_m}]$, est

$$\mathrm{vol}(\mathcal{D}) = \frac{1}{(m-1)!}\ \det\begin{bmatrix}\mathbf{1}^\top \\ \widetilde{\boldsymbol{X}}_{\mathcal{D}}\end{bmatrix}. \tag{5.4}$$

L'algorithme N-Findr [Winter, 1999] opère selon une stratégie par substitution. Après une initialisation aléatoire de l'ensemble des sommets candidats avec les indices $\{\omega_1, \omega_2, \ldots, \omega_m\}$, l'étape suivante est itérée sur chaque échantillon $\widetilde{\boldsymbol{x}}$. Chaque élément de l'ensemble candidat est remplacé par $\widetilde{\boldsymbol{x}}$, et le volume du simplexe correspondant évalué. En plus du volume initial $\mathrm{vol}(\mathcal{D})$, m volumes sont disponibles, $\mathrm{vol}(\mathcal{D}\backslash\{\widetilde{\boldsymbol{x}}_{\omega_k}\} \cup \{\widetilde{\boldsymbol{x}}\})$, pour $k = 1, 2, \ldots, m$, où $\mathcal{D}\backslash\{\widetilde{\boldsymbol{x}}_{\omega_k}\} \cup \{\widetilde{\boldsymbol{x}}\}$ désigne l'ensemble obtenu de $\mathcal{D}$ en remplaçant le sommet $\widetilde{\boldsymbol{x}}_{\omega_k}$ par $\widetilde{\boldsymbol{x}}$. L'ensemble correspondant au plus grand volume, en valeur absolue, est alors retenu.

L'algorithme SGA [Chang et al., 2006] est une version incrémentale de l'algorithme N-Findr, chaque composant pur étant identifié successivement par une approche gloutonne. Les composants purs sont alors extraits tour à tour tel que, à l'étape i, la dimension ayant été préalablement réduite à $i-1$, le i-ème composant pur est identifié par l'indice ω_i qui maximise $|\mathrm{vol}(\mathcal{D} \cup \{\widetilde{\boldsymbol{x}}_{\omega_i}\})|$, où $\mathcal{D}$ correspond à l'ensemble déjà identifié à la précédente étape.

Maximisation des distances

Le simplexe à identifier est aussi celui dont les hauteurs sont les plus grandes. La maximisation de la distance d'un sommet au sous-espace engendré par les autres sommets est exploitée par les algorithmes VCA (pour *vertex component analysis*) [Nascimento and Dias, 2004] et OSP (pour *orthogonal subspace projection*) [Harsanyi and Chang, 1994]. Cette dernière est également utilisée par l'algorithme *Automatic Target Generation Process* [Ren and Chang, 2003]. Principalement, cette approche incrémentale s'intéresse à la distance des échantillons au sous-espace engendré par les composants purs déjà identifiés. Le nouveau composant pur est alors l'élément le plus éloigné. Le lien avec l'algorithme N-Findr est étudié dans [Du et al., 2008].

Une dernière approche consiste à caractériser le plus grand simplexe à partir des longueurs de ses arêtes. C'est le cas de l'algorithme ICE (pour *iterated constrained endmembers*) [Berman et al., 2004] qui considère la maximisation de la distance quadratique totale $\sum_{i,j=1}^{m} \|\widetilde{\boldsymbol{x}}_{\omega_i} - \widetilde{\boldsymbol{x}}_{\omega_j}\|^2$. Cette approche nécessite le calcul des distances de toutes les paires d'échantillons disponibles.

5.1.1.c Estimation des fractions d'abondance

Afin que le résultat se prête à une interprétation physique (complétude de la décomposition, additivité des contributions), comme il a été évoqué précédemment, les fractions d'abondance doivent satisfaire deux contraintes : la somme unité, c'est à dire $\mathbf{1}^\top\boldsymbol{\alpha} = 1$, et la non-négativité, c'est à dire $\boldsymbol{\alpha} \geq \mathbf{0}$. Cette dernière contrainte est illustrée par la géométrie dans la section 5.2.1. En ignorant ces contraintes, la solution optimale au sens des moindres carrés, c'est à dire

$$\boldsymbol{\alpha}^{\mathrm{LS}} = \underset{\boldsymbol{\alpha}}{\mathrm{argmin}}\, \|\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{X}}_{\mathcal{D}}\,\boldsymbol{\alpha}\|^2,$$

est donnée par

$$\boldsymbol{\alpha}^{\mathrm{LS}} = (\widetilde{\boldsymbol{X}}_{\mathcal{D}}^\top \widetilde{\boldsymbol{X}}_{\mathcal{D}})^{-1}\boldsymbol{X}_{\mathcal{D}}^\top\widetilde{\boldsymbol{x}}. \tag{5.5}$$

La contrainte de somme unité est la plus simple à imposer. Deux approches sont souvent considérées.

La première approche consiste à introduire l'expression de la somme unité dans le modèle de démélange linéaire. Ceci est possible grâce à la nature de cette contrainte, produisant ainsi le système augmenté suivant :

$$\begin{bmatrix}1 \\ \widetilde{\boldsymbol{x}}\end{bmatrix} = \begin{bmatrix}\mathbf{1}^\top \\ \widetilde{\boldsymbol{X}}_{\mathcal{D}}\end{bmatrix}\boldsymbol{\alpha} + \begin{bmatrix}\epsilon \\ \boldsymbol{\epsilon}\end{bmatrix}. \tag{5.6}$$

Précisons toutefois que la solution aux moindres carrés de ce système ne permet pas d'imposer exactement la somme unité. Le prix est la présence d'une erreur qui se trouve désormais combinée avec les erreurs associées au modèle de mélange linéaire.

La seconde approche consiste à résoudre le problème d'optimisation avec contraintes en utilisant la méthode des multiplicateurs de Lagrange, pour avoir

$$\boldsymbol{\alpha}^{\text{eqLS}} = \boldsymbol{\alpha}^{\text{LS}} - \frac{1}{\mathbf{1}^\top (\widetilde{\boldsymbol{X}}_{\mathcal{D}}^\top \widetilde{\boldsymbol{X}}_{\mathcal{D}})^{-1}\mathbf{1}} (\widetilde{\boldsymbol{X}}_{\mathcal{D}}^\top \widetilde{\boldsymbol{X}}_{\mathcal{D}})^{-1}\mathbf{1}(\mathbf{1}^\top \boldsymbol{\alpha}^{\text{LS}} - 1). \tag{5.7}$$

Nous pouvons simplifier davantage cette expression. Pour cela, la somme unité permet de réécrire l'identité (5.3) selon $\sum_{i=1}^{m} \alpha_i \, (\widetilde{\boldsymbol{x}}_{\omega_i} - \widetilde{\boldsymbol{x}}) = \mathbf{0}$. En reprenant les deux expressions (5.5) et (5.7) dans ce cas, nous avons

$$\boldsymbol{\alpha}^{\text{eqLS}} = \frac{\boldsymbol{K}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{K}^{-1}\mathbf{1}}, \tag{5.8}$$

où $\boldsymbol{K}$ désigne la matrice de Gram locale, d'élément général $[\boldsymbol{K}]_{ij} = (\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{\omega_i})^\top (\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_{\omega_j})$, pour $i, j = 1, 2, \ldots, m$.

La contrainte de non-négativité ne peut pas être introduite explicitement dans la solution comme ci-dessus, et nécessite l'usage de méthodes de résolution itératives [Lawson and Hanson, 1987, Lantéri et al., 2001, Heinz and Chang, 2001]. Voir les sections 5.3.1 et 5.3.2.

5.1.2 Résumé des principales contributions

Nos contributions dans la résolution du problème de démélange sont essentiellement structurées autour de deux axes : d'une part, la maîtrise du problème de démélange linéaire, aussi bien par la géométrie que par les méthodes statistiques sous contraintes ; d'autre part, la définition de nouveaux modèles non linéaires et le développement de techniques de résolution appropriées. A ces deux axes de recherche s'ajoute l'intégration de l'information spatiale dans la résolution du problème. Les différentes contributions sont résumées dans la suite :

- En démélange par la géométrie, la plupart des méthodes de démélange traite le problème de démélange en deux étapes : d'abord l'identification des composants purs, avec N-Findr, SGA, VCA, OSP, etc. ; ensuite l'estimation des fractions d'abondance qui est souvent réalisée par la résolution d'un problème inverse secondaire. Nous montrons que ces techniques géométriques d'extraction des composants purs permettent d'estimer conjointement les fractions d'abondance, pour un coût calculatoire supplémentaire négligeable. Voir la section 5.2.1.
- Pour faire du démélange non linéaire par la géométrie, nous adaptons ce socle commun pour l'identification des composants purs et l'estimation des fractions d'abondance. Pour cela, nous proposons de l'étendre au problème de démélange non linéaire, en modifiant l'étape de réduction de dimension en une technique non linéaire. Nous montrons notre méthode au travers d'une approche de type géodésique (ISOMAP) et une approche localement linéaire (LLE) . Voir la section 5.2.2.
- En démélange linéaire par méthodes statistiques, notre intention est d'illustrer que les contraintes totales peuvent être facilement imposées en modifiant certaines techniques classiques. D'une part, nous montrons que la méthode de descente de gradient peut être adaptée pour satisfaire les contraintes totales. Voir la section 5.3.1. D'autre part, nous explorons les techniques de projections multiples dont les célèbres techniques de Cimmino et de Kaczmarz. Voir la section 5.3.2.
- En démélange non linéaire, nous décrivons un nouveau paradigme sur l'hypothèse que le mécanisme de mélange peut être décrit par un mélange linéaire avec une fluctuation additive non linéaire. En décrivant cette fluctuation dans le cadre des méthodes à noyaux, nous exploitons divers modèles de non-linéarité, dont les mélanges par noyaux multiples et par mélange post-non-linéaire. Voir la section 5.4. Une étude sur le choix de la non-linéarité et l'interprétation physique de certains noyaux est donnée dans la section 5.4.2.
- Il est judicieux d'intégrer l'information spatiale et ainsi profiter de la dualité spatiale-spectrale pour la résolution du problème de démélange. Nous proposons une méthode pour résoudre ce

problème en considérant une régularisation spatiale. L'approche proposée est étudiée en détail dans la section 5.5 dans le cadre du modèle combinant un mélange linéaire avec une fluctuation non linéaire.

– Dans la section 5.5.2, nous décrivons une nouvelle formulation du problème de démélange non linéaire. Notre travail tire toute son originalité de l'écriture du problème de démélange selon un problème de pré-image. Ainsi, les diverses techniques présentées dans le chapitre 3 peuvent être utilisées pour le démélange spectral. Nous montrons que la méthode de résolution par transformation conforme que nous proposons dans la section 3.2.2 permet d'intégrer la régularisation spatiale ainsi que les contraintes totales.

Les travaux entamés sont résumés tout au long de ce chapitre. Mais avant, nous rappelons que nos contributions en classification ne seront pas présentées dans la suite, en renvoyant le lecteur vers [Honeine and Richard, 2010a] pour le choix d'une représentation adaptée, et vers [Noumir et al., 2011b] pour des algorithmes de classification à complexité réduite (voir aussi le chapitre suivant).

5.2 Contributions au démélange par la géométrie

Par ses performances soutenues avec des interprétations physiques et sa faible complexité de calcul, l'approche géométrique est fréquemment traitée dans la littérature pour extraire les composants purs. Ces derniers sont identifiés comme étant les sommets du simplexe enfermant tous les échantillons. Dans la littérature, l'estimation des fractions d'abondance a toujours été réalisée dans un second temps, par résolution d'un problème inverse. La résolution de ce problème n'a malheureusement pas profité de la simplicité de l'approche géométrique.

Nous montrons dans la suite que les techniques géométriques classiques d'extraction des composants purs permettent d'estimer conjointement les fractions d'abondance, pour un coût calculatoire supplémentaire négligeable. Pour ce faire, un socle commun d'interprétations géométriques du problème est proposé. Nous pouvons le décliner pour mieux l'adapter à la technique retenue pour l'extraction de composants purs. Le caractère géométrique de l'approche lui confère une flexibilité très appréciable dans le cadre de techniques de démélange géométrique. Nous montrons dans [Honeine and Richard, 2010b] l'intérêt de notre approche pour la technique N-Findr, avant de l'étendre aux techniques SGA, VCA, OSP et ICE dans [Honeine and Richard, 2011a, Honeine and Richard, 2012]. La section 5.2.1 ci-après est dédiée à présenter notre approche, Des résultats expérimentaux sont présentés dans le Cadre 17.

Le démélange linéaire, dont les techniques géométriques, hérite de la propriété de linéarité des différentes étapes considérées. Tout en gardant la même approche géométrique, aussi bien pour identifier les composants purs que pour estimer la proportion de ces derniers dans le mélange, nous proposons de l'étendre au domaine non linéaire. Pour ce faire, nous exploitons diverses techniques de réduction de dimensionnalité, dont les approches de type MDS [Cox and Cox, 2000], les distances géodésiques (ISOMAP) [Tenenbaum et al., 2000], ou encore l'approche localement linéaire (LLE) [Roweis and Saul, 2000]. L'approche proposée est étudiée dans [Nguyen et al., 2012, Honeine et al., 2013c], et résumée dans la section 5.2.2, avec des expérimentations présentées dans le Cadre 18 et le Cadre 19.

5.2.1 La « géométrie » des abondances, ou l'estimation des abondances sans coût supplémentaire

Afin de résoudre le problème de démélange sous la contrainte de somme unité, la contrainte de non-négativité étant étudiée plus loin, nous proposons de résoudre le système augmenté [4] suivant :

$$\begin{bmatrix} \mathbf{1}^\top \\ \widetilde{\boldsymbol{X}}_{\mathcal{D}} \end{bmatrix} \boldsymbol{\alpha} = \begin{bmatrix} 1 \\ \widetilde{\boldsymbol{x}} \end{bmatrix}.$$

4. Le système augmenté proposé ici est utilisé à d'autres fins que celles du système étudié dans [Heinz and Chang, 2001], sans les inconvénients de la résolution aux moindres carrés évoqués dans (5.6).

[CADRE 16] Illustration de la géométrie des coefficients d'abondance.

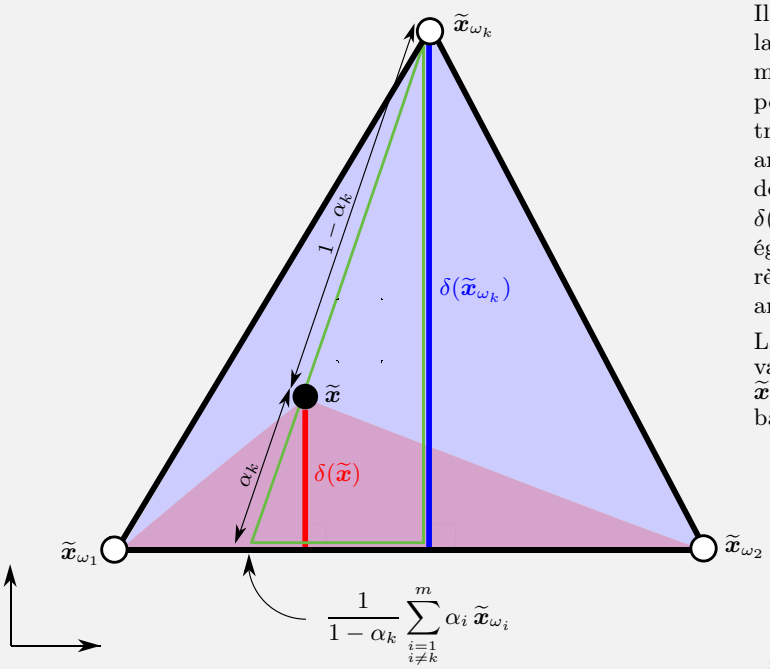
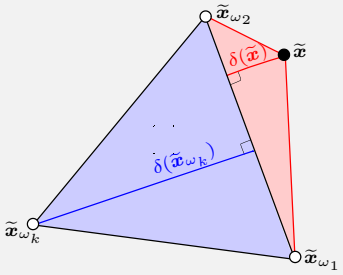


Illustration géométrique de α_k, la contribution de $\widetilde{\boldsymbol{x}}_{\omega_k}$ dans le mélange $\widetilde{\boldsymbol{x}}$. En 2D, α_k correspond au rapport des aires du triangle rouge et du grand triangle bleu, ou encore au rapport des distances à la base commune $\delta(\widetilde{\boldsymbol{x}})/\delta(\widetilde{\boldsymbol{x}}_{\omega_k})$. Cette relation est également donnée par le théorème de Thalès appliqué au triangle rectangle vert.

La figure ci-dessous illustre une valeur négative de α_k, où $\widetilde{\boldsymbol{x}}$ et $\widetilde{\boldsymbol{x}}_{\omega_k}$ sont des deux côtés de la base définie par $\{\widetilde{\boldsymbol{x}}_{\omega_1}, \widetilde{\boldsymbol{x}}_{\omega_2}\}$.



La solution de ce système de m équations à m inconnues est donnée par la règle de Cramer [Strang, 2003] :

$$\alpha_k = \frac{\det\begin{bmatrix}\mathbf{1}^\top \\ \widetilde{\boldsymbol{X}}_{\mathcal{D}\setminus\{\widetilde{\boldsymbol{x}}_{\omega_k}\}\cup\{\widetilde{\boldsymbol{x}}\}}\end{bmatrix}}{\det\begin{bmatrix}\mathbf{1}^\top \\ \widetilde{\boldsymbol{X}}_{\mathcal{D}}\end{bmatrix}}, \tag{5.9}$$

où $\widetilde{\boldsymbol{X}}_{\mathcal{D}\setminus\{\widetilde{\boldsymbol{x}}_{\omega_k}\}\cup\{\widetilde{\boldsymbol{x}}\}}$ désigne la matrice obtenue de $\widetilde{\boldsymbol{X}}_{\mathcal{D}}$ en remplaçant sa k-ème colonne $\widetilde{\boldsymbol{x}}_{\omega_k}$ par $\widetilde{\boldsymbol{x}}$. Le choix d'une telle approche pour résoudre ce problème réside dans l'interprétation géométrique de l'expression (5.9). En reprenant l'expression (5.4) du volume d'un simplexe, il en résulte immédiatement que

$$\alpha_k = \frac{\mathrm{vol}(\mathcal{D}\setminus\{\widetilde{\boldsymbol{x}}_{\omega_k}\}\cup\{\widetilde{\boldsymbol{x}}\})}{\mathrm{vol}(\mathcal{D})}, \tag{5.10}$$

pour tout $k = 1, 2, \ldots, m$. Ces volumes sont souvent déjà calculés par les techniques d'identification de composants purs, comme avec les algorithmes N-Findr et SGA. Ceci permet d'estimer les fractions d'abondance sans coût calculatoire supplémentaire.

L'expression précédente se prête à une simplification puisqu'elle implique deux simplexes qui partagent $m-1$ sommets définissant une base commune et qui diffèrent par l'un de leurs sommets, ici $\widetilde{\boldsymbol{x}}_{\omega_k}$ et $\widetilde{\boldsymbol{x}}$. Or, le volume d'un simplexe est proportionnel au produit d'une hauteur (c'est à dire la distance entre un sommet et la base associée, définie par le sous-espace engendré par les autres sommets) et le volume de la base correspondante (c'est à dire le volume du simplexe engendré par les autres sommets). Au jour de cette définition, l'expression (5.10) se ramène à

$$\alpha_k = \frac{\delta(\widetilde{\boldsymbol{x}})}{\delta(\widetilde{\boldsymbol{x}}_{\omega_k})}, \tag{5.11}$$

pour tout $k = 1, 2, \ldots, m$, où $\delta(\widetilde{\boldsymbol{x}})$ est la distance entre le sommet $\widetilde{\boldsymbol{x}}$ et la base commune formée par les éléments de l'ensemble $\mathcal{D}$ privé de $\widetilde{\boldsymbol{x}}_{\omega_k}$ et $\delta(\widetilde{\boldsymbol{x}}_{\omega_k})$ la distance de $\widetilde{\boldsymbol{x}}_{\omega_k}$ à cette même base.

Cette formulation est particulièrement adaptée aux techniques reposant sur la maximisation d'une distance, telles que VCA et OSP, pour identifier les sommets du simplexe. Voir aussi [Luo et al., 2008] pour une interprétation similaire. Une autre interprétation peut également être obtenue à partir de la décomposition

$$\widetilde{\boldsymbol{x}} = \alpha_k \, \widetilde{\boldsymbol{x}}_{\omega_k} + (1-\alpha_k) \sum_{\substack{i=1 \\ i \neq k}}^{m} \frac{\alpha_i}{\sum_{j=1, j\neq k}^{m} \alpha_j} \, \widetilde{\boldsymbol{x}}_{\omega_i}, \tag{5.12}$$

après avoir noté que le dénominateur dans cette expression vaut $1-\alpha_k$. Il en résulte que $\widetilde{\boldsymbol{x}}$ est le mélange de $\widetilde{\boldsymbol{x}}_{\omega_k}$ et d'une signature spectrale équivalente aux autres composants purs $\widetilde{\boldsymbol{x}}_{\omega_i}$, $i \neq k$, en proportions α_k et $1-\alpha_k$. De manière beaucoup plus synthétique, les deux relations (5.10) et (5.12) reprennent le théorème de Thalès illustré dans le Cadre 16, que nous exprimons à partir du sommet équivalent défini dans (5.12) par $\frac{1}{1-\alpha_k} \sum_{i=1, i\neq k}^{m} \alpha_i \, \widetilde{\boldsymbol{x}}_{\omega_i}$.

Toujours afin de construire le simplexe circonscrivant tous les échantillons, une troisième stratégie est considérée dans la littérature en recherchant le simplexe d'arêtes les plus longues, comme avec la méthode ICE. Cette approche nécessite le calcul des distances de toutes les paires d'échantillons. Celles-ci peuvent alors être aisément utilisées pour déterminer le volume du simplexe correspondant puisque les déterminants de Cayley-Menger [Sommerville, 1958] permettent d'écrire

$$\mathrm{vol}(\mathcal{D})^2 = \frac{(-1)^m}{2^{m-1}((m-1)!)^2} \det \begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & 0 & \delta_{1,2}^2 & \cdots & \delta_{1,m}^2 \\ 1 & \delta_{2,1}^2 & 0 & \cdots & \delta_{2,m}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \delta_{m,1}^2 & \delta_{m,2}^2 & \cdots & 0 \end{bmatrix},$$

où $\delta_{i,j}$ désigne la distance entre deux sommets. Cette expression est la généralisation de la formule de Héron pour le calcul de l'aire d'un triangle quelconque avec $\mathrm{vol}(\mathcal{D})^2 = \frac{1}{16}(\delta_{1,2}+\delta_{1,3}+\delta_{2,3})(\delta_{1,2}+\delta_{1,3}-\delta_{2,3})(\delta_{1,2}-\delta_{1,3}+\delta_{2,3})(-\delta_{1,2}+\delta_{1,3}+\delta_{2,3})$. Une fois encore, ces considérations géométriques peuvent être exploitées afin de déterminer directement les fractions d'abondance correspondantes, sans développement calculatoire lourd.

Remarque sur la non-négativité

Il est à noter que la contrainte de non-négativité n'est pas imposée dans l'approche géométrique considérée. En effet, la violation de cette contrainte signifie que $\widetilde{\boldsymbol{x}}$ est à l'extérieur du simplexe. Ceci témoigne notamment, d'une part de l'inadéquation du modèle de mélange linéaire, et d'autre part des limites de la technique d'identification des signatures des composants purs. La méthode proposée montre que, dans ce cas, les deux déterminants dans (5.9) sont de signes opposés. Ceci s'exprime selon (5.11) par des éléments dans chaque demi-plan défini par la base commune, comme illustré dans l'illustration en bas à droite du Cadre 16.

La violation de la contrainte de non-négativité, c'est à dire la présence d'échantillons à l'extérieur du simplexe, est essentiellement liée à la distribution des échantillons obtenue par la réduction de dimension. Nous étudions dans la suite une possibilité pour remédier éventuellement à cet inconvénient, en considérant diverses techniques de réduction de dimension.

5.2.2 Du linéaire au non linéaire par réduction de dimension

La méthode de démélange étudiée jusqu'à présent hérite la propriété de linéarité des différentes étapes considérées. Tout en gardant la même approche géométrique présentée dans la partie précédente, nous exploitons diverses représentations des données hyperspectrales par des techniques de réduction de dimension.

Les méthodes de type MDS

L'approche MDS [Cox and Cox, 2000, Kruskal, 1964a, Kruskal, 1964b] représente les échantillons dans un espace de dimension réduite à partir des distances (ou dissimilarités) entre eux. Sans perte

de généralité, nous résumons dans la suite la mise en œuvre de la méthode MDS classique de Torgerson [Torgerson, 1952]. Le coût à minimiser repose sur l'écart entre les produits scalaires dans chaque espace, c'est à dire $\sum_{i,j}(\boldsymbol{x}_i^\top \boldsymbol{x}_j - \widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{x}}_j)^2$. L'algorithme MDS classique se résume aux étapes suivantes : à partir des distances $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$, pour $i, j = 1, 2, \ldots, n$, la matrice des produits scalaires est déterminée à partir de l'identité $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \|\boldsymbol{x}_i\|^2 + \|\boldsymbol{x}_j\|^2 - 2\,\boldsymbol{x}_i^\top \boldsymbol{x}_j$. Par diagonalisation de cette matrice, nous avons les vecteurs propres normés $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{m-1}$ associés aux plus grandes valeurs propres $\lambda_1, \ldots, \lambda_{m-1}$. Les coordonnées des échantillons dans l'espace de dimension $m-1$ sont alors :

$$[\widetilde{\boldsymbol{x}}_1 \;\; \widetilde{\boldsymbol{x}}_2 \;\; \cdots \;\; \widetilde{\boldsymbol{x}}_n] = \begin{bmatrix} \sqrt{\lambda_1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sqrt{\lambda_{m-1}} \end{bmatrix} \begin{bmatrix} \boldsymbol{w}_1^\top \\ \vdots \\ \boldsymbol{w}_{m-1}^\top \end{bmatrix}.$$

ISOMAP : une approche de type géodésique

Les méthodes MDS reposent essentiellement sur une métrique euclidienne, perdant ainsi la notion de variété. L'approche ISOMAP [Tenenbaum et al., 2000] consiste à appliquer la technique MDS à des distances géodésiques, c'est à dire mesurées sur la variété. Ces distances sont estimées à partir de l'information sur le voisinage des échantillons. L'algorithme ISOMAP se résume en trois étapes :

1. Le voisinage de chaque échantillon est identifié. L'ensemble des voisinages est représenté sous forme d'arêtes d'un graphe. Chaque arête est pondérée par la distance $\delta_{i,j}$.
2. A partir du graphe de voisinage, le plus court chemin entre chaque couple de sommets est déterminé. Le plus court chemin est estimé par l'algorithme de Djikstra ou de Floyd [5].
3. La matrice des plus courts chemins, constituant ainsi une matrice de distances géodésiques, est alors utilisée avec la technique MDS afin de construire un sous-espace euclidien.

L'usage de distances géodésiques s'adapte facilement, à l'identification de composants purs avec la méthode ICE d'une part, et à l'estimation des volumes des simplexes à partir du déterminant de Cayley-Menger d'autre part [Nguyen et al., 2012]. D'autres récents travaux confirment l'intérêt de notre approche. Voir par exemple [Heylen et al., 2011, Heylen and Scheunders, 2012].

LLE : une approche localement linéaire

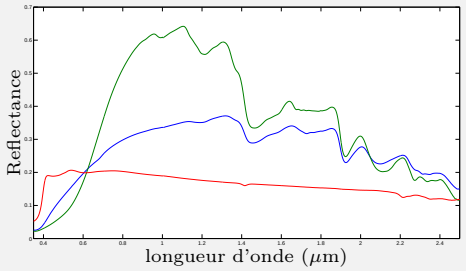
Les approches linéaires telles que l'ACP ne conservent pas nécessairement le voisinage des échantillons dans l'espace de dimension réduite, au risque de produire des distorsions aussi bien locales que globales. Une telle perte dans la géométrie des échantillons peut se traduire par certains éléments à l'extérieur du simplexe, voire une modification de ses sommets. La méthode LLE [Roweis and Saul, 2000] cherche à surmonter cette difficulté, en opérant en trois étapes :

1. Le voisinage de chaque élément est déterminé. Soit $\widetilde{m}$ le nombre de voisins retenus, avec $\widetilde{m} \geq m$.
2. Chaque vecteur spectral $\boldsymbol{x}_\ell$ est représenté par une combinaison linéaire de ses voisins, selon $\boldsymbol{x}_\ell \approx \sum_{i=1}^{\widetilde{m}} w_{\ell,i}\, \boldsymbol{x}_{\ell_i}$, où $\{\boldsymbol{x}_{\ell_1}, \boldsymbol{x}_{\ell_2}, \ldots, \boldsymbol{x}_{\ell_{\widetilde{m}}}\}$ désigne la collection des $\widetilde{m}$ voisins de $\boldsymbol{x}_\ell$. En incluant la contrainte de somme unité, les coefficients sont obtenus selon l'expression (5.8), avec
$$\boldsymbol{w}_\ell = \frac{\boldsymbol{K}_\ell^{-1}\mathbf{1}_{\widetilde{m}}}{\mathbf{1}_{\widetilde{m}}^\top \boldsymbol{K}_\ell^{-1}\mathbf{1}_{\widetilde{m}}},$$
où $\boldsymbol{K}_\ell$ est la matrice de Gram locale de taille $(\widetilde{m} \times \widetilde{m})$ d'élément général $(\boldsymbol{x}_\ell - \boldsymbol{x}_{\ell_i})^\top (\boldsymbol{x}_\ell - \boldsymbol{x}_{\ell_j})$.
3. En utilisant ces coefficients, les coordonnées des échantillons dans l'espace de dimension $m-1$ sont données selon le problème d'optimisation
$$\min_{\substack{\widetilde{\boldsymbol{x}}_\ell \in \mathbb{R}^{m-1} \\ \ell = 1,2,\ldots}} \sum_\ell \|\widetilde{\boldsymbol{x}}_\ell - \sum_{i=1}^{\widetilde{m}} w_{\ell,i}\, \widetilde{\boldsymbol{x}}_{\ell_i}\|^2.$$

5. L'agorithme de Floyd consiste à remplacer, en parcourant tous les sommets avec $k = 1, 2, \ldots$, toutes les valeurs $\delta_{i,j}$ par $\min\{\delta_{i,j}, \delta_{i,k} + \delta_{k,j}\}$. Cette procédure permet de converger vers la matrice des plus courts chemins entre tous les sommets du graphe.

[Cadre 17] Expérimentations sur l'estimation des abondances par la géométrie.

Nous illustrons l'intérêt de la méthode proposée dans la section 5.2.1. Nous nous intéressons au problème de démélange de données hyperspectrales, synthétisées à partir de trois composants purs : « herbe dorée », « cédre » et « asphalte » du *USGS Library* [Clark and Geological Survey (U.S.), 2007], définis par GDS480, GDS357 et GDS368, respectivement. Leurs signatures spectrales sont illustrées dans la figure ci-contre, avec 2 151 bandes couvrant les longueurs d'onde de 0,35 à 2,5 μm. Une image hyperspectrale de taille 64×64 a été générée, selon le modèle linéaire $\widetilde{\boldsymbol{x}}_\ell = \sum_{i=1}^{3} \alpha_{\ell,i}\, \widetilde{\boldsymbol{x}}_{\omega_i} + \widetilde{\boldsymbol{\epsilon}}_\ell$, pour $\ell = 1, 2, \ldots, 4\,096$, et $\widetilde{\boldsymbol{\epsilon}}_\ell$ un bruit blanc gaussien de moyenne nulle et de variance $2{,}5 \cdot 10^{-3}$. La variance du bruit est choisie pour avoir un rapport signal sur bruit de 20 dB en moyenne. Les coefficients d'abondance ont été générés selon une loi uniforme sur le simplexe. Pour cela, il suffit de les générer selon la loi de Dirichlet de paramètres égaux à l'unité, puis de les normaliser pour vérifier la contrainte de somme unité. Les signatures des composants purs ont été estimées à partir de l'image générée selon le modèle bruité précédent.
Suite à une réduction de dimension par une ACP classique, ces composants purs ont été convenablement retrouvés par la plupart des techniques d'extraction. Nous avons comparé l'une de nos approches, celle consistant à associer N-Findr et la relation (5.10), à différentes méthodes de moindres carrés pour l'estimation de fractions d'abondance. Le tableau ci-dessous compare la violation des contraintes de non-négativité et de somme unité, pour les différentes techniques. A ces violations, s'ajoutent des valeurs des fractions d'abondance supérieures à l'unité pour les deux algorithmes de moindres carrés : avec contrainte de somme unité et/ou avec contrainte de non-négativité [Lawson and Hanson, 1987, Heinz and Chang, 2001]. Précisons pour finir que l'approche proposée nécessite nettement moins de ressources que les autres techniques, avec 3×4096 divisions arithmétiques seulement, l'algorithme étant couplé à N-Findr. Les autres techniques ont nécessité au mieux une inversion matricielle.



	$\exists\, i\colon \alpha_i < 0$	$\mathbf{1}^\top\boldsymbol{\alpha} \neq 1$	temps
Moindres carrés (5.5)	0,05%	98%	0,847
- avec somme unité (5.7)	4,95%	0	5,151
- avec non-négativité	0	98%	5,164
- avec contraintes totales	0	0	1,542
Méthode décrite dans 5.2.1	2,19%	0	0,173

La contrainte $\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i = \mathbf{0}$, souvent ajoutée, permet d'annihiler l'invariance par translation, ainsi qu'une contrainte sur l'échelle en imposant une matrice de covariance identité.

5.3 Contributions au démélange linéaire par les méthodes statistiques

Les méthodes des moindres carrés sont le fer de lance de l'approche statistique. L'estimation des fractions d'abondance impose au problème de moindres carrés les contraintes de somme unité et de non-négativité. Dans [Heinz and Chang, 2001], les auteurs proposent un algorithme itératif qui impose la contrainte de somme unité dans un premier temps en utilisant le système augmenté (5.6), puis dans un second temps effectue l'algorithme *active-set* de [Lawson and Hanson, 1987] pour satisfaire la non-négativité. Cette méthode ne satisfait pas la contrainte de somme unité exactement, et l'algorithme *active-set* pour les moindres carrés non-négatifs se heurte toutefois à une complexité de calcul importante. Dans [Theys et al., 2009], une autre classe d'algorithmes est proposée où la contrainte de somme unité est assurée à chaque itération par normalisation, c'est à dire en remplaçant l'estimation courante $\boldsymbol{\alpha}_{t+1}$ par $\boldsymbol{\alpha}_{t+1}/(\mathbf{1}^\top\boldsymbol{\alpha}_{t+1})$. Notons toutefois qu'il ne s'agit pas d'une projection orthogonale, et par la suite cette opération n'est pas non-expansivité. Cela signifie que les distances entre deux éléments projetés ne sont pas garanties d'être inférieures aux distances des éléments originaux.

Nous proposons dans la suite deux méthodes de démélange linéaire avec contraintes totales. La

[CADRE 18] Influence de la réduction de dimension pour l'estimation des abondances.

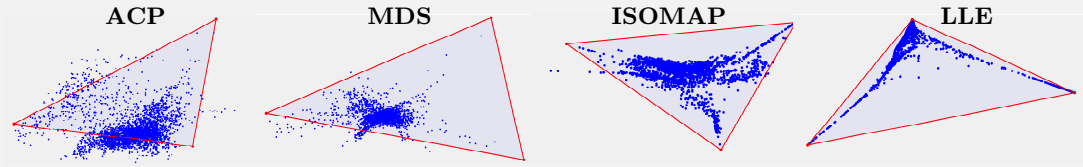
Nous considérons la scène du site minier de Cuprite, au Nevada (Etats-Unis), qui a suscité de nombreuses études. L'image, extraite de la base AVIRIS, est décrite en détail dans [Nascimento and Dias, 2004]. Une portion de l'image est étudiée, avec 30 lignes verticales et 100 échantillons par ligne. Voir [Honeine et al., 2013c]. Celle-ci correspond à 224 canaux spectraux de résolution 10 nm, couvrant les longueurs d'ondes de 400 à 2 500 nm. La résolution spatiale est de 20 m. Afin de corriger l'absorption eau-vapeur ainsi que le faible rapport signal sur bruit, les bandes spectrales 1-2, 104-113, 148-167, et 221-224 ont été supprimées. On obtient alors 188 bandes spectrales.

Afin de représenter les données en 2D, nous nous sommes intéressés à l'extraction de $n = 3$ composants purs à titre d'exemple. Les figures ci-dessous illustrent l'influence de la méthode de réduction de dimension sur la géométrie des données dans l'espace de dimension réduite, à $n - 1 = 2$. Pour cela, nous avons étudié les algorithmes ACP, MDS Classique, ISOMAP et LLE. Pour les deux derniers algorithmes, le nombre de voisins a été fixé à $m = 7$, des valeurs plus grandes ayant conduit à des résultats très proches. Les composants purs ont été extraits par l'algorithme N-Findr, identifiant ainsi le triangle (2D simplexe) renfermant la plupart des données. Il est clair que l'ACP, la méthode la plus utilisée dans la littérature, donne les moins bons résultats en termes d'échantillons circonscrits dans le simplexe.

Nous avons déterminé les cartes d'abondance de ces composants purs, en estimant conjointement les fractions d'abondance avec N-Findr comme décrit dans la section 5.2.1. Le CADRE 19 présente, pour chacune des techniques de réduction de dimension, les trois cartes d'abondance, ainsi que la carte de présence de valeurs négatives des fractions d'abondance. Cette dernière montre la pertinence de l'algorithme LLE pour l'image étudiée, avec des valeurs d'abondance $\alpha_i > -10^{-4}\ \forall i$, ainsi que l'algorithme ISOMAP avec $\alpha_i > -0{,}2\ \forall i$. Une comparaison plus détaillée de la violation de la contrainte de non-négativité est donnée dans le tableau ci-dessous.

Une comparaison du coût calculatoire des trois étapes (réduction de dimension, identification des composants purs et des fractions d'abondance) est donnée dans le tableau ci-dessous en termes de temps de calcul. Ces valeurs ont été déterminées à partir d'une implémentation sous Matlab 7,9 sur un ordinateur portable Macbook Pro Intel Core 2 Duo 2,53 GHz, avec 4 GB de mémoire.

	$\exists i\colon \alpha_i < 0$		temps
	nombre	%	mm:ss
ACP	976	32,5	00 : 01
MDS Classique	79	2,6	1 : 45
ISOMAP	10	0,3	15 : 03
LLE	6	0,2	00 : 05



première est une méthode de descente de gradient adaptée pour imposer ces contraintes, alors que la seconde explore des techniques de projections multiples.

5.3.1 Démélange par descente de gradient avec contraintes totales

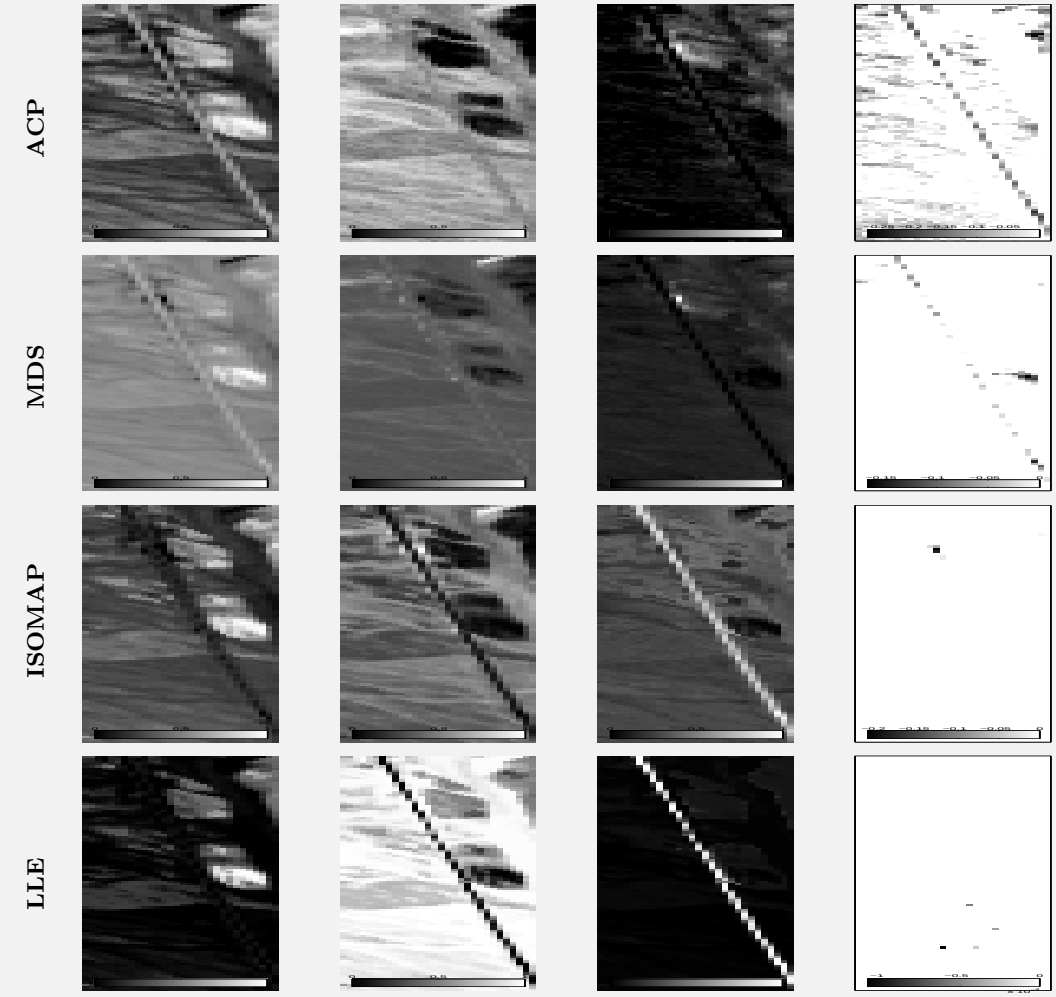
Nous proposons une méthode de descente de gradient qui ne nécessite aucune projection supplémentaire sur l'espace des contraintes. Elle intègre les deux contraintes de somme unité et de non-négativité dans la règle de mise à jour des coefficients. Ainsi les deux contraintes sont-elles toujours satisfaites. Nous présentons dans la suite cette méthode, en renvoyant le lecteur à [Chen et al., 2011e] pour une étude analytique sur la convergence de l'algorithme proposé.

Solution avec contraintes de non-négativité

Considérons la minimisation de la fonction coût quadratique

$$\mathcal{J}(\boldsymbol{\alpha}) = \|\boldsymbol{x} - \boldsymbol{X}_{\mathcal{D}}\,\boldsymbol{\alpha}\|^2,$$

[CADRE 19] **Les cartes d'abondances pour les expérimentations décrites dans le CADRE 18.** Pour chaque méthode de réduction de dimension, les trois cartes d'abondance sont associées aux trois composants purs. La quatrième carte correspond à la distribution spatiale des valeurs négatives des fractions d'abondance.



sous contrainte de non-négativité. Les conditions de Karush-Kuhn-Tucker au minimum sont résumées par l'identité $\alpha_i[-\nabla_{\boldsymbol{\alpha}}\mathcal{J}(\boldsymbol{\alpha})]_i = 0$ pour tout $i = 1, 2, \ldots, m$, où le signe négatif permet d'illustrer la descente de gradient de $\mathcal{J}(\boldsymbol{\alpha})$, le gradient étant désigné par $\nabla_{\boldsymbol{\alpha}}\mathcal{J}(\boldsymbol{\alpha})$. Nous avons montré dans [Chen et al., 2011b] que la solution d'une équation de la forme $f(u) = 0$ peut être donnée, sous certaines conditions sur la fonction $f(\cdot)$, par un algorithme de point fixe de la forme $u = u + f(u)$. La mise en œuvre de cette stratégie du point fixe nous conduit à un algorithme de descente de gradient. La mise à jour à l'itération t est alors

$$\alpha_{i,t+1} = \alpha_{i,t} - \eta_{i,t}\,\alpha_{i,t}[-\nabla_{\boldsymbol{\alpha}_t}\mathcal{J}(\boldsymbol{\alpha}_t)]_i. \tag{5.13}$$

Le pas $\eta_{i,t}$ permet de contrôler la convergence sans sortir du domaine admissible de la non-négativité. Avec $\alpha_{i,t} > 0$, la non-négativité de $\alpha_{i,t+1}$ est garantie si et seulement si $1 + \eta_{i,t}[-\nabla_{\boldsymbol{\alpha}}\mathcal{J}(\boldsymbol{\alpha})]_i > 0$. Si $[\nabla_{\boldsymbol{\alpha}}\mathcal{J}(\boldsymbol{\alpha})]_i < 0$, cette condition est toujours satisfaire. Par contre, si $[\nabla_{\boldsymbol{\alpha}}\mathcal{J}(\boldsymbol{\alpha})]_i > 0$, la non-négativité de $\alpha_{i,t+1}$ est vérifiée si $0 < \eta_{i,t} \leq 1/[\nabla_{\boldsymbol{\alpha}}\mathcal{J}(\boldsymbol{\alpha})]_i$, ou encore en considérant un seule valeur de pas pour toutes les directions :

$$\eta_t \leq \min_i \frac{1}{[\nabla_{\boldsymbol{\alpha}_t}\mathcal{J}(\boldsymbol{\alpha}_t)]_i}.$$

Voir [Chen et al., 2010b] pour une analyse de convergence de cet algorithme qui impose la non-négativité.

Solution avec contraintes totales

Soit le changement de variable suivant qui permet de satisfaire la contrainte de somme unité : $\alpha_i = u_i / \sum_j u_j$. La dérivée partielle de la fonction coût $\mathcal{J}(\cdot)$ par rapport à la nouvelle variable u_i peut être exprimée comme suit :

$$\frac{\partial \mathcal{J}}{\partial u_i} = \sum_{j=1}^{m} \frac{\partial \mathcal{J}}{\partial \alpha_j} \frac{\partial \alpha_j}{\partial u_i} = \sum_{j=1}^{m} \frac{\partial \mathcal{J}}{\partial \alpha_j} \frac{\frac{\partial u_j}{\partial u_i} \sum_{k=1}^{m} u_k - \frac{\partial \sum_{k=1}^{m} u_k}{\partial u_i} u_j}{(\sum_{k=1}^{m} u_k)^2} = \frac{-1}{\sum_{k=1}^{m} u_k} \left(- \frac{\partial \mathcal{J}}{\partial \alpha_i} + \sum_{j=1}^{m} \alpha_j \frac{\partial \mathcal{J}}{\partial \alpha_j} \right).$$

En appliquant la règle (5.13) sur la mise à jour non négative des $u_{i,t}$, nous avons à l'itération t :

$$u_{i,t+1} \quad = \quad u_{i,t} + \eta_t \, u_{i,t} \left(- \frac{\partial \mathcal{J}}{\partial \alpha_{i,t}} + \sum_{j=1}^{m} \alpha_{j,t} \frac{\partial \mathcal{J}}{\partial \alpha_{j,t}} \right),$$

où l'égalité est due à l'absorption du facteur $1/(\sum_j u_{j,t})$ dans η_t, tout en vérifiant l'identité $\sum_j u_{j,t+1} = \sum_j u_{j,t}$ pour tout pas. En divisant par $\sum_j u_{j,t+1}$ et $\sum_j u_{j,t}$ respectivement à gauche et à droite de cette expression, on obtient la mise à jour

$$\alpha_{i,t+1} = \alpha_{i,t} + \eta_t \, \alpha_{i,t} \left(- \frac{\partial \mathcal{J}}{\partial \alpha_{i,t}} + \sum_{j=1}^{m} \alpha_{j,t} \frac{\partial \mathcal{J}}{\partial \alpha_{j,t}} \right),$$

ou encore sous forme vectorielle

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \eta_t \, \mathrm{diag}(\boldsymbol{\alpha}_t) \big(\nabla_{\boldsymbol{\alpha}_t} \mathcal{J}(\boldsymbol{\alpha}_t) - \mathbf{1} \boldsymbol{\alpha}_t^\top \nabla_{\boldsymbol{\alpha}_t} \mathcal{J}(\boldsymbol{\alpha}_t) \big),$$

où le gradient de la fonction coût quadratique est $\nabla_{\boldsymbol{\alpha}} \mathcal{J}(\boldsymbol{\alpha}) = \boldsymbol{X}_{\mathcal{D}}^\top \boldsymbol{x} - \boldsymbol{X}_{\mathcal{D}}^\top \boldsymbol{X}_{\mathcal{D}} \boldsymbol{\alpha}$. Une étude analytique sur la convergence de l'algorithme proposé est présentée dans [Chen et al., 2011e] avec une étude expérimentale sur des images synthétisées et réelles.

5.3.2 Démélange par des projections multiples avec contraintes totales

Les méthodes de projections orthogonales sont appliquées avec succès dans la résolution de nombreux problèmes d'optimisation, y compris dans le domaine du traitement du signal et des images. Ces méthodes remontent aux années 1930 avec les célèbres approches de Cimmino [Cimmino, 1938] et de Kaczmarz [Kaczmarz, 1937], respectivement sur les projections parallèles et sur les projections cycliques. Plus récemment, les projections sur les espaces ont été généralisées à des projections sur des ensembles convexes. Ces méthodes fournissent un cadre unifié pour aborder de nombreux problèmes en traitement du signal, y compris le filtrage adaptatif et l'apprentissage statistique [Yukawa, 2010, Theodoridis et al., 2011]. A notre connaissance, ces méthodes n'ont pas encore été exploitées en traitement d'images hyperspectrales, bien que l'intérêt y soit accru [Heylen et al., ress]. Très récemment, nous avons adapté avec succès les méthodes de projections multiples pour le démélange avec contraintes totales, d'une part avec les projections parallèles de Cimmino dans [Honeine and Lantéri, 2013], et d'autre part avec les projections cycliques de Kaczmarz dans [Honeine et al., 2013a]. Ces travaux sont résumés dans la suite de cette section.

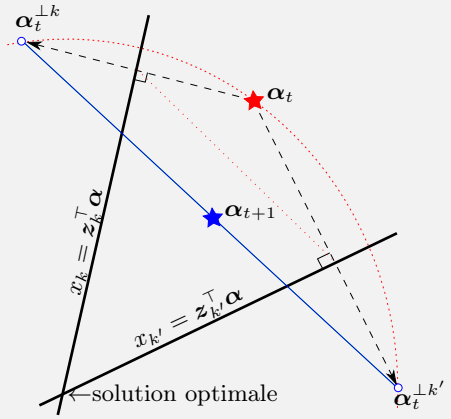
Pour ce faire, le problème d'optimisation est étudié dans l'espace des vecteurs $\boldsymbol{\alpha}$. Le système de démélange linéaire $\boldsymbol{x} = \boldsymbol{X}_{\mathcal{D}} \, \boldsymbol{\alpha}$ est étudié bande par bande, en considérant les d équations linéaires

$$x_k = \boldsymbol{z}_k^\top \boldsymbol{\alpha}, \tag{5.14}$$

pour $k = 1, 2, \ldots, d$, où $\boldsymbol{z}_k$ est le vecteur des signatures spectrales des m composants purs à la k-ème longueur d'onde, c'est à dire la k-ème ligne de $\boldsymbol{X}_{\mathcal{D}}$. Ainsi la solution du problème de démélange (sans

[Cadre 20] Schéma illustratif de la méthode réflexion-puis-agrégation.

Schéma illustrant en 2D la méthode réflexion-puis-agrégation de Cimmino, en considérant deux hyperplans ($x_k = \boldsymbol{z}_k^\top \boldsymbol{\alpha}$ et $x_{k'} = \boldsymbol{z}_{k'}^\top \boldsymbol{\alpha}$). La solution à l'itération t, désignée par $\boldsymbol{\alpha}_{t+1}$ (★), est obtenue en considérant une combinaison convexe des reflections (∘) de la solution initiale $\boldsymbol{\alpha}_t$ (★). En particulier, la pondération $\gamma_k = 1/d$ pour tout $k = 1, 2, \dots d$ permet d'extraire le barycentre des solutions intermédiaires.
Il est clair que toute combinaison convexe des réflexions (—) est une meilleure solution que l'élément initial (★). Elle est aussi préférable à la stratégie projection-puis-agrégation (···) souvent préconisée dans la littérature [Theodoridis et al., 2011].



contraintes) est-elle l'intersection des d hyperplans (affines) définis par ces équations. Notons que la projection d'une solution courant $\boldsymbol{\alpha}_t$ sur le k-ème hyperplan est donnée par

$$\boldsymbol{\alpha}_t + \eta_t \, \frac{x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t}{\|\boldsymbol{z}_k\|^2} \, \boldsymbol{z}_k, \tag{5.15}$$

où nous avons introduit le paramètre de relaxation $\eta_t \in [0, 2]$. La borne inférieure de cet intervalle correspond à un régime inchangé, la borne supérieure réalise une réflexion, alors que $\eta_t = 0$ conduit à une projection sur l'hyperplan en question. Les valeurs strictement supérieures à 1 prolongent la projection au-delà de la frontière, ce qui permet de compenser une convergence lente [Youla and Webb, 1982]. Il s'agit de ladite « sur-relaxation ».

5.3.2.a Méthode réflexion-puis-agrégation de Cimmino

La méthode de Cimmino se compose de deux étapes pour affiner la solution courante $\boldsymbol{\alpha}_t$ à l'itération t. La première génère des solutions intermédiaires, chacune par réflexion de $\boldsymbol{\alpha}_t$ par rapport à chacun des hyperplans définis par (5.14). Soient $\boldsymbol{\alpha}_t^{\perp 1}, \boldsymbol{\alpha}_t^{\perp 2}, \dots, \boldsymbol{\alpha}_t^{\perp d}$ ces solutions intermédiaires, obtenues selon l'expression (5.15) pour le paramètre de relaxation fixé à $\eta_t = 2$. La seconde étape opère par agrégation de ses réflexions selon $\boldsymbol{\alpha}_{t+1} = \sum_{k=1}^{d} \gamma_k \; \boldsymbol{\alpha}_t^{\perp k}$, où les poids γ_k satisfont les contraintes de convexité. La motivation de cette approche de réflexion-puis-agrégation est illustrée dans le Cadre 20. Finalement, ces deux étapes sont fusionnées en une seule étape avec la règle de mise à jour suivante :

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + 2 \, \boldsymbol{X}_{\mathcal{D}}^\top \boldsymbol{\Gamma} (\boldsymbol{x} - \boldsymbol{X}_{\mathcal{D}} \, \boldsymbol{\alpha}_t),$$

où $\boldsymbol{\Gamma}$ est la matrice diagonale de taille $(d \times d)$ d'élément général $\gamma_k / \|\boldsymbol{z}_k\|^2$. Cette méthode possède plusieurs propriétés intéressantes, telles que la parallélisation grâce aux multiples réflexions qui peuvent être traitées séparément, ainsi que le traitement de plusieurs pixels à chaque itération. Pour ce dernier cas, nous pouvons écrire $\boldsymbol{A}_{t+1} = \boldsymbol{A}_t + 2 \, \boldsymbol{X}_{\mathcal{D}}^\top \boldsymbol{\Gamma} (\boldsymbol{X} - \boldsymbol{X}_{\mathcal{D}} \, \boldsymbol{A}_t)$, en rassemblant le spectre $\boldsymbol{x}$ et la fraction d'abondance correspondante $\boldsymbol{\alpha}_t$ de chaque pixel dans les matrices $\boldsymbol{X}$ et $\boldsymbol{A}_t$, respectivement.

Nous proposons l'adaptation de l'approche réflexion-puis-agrégation de Cimmino dans le cadre d'une optimisation avec contraintes. Deux stratégies sont proposées pour la contrainte de somme unité et deux stratégies pour la non-négativité, comme suit :

– **Système augmenté :** le système augmenté (5.6) se conjugue naturellement à l'approche réflexion-puis-agrégation, selon la règle

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + 2 \begin{bmatrix} \mathbf{1} & \boldsymbol{X}_{\mathcal{D}}^\top \end{bmatrix} \boldsymbol{\Gamma} \left(\begin{bmatrix} 1 \\ \boldsymbol{x} \end{bmatrix} - \begin{bmatrix} \mathbf{1}^\top \\ \boldsymbol{X}_{\mathcal{D}} \end{bmatrix} \boldsymbol{\alpha}_t \right),$$

où $\boldsymbol{\Gamma}$ est ici de taille $(d+1)\times(d+1)$. L'interprétation de cette formulation augmentée réside dans une réflexion supplémentaire, par rapport à l'hyperplan défini par l'équation $1 = \mathbf{1}^\top \boldsymbol{\alpha}$.

- **Normalisation :** l'approche la plus simple pour imposer la somme unité consiste à remplacer $\boldsymbol{\alpha}_{t+1}$, obtenu par réflexion-puis-agrégation, par $\boldsymbol{\alpha}_{t+1}/(\mathbf{1}^\top \boldsymbol{\alpha}_{t+1})$. Cette dernière étape est une projection sur l'hyperplan défini par $1 = \mathbf{1}^\top \boldsymbol{\alpha}$. Notons toutefois qu'il ne s'agit pas d'une projection orthogonale, et par la suite elle perd la propriété de non-expansivité. Cela signifie que les distances entre deux éléments projetés ne sont pas garanties d'être inférieures aux distances des éléments originaux. Même s'il s'agit d'un inconvénient qui affecte la convergence, la combinaison de cette normalisation avec les réflexions semble heureusement bien fonctionner.
- **Relaxation :** pour imposer la non-négativité, nous proposons d'inclure une relaxation dans la réflexion, selon

$$\boldsymbol{\alpha}_t^{\perp k} = \boldsymbol{\alpha}_t + 2\,\eta_{k,t}\frac{x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t}{\|\boldsymbol{z}\|^2}\,\boldsymbol{z}_k,$$

où la valeur du pas $\eta_{k,t} \in [0\,;1]$ est choisi afin d'imposer la non-négativité à chaque réflexion. Nous montons que la valeur de $\eta_{k,t}$ est

$$\min_i \frac{-\|\boldsymbol{z}_k\|^2}{x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t}\frac{[\boldsymbol{\alpha}_t]_i}{[\boldsymbol{z}_k]_i}$$

quand celle-ci est dans l'intervalle $[0\,;1]$; sinon $\eta_k = 1$.

- **Rectification à zéro :** l'approche la plus simple pour satisfaire la contrainte de non-négativité consiste à remettre à zéro les valeurs négatives obtenues après chaque itération. Pour cela, il suffit de remplaçant chaque $[\boldsymbol{\alpha}_t]_i$ par $\max\big\{[\boldsymbol{\alpha}_t]_i; 0\big\}$.

En combinant les 2 stratégies qui gèrent la somme unité et les 2 stratégies qui imposent la non-négativité, nous avons 4 variantes pour traiter les contraintes totales. Voir [Honeine and Lantéri, 2013] pour plus de détails.

5.3.2.b Méthode des projections cycliques de Kaczmarz

La méthode de Kaczmarz procède par des projections successives où, à chaque itération, la solution courante est actualisée en considérant sa projection sur un des hyperplan. Soit le k-ème hyperplan sélectionné [6] à l'itération t. La projection de $\boldsymbol{\alpha}_t$ sur cet hyperplan est donnée par (5.15), c'est à dire

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \eta_t\,\frac{x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t}{\|\boldsymbol{z}_k\|^2}\,\boldsymbol{z}_k.$$

Dans la méthode conventionnelle de Kaczmarz, le paramètre de relaxation est fixé à $\eta_t = 1$. Dans la suite, ce paramètre nous permet de satisfaire la non-négativité. Mais avant, nous considérons la contrainte de somme unité.

Nous proposons d'opérer en deux étapes à chaque itération t. La première étape est une itération classique de l'algorithme de Kaczmarz, où la solution courante $\boldsymbol{\alpha}_t$ est projetée sur un hyperplan donnant $\boldsymbol{\alpha}_t^{\perp k}$ selon (5.15) pour $\eta_t = 0$, où k désigne l'hyperplan considéré à l'itération t. Afin de satisfaire la contrainte de somme unité, la seconde étape opère une simple projection sur l'hyperplan défini par l'équation $1 = \mathbf{1}^\top \boldsymbol{\alpha}$. En reprenant l'expression de projection (5.15) au jour de cette équation de l'hyperplan, nous avons

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t^{\perp k} + \tfrac{1}{m}\left(1 - \mathbf{1}^\top \boldsymbol{\alpha}_t^{\perp k}\right)\mathbf{1},$$

6. Plusieurs stratégies ont été proposées pour choisir la séquence des hyperplans, c'est à dire le choix à l'itération t de la k-ème ligne de la matrice $\boldsymbol{X}_{\mathcal{D}}$ [Needell, 2010, Strohmer and Vershynin, 2009, Censor et al., 2009]. La sélection cyclique classique, c'est à dire $k = t \mod d$, est la plus simple mais au prix d'une lente convergence. D'autres stratégies ont été proposées pour remédier à cet inconvénient. La stratégie de la plus grande erreur nécessite l'évaluation du numérateur $x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t$ pour tout k à chaque itération t, ce qui conduit à un coût calculatoire élevé. La stratégie aléatoire consiste à sélectionner k d'une manière aléatoire avec une probabilité proportionnelle au dénominateur $\|\boldsymbol{z}_k\|^2$. Cette dernière stratégie a montré des propriétés très intéressantes [Strohmer and Vershynin, 2009], mais est également critiquée dans [Censor et al., 2009].

où le vecteur $\boldsymbol{z}_k$ a été remplacé par $\mathbf{1}$. En remplaçant $\boldsymbol{\alpha}_t^{\perp k}$ par son expression dans (5.15), nous pouvons combiner ces deux étapes en une seule avec

$$\begin{aligned}\boldsymbol{\alpha}_{t+1} &= \boldsymbol{\alpha}_t + \eta_t \frac{x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t}{\|\boldsymbol{z}_k\|^2}\, \boldsymbol{z}_k + \tfrac{1}{m}\Big(1 - \mathbf{1}^\top \boldsymbol{\alpha}_t - \eta_t \frac{x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t}{\|\boldsymbol{z}_k\|^2}\, \mathbf{1}^\top \boldsymbol{z}_k\Big)\mathbf{1} \\ &= \boldsymbol{\alpha}_t + \eta_t \frac{x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t}{\|\boldsymbol{z}_k\|^2}\, \Big(\mathbf{I} - \tfrac{1}{m}\mathbf{1}\,\mathbf{1}^\top\Big)\, \boldsymbol{z}_k, \end{aligned} \tag{5.16}$$

où la dernière égalité est due à l'identité $\mathbf{1}^\top\boldsymbol{\alpha}_t = 1$ de l'itération précédente et $\mathbf{I}$ est la matrice identité de taille $(m \times m)$.

Notons que la matrice $\mathbf{I} - \frac{1}{m}\mathbf{1}\,\mathbf{1}^\top$ n'est autre que la matrice de centrage utilisée en méthodes à noyaux, comme illustré avec l'expression (4.20) en ACP-à-noyaux. En effet, nous avons $\mathbf{1}^\top\big(\mathbf{I} - \frac{1}{m}\mathbf{1}\,\mathbf{1}^\top\big)\,\boldsymbol{v} = \mathbf{0}$ pour tout $\boldsymbol{v}$. En conséquence, toute règle de mise à jour de la forme

$$\boldsymbol{\alpha}_{t+1} = \boldsymbol{\alpha}_t + \big(\mathbf{I} - \tfrac{1}{m}\mathbf{1}\,\mathbf{1}^\top\big)\,\boldsymbol{v}$$

garantit la conservation du flux pour tout vecteur $\boldsymbol{v}$, c'est à dire la somme des éléments de $\boldsymbol{\alpha}_t$ demeure inchangée.

La règle (5.16) garantit la contrainte de somme unité indépendamment de la valeur du paramètre de relaxation η_t. Nous profitons de cet avantage afin d'imposer la contrainte de non-négativité. Soit $\eta_t = \min_{i=1,\ldots,m} \eta_{i,t}$, où $\eta_{i,t}$ est une valeur valide du paramètre dans la i-ème direction, obtenue selon deux cas :

- si $\frac{x_k - \boldsymbol{z}_k^\top \boldsymbol{\alpha}_t}{\|\boldsymbol{z}_k\|^2}\,\big[\big(\mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\big)\boldsymbol{z}_k\big]_i > 0$, alors aucune contrainte est imposée et donc $\eta_{i,t} = 1$;
- sinon, la valeur du paramètre de relaxation doit être réduite telle que $\eta_{i,t} \leq \frac{[\boldsymbol{\alpha}_t]_i}{\big[\big(\mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\top\big)\boldsymbol{z}_k\big]_i}$.

Une analyse de convergence est proposée dans le Cadre 21 dans deux cas : le premier concerne le modèle sans bruit, et le second prend en considération la présence de bruit comme décrit avec l'expression (5.1). Voir aussi [Honeine et al., 2013a] pour des expérimentations.

5.4 Démélange non linéaire par les méthodes à noyaux

Jusqu'à présent, nous avons supposé une relation linéaire entre x_i et le contenu spectral $\boldsymbol{z}_i$ des composants purs à la longueur d'onde i. Toutefois, d'après [Hapke, 1981] et plusieurs études plus récentes [Ray and Murray, 1996, Guilfoyle et al., 2001, Broadwater and Banerjee, 2009, Raksuntorn and Du, 2010], la lumière pourrait clairement interagir avec plusieurs de ces composants selon un mécanisme non linéaire.

Les méthodes à noyaux proposent un cadre privilégié pour répondre à ce problème. La mise en œuvre de modèles non linéaires de la forme $x_i = \psi(\boldsymbol{z}_i)$, où $\psi(\cdot)$ est un élément d'un espace fonctionnel à définir, se heurte à une difficulté majeure. Il est crucial d'exhiber les abondances $\boldsymbol{\alpha}$ dans le modèle afin de pouvoir imposer les contraintes totales, c'est à dire la non-négativité et la somme unité des α_i. Dans la suite, nous proposons une combinaison d'un modèle linéaire classique avec une fluctuation non linéaire, où la non-linéarité est définie dans un RKHS. En d'autres termes, il s'agit de la forme

$$\psi(\boldsymbol{z}_i, \boldsymbol{\alpha}) = \psi_{\text{lin}}(\boldsymbol{z}_i, \boldsymbol{\alpha}) + \psi_{\text{nlin}}(\cdot),$$

où $\psi_{\text{lin}}(\boldsymbol{z}_i, \boldsymbol{\alpha}) = \boldsymbol{z}_i^\top\boldsymbol{\alpha}$ est le modèle linéaire, et $\psi_{\text{nlin}}(\cdot)$ est un élément d'un espace fonctionnel $\mathbb{H}_{\text{nlin}}$, un espace de Hilbert à noyau reproduisant de fonctions à valeurs réelles sur un compact. Le noyau de cet espace, désigné par $\kappa_{\text{nlin}}(\cdot,\cdot)$, définit l'espace RKHS et en conséquence le modèle en question.

Trois modèles pour définir la fonction non linéaire $\psi_{\text{nlin}}(\cdot)$ sont investis dans cette section. Le premier, résumé dans la section 5.4.1 et étudié en détail dans [Honeine and Richard, 2011a, Chen et al., 2013e], considère une non-linéarité par rapport au contenu spectral $\boldsymbol{z}_i$, c'est à dire $\psi_{\text{nlin}}(\boldsymbol{z}_i)$. Une étude sur le choix du noyau en démélange hyperspectral est présentée dans la section 5.4.2. Le deuxième modèle propose de relâcher l'équilibre rigide entre le modèle linéaire et la fonction non linéaire, en considérant une combinaison convexe optimisée par apprentissage de noyaux multiples. Voir la section 5.4.3 et [Chen et al., 2012a]. Le troisième modèle, résumé dans la section 5.4.4

[CADRE 21] **Convergence de la méthode de Kaczmarz avec contrainte de somme unité**

Soit θ_k l'angle entre l'hyperplan associé à la contrainte de somme unité, et le k-ème hyperplan, sélectionné à l'itération t (et noté aussi $k(t)$). D'une part, le vecteur normal à chaque hyperplan est non négatif, puisqu'il s'agit de $\boldsymbol{z}_k$, *i.e.*, un vecteur de valeurs spectrales selon (5.14). D'autre part, l'hyperplan de somme unité correspond à la bissectrice du premier quadrant. Nous avons alors $\cos(\theta_k) > 1/\sqrt{2}$ pour tout k, que nous pouvons calculer facilement avec $\cos(\theta_\ell) = \frac{\mathbf{1}^\top \boldsymbol{z}_k}{\|\mathbf{1}\|\|\boldsymbol{z}_k\|}$.

Théorème 3 (Méthode de Kaczmarz avec contrainte : convergence du modèle sans bruit)**.**
A partir d'une estimation initiale $\boldsymbol{\alpha}_0$, l'algoritme converge à la solution optimale $\boldsymbol{\alpha}_{\text{opt}}$ à la vitesse

$$\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{\text{opt}}\| = \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_{\text{opt}}\| \prod_{k(t');t'\leq t} \cos^2(\theta_k),$$

où le produit est pris sur tous les hyperplans visités jusqu'à l'itération t.

Démonstration. Comme préconisé par la méthode classique de Kaczmarz, nous supposons que la solution du système linéaire existe, qu'elle est unique, et qu'elle appartient à l'intersection des hyperplans en question. Nous avons alors, où la seconde égalité est due au théorème de Pythagore :

$$\|\boldsymbol{\alpha}_t - \boldsymbol{\alpha}_{\text{opt}}\| = \|\boldsymbol{\alpha}_{t-1}^{\perp k} - \boldsymbol{\alpha}_{\text{opt}}\| \cos(\theta_k) = \|\boldsymbol{\alpha}_{t-1} - \boldsymbol{\alpha}_{\text{opt}}\| \cos^2(\theta_k) = \cdots = \|\boldsymbol{\alpha}_0 - \boldsymbol{\alpha}_{\text{opt}}\| \prod_{k(t');t'\leq t} \cos^2(\theta_k) \blacksquare$$

Théorème 4 (Méthode de Kaczmarz avec contrainte : convergence en présence de bruit additif)**.**
Selon le modèle (5.1)*, la vitesse de convergence est dictée par l'expression suivante, à partir d'une estimation initiale erronée $\widetilde{\boldsymbol{\alpha}}_0$:*

$$\|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_{\text{opt}}\|^2 = \|\widetilde{\boldsymbol{\alpha}}_0 - \boldsymbol{\alpha}_{\text{opt}}\|^2 \prod_{k(t');t'\leq t} \cos^4(\theta_k) + \sum_{k(t');t'\leq t} \frac{\epsilon_k^2}{\|\boldsymbol{z}_k\|^2} \cos^2(\theta_k) \prod_{k(t'');t''<t'} \cos^4(\theta_{k''}).$$

Démonstration. Tout d'abord, nous avons de la projection la relation suivante : $\cos(\theta_\ell) = \|\widetilde{\boldsymbol{\alpha}}_t - \boldsymbol{\alpha}_{\text{opt}}\| / \|\widetilde{\boldsymbol{\alpha}}_{t-1\perp\ell} - \boldsymbol{\alpha}_{\text{opt}}\|$, où le dénominateur peut être décomposé grâce au théorème de Pythagore, selon $\|\widetilde{\boldsymbol{\alpha}}_{t-1}^{\perp\ell} - \boldsymbol{\alpha}_{\text{opt}}\|^2 = \|\boldsymbol{\alpha}_{t-1}^{\perp\ell} - \boldsymbol{\alpha}_{\text{opt}}\|^2 + \|\frac{\epsilon_k}{\|\boldsymbol{z}_k\|^2}\boldsymbol{z}_k\|^2$. Cette égalité est due à l'orthogonalité et à la définition donnée par l'expression encadrée ci-dessus. D'une part, le dernier terme est simplement $\epsilon_k^2/\|\boldsymbol{z}_k\|^2$, et d'autre part $\|\boldsymbol{\alpha}_{t-1}^{\perp k} - \boldsymbol{\alpha}_{\text{opt}}\| = \|\widetilde{\boldsymbol{\alpha}}_{t-1} - \boldsymbol{\alpha}_{\text{opt}}\| \cos(\theta_k)$ d'après la démonstration du théorème 3. Le résultat final est obtenu en combinant ces relations. $\blacksquare$

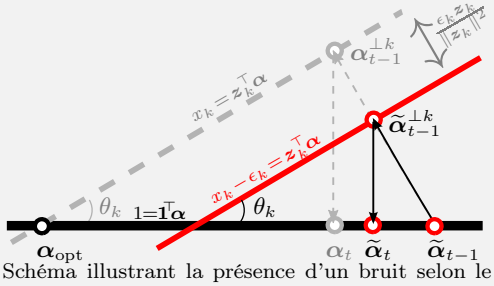


Schéma illustrant la présence d'un bruit selon le modèle (5.1), c'est à dire $\boldsymbol{x} = \boldsymbol{X}^\top \boldsymbol{\alpha} + \boldsymbol{\epsilon}$. Le k-ème hyperplan résultant (en rouge) induit une estimation erronée $\widetilde{\boldsymbol{\alpha}}_t$. L'hyperplan en question, défini par l'ensemble des $\boldsymbol{\alpha}$ vérifiant l'équation $x_k - \epsilon_k = \boldsymbol{z}_k^\top \boldsymbol{\alpha}$, est aussi défini par l'ensemble des

$$\boxed{\boldsymbol{\alpha} + \tfrac{\epsilon_k}{\|\boldsymbol{z}_k\|^2}\boldsymbol{z}_k}$$

où $\boldsymbol{\alpha}$ appartenant à l'hyperplan non erroné (en gris), c'est à dire vérifiant $x_k = \boldsymbol{z}_k^\top \boldsymbol{\alpha}$. Voir [Needell, 2010, Lemma 2.2].

et dans [Chen et al., 2013c], consiste à incorporer l'abondance dans la fonction non linéaire, en considérant un modèle post-non-linéaire selon $\psi_{\text{nlin}}(\boldsymbol{\alpha}^\top \boldsymbol{z}_i)$. Aux trois modèles étudiés, nous complétons ce travail en proposant d'intégrer l'information spatiale dans le modèle de démélange non linéaire, comme étudié dans la section 5.5 avec deux méthodes [Nguyen et al., 2013, Chen et al., 2013d].

5.4.1 Démélange en combinant un modèle linéaire et une fluctuation non linéaire

En sortant du cadre de mélange linéaire, nous considérons dans un premier temps le modèle $x_i = \psi(\boldsymbol{z}_i)$ avec

$$\psi(\boldsymbol{z}_i) = \psi_{\text{lin}}(\boldsymbol{z}_i, \boldsymbol{\alpha}) + \psi_{\text{nlin}}(\boldsymbol{z}_i), \tag{5.17}$$

où $\psi_{\text{nlin}}(\cdot)$ est un élément d'un espace RKHS $\mathbb{H}_{\text{nlin}}$. Le noyau de cet espace est désigné par $\kappa_{\text{nlin}}(\cdot,\cdot)$, vérifiant ainsi la propriété reproduisante $\psi(\boldsymbol{z}) = \langle \psi(\cdot), \kappa(\boldsymbol{z},\cdot)\rangle_{\mathbb{H}_{\text{nlin}}}$ pour tout $\boldsymbol{z}$ et toute fonction $\psi(\cdot)$ de $\mathbb{H}_{\text{nlin}}$. Clairement (comme démontré dans [Haussler, 1999]), l'espace $\mathbb{H}$ des fonctions $\psi(\cdot) =$

$\psi_{\text{lin}}(\cdot,\boldsymbol{\alpha}) + \psi_{\text{nlin}}(\cdot)$, défini par la somme directe $\mathbb{H}_{\text{lin}} \oplus \mathbb{H}_{\text{nlin}}$ de deux espace RKHS de noyaux $\kappa_{\text{lin}}(\cdot,\cdot)$ et $\kappa_{\text{nlin}}(\cdot,\cdot)$, est aussi un RKHS de noyau

$$\begin{aligned}\kappa(\boldsymbol{z}_i,\boldsymbol{z}_j) &= \big(\kappa_{\text{lin}} \oplus \kappa_{\text{nlin}}\big)(\boldsymbol{z}_i,\boldsymbol{z}_j)\\ &= \boldsymbol{z}_i^\top \boldsymbol{z}_j + \kappa_{\text{nlin}}(\boldsymbol{z}_i,\boldsymbol{z}_j).\end{aligned}$$

Soit $\boldsymbol{K}$ la matrice de Gram de taille $(d\times d)$ d'élément général $\kappa(\boldsymbol{z}_i,\boldsymbol{z}_j)$, pour $i,j=1,2,\ldots,d$. Il est fondamental de noter que

$$\boldsymbol{K} = \boldsymbol{K}_{\text{lin}} + \boldsymbol{K}_{\text{nlin}}\,, \tag{5.18}$$

où $\boldsymbol{K}_{\text{lin}}$ est la matrice de Gram du noyau linéaire, c'est à dire $\boldsymbol{X}\boldsymbol{X}^\top$ et $\boldsymbol{K}_{\text{nlin}}$ celle associée à la transformation non linéaire, d'élément général $\kappa_{\text{nlin}}(\boldsymbol{z}_i,\boldsymbol{z}_j)$, pour $i,j=1,2,\ldots,d$.

Le problème d'optimisation considère une fonction coût quadratique, selon :

$$\min_{\psi(\cdot)\in\mathbb{H}} \sum_{i=1}^{d} (x_i - \psi(\boldsymbol{z}_i))^2 + \eta\mathcal{R}(\|\psi(\cdot)\|^2_{\mathbb{H}}),$$

où un terme de régularisation a été ajouté comme préconisé en apprentissage statistique. Il est clair que ce problème d'optimisation est proche de celui des *least-squares SVM* (LS-SVM) [Suykens et al., 2002], avec la fonction de régularisation $\|\psi(\cdot)\|^2_{\mathbb{H}}$:

$$\min_{\psi(\cdot)\in\mathbb{H}} \sum_{i=1}^{d} e_i^2 + \eta\|\psi(\cdot)\|^2_{\mathbb{H}}, \qquad \text{sous contraintes : } e_i = x_i - \psi(\boldsymbol{z}_i),\ \text{pour tout } i=1,\ldots,d.$$

Toutefois, le problème que nous traitons est plus difficile, à cause de la forme particulière de l'espace $\mathbb{H}$ avec $\psi(\cdot) = \psi_{\text{lin}}(\cdot,\boldsymbol{\alpha}) + \psi_{\text{nlin}}(\cdot)$, et des contraintes totales sur $\boldsymbol{\alpha}$.

Nous considérons la résolution du problème d'optimisation convexe suivant :

$$\begin{aligned}\psi^*(\cdot) = \underset{\psi(\cdot)}{\text{argmin}} &\sum_{i=1}^{d} e_i^2 + \eta\left(\|\psi_{\text{lin}}(\cdot)\|^2_{\mathcal{H}_{\text{lin}}} + \|\psi_{\text{nlin}}(\cdot)\|^2_{\mathcal{H}_{\text{nlin}}}\right),\\ \text{où}\quad \psi(\cdot) &= \psi_{\text{lin}}(\cdot) + \psi_{\text{nlin}}(\cdot),\quad \text{avec}\quad \psi_{\text{lin}}(\boldsymbol{z}_i) = \boldsymbol{z}_i^\top\boldsymbol{\alpha},\\ \text{sous contraintes :}\quad & e_i = x_i - \psi(\boldsymbol{z}_i),\\ & \boldsymbol{\alpha}\geq\mathbf{0}\quad\text{et}\quad \mathbf{1}^\top\boldsymbol{\alpha} = 1.\end{aligned}$$

Par la propriété de dualité forte, le problème dual admet la même solution que ce problème primal. En désignant par β_i, γ_j et λ les multiplicateurs de Lagrange, le Lagrangien associé à ce problème d'optimisation avec contraintes est alors

$$\sum_{i=1}^{d} e_i^2 + \eta\left(\|\boldsymbol{\alpha}\|^2 + \|\psi_{\text{nlin}}(\cdot)\|^2_{\mathcal{H}_{\text{nlin}}}\right) - \sum_{i=1}^{d}\beta_i\left(e_i - x_i + \psi(\boldsymbol{z}_i)\right) - \sum_{j=1}^{m}\gamma_j\,\alpha_j + \lambda(\mathbf{1}^\top\boldsymbol{\alpha} - 1),$$

avec $\gamma_j \geq 0$. Dans cette expression, nous avons utilisé l'identité $\|\psi_{\text{lin}}(\cdot)\|_{\mathcal{H}_{\text{lin}}} = \|\boldsymbol{\alpha}\|$. Les conditions d'optimalité du Lagrangien par rapport aux variables primales sont :

$$\boldsymbol{\alpha}^* = \sum_{i=1}^{d}\beta_i^*\,\boldsymbol{z}_i + \boldsymbol{\gamma}^* - \lambda^*\mathbf{1}, \qquad \psi^*_{\text{nlin}}(\cdot) = \sum_{i=1}^{d}\beta_i^*\,\kappa_{\text{nlin}}(\cdot,\boldsymbol{z}_i) \qquad \text{et} \qquad e_i^* = \eta\,\beta_i^*,$$

où nous retrouvons en particulier le résultat du Théorème de Représentation (`Th.Rep`). En remplaçant ces expression dans le Lagrangien, on obtient le problème dual suivant :

$$\max_{\boldsymbol{\beta},\boldsymbol{\gamma},\lambda} -\begin{bmatrix}\boldsymbol{\beta}^\top & \boldsymbol{\gamma}^\top & \lambda\end{bmatrix}\begin{bmatrix}\boldsymbol{K}+\eta\mathbf{I} & \boldsymbol{X} & -\boldsymbol{X}\mathbf{1}\\ \boldsymbol{X}^\top & \mathbf{I} & -\mathbf{1}\\ -\mathbf{1}^\top\boldsymbol{X}^\top & -\mathbf{1}^\top & m\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\\ \lambda\end{bmatrix} + \begin{bmatrix}\boldsymbol{x}^\top & \mathbf{0}^\top & -1\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\\ \lambda\end{bmatrix}, \tag{5.19}$$

sous contraintes : $\boldsymbol{\gamma}\geq\mathbf{0}$,

où l'identité $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top + \boldsymbol{K}_{\text{nlin}}$ est utilisée selon (5.18). Notons qu'une fois le vecteur optimal $\boldsymbol{\beta}^*$ estimé, la reconstruction du vecteur spectral $\boldsymbol{x}$ est donnée par $\boldsymbol{x}^* = [\psi^*(\boldsymbol{z}_1), \psi^*(\boldsymbol{z}_2), \cdots, \psi^*(\boldsymbol{z}_d)]^\top$, c'est à dire

$$\boldsymbol{x}^* = \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{\beta}^* + \boldsymbol{\gamma}^* - \lambda^*\mathbf{1}) + \boldsymbol{K}_{\text{nlin}}\boldsymbol{\beta}^*,$$

où nous retrouvons la tendance linéaire et la fluctuation non linéaire. Ainsi le vecteur des abondances est-il estimé selon $\boldsymbol{\alpha}^* = \boldsymbol{X}^\top\boldsymbol{\beta}^* + \boldsymbol{\gamma}^* - \lambda^*\mathbf{1}$. Il s'agit d'un problème de programmation quadratique. De nombreuses méthodes de résolution existent dans la littérature, comme les méthodes des points intérieurs, de *active set* et de gradient projeté. Voir [Bertsekas, 1999, Luenberger and Ye, 2008] pour plus de détails sur ces méthodes d'optimisation. Ces procédures numériques bien connues se situent au-delà de la portée de ce document. Le CADRE 22 décrit les performances de la méthode proposée.

5.4.2 Choix du noyau en démélange hyperspectral

Le noyau $\kappa_{\text{nlin}}(\cdot,\cdot)$, qui fixe la non-linéarité du modèle de mélange adopté, est le seul élément à définir dans cette approche. Ce choix reste ouvert. Si les possibilités offertes sont infinies, mais devraient idéalement être pilotées par l'application considérée, nous avons généralement recours à des familles de noyaux aux capacités d'apprentissage éprouvées dont le noyau gaussien et le noyau polynomial. Ce dernier, de forme $\left(c + \boldsymbol{z}_1^\top\boldsymbol{z}_2\right)^p$ de degré p avec $c = 0$ pour le noyau polynomial dit homogène.

Considérons le modèle de mélange bilinéaire généralisé présenté dans [Halimi et al., 2011], illustré ici à trois composants purs par souci de clarté :

$$\boldsymbol{x} = \boldsymbol{X}\,\boldsymbol{\alpha} + \gamma_{12}\,\alpha_1\,\alpha_2\,(\boldsymbol{x}_{\omega_1} \otimes \boldsymbol{x}_{\omega_2}) + \gamma_{13}\,\alpha_1\,\alpha_3\,(\boldsymbol{x}_{\omega_1} \otimes \boldsymbol{x}_{\omega_3}) + \gamma_{23}\,\alpha_2\,\alpha_3\,(\boldsymbol{x}_{\omega_2} \otimes \boldsymbol{x}_{\omega_3}) + \boldsymbol{\epsilon},$$

où γ_{12}, γ_{13} et γ_{23} sont des paramètres d'atténuation et $\otimes$ le produit matriciel de Hadamard. La composante non linéaire par rapport à $\boldsymbol{\alpha}$ est étroitement liée au noyau polynomial homogène de degré 2, avec $\kappa_{\text{nlin}}(\boldsymbol{z}_i, \boldsymbol{z}_j) = (\boldsymbol{z}_i^\top\boldsymbol{z}_j)^2$. En effet, ce dernier s'écrit sous forme d'un produit scalaire, selon

$$\kappa_{\text{nlin}}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \boldsymbol{\phi}_{\text{nlin}}(\boldsymbol{z}_i)^\top\boldsymbol{\phi}_{\text{nlin}}(\boldsymbol{z}_j),$$

avec

$$\boldsymbol{\phi}_{\text{nlin}}(\boldsymbol{z}_i) = \begin{bmatrix} z_{i,1}^2 & z_{i,2}^2 & z_{i,3}^2 & \sqrt{2}\,z_{i,1}^2\,z_{i,2}^2 & \sqrt{2}\,z_{i,1}^2\,z_{i,3}^2 & \sqrt{2}\,z_{i,2}^2\,z_{i,3}^2 \end{bmatrix}^\top,$$

où $z_{i,k}$, la k-ème composante de $\boldsymbol{z}_i$, n'est autre que la i-ème valeur spectrale de la signature spectrale $\boldsymbol{x}_{\omega_k}$ d'un composant pur. Cela signifie que, en plus du mélange linéaire $\boldsymbol{X}\alpha$, les termes d'interaction inter-bandes spectrales du noyau utilisé sont de la forme $\boldsymbol{x}_{\omega_i} \otimes \boldsymbol{x}_{\omega_j}$, pour $i, j = 1, 2, \ldots, m$.

En vertu du coup du noyau, nous n'avons pas besoin d'exhiber la transformation non linéaire ci-dessus. De plus, le rapprochement illustré ci-dessus, entre le modèle bilinéaire généralisé et le noyau polynomial de degré 2, ouvre évidemment de nombreuses autres perspectives de couplage. Cela permet d'envisager des mécanismes d'interaction plus complexes, pour un coût calculatoire inchangé puisque seules les valeurs des composantes de $\boldsymbol{K}_{\text{nlin}}$ sont affectées par tout changement de noyau. A titre indicatif, considérons le noyau polynomial $\kappa_{\text{nlin}}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \left(1 + \boldsymbol{z}_i^\top\,\boldsymbol{z}_j\right)^p$. La formule du binôme de Newton permet d'écrire

$$\kappa_{\text{nlin}}(\boldsymbol{z}_i, \boldsymbol{z}_j) = \sum_{k=0}^{p} \binom{p}{k} (\boldsymbol{z}_i^\top\,\boldsymbol{z}_j)^k,$$

où chaque $\left(\boldsymbol{z}_i^\top\boldsymbol{z}_j\right)^k = (z_{i,1}z_{j,1} + \ldots + z_{i,m}\,z_{j,m})^k$ peut être développé en une somme pondérée de monômes de la forme

$$(z_{i,1}z_{j,1})^{k_1}(z_{i,2}z_{j,2})^{k_2}\ \ldots\ (z_{i,m}z_{j,m})^{k_m},$$

avec $\sum_{r=1}^m k_r = k$. Nous retrouvons ainsi avec le noyau polynomial, en plus du terme de mélange linéaire $\boldsymbol{X}\boldsymbol{\alpha}$, des termes d'interaction de la forme $\boldsymbol{x}_{\omega_1}^{k_1} \otimes \boldsymbol{x}_{\omega_2}^{k_2} \otimes \ldots \otimes \boldsymbol{x}_{\omega_m}^{k_m}$ pour chaque ensemble d'exposants dans le sens de Hadamard avec $0 \leq \sum_{r=1}^m k_r \leq p$.

Notons que ce serait coûteux en calcul pour former explicitement ces termes d'interaction. Leur nombre est en effet très large. De plus, en comparaison avec les méthodes introduites dans [Raksuntorn and Du, 2010, Nascimento and Bioucas-Dias, 2009] qui insèrent les produits de signatures de composants purs comme de nouveaux composants purs, nous n'avons pas besoin d'étendre la matrice des signatures spectrales des composants purs en y ajoutant ces termes. L'astuce du noyau rend le calcul beaucoup plus simple. Notons enfin qu'il est possible de combiner des noyaux pour obtenir de nouveaux noyaux valides, d'optimiser ces combinaisons pour améliorer les performances, etc. [Herbrich, 2002].

5.4.3 Démélange non linéaire par apprentissage de noyaux multiples

Le modèle (5.17) proposé ci-dessus repose sur l'hypothèse que le mécanisme de mélange est décrit par un mélange linéaire de vecteurs spectraux, auquel s'ajoute une fluctuation non linéaire $\psi_{\text{nlin}}(\cdot)$ définie par un noyau reproduisant. Ce modèle possède cependant une limitation majeure due à l'équilibre rigide entre les deux composants, linéaire et non linéaire, du modèle.

Afin de surmonter cette difficulté, nous nous intéressons à un modèle plus général que (5.17) en utilisant une forme convexe au lieu d'une somme directe. Nous considérons ici l'adaptation d'une stratégie pour apprendre de noyaux multiples [Rakotomamonjy et al., 2008], avec une matrice de Gram définie selon la forme convexe

$$\boldsymbol{K} = u\,\boldsymbol{K}_{\text{lin}} + (1-u)\,\boldsymbol{K}_{\text{nlin}}, \tag{5.20}$$

avec $\boldsymbol{K}_{\text{lin}} = \boldsymbol{X}\boldsymbol{X}^\top$ et où $u \in [0,\,1]$ est un paramètre réglable. L'espace associé au noyau résultant correspond à la somme directe de deux espaces, $\{\psi_{\text{lin}}(\cdot) \in \mathbb{H}_{\text{lin}} \;:\; \|\psi_{\text{lin}}{}'(\cdot)\|_{\mathbb{H}_{\text{lin}}}/u < \infty\}$ et $\{\psi_{\text{nlin}}(\cdot) \in \mathbb{H}_{\text{nlin}} \;:\; \|\psi_{\text{nlin}}(\cdot)'\|_{\mathbb{H}_{\text{nlin}}}/(1-u) < \infty\}$, où par convention $\frac{x}{0} = 0$ pour $x = 0$. L'estimation conjointe du paramètre u et des coefficients $\boldsymbol{\beta}$ dans un unique problème d'optimisation est connu par le problème d'apprentissage à noyaux multiples. Voir [Rakotomamonjy et al., 2008] et références incluses. Cependant, l'expression (5.20) ne peut pas être remplacée directement dans (5.19) en raison des contraintes imposées à $\boldsymbol{\alpha}$. Le reste de cette partie est consacré à la formulation et à la résolution du problème de l'apprentissage à noyaux multiples sous contraintes totales sur $\boldsymbol{\alpha}$.

Nous proposons la résolution du problème d'optimisation suivant :

$$\begin{aligned}
(u^*, \psi^*(\cdot)) = \underset{\psi(\cdot),u}{\operatorname{argmin}} & \sum_{i=1}^{d} e_i^2 + \eta \left(\frac{1}{u} \|\psi_{\text{lin}}(\cdot)\|^2_{\mathcal{H}_{\text{lin}}} + \frac{1}{1-u} \|\psi_{\text{nlin}}(\cdot)\|^2_{\mathcal{H}_{\text{nlin}}} \right), \\
\text{où} \quad & \psi(\cdot) = \psi_{\text{lin}}(\cdot) + \psi_{\text{nlin}}(\cdot), \quad \text{avec} \quad \psi_{\text{lin}}(\boldsymbol{z}_i) = \boldsymbol{z}_i^\top \boldsymbol{\alpha} \\
\text{sous contraintes :} \quad & e_i = x_i - \psi(\boldsymbol{z}_i), \\
& \boldsymbol{\alpha} \geq \boldsymbol{0} \quad \text{et} \quad 0 \leq u \leq 1.
\end{aligned}$$

Le paramètre u nous permet de contrôler le compromis entre les deux composantes, linéaire et non linéaire, via les normes de $\psi_{\text{lin}}(\cdot)$ et $\psi_{\text{nlin}}(\cdot)$. Notons que la contrainte de somme unité $\boldsymbol{1}^\top\boldsymbol{\alpha} = 1$ n'est pas imposée ici, en raison de sa nature antagoniste avec le paramètre u. Pour obtenir le vecteur des abondances après avoir atteint la convergence, il suffit d'effectuer une normalisation d'une manière explicite.

Il est important de préciser qu'il s'agit bien d'un problème d'optimisation convexe, en vertu de la convexité de ladite fonction perspective $f(u, \psi(\cdot)) = \|\psi(\cdot)\|^2_{\mathbb{H}}/u$ sur $\mathbb{R}_+^* \times \mathbb{H}$, comme indiqué dans [Boyd and Vandenberghe, 2004, chapitre 3] en dimension finie et étendu dans [Rakotomamonjy et al., 2008] pour le cas d'une dimension infinie. Cela nous permet de formuler la procédure d'optimisation en deux étapes, par rapport à $\psi(\cdot)$ et u successivement :

$$\min_u \mathcal{J}(u) \qquad \text{sous contrainte :} \qquad 0 \leq u \leq 1,$$

où

$$\mathcal{J}(u) = \left\{ \begin{array}{l} \min\limits_{\psi(\cdot)} F(u,\psi(\cdot)) = \sum\limits_{i=1}^{d} e_i^2 + \eta\Big(\frac{1}{u}\|\psi_{\text{lin}}(\cdot)\|^2_{\mathbb{H}_{\text{lin}}} + \frac{1}{1-u}\|\psi_{\text{nlin}}(\cdot)\|^2_{\mathbb{H}_{\text{nlin}}}\Big), \\ \qquad \text{sous contraintes :} \quad e_i = x_i - \psi(\boldsymbol{z}_i), \quad \text{avec} \quad \psi(\cdot) = \psi_{\text{lin}}(\cdot) + \psi_{\text{nlin}}(\cdot) \\ \qquad\qquad\qquad\qquad \text{et} \qquad \psi_{\text{lin}}(\boldsymbol{z}_i) = \boldsymbol{z}_i^\top \boldsymbol{\alpha} \quad \text{avec} \quad \boldsymbol{\alpha} \leq \mathbf{0} \end{array} \right. \tag{5.21}$$

La fonction $\mathcal{J}(u)$ dans le sous-problème ci-dessus est définie comme l'optimum [7] de l'ensemble de fonctions convexes dans $(u, \psi(\cdot))$, sous contrainte de convexité sur $\psi(\cdot)$. Comme démontré dans [Boyd and Vandenberghe, 2004, Chapter 3], il s'avère que la fonction $\mathcal{J}(u)$ est convexe en u. En conséquence, le problème d'optimisation sous contraintes sus-mentionné est convexe.

Par la propriété de dualité forte, nous pouvons déduire un problème dual qui admet la même solution $\mathcal{J}(u) = F(u, \psi^*(\cdot))$ que le problème primal ci-dessus. En utilisant le Lagrangien et ses conditions d'optimalité, on obtient une dérivation semblable à celle donnée dans la section 5.4.1, qui conduit au problème dual suivant :

$$\mathcal{J}(u) = \left\{ \begin{array}{l} \max\limits_{\boldsymbol{\beta},\boldsymbol{\gamma}} G(u,\boldsymbol{\beta},\boldsymbol{\gamma}) = -\begin{bmatrix}\boldsymbol{\beta}^\top & \boldsymbol{\gamma}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{K}+\eta\mathbf{I} & u\boldsymbol{X} \\ u\boldsymbol{X}^\top & u\mathbf{I}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{\gamma}\end{bmatrix} + \begin{bmatrix}\boldsymbol{x}^\top & \mathbf{0}^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta} \\ \boldsymbol{\gamma}\end{bmatrix}, \\ \qquad \text{sous contrainte :} \quad \boldsymbol{\gamma} \leq \mathbf{0}. \end{array} \right. \tag{5.22}$$

La reconstruction est alors donnée par $\boldsymbol{x}^* = [\psi^*(\boldsymbol{z}_1) \quad \psi^*(\boldsymbol{z}_2) \quad \cdots \quad \psi^*(\boldsymbol{z}_d)]^\top$ où $\psi^*(\boldsymbol{z}_i) = \boldsymbol{z}_i^\top \boldsymbol{\alpha}^* + \psi^*_{\text{nlin}}(\boldsymbol{z}_i)$. L'estimation finale du vecteur d'abondance est alors obtenue par normalisation, avec

$$\boldsymbol{\alpha}^* = \frac{\boldsymbol{X}^\top\boldsymbol{\beta}^* + \boldsymbol{\gamma}^*}{\mathbf{1}^\top(\boldsymbol{X}^\top\boldsymbol{\beta}^* + \boldsymbol{\gamma}^*)}.$$

Le CADRE 22 étudie les performances de la méthode proposée.

5.4.4 Démélange par modèle de mélange post-non-linéaire

Jusqu'à présent, le modèle de fluctuation non linéaire $\psi_{\text{nlin}}(\boldsymbol{z}_i)$ est indépendant du vecteur d'abondance $\boldsymbol{\alpha}$. En principe, l'ajout de l'abondance dans la fonction non linéaire devrait être avantageuse pour une modélisation plus précise. Nous proposons dans la suite le modèle post-non-linéaire suivant pour $\psi_{\text{nlin}}(\cdot)$:

$$\psi(\boldsymbol{z}_i, \boldsymbol{\alpha}) = \psi_{\text{lin}}(\boldsymbol{z}_i, \boldsymbol{\alpha}) + \psi_{\text{nlin}}(\boldsymbol{\alpha}^\top \boldsymbol{z}_i),$$

où $\psi_{\text{lin}}(\boldsymbol{z}_i, \boldsymbol{\alpha}) = \boldsymbol{z}_i^\top \boldsymbol{\alpha}$. Ce modèle peut caractériser de nombreuses formes de non-linéarités. C'est le cas par exemple avec la transformation quadratique $\psi_{\text{lin}}(\zeta) = \zeta^2$, où le modèle de mélange post-non-linéaire imite les interactions de second ordre entre les signatures des matières, y compris les termes croisés du modèle de mélange bilinéaire abordé dans [Halimi et al., 2011].

Au jour de ce modèle, nous considérons la résolution du problème d'optimisation suivant :

$$\begin{array}{l} \psi^*(\cdot) = \underset{\psi(\cdot)}{\text{argmin}} \sum\limits_{i=1}^{d} e_i^2 + \eta\left(\|\psi_{\text{lin}}(\cdot)\|^2_{\mathcal{H}_{\text{lin}}} + \|\psi_{\text{nlin}}(\cdot)\|^2_{\mathcal{H}_{\text{nlin}}}\right), \\ \qquad \text{où} \quad \psi(\cdot) = \psi_{\text{lin}}(\cdot) + \psi_{\text{nlin}}(\cdot), \quad \text{avec} \quad \psi_{\text{lin}}(\boldsymbol{z}_i) = \boldsymbol{z}_i^\top\boldsymbol{\alpha}, \\ \qquad \text{sous contraintes :} \quad e_i = x_i - \big(\boldsymbol{z}_i^\top\boldsymbol{\alpha} + \psi_{\text{nlin}}(\boldsymbol{z}_i^\top\boldsymbol{\alpha})\big), \\ \qquad\qquad\qquad\qquad \boldsymbol{\alpha} \geq \mathbf{0} \quad \text{et} \quad \mathbf{1}^\top\boldsymbol{\alpha} = 1 \end{array}$$

7. L'existence et le calcul des dérivées de fonctions supremum telles que $\mathcal{J}(u)$ ont été largement discutés dans la littérature. Comme indiqué dans [Rakotomamonjy et al., 2008, Bonnans and Shapiro, 1998], la différentiabilité de $\mathcal{J}(\cdot)$ en tout u_0 est assurée par l'unicité du minimiseur correspondant $(\boldsymbol{\beta}_0^*, \boldsymbol{\gamma}_0^*)$, et par la dérivabilité de la fonction coût $F(u, \psi(\cdot))$ dans (5.21). La dérivée de $\mathcal{J}(\cdot)$ en u_0 peut être calculée comme si $(\boldsymbol{\beta}_0^*, \boldsymbol{\gamma}_0^*)$ est indépendant de u_0, soit $\frac{d\mathcal{J}(u)}{du}\big|_{u=u_0} = \frac{\partial G(u,\boldsymbol{\beta}_0^*,\boldsymbol{\gamma}_0^*)}{\partial u}\big|_{u=u_0}$ où la fonction G est donnée dans (5.22). Cela donne

$$\left.\frac{d\mathcal{J}(u)}{du}\right|_{u=u_0} = -\left(\|\boldsymbol{X}^\top\boldsymbol{\beta}_0^* + \boldsymbol{\gamma}_0^*\|^2 - \boldsymbol{\beta}_0^{*\top}\boldsymbol{K}_{\text{nlin}}\boldsymbol{\beta}_0^*\right).$$

[Cadre 22] Expérimentations et analyse des performances en démélange non linéaire.

Nous considérons deux sous-images hyperspectrales réalisées par l'imageur AVIRIS. La première provient de l'image de Moffett Field, en Californie (Etats-Unis), composée de trois composants purs : eau, sol et végétation. La seconde correspond à la région minière Cuprite, au Névada (Etats-Unis). Pour l'illustration, le nombre de composants purs est fixé à 3 et 5. L'algorithme VCA [Nascimento and Dias, 2004] est utilisé pour extraire les composants purs.
Les performances sont mesurées par l'angle spectral moyen entre le spectre original $\boldsymbol{x}_i$ et sa reconstruction $\boldsymbol{x}_i^*$, selon

$$\frac{1}{n}\sum_{i=1}^{n}\cos^{-1}\left(\frac{\boldsymbol{x}_i^\top \boldsymbol{x}_i^*}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_i^*\|}\right).$$

Le tableau ci-contre montre la pertinence des différentes méthodes proposées dans ce chapitre. Notons que la méthode Kernel-FCLS [Broadwater et al., 2008] ne peut pas être considérée, puisqu'elle ne permet pas d'avoir accès à la reconstruction. La résolution du problème de pré-image est nécessaire, comme préconisé dans le chapitre 3.

	Moffet Field	Cuprite	
	m=3	m=3	m=5
FCLS	0,1416	0,0580	0,0232
ExtM	0,1402	0,0577	0,0211
BilBay	0,1444	0,0657	0,0300
lin+fluct (G)	0,1226	0,0100	0,0075
lin+fluct (P)	0,1236	0,0104	0,0082
multi-κ (G)	0,1255	0,0104	0,0083
multi-κ (P)	0,1286	0,0107	0,0100

Légende

- FCLS : algorithme « *fully constrained least squares* » [Heinz and Chang, 2001]
- ExtM : algorithme « *extended endmember-matrix* » [Raksuntorn and Du, 2010]
- BilBay : algorithme Bayesien pour le modèle bilinéaire généralisé [Halimi et al., 2011]
- lin+fluct : algorithme « linéaire+fluctuation » proposé dans la section 5.4.1
- multi-κ : algorithme « apprentissage de noyaux multiples » proposé dans la section 5.4.3
- G : noyau gaussien
- P : noyau polynomial de degré 2

Malheureusement, la résolution de ce problème est un défi, puisqu'il n'est plus convexe, et les variables duales ne peuvent pas être exprimées sous une forme analytique comme présenté jusqu'à présent. Toutefois, nous pouvons profiter de la forme duale du problème pour déterminer un minimum local d'une manière efficace, par une technique de séparation des variables. Heureusement, une telle approche peut profiter de deux propriétés du modèle de mélange post-non-linéaire. D'une part, les contraintes de non-négativité et de somme unité restreignent d'une manière considérable l'espace des solutions admissibles. D'autre part, il est possible d'initialiser l'algorithme avec une solution appropriée obtenue avec des algorithmes classiques, par exemple avec un algorithme de démélange linéaire.

La méthode proposée consiste à introduire une nouvelle variable $\boldsymbol{v}$ qui remplace $\boldsymbol{\alpha}$ dans la fluctuation non linéaire $\psi_{\text{nlin}}(\cdot)$. Une contrainte d'égalité entre ces deux variables est explicitée, afin d'assurer l'équivalence entre le problème initial et le problème transformé. Ce dernier est alors défini comme suit :

$$\begin{aligned}
\boldsymbol{\alpha}^* = \underset{\psi(\cdot),\boldsymbol{\alpha},\boldsymbol{v},\boldsymbol{e}}{\operatorname{argmin}} \; & \mathcal{J}(\psi(\cdot),\boldsymbol{\alpha},\boldsymbol{v},\boldsymbol{e}),\\
& \text{sous contrainte :} \quad \boldsymbol{v}=\boldsymbol{\alpha},\\
\text{avec} \quad \mathcal{J}(\psi(\cdot),\boldsymbol{\alpha},\boldsymbol{v},\boldsymbol{e}) &= \sum_{i=1}^{d} e_i^2 + \eta\left(\|\psi_{\text{lin}}(\cdot)\|^2_{\mathcal{H}_{\text{lin}}} + \|\psi_{\text{nlin}}(\cdot)\|^2_{\mathcal{H}_{\text{nlin}}}\right),\\
& \text{où} \quad \psi(\cdot)=\psi_{\text{lin}}(\cdot)+\psi_{\text{nlin}}(\cdot), \quad \text{avec} \quad \psi_{\text{lin}}(\boldsymbol{z}_i)=\boldsymbol{z}_i^\top\boldsymbol{\alpha},\\
& \text{sous contraintes :} \quad e_i = x_i - \left(\boldsymbol{z}_i^\top\boldsymbol{\alpha}+\psi_{\text{nlin}}(\boldsymbol{z}_i^\top\boldsymbol{v})\right),\\
& \qquad\qquad \boldsymbol{\alpha}\geq\mathbf{0} \quad \text{et} \quad \mathbf{1}^\top\boldsymbol{\alpha}=1.
\end{aligned}$$

Grâce à cette stratégie, les dérivées par rapport à chacune des deux variables $\boldsymbol{\alpha}$ et $\boldsymbol{v}$ peuvent maintenant être évaluées séparément. D'une part, la variable $\boldsymbol{\alpha}$ est estimée en utilisant un algorithme similaire à celui présenté précédemment dans la section 5.4.1. D'autre part, l'estimation de la variable $\boldsymbol{v}$ est donnée par une approche de descente de gradient. Nous présentons dans la suite cette dernière

approche. Voir [Chen et al., 2013c] pour l'algorithme final, avec des résultats expérimentaux.

La mise à jour de la variable $\boldsymbol{v}$ est donnée par une descente de gradient, selon

$$\boldsymbol{v}_{k+1} = \boldsymbol{v}_k - \eta \nabla_{\boldsymbol{v}} \mathcal{J}_a(\psi_{k+1}, \boldsymbol{\alpha}_{k+1}, \boldsymbol{v}, \boldsymbol{e}_{k+1}, \boldsymbol{d}_k),$$

où $\mathcal{J}_a$ est le Lagrangien augmenté sous sa forme réduite, avec $\boldsymbol{d}$ une variable duale [Boyd et al., 2011], selon

$$\mathcal{J}_a(\psi, \boldsymbol{\alpha}, \boldsymbol{v}, \boldsymbol{e}, \boldsymbol{d}) = \mathcal{J}(\psi, \boldsymbol{\alpha}, \boldsymbol{v}, \boldsymbol{e}) + \frac{1}{2\zeta} \|\boldsymbol{v} - \boldsymbol{\alpha} - \boldsymbol{d}\|^2,$$

où $\zeta > 0$ est un paramètre positif. A cause de la non-linéarité intrinséque de la fonction noyau utilisée, nous proposons une mise à jour selon le gradient :

$$\nabla_{\boldsymbol{v}} \mathcal{J}_a = -\frac{\partial \boldsymbol{\beta}^{*\top} \boldsymbol{K}_{\text{nlin}} \boldsymbol{\beta}^*}{\partial \boldsymbol{v}} + \frac{1}{\zeta}(\boldsymbol{v} - \boldsymbol{\alpha}_{k+1} - \boldsymbol{d}_k).$$

Un exemple de la fonction noyau est le noyau exponentiel $\kappa_{\text{nlin}}(\boldsymbol{z}_i^\top \boldsymbol{\alpha}, \boldsymbol{z}_j^\top \boldsymbol{\alpha}) = \exp(\boldsymbol{\alpha}^\top \boldsymbol{z}_i \, \boldsymbol{z}_j^\top \boldsymbol{\alpha})$, qui est une fonction polynomiale de degré infini. Cela conduit à l'expression suivante pour le gradient :

$$\frac{\partial \boldsymbol{\beta}^\top \boldsymbol{K}_{\text{nlin}} \boldsymbol{\beta}}{\partial \boldsymbol{v}} = \boldsymbol{X}^\top \text{diag}(\boldsymbol{\beta}) \boldsymbol{K}_{\text{nlin}} \text{diag}(\boldsymbol{\beta}) \boldsymbol{X} \boldsymbol{v}.$$

5.5 Démélange non linéaire avec une régularisation spatiale

L'acquisition de cubes de données en imagerie hyperspectrale montre l'intérêt de la dualité spatiale-spectrale pour tout traitement. Une approche de démélange linéaire avec régularisation de type ℓ_1 selon le voisinage de chaque pixel est proposée dans [Zymnis et al., 2007], avec l'utilisation d'une descente de sous-gradient projeté. Dans [Dobigeon and Tourneret, 2011], les auteurs considèrent, d'une part les champs aléatoires de Markov pour modéliser les corrélations spatiales des pixels, et d'autre part l'inférence Bayésienne pour estimer les paramètres du modèle résultant. Dans [Iordache et al., 2012], une approche de type variation totale est utilisée pour la régularisation spatiale. Tous les travaux effectués ont montré que l'intégration de l'information spatiale a un effet positif sur la précision des abondances estimées. Voir aussi [Rogge et al., 2007, Zortea and Plaza, 2009].

Le démélange non linéaire est en soi un problème difficile. Il semble très ambitieux d'y inclure l'information spatiale. Sur la base des avancées prometteuses en démélange non linéaire présentées dans la section précédente, nous proposons de prendre en compte l'information spatiale en utilisant une régularisation de type ℓ_1. La non-linéarité et la norme ℓ_1 rendent le problème difficile à résoudre. Nous proposons de résoudre ce problème avec une méthode itérative *split Bregman*. Nous illustrons dans la suite le principe de notre approche, en renvoyant le lecteur à [Chen et al., 2013d] pour plus de détails avec une étude expérimentale.

Considérons une image hyperspectrale, avec w pixels par ligne et h lignes. Soit $N = wh$ le nombre total de pixels. Soit $\boldsymbol{A} = [\boldsymbol{\alpha}_1 \quad \boldsymbol{\alpha}_2 \quad \cdots \quad \boldsymbol{\alpha}_N]$ la matrice de tous les vecteurs d'abondance $\boldsymbol{\alpha}_i$, pour $i = 1, 2, \ldots, N$. Les contraintes sont alors la non-négativité de chaque élément de $\boldsymbol{A}$ et la somme unité de chaque colonne de $\boldsymbol{A}$, c'est à dire $\boldsymbol{1}_m^\top \boldsymbol{A} = \boldsymbol{1}_N^\top$. Pour résoudre le problème de démélange, nous considérons la fonction coût de la forme

$$\mathcal{J}(\boldsymbol{A}) = \mathcal{J}_{\text{err}}(\boldsymbol{A}) + \eta_1 \, \mathcal{J}_{\text{sp}}(\boldsymbol{A}),$$

où la fonction coût $\mathcal{J}_{\text{err}}(\boldsymbol{A})$ représente l'erreur de modélisation, comme étudié dans la suite selon deux formulations différentes, et $\mathcal{J}_{\text{sp}}(\boldsymbol{A})$ est un terme de régularisation pour favoriser la similarité des abondances dans les pixels voisins. Le paramètre η_1 permet de contrôler le compromis entre la fidélité du modèle et la similarité entre pixels voisins. Afin de privilégier l'homogénéité spatiale, nous proposons le terme de régularisation suivant :

$$\mathcal{J}_{\text{sp}}(\boldsymbol{A}) = \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_i} \|\boldsymbol{\alpha}_i - \boldsymbol{\alpha}_j\|_1, \tag{5.23}$$

où $\|\cdot\|_1$ désigne la norme vectorielle ℓ_1 et $\mathcal{N}_i$ l'ensemble des voisins du pixel i. Par souci de clarté, le voisinage est limité aux quatre pixels adjacents. Dans ce cas, soit la matrice $\boldsymbol{H}_{\leftarrow}$ de taille $(N \times N)$ qui correspond à l'opérateur linéaire de différence entre un vecteur d'abondance et celui de son voisin à gauche. De même pour les matrices $\boldsymbol{H}_{\rightarrow}$, $\boldsymbol{H}_{\uparrow}$ et $\boldsymbol{H}_{\downarrow}$, avec respectivement le voisin à droite, en haut et en bas. En regroupant ces quatre matrices en une seule matrice $\boldsymbol{H} = \begin{bmatrix}\boldsymbol{H}_{\leftarrow} & \boldsymbol{H}_{\rightarrow} & \boldsymbol{H}_{\uparrow} & \boldsymbol{H}_{\downarrow}\end{bmatrix}$ de taille $(N \times 4N)$, le terme de régularisation s'écrit selon

$$\mathcal{J}_{\text{sp}}(\boldsymbol{A}) = \|\boldsymbol{A}\boldsymbol{H}\|_{1,1},$$

où $\|\cdot\|_{1,1}$ est la somme des normes vectorielles ℓ_1 des colonnes de la matrice.

Dans la suite, nous étudions deux expressions différentes de la fonction coût $\mathcal{J}_{\text{err}}(\boldsymbol{A})$. La première expression est l'erreur de modélisation donnée par l'approche présentée dans la section 5.4.1 avec le modèle combinant un mélange linéaire et une fluctuation non linéaire. La seconde expression est due à une écriture du problème de démélange selon un problème de pré-image, et ainsi profite du cadre présenté dans le chapitre 3.

5.5.1 Démélange par le modèle à fluctuation non linéaire

Dans cette section, nous considérons le modèle non linéaire décrit dans la partie 5.4.1 avec un mélange linéaire et une fluctuation non linéaire. Bien que le problème résultant $\mathcal{J}(\boldsymbol{A})$ soit convexe, sa résolution demeure difficile compte tenu du fait qu'il associe un problème de moindres carrés non linéaire LS-SVM avec un terme de régularisation de type ℓ_1. Pour surmonter cette difficulté, nous considérons une approche *split* de séparation des variables, selon

$$\begin{aligned}\min_{\boldsymbol{A},\boldsymbol{\psi}(\cdot)} \sum_{i=1}^{N} \left(\tfrac{1}{\eta}\|\boldsymbol{e}_i\|^2 + \|\boldsymbol{\alpha}_i\|^2 + \|\psi_i(\cdot)\|^2_{\mathrm{H}}\right) + \eta_1 \|\boldsymbol{U}\|_{1,1},\\ \text{sous contraintes :} \quad \boldsymbol{V} = \boldsymbol{A} \quad \text{et} \quad \boldsymbol{U} = \boldsymbol{V}\boldsymbol{H}.\end{aligned}$$

Ainsi la matrice $\boldsymbol{U}$ permet-elle de découpler la norme ℓ_1 du problème de moindres carrés LS-SVM, alors que la matrice $\boldsymbol{V}$ rend possible la résolution du problème LS-SVM en relaxant les relations entre les pixels.

Comme étudié dans [Goldstein and Osher, 2009], l'approche itérative *split Bregman* permet d'une manière efficace de traiter des problèmes à régularisation ℓ_1. En appliquant ces travaux dans le présent cadre, le problème d'optimisation se ramène à

$$\begin{aligned}\boldsymbol{A}_{t+1}, \boldsymbol{\psi}_{t+1}(\cdot), \boldsymbol{V}_{t+1}, \boldsymbol{U}_{t+1} = \underset{\boldsymbol{A},\boldsymbol{\psi}(\cdot),\boldsymbol{V},\boldsymbol{U}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\frac{1}{\eta}\|\boldsymbol{e}_i\|^2 + \|\boldsymbol{\alpha}_i\|^2 + \|\psi_i(\cdot)\|^2_{\mathrm{H}}\right) + \eta_1 \|\boldsymbol{U}\|_{1,1},\\ + \frac{\zeta}{2}\|\boldsymbol{A} - \boldsymbol{V} - \boldsymbol{D}_{1,t}\|_F^2 + \frac{\zeta}{2}\|\boldsymbol{U} - \boldsymbol{V}\boldsymbol{H} - \boldsymbol{D}_{2,t}\|_F^2,\\ \text{où} \qquad \boldsymbol{D}_{1,t+1} = \boldsymbol{D}_{1,t} + (\boldsymbol{V}_{t+1} - \boldsymbol{A}_{t+1}) \qquad \text{et} \qquad \boldsymbol{D}_{2,t+1} = \boldsymbol{D}_{2,t} + (\boldsymbol{V}_{t+1}\boldsymbol{H} - \boldsymbol{U}_{t+1}),\end{aligned} \tag{5.24}$$

où $\|\cdot\|_F^2$ désigne la norme matricielle de Frobenius et ζ un paramètre positif. Le découplage entre les variables nous permet alors de résoudre le problème en minimisant, séparément, par rapport à $(\boldsymbol{A}, \boldsymbol{\psi}(\cdot))$, $\boldsymbol{V}$ et $\boldsymbol{U}$. Ceci donne lieu aux trois étapes suivantes :

1. Optimiser $\boldsymbol{A}$ et $\boldsymbol{\psi}(\cdot)$. Le problème d'optimisation (5.24) se simplifie selon

$$\boldsymbol{A}_{t+1}, \boldsymbol{\psi}_{t+1}(\cdot) = \underset{\boldsymbol{A},\boldsymbol{\psi}(\cdot)}{\operatorname{argmin}} \sum_{i=1}^{N} \left(\tfrac{1}{\eta}\|\boldsymbol{e}_i\|^2 + \|\boldsymbol{\alpha}_i\|^2 + \|\psi_i(\cdot)\|^2_{\mathrm{H}} + \zeta\|\boldsymbol{\alpha}_i - \boldsymbol{\xi}_i^{(t)}\|^2\right), \tag{5.25}$$

   où $\boldsymbol{\xi}_{i,t} = \boldsymbol{V}_{i,t} - \boldsymbol{D}_{1,i,t}$ et où $\boldsymbol{V}_i$ et $\boldsymbol{D}_{1,i}$ désignent la i-ème colonne de $\boldsymbol{V}$ et $\boldsymbol{D}_1$, respectivement. Le problème obtenu est similaire aux problèmes traités dans la section 5.4, avec une solution

de la forme

$$\max_{\boldsymbol{v}} -\tfrac{\rho}{2\zeta}\,\boldsymbol{w}^\top \begin{bmatrix} \boldsymbol{K}_\psi & \boldsymbol{X} & -\boldsymbol{X}\mathbf{1}_m \\ \boldsymbol{X}^\top & \mathbf{I} & -\mathbf{1}_m \\ -\mathbf{1}_m^\top \boldsymbol{X}^\top & -\mathbf{1}_m^\top & m \end{bmatrix} \boldsymbol{w} + \boldsymbol{w}^\top \begin{bmatrix} \boldsymbol{x}_i - \rho\,\boldsymbol{X}\,\boldsymbol{\xi}_{i,t} \\ -\rho\,\boldsymbol{\xi}_{i,t} \\ \rho\,\boldsymbol{\xi}_{i,t}^\top\,\mathbf{1}_m - 1 \end{bmatrix},$$
$$\text{sous contraintes :} \quad \boldsymbol{\gamma}_i \geq \mathbf{0},$$

où $\boldsymbol{K}_\psi = \boldsymbol{X}\boldsymbol{X}^\top + \frac{1}{\zeta}\,(\boldsymbol{K} + \eta\,\mathbf{I})$, $\rho = \frac{\zeta}{1+\zeta}$ et le vecteur $\boldsymbol{w} = [\boldsymbol{\beta}_i^\top \quad \boldsymbol{\gamma}_i^\top \quad \lambda_i^\top]^\top$ contient les multiplicateurs de Lagrange à l'itération t. Une fois les valeurs optimales de ces variables duales déterminées, désignées par $\boldsymbol{\beta}_i^*$, $\boldsymbol{\gamma}_i^*$ et λ_i^*, le vecteur des abondances $\boldsymbol{\alpha}_{i,t+1}$ est estimé selon $\boldsymbol{\alpha}_i^* = \frac{1}{\zeta+1}\left(\boldsymbol{X}^\top\boldsymbol{\beta}_i^* + \boldsymbol{\gamma}_i^* - \lambda_i^*\mathbf{1} + \zeta\boldsymbol{\xi}_{i,k}\right)$. Ce processus est répété pour $i = 1, 2, \ldots, N$ pixels afin de déterminer la matrice $\boldsymbol{A}_{t+1}$.

2. Optimiser $\boldsymbol{V}$. Le problème d'optimisation (5.24) devient

$$\boldsymbol{V}_{t+1} = \underset{\boldsymbol{V}}{\operatorname{argmin}} \|\boldsymbol{A}_t - \boldsymbol{V} - \boldsymbol{D}_{1,t}\|_F^2 + \|\boldsymbol{U}, t - \boldsymbol{V}\boldsymbol{H} - \boldsymbol{D}_{2,t}\|_F^2.$$

En annulant la dérivée de la fonction coût ci-dessus par rapport à $\boldsymbol{V}$, la solution est

$$\boldsymbol{V}_{t+1} = \left(\boldsymbol{A}_t - \boldsymbol{D}_{1,t} + (\boldsymbol{U}_t - \boldsymbol{D}_{2,t})\boldsymbol{H}^\top\right)(\mathbf{I} + \boldsymbol{H}\boldsymbol{H}^\top)^{-1}.$$

3. Optimiser $\boldsymbol{U}$. Le problème d'optimisation est

$$\boldsymbol{U}_{t+1} = \underset{\boldsymbol{U}}{\operatorname{argmin}}\, \eta_1\|\boldsymbol{U}\|_{1,1} + \frac{\zeta}{2}\|\boldsymbol{U} - \boldsymbol{V}_t\,\boldsymbol{H} - \boldsymbol{D}_{2,t}\|_F^2.$$

La solution peut être exprimée par la fonction $\text{Thresh}(x, \tau) = \text{sign}(x)\,\max(|x| - \tau, 0)$, selon

$$\boldsymbol{U}_{t+1} = \text{Thresh}\left(\boldsymbol{V}_t\,\boldsymbol{H} + \boldsymbol{D}_{2,t}\,,\, \sigma/\zeta\right).$$

Pour conclure, nous pouvons montrer que, si le problème (5.24) a une solution $\boldsymbol{A}^*$ donnée pour une valeur fixée de $\zeta > 0$, alors la séquence générée $\boldsymbol{A}_t$ converge vers l'optimum $\boldsymbol{A}^*$ [Eckstein and Bertsekas, 1992].

5.5.2 Démélange supervisé par la résolution du problème de pré-image

Nous concluons ces différentes méthodes de démélange en étudiant le problème de démélange supervisé. Dans le cadre de démélange supervisé, on dispose d'un ensemble d'apprentissage avec des couples abondance-spectre. Soit un ensemble d'échantillons d'apprentissage $\{(\boldsymbol{\alpha}_1, \boldsymbol{x}_1), (\boldsymbol{\alpha}_2, \boldsymbol{x}_2), \ldots, (\boldsymbol{\alpha}_n, \boldsymbol{x}_n)\}$, où $\boldsymbol{\alpha}_i$ désigne les fractions d'abondances associées au vecteur spectral $\boldsymbol{x}_i$. Nous retrouvons ainsi le cas particulier des composants purs connus, comme présenté dans la section 5.1, où les $\boldsymbol{\alpha}_i$ sont les colonnes de la matrice identité de taille $(m \times m)$ avec $n = m$. Le problème consiste à estimer les fractions d'abondances d'un vecteur spectral donné $\boldsymbol{x}$.

Dans la suite, nous proposons d'écrire le problème de démélange comme un problème de pré-image. La formulation que nous présentons permet de profiter des avancées décrites dans le chapitre 3. Nous proposons de résoudre le problème de pré-image en utilisant la méthode de transformation conforme étudiée dans la section 3.2.2. Pour satisfaire une interprétation physique, nous introduisons les deux contraintes de somme unité et de non-négativité de la pré-image. Nous proposons aussi d'intégrer une régularisation spatiale, en utilisant le même principe décrit ci-avant. La méthode de démélange proposée est résumée dans la suite. Voir [Nguyen et al., 2013] pour plus de détails, avec une étude expérimentale sur ses performances.

5.5.2.a Démélange non linéaire comme un problème de pré-image

Pour commencer, considérons le modèle linéaire. Dans ce cas, chaque vecteur d'abondance $\boldsymbol{\alpha}_i$ produit un vecteur spectral $\boldsymbol{x}_i$. Il s'agit d'une transformation linéaire de l'espace des abondances

de dimension m à l'espace spectral de plus grande dimension, notée d. Le problème de démélange consiste à déterminer le vecteur d'abondance d'un vecteur spectral donné. Ceci correspond à un problème de réduction de dimension.

Nous proposons d'utiliser le formalisme des noyaux reproduisants pour définir un modèle de mélange non linéaire. Dans ce cas, la transformation non linéaire est telle que $\boldsymbol{\alpha}_i \mapsto \kappa(\boldsymbol{x}_i, \cdot)$. Le but est d'estimer la transformation inverse, c'est à dire la pré-image $\boldsymbol{\alpha}$ d'un arbitraire élément $\kappa(\boldsymbol{x}_i, \cdot)$. Pour ce faire, nous exploitons l'approche de transformation conforme présentée dans la section (3.2.2). Il s'agit d'une résolution en deux étapes : d'abord, apprendre la transformation inverse ; ensuite, estimer la pré-image. Ces deux étapes sont présentées succinctement dans la suite, en vue de la résolution d'un problème de démélange.

La première étape détermine n fonctions de l'espace RKHS, de la forme $\psi_\ell(\cdot) = \sum_{i=1}^n \theta_{i,\ell}\, \kappa(\boldsymbol{x}_i, \cdot)$. La représentation d'un $\kappa(\boldsymbol{x}_j, \cdot)$ dans ce repère est alors donnée par $\boldsymbol{\psi}_{\boldsymbol{x}_j} = [\psi_1(\boldsymbol{x}_j) \;\; \cdots \;\; \psi_k(\boldsymbol{x}_j)]^\top$. Les paramètres $\theta_{i,\ell}$ sont déterminés selon le modèle de transformation conforme $\boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_j = \boldsymbol{\psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\psi}_{\boldsymbol{x}_j} + \epsilon_{i,j}$, c'est à dire par la minimisation de la variance de l'erreur avec

$$\min_{\psi_1,\ldots,\psi_n} \sum_{i,j=1}^{n} \big|\; \boldsymbol{\alpha}_i^\top \boldsymbol{\alpha}_j - \boldsymbol{\psi}_{\boldsymbol{x}_i}^\top \boldsymbol{\psi}_{\boldsymbol{x}_j} \big|^2 + \eta \sum_{\ell=1}^{n} \|\psi_\ell(\cdot)\|^2.$$

En regroupant les paramètres $\theta_{i,\ell}$ dans une matrice $\boldsymbol{\Theta}$, nous avons sous forme matricielle

$$\boldsymbol{\Theta}^* = \underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} - \boldsymbol{K}\, \boldsymbol{\Theta}^\top \boldsymbol{\Theta}\, \boldsymbol{K}\|_F^2 + \eta \operatorname{tr}(\boldsymbol{\Theta}^\top \boldsymbol{\Theta} \boldsymbol{K}),$$

où $\boldsymbol{\Lambda} = [\boldsymbol{\alpha}_1 \;\; \cdots \;\; \boldsymbol{\alpha}_n]$ et $\boldsymbol{K}$ est la matrice de Gram d'éléments $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, pour $i, j = 1, \ldots, n$. La solution est donnée dans (3.12), avec $\boldsymbol{\Theta}^{*\top} \boldsymbol{\Theta}^* = \boldsymbol{K}^{-1} \big(\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} - \eta\, \boldsymbol{K}^{-1}\big)\, \boldsymbol{K}^{-1}$.

La seconde étape consiste à faire la pré-image de $\kappa(\boldsymbol{x}, \cdot)$, la fonction noyau associée au vecteur spectral à démélanger. Pour cela, elle est projetée sur le sous-espace défini précédemment, résultant en une représentation selon $\sum_{i=1}^n \beta_i^*\, \kappa(\boldsymbol{x}_i, \cdot)$ où les coefficients sont obtenus par $\boldsymbol{\beta}^* = \boldsymbol{K}^{-1} \boldsymbol{\kappa}(\boldsymbol{x})$ et $\boldsymbol{\kappa}(\boldsymbol{x})$ est le vecteur de taille $(n \times 1)$ d'éléments $\kappa(\boldsymbol{x}_i, \boldsymbol{x})$, pour $i = 1, 2, \ldots, n$. En reprenant la transformation conforme présentée ci-avant, on obtient le problème d'optimisation

$$\boldsymbol{\alpha}^* = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha} - \big(\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} - \eta\, \boldsymbol{K}^{-1}\big) \boldsymbol{K}^{-1} \boldsymbol{\kappa}(\boldsymbol{x})\|^2,$$

sous contraintes de somme unité et de non-négativité, comme préconisé pour une interprétation physique du résultat.

5.5.2.b Résolution du problème de pré-image avec une régularisation spatiale

Pour le démélange d'une image hyperspectrale, nous considérons une régularisation spatiale, à l'instar du principe décrit au début de la section 5.5. Dans la suite, nous désignons les vecteurs spectraux à démélanger par $\boldsymbol{x}_{0,j}$, pour $j = 1, 2, \ldots, N$ et la matrice des fractions d'abondance correspondantes par $\boldsymbol{A} = [\boldsymbol{\alpha}_{0,1} \;\; \cdots \;\; \boldsymbol{\alpha}_{0,N}]$, à estimer. Le problème d'optimisation sous contraintes est alors

$$\begin{aligned} \min_{\boldsymbol{A}} \sum_{i=1}^{N} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha}_{0,i} - \big(\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} - \eta\, \boldsymbol{K}^{-1}\big) \boldsymbol{K}^{-1} \boldsymbol{\kappa}(\boldsymbol{x}_{0,i})\|^2 + \eta_1 \, \|\boldsymbol{A}\boldsymbol{H}\|_{1,1}, \\ \text{sous contraintes :} \quad \boldsymbol{A} \geq \boldsymbol{0} \quad \text{et} \quad \boldsymbol{A} \boldsymbol{1}_m^\top = \boldsymbol{1}_N. \end{aligned}$$

Bien qu'il s'agisse d'un problème d'optimisation convexe, sa résolution demeure difficile à cause du terme de la régularisation spatiale. Pour surmonter cette difficulté, nous introduisons deux matrices de découplage, $\boldsymbol{U}$ et $\boldsymbol{V}$, par l'utilisation de l'approche *split* sus-

mentionnée [Goldstein and Osher, 2009]. Le problème résultant est alors

$$\min_{\boldsymbol{A}} \sum_{i=1}^{N} \frac{1}{2} \|\boldsymbol{\Lambda}^\top \boldsymbol{\alpha}_{0,i} - \left(\boldsymbol{\Lambda}^\top \boldsymbol{\Lambda} - \eta\, \boldsymbol{K}^{-1}\right) \boldsymbol{K}^{-1} \boldsymbol{\kappa}(\boldsymbol{x}_{0,i})\|^2 + \eta_1 \|\boldsymbol{U}\|_{1,1},$$
$$\text{sous contraintes :} \quad \boldsymbol{A} \geq \boldsymbol{0}, \qquad \boldsymbol{1}_m^\top \boldsymbol{A} = \boldsymbol{1}_N^\top,$$
$$\boldsymbol{V} = \boldsymbol{A} \quad \text{et} \quad \boldsymbol{U} = \boldsymbol{V}\boldsymbol{H}.$$

L'approche itérative *split Bregman*, proposée [Goldstein and Osher, 2009] et étudiée dans la section précédente, permet d'une manière efficace de résoudre ce problème d'optimisation. Nous renvoyons le lecteur à [Nguyen et al., 2013] pour plus de détails sur l'algorithme final, avec une étude expérimentale des performances de la méthode proposée.

5.6 Conclusion et perspectives

Le problème de démélange en imagerie hyperspectrale suit naturellement la thématique de recherche développée dans le chapitre précédent. Il s'agit bien d'un problème d'approximation parcimonieuse, où le dictionnaire est constitué des signatures spectrales de composants purs, et le problème consiste à estimer leurs contributions. Les défis se résument par des contraintes associées au modèle physique, linéaire ou non linéaire, avec des contraintes sur le mélange spectral ou encore sur la régularité spatiale. Les contributions décrites ici ont eu pour cible les 3 axes de recherches historiques :

- Le démélange par la géométrie, souvent plébiscité pour sa faible complexité de calcul. La méthodologie présentée pour l'estimation des abondances permet de réduire davantage cette complexité ;
- Le démélange par les méthodes statistiques, où nous avons montrer la simplicité d'imposer les contraintes dans les approches de descente de gradient et de projections multiples ;
- Le démélange non linéaire. Dans ce cadre, en combinant un modèle linéaire et une fluctuation non linéaire, nous avons proposé divers modèles de non-linéarité aussi que des méthodes de résolution.

A ces développements s'ajoute une contribution principale qui est l'utilisation de la dualité spectrale-spatiale dans le problème de démélange. Ainsi retrouve-t-on dans ce chapitre divers principes empruntés de l'approximation parcimonieuse présentée dans le chapitre 4, ou encore du problème de pré-image étudié en détail dans le chapitre 3.

Nous sommes persuadés qu'une fertilisation croisée des différents axes de recherche devra fournir des outils adaptés à la résolution du problème de démélange. C'est le cas d'une part avec l'approche géométrique. En suivant la méthodologie exploitée à la fin de ce chapitre, l'information spatiale peut être intégrée dans le cadre proposé pour l'estimation des fractions d'abondance par la géométrie. Cette vision globale permettra d'exploiter les techniques classiques, telles que N-Findr, SGA, VCA, et OSP. D'autre part, les méthodes de démélange non linéaire devront profiter du point de vue illustré par la géométrie, en proposant des algorithmes à faible complexité calculatoire.

Au delà du problème du démélange étudié dans ce chapitre, se dressent plusieurs défis. Nous retrouvons ainsi la fusion de l'information, où deux (voire plusieurs) imageurs sont utilisés avec des résolutions et des propriétés physiques différentes. C'est le cas en particulier de la fusion d'une image hyperspectrale avec une image radar à synthèse d'ouverture (ou SAR pour *synthetic aperture radar*) [Hsu and hua K. Burke, 2003]. Un autre défi réside dans le problème de classification en imagerie hyperspectrale, permettant ainsi une segmentation automatique des images. Le chapitre suivant est dédié au problème de la classification, en proposant, d'une manière générique, une méthodologie pour définir des classifieurs très performants à faible complexité de calcul. Voir [Noumir et al., 2011b] pour une illustration sur la classification en imagerie hyperspectrale. A tous ces axes de recherches, le choix de l'espace de représentation demeure un problème ouvert. Des éléments de solution sont présentés dans [Honeine and Richard, 2010a] qu'il est nécessaire de compléter par des études plus approfondies.

Il semble que la perfection soit atteinte non quand il n'y a plus rien à ajouter, mais quand il n'y a plus rien à retrancher.

[Antoine de Saint-Exupéry]

6

Méthodes à noyaux mono-classe et multi-classes, « simplifiées »

Sommaire

L'apprentissage statistique a toujours été principalement motivé dans le cadre de classification en discriminant deux ou plusieurs hypothèses. Bien que cette thématique n'ait pas été traitée tout au long des chapitres précédents, le présent chapitre vise à combler ce manque en vue de satisfaire la curiosité des lecteurs qui souvent associent méthodes à noyaux et SVM. Malheureusement, les méthodes « à la SVM », aussi bien en classification qu'en régression, se ramènent à des problèmes d'optimisation qui requièrent des bibliothèques d'optimisation dédiées. Les bibliothèques « sur étagére » sont souvent propriétaires et coûteuses. De plus, elles sont loin d'être adaptées à une implémentation industrielle à grande échelle, un défi de plus en plus plébiscité.

Le travail présenté dans ce chapitre s'efforce de proposer des méthodes à noyaux « simplifiées », par opposition aux méthodes classiques. Par « simplification », nous entendons une simplicité de mise en œuvre et d'implémentation avec une efficacité prouvée. Ces propriétés sont essentielles pour une capacité à exploiter ces algorithmes dans des contextes applicatifs issus du monde industriel, tout en quittant notre terrain de jeu favorable qu'est MATLAB. Les algorithmes de méthodes à noyaux devront être à la hauteur des défis envisagés.

J'ai commencé cette activité de recherche depuis que je suis Maître de Conférences, en découvrant l'énorme écart entre le monde académique et le monde industriel. Cette activité a démarré dans le cadre du projet Vigires'eau (ANR, programme CSOSG, 2009-2012) sur la qualité de l'eau dans un réseau d'eau potable, avec notre partenaire Suez Environnement (équipe Ondéo Systems). Ce partenariat industriel est consolidé dans le cadre du nouveau projet franco-allemand SCALA (ANR,

programme CSOSG, 2012-2015) sur les systèmes cyber-physiques SCADA (*Supervisory Control And Data Acquisition*). Les doctorants Zineb Noumir (projet Vigires'eau) et Patric Nader (projet SCALA) y ont participé. Les travaux résumés dans ce chapitre ont fait l'objet d'une publication dans une revue internationale [1], et de plusieurs conférences dans le domaine théorique [2]et applicatif [3]. La poursuite de ces travaux est évidemment d'actualité, à la hauteur des défis envisagés en vue de démocratiser les méthodes à noyaux dans le domaine industriel.

Sommaire

Le présent chapitre s'efforce de présenter succinctement les approches proposées. Le caractère générique des méthodes est mis en avant, avec une présentation assez générale et indépendante de l'application investie. En vue d'apporter des solutions originales à la classification par méthodes à noyaux, nos travaux sont structurés autour de deux axes : la classification mono-classe et la classification multi-classes.

Le premier axe vise une dégénérescence de la classification classique avec le problème de classification mono-classe. Il permet d'aborder le problème de détection à partir d'échantillons disponibles d'une seule classe, celle du bon fonctionnement du système sous surveillance. Les développements scientifiques proposés dans le cadre du problème mono-classe ont été conçus dans les contextes suivants :

- définition du problème mono-classe en proposant une solution optimale au sens des moindres carrés ;
- description de divers critères de parcimonie, en s'inspirant du travail présenté dans le chapitre 4 ;
- mise en œuvre de solutions adaptatives, en vue d'une tâche de détection en ligne ;
- développement d'un cadre théorique pour l'analyse du problème, dont une étude sur le risque de première espèce.

Ces travaux ont pour vocation de faire un pont entre la classification mono-classe par méthodes à noyaux et la détection séquentielle de rupture avec la théorie qui l'accompagne. Précisons toutefois que cette dernière ne peut pas être appliquée en pratique, puisqu'elle nécessite une modélisation exacte du système traité [Basseville and Nikiforov, 1993]. Les méthodes d'apprentissage proposées permettent ainsi de combler une telle pénurie d'information.

Le second axe concerne la classification multi-classes. Les méthodes de classification les plus connues en méthodes à noyaux, dont les SVM sont le fer de lance, définissent le problème d'optimisation avec un très grand nombre d'inconnues. Il s'agit au moins de $n \times \ell$ inconnues, où n désigne le nombre d'échantillons d'apprentissage et ℓ le nombre de classes en compétition. Les techniques de résolution les plus utilisées ont une complexité calculatoire très importante. Les développements que nous proposons se basent sur une réduction du nombre d'inconnues à n, et en conséquence indépendamment du nombre de classes en compétition. Nous décrivons plusieurs techniques pour estimer ces inconnues avec une complexité calculatoire réduite. Ces travaux sont étudiés comme suit :

- développement de classifieurs multi-classes, en explorant l'apprentissage multi-tâches avec des modèles à sortie vectorielle ;
- étude de la classification multi-classes à moindres carrés et en particulier avec une stratégie de type un-contre-tous ;
- analyse théorique du codage des étiquettes et de leurs équivalences dans certains cas.

La pertinence des simplifications proposées est illustrée sur des jeux de données bien connus dans la littérature, aussi bien au niveau des performances en classification qu'au niveau du temps de calcul. Des tests statistiques sont aussi menés pour compléter l'analyse.

1. [Honeine et al., 2013b]
2. [Noumir et al., 2012b, Noumir et al., 2012e, Noumir et al., 2012d, Noumir et al., 2012c, Noumir et al., 2011b, Noumir et al., 2011a]
3. [Nader et al., 2013, Noumir et al., 2012a, Deveughèle et al., 2012, Yin et al., 2012, Fillatre et al., 2011, Fillatre et al., 2010]

6.1 Introduction à la classification mono-classe

A l'opposé des problèmes conventionnels en classification où plusieurs classes sont en compétition, on est souvent confronté à la disponibilité d'échantillons provenant d'une seule classe, la classe du bon fonctionnement du système étudié. Ce problème apparaît naturellement quand on s'intéresse à la détection de changement en traitement du signal, comme c'est le cas avec la détection de rupture, de cyber-attaque, d'intrusion dans un réseau ou encore d'injection de polluants. Dans tous ces domaines d'application, nous avons souvent des échantillons de la classe du bon fonctionnement et peu d'information sur un fonctionnement anormal. De plus, il existe une infinité d'états possibles en dysfonctionnement.

Il s'agit d'une classification dite mono-classe, où un détecteur d'anomalie est déterminé à partir d'échantillons d'apprentissage. En absence d'information sur les diverses classes d'anomalies possibles, la classification mono-classe fournit une règle de détection motivée par les récentes avancées en théorie d'apprentissage statistique. Elle est d'un grand intérêt dans des domaines applicatifs, comme déjà illustré avec succès en science criminalistique [Ratle et al., 2007], réseaux de capteurs sans fil [Zhang et al., 2009], détection de chiffres manuscrits [Tax and Juszczak, 2003], reconnaissance d'objets [Kemmler et al., 2010] et segmentation de la parole en traitement du signal [Davy and Godsill, 2002, Gretton and Désobry, 2003], pour n'en citer que quelques uns. De plus, il est aisé de l'étendre pour la résolution d'une tâche multi-classes, par une combinaison de plusieurs règles de décision où chacune est obtenue à partir d'un classifieur mono-classe défini par classe [Hempstalk and Frank, 2008] (voir aussi le Cadre 24).

La classification mono-classe est devenue une thématique de recherche très active en apprentissage statistique. Il s'agit essentiellement de définir le domaine de description des échantillons dans une seule et unique classe. Dans le cadre des méthodes à noyaux, la frontière de discrimination en classification mono-classe est la sphère de volume minimal qui englobe les échantillons d'apprentissage. La règle de décision consiste à désigner comme nouveauté tout échantillon à l'extérieur de la sphère. C'est l'essence des SVM mono-classe [Schölkopf et al., 2001, Tax, 2001] (voir aussi [Shawe-Taylor and Cristianini, 2004, chapitre 5]), qui estime conjointement le centre de la sphère et son rayon. L'analogie avec les SVM pour la classification permet d'exploiter de nombreuses fonctionnalités, dont l'extension non linéaire grâce au coup du noyau et la parcimonie de la solution dans l'estimation du centre permettant une robustesse aux échantillons aberrants. Le prix à payer est un problème d'optimisation quadratique sous contraintes. Nous reviendrons par la suite sur la formulation du problème mono-classe.

Divers efforts ont été menés pour développer des classifieurs mono-classe à faible complexité de calcul [Liu et al., 2010]. Peu de tentatives ont été faites pour la mise en œuvre de classifieurs mono-classe à moindres carrés, à l'instar des SVM à moindres carrés pour la classification avec LS-SVM [Suykens and Vandewalle, 1999a, Rifkin et al., 2003]. C'est le cas en particulier de la méthode proposée dans [Choi, 2009], toutefois inappropriée pour une tâche de détection puisqu'elle ne dispose pas de règle de décision. De plus, une telle approche n'est pas naturelle pour un traitement en ligne, comme c'est le cas en surveillance avec des échantillons disponibles en continu ou encore lorsqu'il s'agit d'un très grand nombre d'échantillons d'apprentissage.

Afin de développer des méthodes mono-classe adaptatives en vue d'un traitement en ligne, les travaux se sont orientés vers une exploitation des recherches réalisées en classification binaire. Malheureusement, l'approche classique de type SVM-incrémental [Cauwenberghs and Poggio, 2001] n'est pas adéquate dans le cadre du problème mono-classe, comme indiqué dans [Davy et al., 2006]. Une autre approche proposée dans [Gómez-Verdejo et al., 2011] consiste à utiliser une fenêtre exponentielle appliquée aux échantillons (voir par exemple [Kivinen et al., 2004]), ce qui résulte en une modification de la formulation conventionnelle avec la nécessité de plusieurs approximations et hypothèses dont les faibles fluctuations. Il est clair que toutes ces tentatives sont inspirées des approches dérivées en classification binaire, tout en oubliant les particularités du mono-classe dont le principe de sphère qui englobe les échantillons.

Avant de continuer et présenter nos contributions, la formulation classique du problème mono-classe est rappelée dans la suite.

Formulation du problème

Soit $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ l'ensemble des échantillons disponibles pour l'apprentissage. Un classifieur mono-classe définit la sphère de volume minimum qui englobe (presque) tous les échantillons d'apprentissage dans l'espace RKHS $\mathbb{H}$. Cette sphère est décrite complètement par son centre $\mu(\cdot)$ et son rayon ρ. Ces derniers sont alors obtenus par la résolution du problème d'optimisation sous contraintes suivant :

$$\min_{\rho,\mu,\boldsymbol{\zeta}} \ \rho^2 + \frac{1}{\eta n}\sum_{i=1}^{n}\zeta_i \qquad \text{sous contraintes } \|\kappa(\boldsymbol{x}_i,\cdot) - \mu(\cdot)\|_{\mathbb{H}}^2 \leq \rho^2 + \zeta_i, \qquad \text{pour } i = 1, 2, \ldots, n,$$
$$\text{et} \qquad \zeta_i \geq 0, \qquad \text{pour } i = 1, 2, \ldots, n.$$

Cette formulation permet de tolérer la présence d'une fraction d'échantillons à l'extérieur de la sphère. Cette fraction est bornée par le paramètre $\eta \in [0 \ \ 1]$ qui permet de spécifier le compromis entre le volume de la sphère et le nombre d'échantillons aberrants. Les distances de ces derniers à la sphère sont données par les valeurs non-nulles des variables d'écart non-négatives $\zeta_1, \zeta_2, \ldots, \zeta_n$.

Soit le Lagrangien correspondant, à savoir

$$\rho^2 + \frac{1}{\eta n}\sum_{i=1}^{n}\zeta_i - \sum_{i=1}^{n}\alpha_i\big(\rho^2 + \zeta_i - \|\kappa(\boldsymbol{x}_i,\cdot) - \mu(\cdot)\|_{\mathbb{H}}^2\big) - \sum_{i=1}^{n}\gamma_i\zeta_i,$$

où α_i et γ_i désignent les multiplicateurs de Lagrange. En annulant la dérivée de cette expression par rapport aux différentes inconnues, nous obtenons les conditions de Karush-Kuhn-Tucker suivantes :

$$\sum_{i=1}^{n}\alpha_i = 1, \qquad \mu(\cdot) = \sum_{i=1}^{n}\alpha_i\,\kappa(\boldsymbol{x}_i,\cdot) \qquad \text{et} \qquad \gamma_i = \frac{1}{\eta n} - \alpha_i, \ \text{pour } i = 1, 2, \ldots, n.$$

Ces conditions d'optimalité permettent de définir le problème dual suivant :

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n}\alpha_i\,\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i) - \sum_{i,j=1}^{n}\alpha_i\,\alpha_j\,\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) \quad \text{sous contraintes } \sum_{i=1}^{n}\alpha_i = 1,$$
$$\text{et } 0 \leq \alpha_i \leq \frac{1}{\eta n}, \quad \text{pour } i = 1, 2, \ldots, n.$$

Dans le cas particulier d'un noyau de norme unité et grâce à la contrainte de somme unité, le problème se ramène à $\min_{\boldsymbol{\alpha}} \sum_{i,j=1}^{n}\alpha_i\,\alpha_j\,\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$ sous les contraintes ci-dessus.

La résolution de ce problème d'optimisation permet de définir les coefficients α_i qui définissent le centre selon

$$\mu(\cdot) = \sum_{i=1}^{n}\alpha_i\,\kappa(\boldsymbol{x}_i,\cdot).$$

La plupart des coefficients α_i sont nuls, correspondant à des échantillons à l'intérieur de la sphère. Une petite fraction d'échantillons contribue à cette expression. A l'instar de l'expression (4.1) et des développements réalisés dans le chapitre 4, nous écrivons

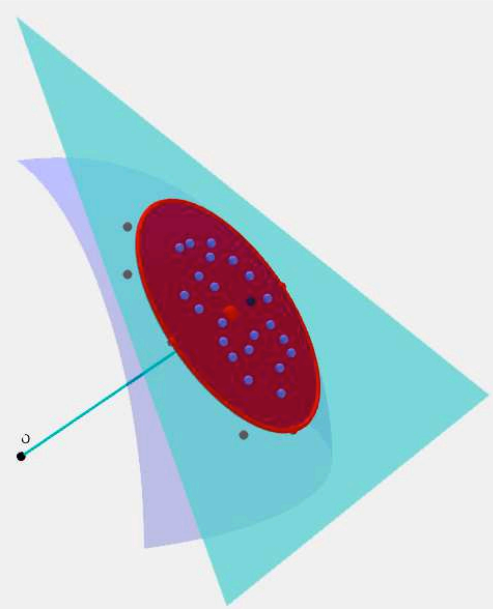
$$\mu(\cdot) = \sum_{j=1}^{m}\alpha_j\,\kappa(\boldsymbol{x}_{\omega_j},\cdot). \tag{6.1}$$

où $\mathcal{D} = \{\boldsymbol{x}_{\omega_1}, \boldsymbol{x}_{\omega_2}, \ldots, \boldsymbol{x}_{\omega_m}\}$ désigne le dictionnaire des m fonctions noyau sélectionnées parmi celles disponibles. Selon le principe des SVM mono-classe, seuls les échantillons du dictionnaire ne sont pas à l'intérieur de la sphère. En conséquence, le rayon est défini par

$$\rho = \min_{j=1,\ldots,m}\|\kappa(\boldsymbol{x}_{\omega_j},\cdot) - \mu(\cdot)\|_{\mathbb{H}}. \tag{6.2}$$

[Cadre 23] Illustration en 3D de la formulation du problème mono-classe.

Avec un noyau de norme unité, les échantillons appartiennent à la sphère (illustrée en bleu) de rayon unité et centrée à l'origine, puisque $\|\kappa(\boldsymbol{x}_i, \cdot) - \mathrm{O}\|_{\mathrm{H}}^2 = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) = 1$ pour tout $\boldsymbol{x}_i$.
Le formalisme proposé par Schölkopf *et coll.* dans [Schölkopf et al., 2001] consiste à définir le plan (illustré en cyan) qui soit le plus éloigné de l'origine et qui la sépare des échantillons. L'intersection entre le plan et la sphère de rayon unité est donnée par le cercle (illustré en rouge). Ce cercle correspond à la solution obtenue selon le formalisme proposé par Tax *et coll.* dans [Tax, 2001, Tax and Duin, 2004], à savoir le cercle qui englobe la plupart des échantillons. Les deux problèmes qui en résultent sont similaires, avec un problème d'optimisation comparable aux SVM pour la classification binaire.
Notre approche consiste à estimer le centre (c'est à dire barycentre) (illustré par le point rouge) des échantillons. Le cercle est alors défini par ce centre et le rayon, ce dernier étant obtenu en fixant la fraction d'échantillons d'apprentissage à l'extérieur du domaine.



La règle de décision pour tout nouvel échantillon $\boldsymbol{x}$ est la suivante : si $\|\kappa(\boldsymbol{x}, \cdot) - \mu(\cdot)\|_{\mathrm{H}} < \rho$, c'est à dire à l'intérieur de la sphère, alors il ne s'agit pas d'un échantillon aberrant. Dans le cas contraire, une anomalie est détectée. Dans cette expression, la distance est obtenue par

$$\|\kappa(\boldsymbol{x}, \cdot) - \mu(\cdot)\|_{\mathrm{H}}^2 = \sum_{i,j=1}^{m} \alpha_i \alpha_j \, \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j}) - 2 \sum_{i=1}^{m} \alpha_i \, \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}) + \kappa(\boldsymbol{x}, \boldsymbol{x}).$$

La formulation présentée ici est celle de *support vector domain description* proposée par Tax *et coll.* dans [Tax, 2001, Tax and Duin, 2004]. Il existe un autre point de vue sur le problème du mono-classe. Il s'agit de l'estimation du support d'une distribution à partir de noyaux radiaux, comme étudié par Schöelkopf *et coll.* dans [Schölkopf et al., 2001]. Ceci correspond à définir le plan le plus éloigné de l'origine qui la sépare des échantillons. Les deux problèmes d'optimisation, de Tax *et coll.* et de Schöelkopf *et coll.*, sont similaires. Le lien entre les deux formulations du problème mono-classe est illustré en 3D dans le Cadre 23.

6.2 Résumé des contributions en classification mono-classe

Indépendamment du point de vue selon lequel le problème du mono-classe est étudié, il s'agit du même problème d'optimisation avec les mêmes ingrédients que les SVM pour la classification binaire. C'est le cas avec la fonction coût charnière, ou encore la présence d'une fraction aberrante parmi les échantillons d'apprentissage. Celle-ci contribue à la forme parcimonieuse de la solution. Comme démontré avec l'approche *support vector domain description* proposée par Tax *et coll.* dans [Tax, 2001, Tax and Duin, 2004], il s'agit d'estimer le centre de la sphère qui englobe la plupart des échantillons. Ce centre est estimé conjointement avec le dictionnaire qui le définit. Les éléments du dictionnaire correspondent aux fonctions noyau $\kappa(\boldsymbol{x}_{\omega_j}, \cdot)$ les plus éloignées du centre $\mu(\cdot)$.

Il semble peu naturel de considérer les éléments les plus éloignés pour obtenir une approximation parcimonieuse. Ce principe est aussi en contradiction avec les approches de type local, comme par

exemple la technique LLE présentée dans le chapitre précédente à la page 100. Il serait judicieux de profiter de la batterie de critères de parcimonie présentés dans le chapitre 4 et qui permettent de borner l'erreur d'approximation tout en garantissant une bonne couverture de l'espace. Pour ce faire, nous proposons de résoudre le problème de classification mono-classe, en découplant les deux problèmes d'estimation, du centre et des éléments du dictionnaire. En se basant sur ce découplage, nos contributions sont multiples comme résumé dans la suite.

Nous proposons d'estimer le centre par la résolution d'un problème de moindres carrés. L'algorithme résultant ne nécessite qu'une simple inversion matricielle, une version adaptative pour la détection en ligne est aussi présentée. Le cadre que nous établissons permet de développer plusieurs résultats théoriques. Nous décrivons en particulier des bornes supérieures sur la probabilité de fausse détection (i.e., risque de première espèce), c'est à dire la probabilité qu'un nouvel échantillon, issu de la même distribution que l'ensemble d'apprentissage, soit à l'extérieur de la sphère définie par notre méthode.

Ce découplage permet d'avoir une liberté sur le choix des éléments du dictionnaire. Nous pouvons appliquer tous les critères de parcimonie étudiés dans le chapitre 4, dont la règle d'approximation linéaire (4.2) et le critère de cohérence (4.4). Pour ce dernier, nous présentons une borne supérieure sur l'erreur d'approximation du centre par le modèle parcimonieux qui en résulte. Nous complétons cette batterie de critères par un nouveau critère de parcimonie qui utilise la distance au centre, à l'instar de l'approche SVM mono-classe classique.

Par opposition à l'optimisation conjointe avec l'approche SVM mono-classe classique, notre stratégie de découplage fournit ainsi une solution sous-optimale au sens des SVM. Par conséquent, l'approche proposée risque de dégrader les performances. Les expérimentations menées montrent que les performances sont essentiellement équivalentes à la technique classique, avec une complexité de calcul plus faible. Ceci est illustré par des expériences sur des jeux de données bien connus et de référence en classification mono-classe. Voir le Cadre 24.

6.2.1 Méthodes de classification mono-classe à moindres carrés

Pour commencer, reprenons le résultat du classifieur SVM mono-classe pour le cas limite $\eta = 1$. Dans ce cas, et comme les conditions $0 \leq \alpha_i \leq \frac{1}{\eta n}$ et $\sum_{i=1}^{n} \alpha_i = 1$ sont toujours satisfaites, on obtient $\alpha_i = 1/n$ pour tout $i = 1, 2, \ldots, n$. Nous avons alors

$$\mu_n(\cdot) = \frac{1}{n} \sum_{i=1}^{n} \kappa(\boldsymbol{x}_i, \cdot), \tag{6.3}$$

ce qui n'est autre que l'estimateur de la densité de probabilité par la méthode de Parzen-Rozenblatt (dit souvent estimateur par noyau), avec $\kappa(\boldsymbol{x}_i, \cdot)$ une fonction radiale centrée sur $\boldsymbol{x}_i$. Pour $\eta < 1$, un sous-ensemble des échantillons d'apprentissage est ainsi utilisé pour construire cet estimateur. Ceci permet d'approcher (6.3) par le modèle d'ordre réduit de la forme

$$\mu_{\mathcal{D}}(\cdot) = \sum_{j=1}^{m} \alpha_j \, \kappa(\boldsymbol{x}_{\omega_j}, \cdot). \tag{6.4}$$

L'approche SVM mono-classe classique détermine conjointement le sous-ensemble $\mathcal{D}$ et les coefficients de pondération α_j. La méthode que nous proposons consiste à scinder le problème en deux parties : (a) d'une part, identifier le dictionnaire selon un critère de parcimonie, soit comme étudié dans le chapitre 4 ou encore en s'inspirant du principe des SVM mono-classe avec la distance au centre ; (b) et d'autre part, estimer les coefficients optimaux α_i dans la solution parcimonieuse $\mu_{\mathcal{D}}(\cdot)$, avec optimalité au sens des moindres carrés, c'est à dire en minimisant $\|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|^2_{\mathbb{H}}$. Ces deux étapes sont décrites dans la suite, tout en proposant un algorithme adaptatif pour la classification mono-classe en ligne. Les performances des méthodes de classification mono-classe proposée sont étudiées dans le Cadre 24.

Règles de parcimonie

Le chapitre 4 étudie une batterie de critères de parcimonie afin de construire un dictionnaire pertinent à partir de l'ensemble d'apprentissage. Tous les critères de parcimonie présentés dans la section 4.1 peuvent être appliqués au problème mono-classe. C'est le cas de la règle d'approximation linéaire (4.2) et du critère de cohérence (4.4) qui opèrent en pré-traitement, c'est à dire sans connaissance de la qualité de l'estimation. Les critères de parcimonie par post-traitement permettent également d'élaborer un modèle d'ordre réduit en mono-classe, comme avec le critère de la cohérence fonctionnelle selon la forme (4.7). Précisons que tous ces critères fournissent une procédure à faible exigence en temps de calcul.

Un nouveau critère de parcimonie par post-traitement émerge en classification mono-classe. En effet, la parcimonie donnée par l'approche SVM mono-classe est définie par la distance au centre de la sphère : les échantillons contribuent au modèle si et seulement si ils sont les plus éloignés du centre. En s'inspirant de ce principe, nous proposons le critère de parcimonie suivant : en fixant à l'avance leur nombre, les éléments du dictionnaire sont identifiés selon

$$\{\omega_1, \omega_2, \ldots, \omega_m\} = \arg \max_{i=1,\ldots,n} \|\kappa(\boldsymbol{x}_i, \cdot) - \mu_n(\cdot)\|^2_{\mathbb{H}},$$

soit $\{\omega_1, \omega_2, \ldots, \omega_m\} = \arg\max_{i=1,\ldots,n} n\kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - 2\sum_{k=1}^n \kappa(\boldsymbol{x}_k, \boldsymbol{x}_i)$. Ce schéma d'élagage peut être modifié pour un traitement en ligne où $\boldsymbol{x}_t$ est ajouté au dictionnaire si $\|\kappa(\boldsymbol{x}_t, \cdot) - \mu_{\mathcal{D}}(\cdot)\|^2_{\mathbb{H}} \geq \rho$. Le niveau de parcimonie est déterminé à l'avance, en fixant soit le rayon ρ ou bien l'ordre du modèle m .

Estimation du centre par moindres carrés

Indépendamment du choix du critère de parcimonie, nous proposons d'estimer les coefficients optimaux α_i dans la solution parcimonieuse $\mu_{\mathcal{D}}(\cdot)$, avec optimalité au sens des moindres carrés entre $\mu_{\mathcal{D}}(\cdot)$ et $\mu_n(\cdot)$, c'est à dire

$$\underset{\alpha_1,\ldots,\alpha_m}{\operatorname{argmin}} \Big\| \frac{1}{n}\sum_{i=1}^n \kappa(\boldsymbol{x}_i, \cdot) - \sum_{j=1}^m \alpha_j \kappa(\boldsymbol{x}_{\omega_j}, \cdot) \Big\|^2_{\mathbb{H}}.$$

En annulant la dérivée de cette fonction coût par rapport à chaque α_k, on obtient $\mu_n(\boldsymbol{x}_k) = \mu_{\mathcal{D}}(\boldsymbol{x}_k)$, c'est à dire $\frac{1}{n}\sum_{i=1}^n \kappa(\boldsymbol{x}_k, \boldsymbol{x}_i) = \sum_{j=1}^m \alpha_j \kappa(\boldsymbol{x}_k, \boldsymbol{x}_{\omega_j})$, pour tout $k = 1, \ldots, m$. Nous avons sous forme matricielle

$$\boldsymbol{\alpha} = \boldsymbol{K}_{\mathcal{D}}^{-1} \boldsymbol{\kappa}_{\mathcal{D}}(\mu_n(\cdot)), \tag{6.5}$$

où $\boldsymbol{K}_{\mathcal{D}}$ est la matrice noyau d'éléments $\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$ et $\boldsymbol{\kappa}_{\mathcal{D}}(\mu_n(\cdot))$ est le vecteur d'éléments $\frac{1}{n}\sum_{k=1}^n \kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_k)$, pour $i, j = 1, \ldots, m$. L'erreur quadratique d'une telle approximation est alors

$$\begin{aligned} \|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|^2_{\mathbb{H}} &= \|\mu_n(\cdot)\|^2_{\mathbb{H}} - 2\,\boldsymbol{\alpha}^\top \boldsymbol{\kappa}_{\mathcal{D}}(\mu_n(\cdot)) + \boldsymbol{\alpha}^\top \boldsymbol{K}_{\mathcal{D}} \boldsymbol{\alpha} \\ &= \frac{1}{n^2}\sum_{i,j=1}^n \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) - \boldsymbol{\kappa}_{\mathcal{D}}(\mu_n(\cdot))^\top \boldsymbol{K}_{\mathcal{D}}^{-1} \boldsymbol{\kappa}_{\mathcal{D}}(\mu_n(\cdot)). \end{aligned} \tag{6.6}$$

Algorithme adaptatif mono-classe pour la détection en ligne

Nous proposons une version adaptative de l'approche mono-classe proposée ci-dessus, ce qui permet la mise en œuvre de détecteur en ligne. Soit $\boldsymbol{x}_t$ l'échantillon présent à l'instant t, avec $t = 1, 2, \ldots$ En suivant une démarche semblable à celle présentée dans la section 4.3.1, la matrice $\boldsymbol{K}_{\mathcal{D}}$ et le vecteur $\boldsymbol{\kappa}_{\mathcal{D}}(\mu_t(\cdot))$ sont actualisés selon le critère de parcimonie :

– **Cas 1. Dictionnaire inchangé**
La matrice $\boldsymbol{K}_{\mathcal{D}}$ reste inchangée et le vecteur $\boldsymbol{\kappa}_{\mathcal{D}}(\mu_t(\cdot))$ dans (6.5) est actualisé selon

$$\boldsymbol{\kappa}_{\mathcal{D}}(\mu_{t+1}(\cdot)) = \frac{1}{n}\big((n-1)\,\boldsymbol{\kappa}_{\mathcal{D}}(\mu_t(\cdot)) + \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t)\big),$$

avec $\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t)$ le vecteur d'éléments $\kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_t)$ pour $j = 1, \ldots, m$. La règle de mise à jour de $\boldsymbol{\alpha}_{t+1}$ à partir de $\boldsymbol{\alpha}_t$ est alors donnée par

$$\boldsymbol{\alpha}_{t+1} = \frac{n-1}{n}\,\boldsymbol{\alpha}_t + \frac{1}{n}\,\boldsymbol{K}_{\mathcal{D}}^{-1}\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t).$$

– **Cas 2. Dictionnaire augmenté**
Le dictionnaire est augmenté selon $\mathcal{D} \cup \{\boldsymbol{x}_t\}$, en y incluant l'échantillon courant. Cela conduit à l'expression suivante de la nouvelle matrice de Gram

$$\begin{bmatrix} \boldsymbol{K}_{\mathcal{D}} & \boldsymbol{\kappa}(\boldsymbol{x}_t) \\ \boldsymbol{\kappa}(\boldsymbol{x}_t)^\top & \kappa(\boldsymbol{x}_t, \boldsymbol{x}_t) \end{bmatrix}.$$

Le lemme d'inversion matricielle par bloc permet de déterminer d'une manière récursive, comme donnée dans l'expression (4.17) où $\boldsymbol{\beta}_t$ vaut $\boldsymbol{K}_{\mathcal{D}}^{-1}\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t)$ dans ce cas. Le vecteur $\boldsymbol{\kappa}_{\mathcal{D}}(\mu_t(\cdot))$ est actualisé selon

$$\boldsymbol{\kappa}_{\mathcal{D}}(\mu_{t+1}(\cdot)) = \frac{1}{n}\begin{bmatrix} (n-1)\,\boldsymbol{\kappa}_{\mathcal{D}}(\mu_t(\cdot)) + \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t) \\ n\,\mu_t(\boldsymbol{x}_t) \end{bmatrix}.$$

En combinant ces expressions, la mise à jour de $\boldsymbol{\alpha}_t$ est donnée par

$$\begin{aligned} \boldsymbol{\alpha}_{t+1} \;=\; & \frac{1}{n}\begin{bmatrix}(n-1)\boldsymbol{\alpha}_t + \boldsymbol{K}_{\mathcal{D}}^{-1}\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t) \\ 0\end{bmatrix} \\ & -\frac{(n-1)\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t)^\top\boldsymbol{\alpha}_t + \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t)^\top\boldsymbol{K}_{\mathcal{D}}^{-1}\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t) - \mu_t(\boldsymbol{x}_t)}{n\,(\kappa(\boldsymbol{x}_t,\boldsymbol{x}_t) - \boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t)^\top\boldsymbol{K}_{\mathcal{D}}^{-1}\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t))}\begin{bmatrix}\boldsymbol{K}_{\mathcal{D}}^{-1}\boldsymbol{\kappa}_{\mathcal{D}}(\boldsymbol{x}_t) \\ 1\end{bmatrix}. \end{aligned}$$

6.2.2 Résultats théoriques sur la classification mono-classe

Dans la suite, nous étudions l'erreur d'approximation $\|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}}$ liée au choix du critère de parcimonie, ainsi que le risque de première espèce en détection. Soit l'espérance mathématique $\mu_\infty(\cdot) = \mathbb{E}[\kappa(\boldsymbol{x}, \cdot)]$, c'est à dire

$$\mu_\infty(\cdot) = \int_{\mathbb{X}} \kappa(\boldsymbol{x}, \cdot)\, dP(\boldsymbol{x}),$$

où $P(\boldsymbol{x})$ est la distribution de probabilité inconnue qui génère les échantillons $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{X}$. A partir de ces échantillons disponibles, le moment empirique $\mu_n(\cdot)$ défini dans (6.3) permet d'estimer $\mu_\infty(\cdot)$. La distance dans $\mathbb{H}$ permet de mesurer la précision de cette approximation, avec

$$\epsilon_0 = \|\mu_n(\cdot) - \mu_\infty(\cdot)\|_{\mathbb{H}}.$$

L'inégalité de Hoeffding montre qu'avec une probabilité d'au moins $1 - \delta$ sur le choix de l'ensemble de n échantillons aléatoires, nous avons l'inégalité de concentration suivante [Shawe-Taylor and Cristianini, 2004] (voir aussi [Bousquet et al., 2004]) :

$$n\,\epsilon_0^2 \leq \sup_{\boldsymbol{x}\in\mathbb{X}} \kappa(\boldsymbol{x}, \boldsymbol{x})\left(2 + \sqrt{-2\ln\delta}\right)^2.$$

Cette inégalité devient pour les noyaux de norme unité : $\epsilon_0 \leq \left(2 + \sqrt{-2\ln\delta}\right)/\sqrt{n}$.

6.2.2.a Erreur d'approximation liée au critère de cohérence

Théorème 5 (Mono-classe : erreur d'approximation pour le critère de coherence)**.**
Le critère de cohérence permet de borner l'erreur d'approximation selon

$$\|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} \leq \left(1 - \frac{m}{n}\right)\sqrt{\max_{i=1,\ldots,n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \nu_0},$$

pour un dictionnaire ν_0-cohérent de m éléments. Cette limite supérieure devient $(1 - m/n)\sqrt{1-\nu_0}$ pour des noyaux de norme unité.

Démonstration. Soit $\mathbf{P}_{\mathcal{D}}$ l'opérateur de projection dans l'espace engendré par les éléments $\kappa(\boldsymbol{x}_{\omega_j}, \cdot)$ pour $j = 1, \ldots, m$ et $\mathbf{I}$ l'application identité. Ainsi la solution optimale, au sens des moindres carrés, vérifie-t-elle $\mu_{\mathcal{D}}(\cdot) = \mathbf{P}_{\mathcal{D}}\mu_n(\cdot)$. Donc, nous avons grâce à l'inégalité triangulaire généralisée

$$\begin{aligned}
\|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} &= \Big\|\frac{1}{n}\sum_{i=1}^{n}(\mathbf{I} - \mathbf{P}_{\mathcal{D}})\kappa(\boldsymbol{x}_i, \cdot)\Big\|_{\mathbb{H}} \\
&\leq \sum_{i=1}^{n}\frac{1}{n}\|\kappa(\boldsymbol{x}_i, \cdot) - \mathbf{P}_{\mathcal{D}}\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}} \\
&= \frac{1}{n}\sum_{i \notin \{\omega_1, \ldots, \omega_m\}}\|\kappa(\boldsymbol{x}_i, \cdot) - \mathbf{P}_{\mathcal{D}}\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}},
\end{aligned}$$

où la dernière égalité est due à l'annulation de $\|\kappa(\boldsymbol{x}_i, \cdot) - \mathbf{P}_{\mathcal{D}}\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}}$ pour tout $\kappa(\boldsymbol{x}_i, \cdot)$ appartenant au développement dans $\mu_{\mathcal{D}}$, c'est à dire pour $i \in \{\omega_1, \omega_2, \ldots, \omega_m\}$. De plus, nous avons

$$\begin{aligned}
\|\kappa(\boldsymbol{x}_i, \cdot) - \mathbf{P}_{\mathcal{D}}\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}}^2 &= \|\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}}^2 - \|\mathbf{P}_{\mathcal{D}}\kappa(\boldsymbol{x}_i, \cdot)\|_{\mathbb{H}}^2 \\
&= \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \max_{\boldsymbol{\gamma}}\frac{\sum_{j\in\mathcal{D}}\gamma_j\kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_i)}{\|\sum_{j\in\mathcal{D}}\gamma_j\kappa(\boldsymbol{x}_{\omega_j}, \cdot)\|_{\mathbb{H}}} \\
&\leq \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \max_{k=1,\ldots,m}\frac{|\kappa(\boldsymbol{x}_{\omega_k}, \boldsymbol{x}_i)|}{\kappa(\boldsymbol{x}_{\omega_k}, \boldsymbol{x}_{\omega_k})} \\
&\leq \kappa(\boldsymbol{x}_i, \boldsymbol{x}_i) - \nu_0,
\end{aligned}$$

où la première égalité est due au théorème de Pythagore et la seconde égalité résulte du fait que le carré de la norme de la projection de $\kappa(\boldsymbol{x}_i, \cdot)$ correspond au plus grand produit scalaire $\langle\kappa(\boldsymbol{x}_i, \cdot), \varphi(\cdot)\rangle_{\mathbb{H}}$ sur toutes les fonctions de norme unité $\varphi(\cdot)$, à savoir $\varphi(\cdot) = \sum_{j\in\mathcal{D}}\gamma_j\kappa(\boldsymbol{x}_{\omega_j}, \cdot)/\|\sum_{j\in\mathcal{D}}\gamma_j\kappa(\boldsymbol{x}_{\omega_j}, \cdot)\|_{\mathbb{H}}$. La première inégalité découle d'une distribution spécifique des coefficients, avec $\gamma_j = 0$ pour tout $\boldsymbol{x}_{\omega_j} \in \mathcal{D}$ sauf pour un seul indice k avec $\gamma_k = \pm 1$, selon le signe de $\kappa(\boldsymbol{x}_k, \boldsymbol{x}_i)$. La dernière inégalité découle du critère de cohérence. ■

6.2.2.b Risque de première espèce

Afin d'étudier le risque de première espèce, nous proposons de borner la probabilité qu'un nouvel échantillon $\boldsymbol{x}$, généré à partir de la même distribution de probabilité que les échantillons $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ d'apprentissage, soit à l'extérieur de la sphère définie par la méthode de classification mono-classe.

Théorème 6 (Mono-classe : risque de première espèce)**.**
Soit la sphère de centre $\mu_{\mathcal{D}}(\cdot)$ et de rayon $2\epsilon_0 + 2\|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} + \max_{i=1,\ldots,n}\|\kappa(\boldsymbol{x}_i, \cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}}$. Avec une probabilité d'au moins $1 - \delta$ sur le choix de l'ensemble de n échantillons aléatoires, la probabilité qu'un nouvel échantillon $\boldsymbol{x}$ soit à l'extérieur de la sphère est bornée selon

$$P\big(\|\kappa(\boldsymbol{x}, \cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} > 2\epsilon_0 + 2\|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} + \max_{i=1,\ldots,n}\|\kappa(\boldsymbol{x}_i, \cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}}\big) \leq \frac{1}{n+1}.$$

Démonstration. La preuve est en deux parties. D'une part, nous avons :

$$\begin{aligned}
\|\kappa(\boldsymbol{x}, \cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} &\leq \|\kappa(\boldsymbol{x}, \cdot) - \mu_n(\cdot)\|_{\mathbb{H}} + \|\mu_n(\cdot) - \mu_{\mathcal{D}}\|_{\mathbb{H}} \\
&\leq \|\kappa(\boldsymbol{x}, \cdot) - \mu_\infty(\cdot)\|_{\mathbb{H}} + \epsilon_0 + \|\mu_n(\cdot) - \mu_{\mathcal{D}}\|_{\mathbb{H}},
\end{aligned}$$

où l'inégalité triangulaire est appliquée à deux reprises, la première fois pour l'approximation du modèle complet par une combinaison d'un sous-ensemble d'échantillons, et la seconde fois pour l'estimation du centre par un ensemble d'échantillons fini de taille n. D'autre part, nous avons pour tout $\boldsymbol{x}_i$:

$$\begin{aligned}
\|\kappa(\boldsymbol{x}_i, \cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} &\geq \|\kappa(\boldsymbol{x}_i, \cdot) - \mu_n(\cdot)\|_{\mathbb{H}} - \|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} \\
&\geq \|\kappa(\boldsymbol{x}_i, \cdot) - \mu_\infty(\cdot)\|_{\mathbb{H}} - \epsilon_0 - \|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}}.
\end{aligned}$$

[CADRE 24] Performances des méthodes de classification mono-classe proposées.

La plupart des jeux de données existants sont ceux de la classification multi-classes. Rien n'empêche d'aborder une classification multi-classes en utilisant les machines mono-classe [Tohmé and Lengellé, 2011] : chaque classe est définie par un classifieur mono-classe, dite classe cible, et la règle de décision finale est donnée par la combinaison de ces classifieurs. Leurs paramètres sont estimés en considérant un sous-ensemble de la classe cible (n_{app} échantillons), et l'erreur de classification est estimée sur les échantillons restants (n_{test}), dont certains de la classe cible et tout le reste des autres classes.

Nous avons utilisé des jeux de données multi-classes bien connus dans la littérature des machines mono-classe [Wang et al., 2006] : *iris*, *wine*, et *breast cancer*, disponibles à partir du référentiel UCI. Pour les expériences sur les données *iris* et comme préconisé dans la littérature, la troisième et quatrième caractéristiques sont utilisées.

Afin d'estimer l'erreur de classification, une validation croisée de 10 partitions a été utilisée, avec des paramètres optimisés par une recherche sur une grille : $\{2^{-5}; 2^{-4}; \ldots; 2^{5}\}$, dont la largeur de bande du noyau gaussien. Afin de favoriser la méthode classique SVM mono-classe à nos méthodes, la taille du dictionnaire a été fixée pour toutes les méthodes en tenant compte de la configuration optimale de l'algorithme SVM classique.

Le tableau suivant donne les résultats de classification de la méthode proposée, couplée avec le critère de distance et le critère de cohérence, en les comparant avec les SVM mono-classe. Nous avons également inclus le pourcentage communs des dictionnaires, entre cette dernière et chacune des méthodes proposées. Les coûts de calcul de ces machines, selon la meilleure configuration, sont donnés à titre indicatif (en seconde).

		SVM classique (ordre optimal)	**Comparaisons entre les méthodes proposées** (même ordre que dans la méthode SVM classique)					
			modèle complet		cohérence		distance	
iris	$n_{\text{app}}\|n_{\text{test}}$	erreur	erreur	dict.	erreur	dict.	erreur	dict.
classe 0	25|125	1,84	1,20	65%	0,49	60%	0,16	50%
classe 1	25|125	5,28	4,50	55%	4,40	45%	3,12	56%
classe 2	25|125	4,80	7,01	55%	4,10	60%	3,01	60%
	moyenne :	**3,97**	**4,23**	60%	**2,99**	55%	**2,09**	55%
	temps :	*(2,5)*	*(0,3)*		*(0,6)*		*(0,20)*	
wine								
classe 0	29|148	15,48	17,18	66%	8,89	80%	16,32	42%
classe 1	35|142	18,26	21,30	50%	22,60	70%	18,24	45%
classe 2	24|154	14,47	14,18	55%	17,70	86%	13,20	45%
	moyenne :	**16,07**	**17,55**	55%	**16,39**	79%	**15,92**	44%
	temps :	*(22,2)*	*(0,22)*		*(0,3)*		*(0,26)*	
cancer								
classe 0	222|461	2,30	6,01	65%	3,03	50%	4,7	55%
classe 1	119|563	5,21	7,86	55%	4,78	46%	4,90	46%
	moyenne :	**4,25**	**6,94**	60%	**3,90**	48%	**4,8**	50%
	temps :	*(42)*	*(1,26)*		*(2,4)*		*(1,30)*	

Par conséquent, nous obtenons en combinant ces deux résultats :

$$\begin{aligned} P\big(\|\kappa(\boldsymbol{x},\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} > 2\epsilon_0 + 2\|\mu_n(\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} + \max_{i=1,\ldots,n} \|\kappa(\boldsymbol{x}_i,\cdot) - \mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}}\big) \\ \leq P\big(\|\kappa(\boldsymbol{x},\cdot) - \mu_{\infty}(\cdot)\|_{\mathbb{H}} > \max_{i=1,\ldots,n} \|\kappa(\boldsymbol{x}_i,\cdot) - \mu_{\infty}(\cdot)\|_{\mathbb{H}}\big) \\ \leq \frac{1}{n+1}, \end{aligned}$$

où la dernière inégalité est due à la symétrie de l'hypothèse *i.i.d.* sur les $n+1$ échantillons générés de la même distribution. ■

Ce théorème généralise des résultats bien connus dans la littérature. C'est le cas en particulier avec le centre empirique défini par le modèle complet, c'est à dire $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$. Nous retrouvons ainsi la relation donnée dans [Shawe-Taylor and Cristianini, 2004, chapitre 5], à savoir :

$$P\big(\|\kappa(\boldsymbol{x},\cdot) - \mu_n(\cdot)\|_{\mathbb{H}} > 2\epsilon_0 + \max_{i=1,\ldots,n} \|\kappa(\boldsymbol{x}_i,\cdot) - \mu_n(\cdot)\|_{\mathbb{H}}\big) \leq \frac{1}{n+1}.$$

Nous proposons d'étendre le théorème 6 au cas où les m échantillons définis par $\mathcal{D}$ sont à l'extérieur ou sur la sphère, comme préconisé par les SVM et le critère de distance au centre.

Théorème 7 (Mono-classe : risque de première espèce)**.**
Considérons le même problème que dans le théorème 6, où les échantillons dans $\mathcal{D}$ sont les plus distants du centre. Avec une probabilité d'au moins $1-\delta$ sur le choix de l'ensemble de n échantillons aléatoires, la probabilité qu'un nouvel échantillon $\boldsymbol{x}$ soit à l'extérieur de la sphère est bornée selon

$$P\big(\|\kappa(\boldsymbol{x},\cdot)-\mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} > 2\epsilon_0 + 2\|\mu_n(\cdot)-\mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}} + \min_{j=1,\ldots,m}\|\kappa(\boldsymbol{x}_{\omega_j},\cdot)-\mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}}\big) \leq \frac{m}{n+1}.$$

La démonstration découle de celle du théorème précédent. Il est à noter que dans les deux théorèmes, l'erreur $\|\mu_n(\cdot)-\mu_{\mathcal{D}}(\cdot)\|_{\mathbb{H}}$ est minimisée par l'approche à moindres carrés que nous proposons à partir de l'expression (6.6).

6.3 Introduction à la classification multi-classes

L'apprentissage automatique/statistique a pris son essor grâce aux méthodes connexionnistes dites réseaux de neurones artificiels [Rosenblatt, 1958, Rumelhart et al., 1986]. Inspirées du fonctionnement du cerveau, la force de ces méthodes réside dans un grand nombre de paramètres qui permet une remarquable capacité d'apprentissage. Le prix à payer est le problème d'optimisation non-convexe, la présence de minima locaux et le surapprentissage. Ces inconvénients ont été surmontés par les méthodes à noyaux, dont les SVM sont les fers de lance, avec la théorie qui les accompagne. Le principe sous-jacent réside dans la résolution d'un problème d'optimisation régularisé. Comme démontré par le Théorème de Représentation (`Th.Rep`), le nombre de paramètres est égal au nombre d'échantillons disponibles pour l'apprentissage.

Le problème de classification binaire, *i.e.,* pour une tâche de discrimination à deux classes, a été largement étudié dans la littérature, dans divers domaines d'application, et les performances des diverses techniques de classification ont été bien établies. C'est le cas en particulier des SVM [Vapnik, 1998] et des machines à moindres carrés (LSM pour *Least Squares Machines* ou aussi LS-SVM) [Rifkin et al., 2003]. Une généralisation de ces résultats est souhaitée pour des tâches de discrimination multi-classes, avec la multiplication du nombre de classes en compétition. Afin d'étendre les méthodes de classification binaire pour des tâches à plusieurs classes, deux axes de recherche sont souvent exploitées : l'approche « machine unique » et l'approche « diviser-pour-conquérir » [Guermeur, 2008].

Le premier axe de recherche vise à intégrer simultanément toutes les contraintes de discrimination au sein d'une « machine unique », par exemple en maximisant simultanément toutes les marges entre les différentes classes. Parmi ces méthodes, la plus connue est la méthode de Weston et Watkins [Weston and Watkins, 1999] qui consiste à étendre le principe des SVM au cas multi-classes par une stratégie de type un-contre-tous, en y incluant autant de plans séparateurs que nécessaire. L'inconvénient majeur de cette méthode est sa complexité de calcul, avec un nombre important de contraintes à satisfaire. Une tentative pour réduire cette complexité est proposée par la méthode de Crammer et Singer [Crammer and Singer, 2002] qui considère le même problème mais avec des modèles sans biais, réduisant ainsi le nombre de contraintes. Dans tous les cas, une machine unique « à la SVM » augmente de manière très significative la complexité du problème traité et donc nécessite des techniques d'optimisation avancées [Bredensteiner and Bennett, 1999].

Le second axe de recherche, « diviser-pour-conquérir », permet de surmonter la complexité du problème abordé en le décomposant en un ensemble de sous-problèmes binaires. Chaque sous-problème est traité séparément et les résultats ainsi obtenus sont fusionnés pour donner la solution multi-classes finale [Crammer and Singer, 2002, Dietterich and Bakiri, 1995, Ou and Murphey, 2007, Allwein et al., 2001]. Les stratégies de décomposition les plus connues sont les stratégies un-contre-tous (OvA pour *one versus all the rest classes*) [Rifkin and Klautau, 2004], un-contre-un (OvO pour *one versus one*) [Fürnkranz, 2002], ainsi que l'utilisation d'un graphe orienté acyclique (DAG pour *directed acyclic graph*) [Platt et al., 2000].

De multiples études comparatives ont été effectuées en passant en revue la plupart des méthodes de classification multi-classes proposées dans la littérature. Voir par exemple [Rifkin et al., 2003, Suykens et al., 2002, Rifkin and Klautau, 2004]. Ces résultats montrent que les stratégies de décomposition, comme OvO, OvA et DAG, ont essentiellement la même précision que les machines uniques. En outre, les solutions aux moindres carrés ont essentiellement les mêmes performances que celles obtenues par la fonction coût charnière des SVM. Par conséquent et comme préconisé par Rifkin et Klautau dans [Rifkin and Klautau, 2004], une machine simple telle qu'en combinant une stratégie « diviser-pour-conquérir » avec LSM, est préférable à une machine plus complexe à mettre en œuvre.

Dans la suite, nous rappelons les deux stratégies « diviser-pour-conquérir » avec OvA et OvO, ainsi que les « machines uniques » les plus connues avec la méthode de Weston et Watkins et la méthode de Crammer et Singer. Soit un problème de classification à ℓ classes. L'appartenance de chaque échantillon $\boldsymbol{x}_i$ de l'ensemble d'apprentissage à une classe est donnée par son étiquette y_i.

Stratégies « diviser-pour-conquérir »

En stratégie « diviser-pour-conquérir », le problème multi-classes est décomposé en un ensemble de sous-problèmes de classification binaire, les fonctions de décision ainsi obtenues sont alors fusionnées pour donner la solution finale. Soit $\psi^{(k)}(\cdot)$ la fonction de décision du k-ème classifieur binaire obtenue en utilisant les mêmes échantillons d'apprentissage, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$, et en leur attribuant ses propres étiquettes, $y_1^{(k)}, y_2^{(k)}, \ldots, y_n^{(k)}$, comme illustré dans la suite.

En stratégie un-contre-tous (OvA), ℓ classifieurs binaires sont nécessaires, chacun confrontant une classe à toutes les autres classes. Les étiquettes sont alors $y_j^{(k)} \in \{-1; +1\}$ selon l'appartenance de $\boldsymbol{x}_j$ pour le k-ème classifieur binaire. En appliquant le Théorème de Représentation (`Th.Rep`) à chacun des ℓ classifieurs binaires, il faut estimer en total $\ell \times n$ coefficients inconnus $\alpha_j^{(k)}$, pour $j = 1, 2, \ldots, n$ et $k = 1, 2, \ldots, \ell$. En stratégie un-contre-un (OvO), $\ell(\ell-1)/2$ classifieurs binaires sont considérés, chacun traitant une paire de classes différentes. Nous avons dans ce cas $y_j^{(k)} \in \{-1; 0; 1\}$, où la valeur zéro correspond à des échantillons qui n'appartiennent à aucune des deux classes en cours de discrimination. Bien que $\ell(\ell-1)/2$ classifieurs binaires sont nécessaires, les sous-problèmes sont « plus petits » que dans le cas de la stratégie OvA, puisque chaque échantillon est utilisé par $\ell - 1$ fonctions de décision.

En terme de performances, les différentes stratégies (dont OvA et OvO, ainsi que le graphe orienté acyclique, les codes complets, ...) sont essentiellement équivalentes. Voir ci-dessus pour les références. Par conséquent, les stratégies les plus simples sont préconisées. Ainsi les stratégies OvA et OvO sont-elles des candidates naturelles, mais au prix de résoudre plusieurs sous-problèmes de classification binaire.

Machines uniques « à la SVM »

Les performances des SVM en classification binaire ont motivé la communauté pour proposer des machines *à la SVM* pour la classification multi-classes [Guermeur, 2008]. Les extensions les plus connues sont la méthode de Weston et Watkins avec un modèle affine et la méthode de Crammer et Singer avec un modèle sans biais.

L'idée est essentiellement une stratégie OvA avec des classifieurs binaires de type SVM classiques, tout en visant à traiter simultanément toutes les classes par la résolution d'un seul problème d'optimisation. Dans [Weston and Watkins, 1999], Weston et Watkins reprennent les principe des SVM avec ℓ plans séparateurs, chacun défini par les paramètres $\boldsymbol{w}_k$ et b_k, pour $k = 1, 2, \ldots, \ell$. En considérant aussi des variables de relaxation $\xi_{i,j}$, le problème d'optimisation est le suivant :

$$\min_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_\ell,b_1,\ldots,b_\ell,\boldsymbol{\xi}} \frac{1}{2}\sum_{i=1}^{\ell} \|\boldsymbol{w}_i\|^2 + \eta \sum_{i=1}^{n} \sum_{j \neq y_i} \xi_{i,j},$$

sous contraintes

$$\begin{aligned} \boldsymbol{w}_{y_i}^\top \boldsymbol{x}_i + b_{y_i} &\geq \boldsymbol{w}_j^\top \boldsymbol{x}_i + b_j + 1 - \xi_{i,j}, && \text{pour } i = 1, \ldots, n \text{ et } j = 1, \ldots, \ell, \\ \text{et} \qquad \xi_{i,j} &\geq 0, && \text{pour } i = 1, \ldots, n \text{ et } j = 1, \ldots, \ell. \end{aligned}$$

La résolution de ce problème d'optimisation est donnée par le point-selle du Lagrangien [Weston and Watkins, 1999]. Une fois les paramètres des ℓ plans séparateurs définis, la fonction de décision pour un échantillon $\boldsymbol{x}$ est

$$\underset{j=1,\ldots,\ell}{\operatorname{argmax}} \boldsymbol{w}_j^\top \boldsymbol{x} + b_j.$$

L'inconvénient majeur de cette méthode est sa complexité calculatoire, avec un nombre important de contraintes à satisfaire.

Afin de réduire la complexité de calcul, Crammer et Singer proposent dans [Crammer and Singer, 2002] d'utiliser un modèle linéaire sans biais pour chaque plan séparateur, à l'opposé du modèle affine décrit ci-avant. Le problème d'optimisation est alors :

$$\min_{\boldsymbol{w}_1,\ldots,\boldsymbol{w}_\ell,\boldsymbol{\xi}} \frac{1}{2} \sum_{i=1}^{\ell} \|\boldsymbol{w}_i\|^2 + \eta \sum_{i=1}^{n} \xi_i,$$

sous contraintes

$$\begin{aligned} \boldsymbol{w}_{y_i} \boldsymbol{x}_i - \boldsymbol{w}_j^\top \boldsymbol{x}_i &\geq 1 - \delta_{y_i,j} - \xi_i && \text{pour } i = 1, \ldots, n \text{ et } j = 1, \ldots, \ell, \\ \text{et} \qquad \xi_i &\geq 0 && \text{pour } i = 1, \ldots, n, \end{aligned}$$

où δ désigne le symbole de Kronecker. La fonction de décision est alors : $\operatorname{argmax}_{j=1,\ldots,\ell} \boldsymbol{w}_j^\top \boldsymbol{x}$. Ce problème d'optimisation est résolu de la même manière que celui de Weston et Watkins, et toujours avec une complexité de calcul élevée.

6.4 Résumé des contributions à la classification multi-classes

Comme montré ci-avant, toute machine multi-classes est définie par un ensemble de (au moins) ℓ fonctions de décision $\psi^{(k)}(\cdot)$, complètement décrites par les coefficients à déterminer $\alpha_j^{(k)}$, pour $j = 1, 2, \ldots, n$ et $k = 1, 2, \ldots, \ell$. Dans les meilleurs cas, il s'agit d'un problème d'estimation avec $n \times \ell$ paramètres inconnus. C'est le cas à titre indicatif de la stratégie OvA qui résout ℓ sous-problèmes de classification binaire, chaque sous-problème admet n inconnues comme démontré par le Théorème de Représentation (`Th.Rep`).

Nos principales contributions se résument par le fait de montrer que le nombre d'inconnues peut être réduit de manière significative, sans sacrifier les performances. Nous proposons de repousser les limites aux extrêmes, en considérant autant de paramètres que d'échantillons d'apprentissage, et ainsi indépendamment du nombre de classes. Ces considérations peuvent sembler étonnantes, puisqu'elles sont à l'opposé des méthodes de classification multi-classes connues dans la littérature. Pourtant, nous rappelons que le principe sous-jacent du Théorème de Représentation stipule qu'il y ait autant de paramètres α_i que d'échantillons $\boldsymbol{x}_i$, et cela indépendamment du type (binaire, réel ou vectoriel) des étiquettes $\boldsymbol{y}_i$.

Nous proposons des machines de classification multi-classes qui ont essentiellement la même complexité de calcul qu'un seul classifieur binaire. Pour ce faire, deux approches sont investies.

– D'une part, nous proposons dans la section 6.4.1 un nouveau cadre de classifieurs multi-classes en développant des machines à sortie vectorielle. Pour cela, nous revisitons des méthodes classiques de classification binaire afin d'en proposer des versions à sortie vectorielle, sur la base des mêmes routines d'optimisation classiques, et sans en augmenter véritablement le coût calculatoire. Plus précisément, nous nous attardons sur trois méthodes largement reconnues dans la littérature : SVM [Vapnik, 1998], LSM (ou LS-SVM) [Suykens and Vandewalle, 1999b] et RLSC (pour *regularized least-squares classification*) [Rifkin, 2002].

– D'autre part, nous étudions dans la section 6.4.2 le couplage entre la stratégie de décomposition OvA et la résolution de sous-problèmes avec optimalité au sens des moindres carrés. Nous décrivons des relations inhérentes entre les inconnues des différents sous-problèmes. Nous montrons qu'il s'agit bien d'un problème d'estimation de n inconnues. La solution du problème multi-classes est ainsi obtenue par l'inversion d'une seule matrice de taille $(n \times n)$.

Nous complétons ces deux approches par une étude du codage des étiquettes en classification multi-classes, décrite dans la section 6.4.3. Il existe divers codages d'étiquettes dans la littérature, dont le codage ± 1 [Dietterich and Bakiri, 1995, Allwein et al., 2001], le codage standard par les fonctions indicatrices [Bishop, 1995, Huang et al., 2012], le codage par alignement noyau-cible [Guermeur, 2008], le codage par principe inductif [Lee et al., 2004] et le codage par corrélation minimale [Szedmak et al., 2006]. Avec le foisonnement des différents codages, notre contribution dans cette thématique n'est pas de proposer « encore » un autre codage des étiquettes. Paradoxalement, nous démontrons que ces différentes étiquettes sont équivalentes dans le cadre d'un problème de moindres carrés. Ces résultats théoriques sur le problème LSM sont complétés par des tests statistiques sur les autres méthodes multi-classes sus-mentionnées. Nous montrons dans le Cadre 27 qu'il n'existe pas de différences significatives entre les codages d'étiquettes.

La pertinence des deux approches proposées est étudiée dans la section 6.4.4 au travers de jeux de données bien connus dans la littérature. Voir le Cadre 26 et le Cadre 28 pour une analyse comparative, en terme de taux d'erreur de classification et aussi en temps de calcul.

6.4.1 Développement de classifieurs multi-classes à sortie vectorielle

Afin de simplifier la présentation, le cas de classification linéaire est avancé. Une extension au cas non linéaire est directe grâce au coup du noyau, en remplaçant le produit scalaire linéaire $\boldsymbol{x}_i^\top \boldsymbol{x}_j$ par un noyau non linéaire $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, ou encore la matrice $\boldsymbol{X}^\top \boldsymbol{X}$ par la matrice de Gram correspondante $\boldsymbol{K}$.

Approche proposée

Pour un ensemble d'apprentissage $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ à ℓ classes, soit la fonction vectorielle (c'est à dire un ℓ-uplet de fonctions)

$$\boldsymbol{\psi}(\cdot) = [\psi^{(1)}(\cdot) \quad \psi^{(2)}(\cdot) \quad \cdots \quad \psi^{(\ell)}(\cdot)]^\top.$$

Dans cette définition, les ℓ fonctions de décision sont déterminées dans le même esprit que la stratégie OvA [4], tout en ne se limitant pas à celle-ci. Pour le k-ème sous-problème de classification binaire, défini par la fonction de décision $\psi^{(k)}(\cdot)$, nous associons une étiquette vectorielle $\boldsymbol{y}^{(k)}$ de taille $(n \times 1)$, par exemple avec des valeurs ± 1 comme présenté dans le Cadre 25. Voir plus loin la section 6.4.3 pour une étude sur les différents codages des étiquettes. Soit $\boldsymbol{Y}$ la matrice de taille $(\ell \times n)$ où chaque ligne correspond à une étiquette vectorielle, c'est à dire $\boldsymbol{Y} = [\boldsymbol{y}^{(1)} \quad \boldsymbol{y}^{(2)} \quad \cdots \quad \boldsymbol{y}^{(\ell)}]^\top$. En d'autres termes, $\boldsymbol{Y} = [\boldsymbol{y}_1 \quad \boldsymbol{y}_2 \quad \cdots \quad \boldsymbol{y}_n]$ où $\boldsymbol{y}_i$ désigne le vecteur des étiquettes associées à $\boldsymbol{x}_i$ par chacun des sous-problèmes de classification binaire. Ainsi, le i-ème élément de $\boldsymbol{y}^{(k)}$ n'est autre que le k-ème élément de $\boldsymbol{y}_i$, c'est à dire $[\boldsymbol{y}^{(k)}]_i = [\boldsymbol{y}_i]_k$. Désignons par $y_i^{(k)}$ cet élément.

En s'inspirant des récents travaux en apprentissage multi-tâches [Evgeniou et al., 2005, Szedmak and Shawe-Taylor, 2005], nous proposons le modèle suivant :

$$\psi^{(k)}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^{(k)} \, y_i^{(k)} \, \boldsymbol{x}_i^\top \boldsymbol{x}. \tag{6.7}$$

Cette formulation est différente de la formulation classique (`Th.Rep`), bien que la relation entre elles est souvent évidente avec une étiquette absorbée par le coefficient de pondération, notamment en

4. En principe, ℓ fonctions permettent de coder jusqu'à 2^ℓ classes différentes. Cependant, afin d'atteindre cette limite supérieure, une information *a priori* sur la distribution des échantillons est nécessaire pour définir la matrice de codage optimale. Malheureusement, une telle information n'est pas disponible en pratique. Cette étude est hors de portée du présent document. Le lecteur pourra se référer à [Voloshynovskiy et al., 2011] pour plus de détails.

classification binaire avec $y_i^{(k)} \in \{-1, +1\}$. Il s'agit alors d'un problème d'optimisation à $n\ell$ inconnues à déterminer. Dans la suite, nous proposons de réduire significativement le nombre d'inconnues en imposant une relation entre ces ℓ fonctions, c'est à dire entre les coefficients qui les caractérisent. Bien qu'une relation soit prescrite ici entre les différents classifieurs binaires, nous démontrons dans la section 6.4.2 qu'une relation semblable apparaît naturellement en classification multi-classes à moindres carrés par stratégie OvA.

Nous proposons dans la suite le modèle suivant :

$$\boldsymbol{\psi}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \, \boldsymbol{y}_i \, \boldsymbol{x}_i^\top \boldsymbol{x},$$

pour tout $\boldsymbol{x} \in \mathbb{X}$. Il est clair que le vecteur $\boldsymbol{\psi}(\boldsymbol{x})$ appartient à l'espace engendré par les vecteurs des étiquettes $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$. Une propriété intéressante de ce modèle est que les fonctions $\psi^{(k)}(\cdot)$ de la fonction vectorielle $\boldsymbol{\psi}(\cdot)$ partagent les mêmes coefficients, à savoir

$$\alpha_j = \alpha_j^{(k)},$$

pour tout $k = 1, 2, \ldots, \ell$. La seule différence entre les fonctions dans $\boldsymbol{\psi}(\cdot)$ est que ces coefficients sont pondérés par l'étiquette associée à $\boldsymbol{x}_i$ par le k-ème classifieur, par exemple ± 1. Cette astuce permet de réduire significativement le nombre d'inconnues à n, qui est exactement le nombre d'échantillons d'apprentissage. Il s'avère que c'est exactement comme avec un seul classifieur binaire, où le Théorème de Représentation impose l'équivalence entre le nombre de paramètres et le nombre d'échantillons disponibles.

La suite de cette section est consacrée à l'estimation de ces coefficients. Pour cela, nous adaptons plusieurs algorithmes classiques afin de fonctionner avec des sorties vectorielles, en remplaçant la norme vectorielle euclidienne $\|\boldsymbol{w}\|$ dans la formulation binaire par la norme matricielle de Frobenius $\|\boldsymbol{W}\|_F$ dans la nouvelle formulation, avec $\|\boldsymbol{W}\|_F^2 = \mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{W})$. A cette fin, nous écrivons

$$\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x},$$

avec $\boldsymbol{W}$ la matrice de taille $d \times \ell$ définie par

$$\boldsymbol{W}^\top = \sum_{i=1}^{n} \alpha_i \, \boldsymbol{y}_i \, \boldsymbol{x}_i^\top.$$

Un biais peut être considéré dans le modèle, selon l'expression $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b}$. L'impact de la présence d'un biais dans le modèle est étudié dans la section 6.4.1.d. Enfin, une fois les paramètres α_i estimés, la règle de décision est définie pour tout $\boldsymbol{x}$ selon

$$\underset{\boldsymbol{y}}{\mathrm{argmax}} \, \boldsymbol{y}^\top \boldsymbol{\psi}(\boldsymbol{x}). \tag{6.8}$$

Dans la suite, nous nous attardons sur trois méthodes largement reconnues en classification binaire :

- SVM : en se reposant sur l'idée de maximisation de marge à l'aide d'une fonction coût charnière, seule une petite fraction des échantillons d'apprentissage contribue à la solution. Le prix à payer est un problème d'optimisation de type programmation quadratique [Vapnik, 1998].
- LSM (ou LS-SVM) : en considérant une fonction coût quadratique avec des contraintes d'égalité, le problème d'optimisation correspond à un système d'équations linéaires, résolu par une procédure d'inversion matricielle [Suykens and Vandewalle, 1999b].
- RLSC (pour *regularized least-squares classification*) : dans [Rifkin, 2002], une fonction coût quadratique est considérée avec une régularisation de type Tikhonov, la solution étant obtenue par une inversion matricielle.

6.4.1.a Méthode SVMvect

L'algorithme SVM est basé sur une fonction coût charnière qui confère un caractère parcimonieux à la solution, au prix d'une optimisation par programmation quadratique. Revisité ici pour définir une version à sortie vectorielle désignée par SVMvect, le problème d'optimisation est :

$$\min_{\boldsymbol{W},\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{W}\|_F^2 + \eta \sum_{i=1}^{n} \xi_i, \qquad \text{sous contraintes : } \boldsymbol{y}_i^\top(\boldsymbol{W}^\top \boldsymbol{x}_i) \geq 1 - \xi_i, \text{et } \xi_i \geq 0 \text{ pour tout } i,$$

où η désigne le paramètre de régularisation et ξ_i est une variable d'écart autorisant à une fraction d'échantillons d'apprentissage à violer la règle de large marge. La solution de ce problème d'optimisation sous contrainte est donnée par le point-selle du Lagrangien. Les conditions d'optimalité conduisent au problème dual suivant :

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j \, \boldsymbol{y}_i^\top \boldsymbol{y}_j \, \boldsymbol{x}_i^\top \boldsymbol{x}_j, \qquad \text{sous contraintes : } 0 \leq \alpha_i \leq \eta, \text{ pour tout } i.$$

Sous forme matricielle, nous avons $\max_{\boldsymbol{\alpha}} \mathbf{1}_\ell^\top \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{K}'\boldsymbol{\alpha}$ sous contraintes $\mathbf{0}_\ell \leq \boldsymbol{\alpha} \leq \eta \mathbf{1}_\ell$, où $\boldsymbol{K}'$ désigne la matrice d'éléments $\boldsymbol{y}_i^\top \boldsymbol{y}_j \, \boldsymbol{x}_i^\top \boldsymbol{x}_j$ et $\mathbf{0}_\ell$ (resp. $\mathbf{1}_\ell$) le vecteur nul (resp. unité) à ℓ éléments.

Une extension au modèle avec biais $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b}$ est directe, où on obtient la contrainte supplémentaire $\boldsymbol{Y}\boldsymbol{\alpha} = \mathbf{0}_\ell$ au problème ci-dessus. Par analogie avec le cas binaire des SVM, la valeur optimale du biais est obtenue à partir de la moyenne, sur tous les vecteurs de support $\boldsymbol{x}_i$, selon

$$\boldsymbol{b}_i = \boldsymbol{y}_i - \sum_{j=1}^{n} \alpha_j \, \boldsymbol{y}_j \, \boldsymbol{x}_j^\top \boldsymbol{x}_i.$$

Notons que l'approche avec le modèle à biais est similaire à celui proposé récemment par Szedmak et Shawe-Taylor dans [Szedmak and Shawe-Taylor, 2005, Szedmak et al., 2006].

Dans tous les cas, il s'agit essentiellement du même problème de programmation quadratique que celui des SVM binaires. Ainsi les mêmes routines d'optimisation peuvent-elles être utilisées pour les deux types de tâches de classification, binaire avec les SVM classiques et multi-classes avec les SVMvect proposées.

6.4.1.b Méthode LSMvect

En considérant à présent une fonction coût quadratique plutôt que charnière, nous aboutissons au problème d'optimisation avec contraintes d'égalité suivant :

$$\min_{\boldsymbol{W},\boldsymbol{b},\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{W}\|_F^2 + \frac{\eta}{2} \sum_{i=1}^{n} \xi_i^2, \qquad \text{sous contraintes : } \boldsymbol{y}_i^\top(\boldsymbol{W}^\top \boldsymbol{x}_i) = 1 - \xi_i, \text{ pour tout } i,$$

où $\boldsymbol{y}_i^\top \boldsymbol{\psi}(\boldsymbol{x}_i) = \boldsymbol{y}_i^\top(\boldsymbol{W}^\top \boldsymbol{x}_i)$. Le Lagrangien associé est alors

$$\frac{1}{2}\|\boldsymbol{W}\|_F^2 + \frac{\eta}{2} \sum_{i=1}^{n} \xi_i^2 - \sum_{i=1}^{n} \alpha_i \big(\boldsymbol{y}_i^\top(\boldsymbol{W}^\top \boldsymbol{x}_i) - 1 + \xi_i\big),$$

où les α_i sont les multiplicateurs de Lagrange. Les conditions d'optimalité permettent d'écrire le système linéaire suivant :

$$\big(\boldsymbol{K}' + \eta^{-1}\mathbf{I}_{n,n}\big)\,\boldsymbol{\alpha} = \mathbf{1}_n,$$

où $\mathbf{I}_{n,n}$ est la matrice identité de taille $(n \times n)$.

Il est facile d'étendre cette formulation dans le cadre d'un modèle avec biais, selon $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{W}^\top \boldsymbol{x} + \boldsymbol{b}$. Dans ce cas, le système à résoudre devient

$$\begin{bmatrix} \mathbf{0}_{\ell,\ell} & \boldsymbol{Y} \\ \boldsymbol{Y}^\top & \boldsymbol{K}' + \eta^{-1}\mathbf{I}_{n,n} \end{bmatrix} \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{0}_\ell \\ \mathbf{1}_n \end{bmatrix}, \tag{6.9}$$

où $\mathbf{0}_{\ell,\ell}$ est la matrice nulle de taille $\ell \times \ell$.

La résolution de ce système nécessite l'inversion d'une matrice de taille $(n+\ell) \times (n+\ell)$, voire $(n \times n)$ pour le modèle sans biais. Puisque le nombre de classes est nettement plus petit que le nombre d'échantillons d'apprentissage, le coût calculatoire requis pour la méthode LSMvect reste comparable à celui du problème bi-classes correspondant, avec $\mathcal{O}(n^3)$. Ces grandeurs sont à mettre en perspective avec celles de l'algorithme LS-SVM multi-classes proposé par Suykens et Vanderwalle dans [Suykens and Vandewalle, 1999b], où une matrice est obtenue par une stratégie un-contre-tous qui conduit à la concaténation de ℓ sous-problèmes, chacun de la forme (6.9). L'algorithme résultant nécessite l'inversion d'une matrice de taille $(n\ell+\ell) \times (n\ell+\ell)$, avec une complexité de calcul cubique avec le nombre de classes, avec $\mathcal{O}((n\ell)^3)$.

6.4.1.c Méthode RLSvect

En considérant toujours la fonction coût quadratique, d'autres formes sont possibles. A l'opposé du produit scalaire $\boldsymbol{y}_i^\top \boldsymbol{\psi}(\boldsymbol{x}_i)$ investi dans la méthode LSMvect décrite ci-avant, nous utilisons dans la suite une fonction coût de type distance $\|\boldsymbol{\psi}(\boldsymbol{x}_i) - \boldsymbol{y}_i\|$. En s'inspirant de [Rifkin, 2002] pour la classification binaire où le biais n'est pas utilisé, nous considérons le problème d'optimisation avec régularisation de type Tikhonov :

$$\min_{\boldsymbol{W}} \sum_{j=1}^{n} \|\boldsymbol{W}^\top \boldsymbol{x}_i - \boldsymbol{y}_i\|^2 + \eta\, \mathcal{R}(\boldsymbol{W}),$$

où le terme de régularisation est donné par

$$\mathcal{R}(\boldsymbol{W}) = \sum_{i,j=1}^{n} \alpha_i \alpha_j\, \boldsymbol{y}_i^\top \boldsymbol{y}_j\, \boldsymbol{x}_i^\top \boldsymbol{x}_j. \tag{6.10}$$

Sous forme matricielle, nous avons le problème d'optimisation

$$\min_{\boldsymbol{\alpha}} \|\boldsymbol{Y}\|_F^2 - 2\,\boldsymbol{d}^\top \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \boldsymbol{G} \boldsymbol{\alpha} + \eta\, \boldsymbol{\alpha}^\top \boldsymbol{K}' \boldsymbol{\alpha},$$

avec $\boldsymbol{d}$ le vecteur d'éléments $\sum_i \boldsymbol{y}_i^\top \boldsymbol{y}_j\, \boldsymbol{x}_i^\top \boldsymbol{x}_j$, pour $j = 1, 2, \ldots, n$, et $\boldsymbol{G}$ la matrice d'éléments $\boldsymbol{y}_i^\top \boldsymbol{y}_j \sum_k (\boldsymbol{x}_i^\top \boldsymbol{x}_k)(\boldsymbol{x}_j^\top \boldsymbol{x}_k)$, pour $i, j = 1, 2, \ldots, n$. En annulant le gradient de la fonction coût ci-dessus par rapport à $\boldsymbol{\alpha}$, nous obtenons la solution finale

$$(\boldsymbol{G} + \eta\, \boldsymbol{K}')\, \boldsymbol{\alpha} = \boldsymbol{d}.$$

Il est clair que le problème de classification multi-classes est donné par un système linéaire de n équations à n inconnues. La complexité calculatoire est alors cubique en n, *i.e.*, le nombre d'échantillons d'apprentissage, mais demeure toutefois indépendante du nombre de classes.

Avec le terme de régularisation (6.10), la formulation ci-dessus est notée dans la suite RLSvect($\boldsymbol{W}$). Nous proposons aussi une autre formulation, où le terme de régularisation $\mathcal{R}(\boldsymbol{W})$ est remplacé par $\|\boldsymbol{\alpha}\|^2$. Dans ce cas, on obtient le problème d'optimisation suivant

$$(\boldsymbol{G} + \eta\, \mathbf{I})\, \boldsymbol{\alpha} = \boldsymbol{d}.$$

Cette dernière formulation est notée RLSvect($\boldsymbol{\alpha}$).

6.4.1.d Discussions sur le biais

Nous avons revisiter plusieurs algorithmes classiques de la littérature pour les transformer en algorithmes à sortie vectorielle. Nous précisons que les algorithmes binaires SVM et LSM sont souvent considérés avec des modèles à biais, alors que l'approche RLSC est généralement sans biais.

L'utilisation du biais en apprentissage statistique reste une question ouverte, aussi bien pour la classification binaire [Poggio et al., 2002] que pour la classification multi-classes [Abril et al., 2008].

Dans [Schölkopf and Smola, 2001, page 203], il est conseillé de ne pas utiliser une valeur optimale, mais de la modifier afin d'ajuster le nombre de faux positifs et faux négatifs. De nombreux auteurs éliminent complètement le terme de biais [Ertekin et al., 2011, Poggio et al., 2002]. Il est intéressant de noter que, dans ce cas, la contrainte linéaire $\sum_{i=1}^{n} \alpha_i \, \boldsymbol{y}_i = \boldsymbol{0}_\ell$ dans la méthode SVMvect est supprimée du problème d'optimisation. Pour les approches des moindres carrés, de nombreuses études motivent l'utilisation de la version sans biais [Poggio et al., 2002, Rifkin, 2002].

En pratique, il s'avère que le modèle sans biais surpasse très souvent celui avec biais. Ceci est illustré dans le CADRE 28, en comparant notre algorithme SVMvect avec celui proposé par Szedmak et Shawe-Taylor dans [Szedmak et al., 2006]. Un test statistique est également mené sur les performances de ce dernier, montrant qu'il dépend fortement du choix du codage de l'étiquette, mais toujours avec des performances moins bonnes que notre approche.

6.4.2 Développement de classifieurs multi-classes à moindres carrés

Dans cette partie, nous revenons au modèle conventionnel issu du Théorème de Représentation (`Th.Rep`), en conjuguant d'une part la stratégie OvA pour construire une machine multi-classes, et d'autre part la solution de chaque sous-problème binaire optimale au sens des moindres carrés. Chacun des ℓ sous-problèmes binaires définit une fonction de décision de la forme

$$\psi^{(k)}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i^{(k)} \, \kappa(\boldsymbol{x}_i, \boldsymbol{x}), \tag{6.11}$$

pour $k = 1, 2, \ldots, \ell$, avec $\kappa(\boldsymbol{x}_i, \boldsymbol{x}) = \boldsymbol{x}_i^\top \boldsymbol{x}$ pour le noyau linéaire. En d'autres termes, $\psi^{(k)}(\boldsymbol{x}) = \boldsymbol{\alpha}^{(k)\top} \boldsymbol{\kappa}(\boldsymbol{x})$ où $\boldsymbol{\alpha}^{(k)}$ désigne le vecteur associé au k-ème classifieur binaire, ayant les éléments $\alpha_i^{(k)}$ pour $i = 1, 2, \ldots, n$. Comme démontré dans la section 2.1.3, la solution au sens des moindres carrés est alors $\boldsymbol{\alpha}^{(k)} = (\boldsymbol{K} + \eta \mathbf{I})^{-1} \boldsymbol{y}^{(k)}$, c'est à dire

$$\psi^{(k)}(\boldsymbol{x}) = \boldsymbol{y}^{(k)\top} (\boldsymbol{K} + \eta \mathbf{I})^{-1} \boldsymbol{\kappa}(\boldsymbol{x}),$$

où $\boldsymbol{y}^{(k)}$ désigne l'étiquette vectorielle associée au k-ème sous-problème binaire. Voir le CADRE 25 et la section 6.4.3 pour une étude sur le codage des étiquettes vectorielles et son impact sur la solution.

Il est évident que le terme $(\boldsymbol{K} + \eta \mathbf{I})^{-1} \boldsymbol{\kappa}(\boldsymbol{x})$ n'a besoin d'être évalué qu'une seule fois pour tous les ℓ classifieurs binaires, et ainsi ne nécessitant qu'une seule inversion matricielle de complexité calculatoire $\mathcal{O}(n^3)$. La seule différence entre les fonctions de décision des classifieurs binaires réside dans les étiquettes vectorielles assignées à chaque sous-problème, c'est à dire $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, \ldots, \boldsymbol{y}^{(\ell)}$. De plus, l'algèbre linéaire permet de simplifier encore plus le problème, puisque nous pouvons écrire sous forme matricielle

$$(\boldsymbol{K} + \eta \mathbf{I})[\boldsymbol{\alpha}^{(1)} \; \boldsymbol{\alpha}^{(2)} \; \cdots \; \boldsymbol{\alpha}^{(\ell)}] = \boldsymbol{Y}^\top, \tag{6.12}$$

où $\boldsymbol{Y}^\top = [\boldsymbol{y}^{(1)} \;\; \boldsymbol{y}^{(2)} \;\; \cdots \;\; \boldsymbol{y}^{(\ell)}]$. Par conséquent, nous avons

$$\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{Y} (\boldsymbol{K} + \eta \mathbf{I})^{-1} \boldsymbol{\kappa}(\boldsymbol{x}), \tag{6.13}$$

avec $\boldsymbol{\psi}(\cdot) = [\psi^{(1)}(\cdot) \;\; \psi^{(2)}(\cdot) \;\; \cdots \;\; \psi^{(\ell)}(\cdot)]^\top$. Enfin, l'étiquette vectorielle associée à tout échantillon $\boldsymbol{x}$ est

$$\operatorname*{argmax}_{\boldsymbol{y} \in \mathbb{Y}} \boldsymbol{y}^\top \boldsymbol{Y} (\boldsymbol{K} + \eta \mathbf{I})^{-1} \boldsymbol{\kappa}(\boldsymbol{x}). \tag{6.14}$$

Cette approche, notée 1LSM dans la suite, n'est autre que la mise en œuvre ingénieuse de la stratégie OvA avec le problème de moindres carrés. Elle montre qu'il n'est pas nécessaire d'implémenter explicitement ℓ classifieurs binaires, mais l'astuce proposée permet d'opérer une seule inversion de matrice de taille $n \times n$. Ainsi la complexité calculatoire est-elle essentiellement indépendante du nombre de classes en compétition.

6.4.2.a 1LSM comme une machine unique

Le théorème suivant permet de faire le lien entre l'approche machine unique et la méthode 1LSM. Pour cela, nous considérons une formulation linéaire de l'expression (6.13), selon $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{A}^\top \boldsymbol{x}$, où $\boldsymbol{A} = [\boldsymbol{\alpha}^{(1)}\ \boldsymbol{\alpha}^{(2)}\ \cdots\ \boldsymbol{\alpha}^{(\ell)}]$ avec $\boldsymbol{\alpha}^{(k)}$ associé au k-ème classifieur binaire. Une extension au cas non linéaire est directe en remplaçant $\boldsymbol{X}^\top\boldsymbol{X}$ par $\boldsymbol{K}$, *i.e.*, la matrice de Gram du noyau utilisé.

Théorème 8 (Multi-classes : 1LSM comme une machine unique)**.**
Le problème multi-classes 1LSM est équivalent au problème d'optimisation

$$\min_{\boldsymbol{A}} \sum_{i=1}^{n} \|\boldsymbol{A}^\top \boldsymbol{x}_i - \boldsymbol{y}_i\|^2 + \eta \|\boldsymbol{A}\|_F^2,$$

dont la solution $\boldsymbol{A}^\top = \boldsymbol{Y}(\boldsymbol{X}^\top\boldsymbol{X} + \eta\mathbf{I})^{-1}\boldsymbol{X}^\top$ *définit la fonction de décision* $\boldsymbol{\psi}(\boldsymbol{x}) = \boldsymbol{A}^\top\boldsymbol{x}$.

Démonstration. Pour démontrer ce théorème, il suffit de dériver par rapport à $\boldsymbol{A}^\top$ la fonction coût en question, pour obtenir $\sum_{i=1}^{n}\left(\boldsymbol{A}^\top\boldsymbol{x}_i^\top\boldsymbol{x}_i - \boldsymbol{y}_i\boldsymbol{x}_i^\top\right) + \eta\boldsymbol{A}^\top$. En annulant cette expression au minimum, nous avons le système d'équations linéaires

$$\boldsymbol{A}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \eta\mathbf{I}) = \boldsymbol{Y}\boldsymbol{X}^\top.$$

L'identité $\boldsymbol{X}^\top(\boldsymbol{X}\boldsymbol{X}^\top + \eta\mathbf{I})^{-1} = (\boldsymbol{X}^\top\boldsymbol{X} + \eta\mathbf{I})^{-1}\boldsymbol{X}^\top$ permet de déterminer le résultat final $\boldsymbol{A}^\top = \boldsymbol{Y}(\boldsymbol{X}^\top\boldsymbol{X} + \eta\mathbf{I})^{-1}\boldsymbol{X}^\top$. ■

6.4.2.b Relations entre les coefficients

La stratégie OvA nécessite en général l'estimation de $n \times \ell$ inconnues, comme illustré dans l'expression (6.11) avec les $\alpha_i^{(k)}$, pour $i = 1, 2, \ldots, n$ et $k = 1, 2, \ldots, \ell$. En conjuguant cette stratégie avec des sous-problèmes aux moindres carrés, nous montrons au travers de l'expression (6.12) qu'il s'agit de la résolution d'un système de n équations. En conséquence, il existe une relation entre ces inconnues, comme établi par le théorème suivant.

Théorème 9 (Relations entre les coefficients en classification multi-classes)**.**
En stratégie OvA avec des sous-problèmes binaires à moindres carrés (c'est à dire LSM, et en particulier 1LSM), les coefficients associés aux sous-problèmes binaires sont reliés entre eux avec la relation suivante :

$$\boldsymbol{\alpha}^{(k)^\top}\boldsymbol{y}^{(k')} = \boldsymbol{\alpha}^{(k')^\top}\boldsymbol{y}^{(k)},$$

pour toutes les paires $k, k' = 1, 2, \ldots, \ell$.

Démonstration. Pour démontrer ce résultat, il suffit de reprendre l'expression (6.12), qui permet d'écrire

$$\boldsymbol{\alpha}^{(k)^\top}\boldsymbol{y}^{(k')} = \boldsymbol{y}^{(k)^\top}(\boldsymbol{K} + \eta\mathbf{I})^{-1}\boldsymbol{y}^{(k')} = \boldsymbol{y}^{(k)^\top}\boldsymbol{\alpha}^{(k')}.$$

■

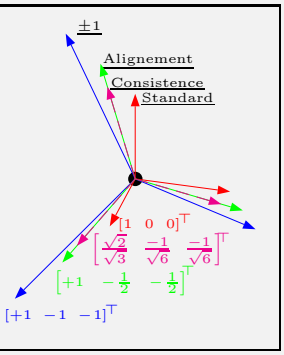
Ce théorème conduit à d'autres résultats, et en particulier $[\boldsymbol{\alpha}^{(1)}\ \cdots\ \boldsymbol{\alpha}^{(\ell)}]^\top\boldsymbol{y}^{(k)} = [\boldsymbol{y}^{(1)}\ \cdots\ \boldsymbol{y}^{(\ell)}]^\top\boldsymbol{\alpha}^{(k)}$. Ces résultats permettent d'exprimer des liens entre les sous-problèmes de classification binaire et peuvent être illustrés sur des types spécifiques d'étiquettes vectorielles. Voir le Cadre 25.

Soit $\mathcal{C}^{(k)}$ la k-ème classe. Dans la suite, nous étudions le codage standard (ou indicatrices), avec $[\boldsymbol{y}^{(k)}]_j = 1$ si $\boldsymbol{x}_j \in \mathcal{C}^{(k)}$; et 0 sinon. Nous avons dans ce cas :

$$\sum_{\boldsymbol{x}_j \in \mathcal{C}^{(k')}} \alpha_j^{(k)} = \sum_{\boldsymbol{x}_i \in \mathcal{C}^{(k)}} \alpha_i^{(k')},$$

pour tout $k, k' = 1, 2, \ldots, \ell$. En considérant le k-ème sous-problème de classification binaire, l'équation ci-dessus permet de relier la classe $\mathcal{C}^{(k)}$ aux autres classes, selon

$$\sum_{k' \neq k} \sum_{\boldsymbol{x}_j \in \mathcal{C}^{(k')}} \alpha_j^{(k)} = \sum_{\boldsymbol{x}_i \in \mathcal{C}^{(k)}} \sum_{k' \neq k} \alpha_i^{(k')},$$

Codage	$[\boldsymbol{y}^{(k)}]_i = [\boldsymbol{y}_i]_k$	$\boldsymbol{y}_i^\top \boldsymbol{y}_j$	Illustration
± 1	$\begin{cases} +1 & \text{si } \boldsymbol{x}_i \in \mathcal{C}^{(k)}; \\ -1 & \text{ailleurs} \end{cases}$	$\begin{cases} \ell & \text{si } \boldsymbol{y}_i = \boldsymbol{y}_j; \\ -1 & \text{ailleurs} \end{cases}$	
Standard (ou indicatrices)	$\begin{cases} 1 & \text{si } \boldsymbol{x}_i \in \mathcal{C}^{(k)}; \\ 0 & \text{ailleurs} \end{cases}$	$\begin{cases} 1 & \text{si } \boldsymbol{y}_i = \boldsymbol{y}_j; \\ 0 & \text{ailleurs} \end{cases}$	
Codage par alignement	$\begin{cases} \sqrt{\frac{\ell-1}{\ell}} & \text{si } \boldsymbol{x}_i \in \mathcal{C}^{(k)}; \\ \frac{-1}{\sqrt{\ell(\ell-1)}} & \text{ailleurs} \end{cases}$	$\begin{cases} 1 & \text{si } \boldsymbol{y}_i = \boldsymbol{y}_j; \\ \frac{-1}{\ell-1} & \text{ailleurs} \end{cases}$	
Codage par principe inductif	$\begin{cases} 1 & \text{si } \boldsymbol{x}_i \in \mathcal{C}^{(k)}; \\ \frac{-1}{\ell-1} & \text{ailleurs} \end{cases}$	$\begin{cases} \frac{\ell}{\ell-1} & \text{si } \boldsymbol{y}_i = \boldsymbol{y}_j; \\ \frac{-\ell}{(\ell-1)^2} & \text{ailleurs} \end{cases}$	

[CADRE 25] **Expressions et illustration des étiquettes vectorielles.**
Expressions des étiquettes vectorielles de codages bien connus dans la littérature, et illustration pour $\ell = 3$ classes. Nous présentons également le produit scalaire de toute paire d'étiquettes vectorielles pour un codage donné, puisqu'il s'agit d'une information suffisante pour la décision. En effet, ceci est illustré dans l'estimation des coefficients et la règle de décision, avec $\operatorname{argmax}_{\boldsymbol{y} \in \mathbb{Y}} \boldsymbol{y}^\top \boldsymbol{Y} (\boldsymbol{K} + \eta \mathbf{I})^{-1} \boldsymbol{\kappa}(\boldsymbol{x})$, où seuls les produits scalaires entre les étiquettes sont nécessaires.

où la double somme à gauche n'est autre que $\sum_{\boldsymbol{x}_i \notin \mathcal{C}^{(k)}} \alpha_j^{(k)}$. Cette identité se prête à l'interprétation suivante : la somme des contributions (en terme des coefficients α) au k-ème classifieur des échantillons appartenant à toutes les autres classes est égale à la somme des contributions à tous les autres classifieurs des échantillons appartenant à la k-ème classe.

A notre connaissance, les relations démontrées ci-dessus n'ont pas été établies avant. Ces résultats liant les paramètres aux étiquettes sont à comparer avec la contrainte d'égalité dans les machines de type SVM et LSM, comme indiqué respectivement aux sections 6.4.1.a et 6.4.1.b pour SVMvect et LSMvect. En effet, dans le cas de la bien connue classification binaire avec le codage ± 1 des étiquettes, nous avons $\sum_{i=1}^n \alpha_i \, y_i = 0$, et par conséquent pour toute classe $\mathcal{C}^{(k)}$

$$\sum_{\boldsymbol{x}_j \in \mathcal{C}^{(k)}} \beta_j = \sum_{\boldsymbol{x}_i \notin \mathcal{C}^{(k)}} \beta_i .$$

6.4.3 Analyse des étiquettes en classification multi-classes

Soit $\mathbb{Y}$ l'ensemble des ℓ étiquettes vectorielles, c'est à dire le livre-étiquette par analogie au livre-code en codage de l'information. A chaque échantillon $\boldsymbol{x}_i$ est associé une et une seule étiquette vectorielle $\boldsymbol{y}_i \in \mathbb{Y}$ telle que $\boldsymbol{y}_i \neq \boldsymbol{y}_j$ si et seulement si $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ proviennent de deux classes différentes. Divers codages ont été proposés pour étendre l'étiquetage binaire au cas multi-classes, les plus connus sont résumés dans le CADRE 25 et rappelés dans la suite :

- Le codage ± 1 est une généralisation directe du cas binaire $\{-1, +1\}$. Il définit l'étiquette vectorielle associée à la k-ème classe par un vecteur aux éléments -1 à l'exception du k-ème élément qui vaut $+1$. Ce codage est souvent préconisé dans la littérature [Dietterich and Bakiri, 1995, Allwein et al., 2001] ;
- Le codage standard, également connu par indicatrices ou encore variables muettes par les statisticiens, définit les étiquettes vectorielles par les colonnes de la matrice identité. Il est utilisé en réseaux de neurones artificiels, et en particulier les réseaux de Boltzmann [Bishop, 1995] et les machines extrêmes [Huang et al., 2012] ;
- Le codage par alignement est construit par optimalité au sens du critère d'alignement noyau-cible [Guermeur, 2008]. Il s'agit aussi d'une simple généralisation au cas multi-classes de l'étiquette binaire $\{-1, +1\}$, choisie de sorte que la norme de l'étiquette vectorielle $\boldsymbol{y}_i$ soit égale à 1, et que la moyenne de ses composantes soit nulle. Ce codage a été systématiquement appliqué dans de nombreuses machines multi-classes [Noumir et al., 2011b, Szedmak et al., 2006] ;

– Le codage par principe inductif est basé sur le principe inductif dans les SVM multi-classes [Lee et al., 2004]. Cette propriété théorique est montrée sous condition d'avoir un ensemble d'étiquettes vectorielles avec une moyenne nulle de leurs composantes.

Tous ces codages partagent la même forme. En effet, pour un codage donné, tous les éléments d'une étiquette vectorielle sont identiques sauf un seul élément qui indique la classe d'appartenance. Nous désignons dans la suite cette catégorie de codages par « un-par-classe ». Avec des valeurs arbitraires, cette définition généralise la notion de matrice à classe-symétrique définie dans [Rifkin and Klautau, 2004] pour le codage ± 1. Le résultat suivant montre le lien entre ces différents codages.

Lemme 10 (Multi-classes : codage un-par-classe)**.**
Les codages un-par-classe sont liés par une transformation affine de la forme $a\boldsymbol{y}_i + b\mathbf{1}_\ell$ pour des valeurs arbitraires a et b, où $\boldsymbol{y}_i$ provient d'un codage un-par-classe donné et $\mathbf{1}_\ell$ est le vecteur unité de ℓ éléments.

La preuve est simple, en générant tout codage un-par-classe à partir du codage standard, et vice-versa. Par conséquent, et en utilisant implicitement le codage standard, nous pouvons définir l'un des codages un-par-classe avec des éléments $\boldsymbol{y}_i$ tels que le k-ème élément $[\boldsymbol{y}_i]_k$ vaut $a+b$ si $\boldsymbol{x}_i \in \mathcal{C}^{(k)}$; et b ailleurs.

Pour un problème de classification à ℓ classes, tout codage un-par-classe définit ℓ étiquettes vectorielles distinctes dans un espace vectoriel de dimension ℓ. Il est également envisageable de définir un codage de ℓ étiquettes vectorielles dans un espace de dimension $\ell - 1$. Il suffit dans ce cas d'imposer une corrélation égale entre les différentes étiquettes vectorielles, et de valeur minimale. Comme montré dans [Szedmak et al., 2006], il s'agit d'une résolution d'un problème de valeurs propres / vecteurs propres, nécessitant ainsi un coût calculatoire supplémentaire. La proposition suivante montre l'équivalence entre ce codage, et les codages par alignement et par principe inductif.

Proposition 11 (Multi-classes : équivalence entre les codages)**.**
Le codage par corrélation minimale, le codage par alignement et le codage par principe inductif sont équivalents.

Démonstration. D'une part, nous avons l'équivalence entre le codage par corrélation minimale et le codage par alignement. En effet, le codage par corrélation minimale définit ℓ vecteurs dans un espace de dimension $\ell - 1$, le produit scalaire entre toute paire de vecteurs étant égal à $-1/(\ell - 1)$ (voir [Szedmak et al., 2006, Proposition 2]). Il s'avère que c'est exactement la valeur du produit scalaire entre les étiquettes vectorielles du codage par alignement (voir la troisième colonne du Cadre 25). D'autre part, le codage par principe inductif et le codage par alignement sont identiques à une constante multiplicative, ici $\sqrt{(\ell-1)/\ell}$ (voir la figure dans le Cadre 25). Ainsi ces codages donnent-ils des résultats comparables en classification multi-classes. ■

Dans la suite, nous étudions les machines 1LSM, c'est à dire dans le cas du problème d'apprentissage (6.13)-(6.14). Une contribution majeure est le théorème suivant, qui démontre que tous les codages un-par-classe sont équivalents. En combinant ce résultat avec la proposition 11, il est facile de faire le lien avec le codage par corrélation minimale comme donné par le corollaire 13.

Théorème 12 (Multi-classes : codages un-par-classe pour le 1LSM)**.**
Tous les codages un-par-classe sont équivalents dans le cadre du problème multi-classes 1LSM.

Démonstration. Considérons la transformation affine donnée par le lemme 10, d'une part sur toute étiquette vectorielle $\boldsymbol{y}^\top$ et d'autre part sur celles de la matrice $\boldsymbol{Y}$, pour obtenir respectivement $a\boldsymbol{y} + b\mathbf{1}_\ell$ et $a\boldsymbol{Y} + b\mathbf{1}_\ell\mathbf{1}_\ell^\top$. Nous montrons dans la suite que la règle de décision (6.14) est invariante par rapport à cette transformation :

$$\begin{aligned}
\operatorname*{argmax}_{\boldsymbol{y}\in\mathbb{Y}} (a\boldsymbol{y} + b\mathbf{1}_\ell)^\top (a\boldsymbol{Y} + b\mathbf{1}_\ell\mathbf{1}_\ell^\top)(\boldsymbol{K} + \eta\mathbf{I})^{-1}\boldsymbol{\kappa}(\boldsymbol{x}) &= \operatorname*{argmax}_{\boldsymbol{y}\in\mathbb{Y}} \left(a^2\boldsymbol{y}^\top\boldsymbol{Y} + ab\boldsymbol{y}^\top\mathbf{1}_\ell\mathbf{1}_\ell^\top\right)(\boldsymbol{K} + \eta\mathbf{I})^{-1}\boldsymbol{\kappa}(\boldsymbol{x}) \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in\mathbb{Y}} a^2\, \boldsymbol{y}^\top\boldsymbol{Y}(\boldsymbol{K} + \eta\mathbf{I})^{-1}\boldsymbol{\kappa}(\boldsymbol{x}) \\
&= \operatorname*{argmax}_{\boldsymbol{y}\in\mathbb{Y}} \boldsymbol{y}^\top\boldsymbol{Y}(\boldsymbol{K} + \eta\mathbf{I})^{-1}\boldsymbol{\kappa}(\boldsymbol{x}),
\end{aligned}$$

[CADRE 26] Etude comparative des méthodes de classification multi-classes.

Nous avons considéré des jeux de données bien connus, et résumés dans le tableau suivant :

	n échantillons	ℓ classes	d attributs
iris	150	3	4
wine	178	3	13
glass	214	6	13
vowel	528	11	10
yeast	1 484	10	8
letters	jusqu'à 2 000	26	16
USPS	1 000	10	jusqu'à 64

La configuration, similaire à [Hsu and Lin, 2002], est résumée comme suit. Les échantillons ont été normalisés à l'échelle $[-1, 1]$, et le noyau gaussien utilisé. Une validation croisée à dix partitions a été utilisée, avec les paramètres η et σ optimisés par une recherche sur la grille $\{2^{-4}; 2^{-3}; \cdots; 2^{3}; 2^{4}\}$. Il s'agit d'une optimisation conjointe, à l'opposé de l'optimisation séparée des 2 paramètres décrite dans [Rifkin and Klautau, 2004].

Le CADRE 28 présente les taux d'erreur selon 10 simulations de Monte Carlo. Pour une étude comparative entre les différentes méthodes, les mêmes partitions de la validation croisée ont été appliquées. Le temps de calcul [a] a été estimé sur la validation croisée avec la recherche des paramètres optimaux, comme illustré dans le tableau en bas. Il est évident que les méthodes proposées renforcent fortement la vitesse de la machine multi-classes sans en compromettre les performances.

Pour les jeux de données images, nous avons considéré la même configuration que dans [Rasmussen and Williams, 2006], avec des intensités de pixels normalisées à l'échelle $[-1, 1]$. Comme indiqué dans [Huang et al., 2012], ces jeux de données exigent des ordinateurs dédiés. Afin de mener une étude comparative, nous avons considéré $n = 1\,000$ et $n = 2\,000$ images, où les images *USPS* ont été sous-échantillonnées à $d = 8 \times 8$. Le tableau ci-dessous montre l'erreur de classification et le temps de calcul. Il est facile de voir que les stratégies OvA et OvO sont inappropriées à ces jeux de données à grande taille. L'algorithme SVM de Szedmak *et coll.* [Szedmak et al., 2006] est plus approprié pour cette tâche, toutefois ses performances sont surpassées par toutes les méthodes que nous proposons.

Erreur de classification et temps de calcul (*hh:mm:ss*).

		USPS		*letters*
	$n =$	1 000	2 000	2 000
	$d =$	8×8	8×8	16
Méthodes proposées	1LSM	6,50 *(7:42)*	4,59 *(46:32)*	10,69 *(25:27)*
	LSMvect	8,41 *(7:54)*	6,19 *(46:02)*	14,08 *(29:08)*
	SVMvect	8,80 *(10:02)*	6,54 *(41:23)*	14,08 *(27:15)*
	LS+OvA	6,50 *(1:06:36)*	4,59 *(7:04:15)*	10,69 *(10:29:03)*
	LS+OvO	6,50 *(4:56:37)*	4,59 *(31:06:11)*	10,69 *(133:10:22)*
	Szedmak *et coll.*	16,61 *(9:42)*	12,69 *(41:52)*	15,04 *(28:23)*

Temps de calcul estimé (en *hh:mm:ss*).

	machines à moindres carrés			machines à sortie vectorielle		
	LS+OvA	LS+OvO	**1LSM**	**LSMvect**	**RLSvect**	**SVMvect**
iris	*1:42*	*1:44*	*38*	*43*	*1:06*	*1:00*
wine	*3:41*	*3:39*	*1:14*	*1:19*	*2:31*	*9:32*
glass	*7:10*	*18:22*	*1:22*	*1:27*	*2:45*	*12:12*
vowel	*1:48:28*	*8:56:00*	*10:32*	*11:48*	*40:51*	*45:35*
yeast	*8:26:45*	*37:36:24*	*54:29*	*1:07:09*	*3:21:06*	*2:40:40*
	voir la sections :		6.4.2	6.4.1.a	6.4.1.b	6.4.1.c

a. Le temps de calcul a été estimé sur un ordinateur portable Macbook Pro Intel Core 2 Duo (2.53 GHz, 4 GB RAM) équipé d'un Matlab 64-bits.

où la première égalité résulte de l'élimination des termes indépendants de $\boldsymbol{y}$, et la seconde égalité est due au fait que, pour un codage donné, le produit scalaire $\boldsymbol{y}^\top \mathbf{1}_\ell$ est constant pour tout $\boldsymbol{y} \in \mathbb{Y}$. ■

Corollaire 13 (Multi-classes : codage par corrélation minimale pour le 1LSM)**.**
Le codage par corrélation minimale est équivalent aux codages un-par-classe pour le 1LSM.

Au-delà de l'équivalence entre les différents codages, le théorème 12 établit également une équivalence entre la règle de décision (6.8) et la règle de la valeur maximale utilisée dans la stratégie classique OvA. La preuve est simple pour le codage standard, puisque $\operatorname{argmax}_{\boldsymbol{y} \in \mathbb{Y}} \boldsymbol{y}^\top \boldsymbol{\psi}(\boldsymbol{x})$ implique $\operatorname{argmax}_{1 \leq k \leq \ell} \psi^{(k)}(\boldsymbol{x})$. L'extension de ce résultat à d'autres codages est assurée par le théorème 12.

6.4.4 Analyse comparative en classification multi-classes

Nous avons mené une étude comparative sur la pertinence de la simplification des méthodes de classification multi-classes comme nous préconisons dans ce chapitre. Nous utilisons plusieurs jeux

[Cadre 27] Test statistique sur les performances des méthodes de classification.

Plusieurs tests statistiques ont été développés dans la littérature pour comparer deux algorithmes de classification. Dans [Dietterich, 1998], Dietterich propose le test « 5×2 cv test t » qui correspond à un test de Student basé sur 5 itérations de validations croisées à 2 partitions. Ce test a une faible erreur du premier type et admet une puissance plus importante que le test de McNemar et le test t bilatéral classique.

Dans [Alpaydin, 1999], cette étude est étendue au test de Fisher « 5×2 cv », que nous résumons dans la suite. Cinq répétitions de validations croisées à 2 partitions sont effectuées, où chacune des répétitions partage les données en deux moitiés : une moitié est utilisée pour l'apprentissage [a] et l'autre moitié pour le test, et vice-versa. Soit $p_i^{(j)}$ la différence dans les taux d'erreur des deux classifieurs dans la partition j de la i-ème itération, pour $j = 1, 2$ et $i = 1, 2, \ldots, 5$. En notant la moyenne par $\overline{p}_i = \frac{1}{2}(p_i^{(1)} + p_i^{(2)})$ et la variance par $s_i^2 = (p_i^{(1)} - \overline{p}_i)^2 + (p_i^{(2)} - \overline{p}_i)^2$, la statistique

$$f = \frac{\sum_{i=1}^{5} \sum_{j=1}^{2} \left(p_i^{(j)}\right)^2}{2 \sum_{i=1}^{5} s_i^2}$$

suit une distribution de Fisher avec 10 et 5 degrés de liberté. L'hypothèse nulle est que les deux algorithmes ont les mêmes performances. Cette hypothèse est rejetée si $f > 4.74$, avec un seuil significatif au niveau de 5%.

Nous avons analysé l'influence du codage de l'étiquette sur les performances en comparant toutes les méthodes proposées dans le présent chapitre. En considérant l'application du test de Fisher, nous avons constaté qu'il n'y a pas de différences significatives (au niveau de 5%).

Nous avons également comparé avec la méthode SVM multi-classes proposée par Szedmak *et coll.*, où un modèle avec biais est considéré [Szedmak et al., 2006]. Nous avons constaté que les performances de leur algorithme dépendent fortement du codage de l'étiquette, une propriété confirmée par le test de Fisher. Les différentes expérimentations montrent que leur méthode possède, dans presque tous les cas, de moins bonnes performances que toutes les machines proposées dans le présent document.

a. Comme préconisé dans [Cantú-Paz and Kamath, 2005], l'ensemble d'apprentissage est également utilisé pour l'estimation des paramètres optimaux. A cette fin, nous avons utilisé une validation croisée de dix partition sur l'ensemble d'apprentissage, avec les paramètres η et σ optimisés par une recherche sur la grille $\{2^{-4}; 2^{-3}; \cdots; 2^3; 2^4\}$. L'ensemble test n'est pas utilisé à l'étape d'apprentissage.

de données bien connus dans la littérature et disponibles à partir du référentiel UCI : *iris*, *wine*, *glass*, *vowel*[5] et *yeast*. Nous avons également examiné deux jeux de données à grande échelle, c'est à dire grande taille et/ou dimension et/ou nombre de classes : *letters* et *USPS* [Hull, 1994].

Afin de fournir une étude comparable avec des travaux antérieurs, nous avons considéré une configuration similaire à celles étudiées dans [Hsu and Lin, 2002, Rifkin and Klautau, 2004]. Le Cadre 26 décrit les expérimentations. Pour résumer, les échantillons ont été normalisés à l'échelle $[-1, 1]$ et les paramètres η et σ ont été optimisés par une recherche sur la grille $\{2^{-4}; 2^{-3}; \cdots; 2^3; 2^4\}$. Les taux d'erreur sont obtenus à partir de 10 simulations de Monte Carlo. Les taux d'erreur sont détaillés dans le Cadre 28.

A titre indicatif, nous interprétons le cas où nos méthodes sont surpassées par une autre méthode. C'est le cas avec le jeu de données *vowel*, où LS+OvO produit un taux d'erreur moyen de 0,58, avec les meilleure et pire performances 0,37 et 0,93, respectivement. En terme de performances, toutes les méthodes que nous proposons conduisent à un taux d'erreur très proche, égal à 0,60 ou 0,62. De plus, les meilleures et pires performances sont exactement celles du LS+OvO. La grande différence réside dans la complexité de calcul. Le Cadre 28 (tableau en bas) montre que la méthode LS+OvO a nécessité presque 9 heures, alors que les temps de calcul de nos méthodes ont été entre 10 minutes (pour 1LSM) et 45 minutes (pour SVMvect).

L'analyse statistique menée et décrite dans le Cadre 27, a montré qu'il n'y a pas de différences significatives (au niveau de 5%) entre les méthodes de classification proposées, ni une influence du codage des étiquettes sur les performances. Il est évident que les différentes approches proposées

5. Dans [Huang et al., 2012], le jeu de données *vowel* comprend 528 + 462 instances pour l'apprentissage+test, par opposition à l'ensemble de données traité dans plusieurs études [Hsu and Lin, 2002, Rifkin and Klautau, 2004], y compris le présent document.

renforcent fortement la vitesse en classification multi-classes, avec d'excellentes performances en terme de classification.

6.5 Conclusion et perspectives

Historiquement, les méthodes à noyaux ont pris leur essor grâce à leurs performances en classification. Dans ce chapitre, nous avons cherché à étudier le problème de classification, en vue d'une simplification des approches classiques. Les contributions détaillées dans ce chapitre ont eu pour cible :

- d'une part la classification mono-classe, avec une méthodologie permettant de développer des représentations parcimonieuses, avec des algorithmes en ligne, ainsi que des résultats théoriques sur la probabilité de fausse alarme ;
- d'autre part la classification multi-classes, avec une méthodologie permettant de définir des classifieurs avec une complexité calculatoire indépendante du nombre de classes en compétition. Une étude détaillée du codage des étiquetages des classes est aussi présentée.

Les travaux entamés sur la classification mono-classe en ligne, et la théorie qui les accompagne, ouvrent la voie à une nouvelle classe de méthodes à noyaux pour la détection séquentielle. Nous sommes convaincus qu'il reste beaucoup à faire, notamment en s'appropriant la vaste littérature sur la détection séquentielle de rupture et la théorie qui l'accompagne [Basseville and Nikiforov, 1993]. Précisons que la pertinence de l'approche étudiée dans ce chapitre a été récemment démontrée sur des domaines d'application divers, dont les attaques cyber-physiques ou encore la détection de changement dans des séquences vidéo.

Dans le cadre de la classification multi-classes, les travaux présentés dans ce chapitre ouvrent la voie à de nouvelles questions, notamment sur l'utilisation d'un noyau sur les étiquettes à l'instar du noyau défini sur les échantillons. En outre, la complexité calculatoire des classifieurs peut être davantage réduite, notamment avec le principe de parcimonie décrit en détail dans le chapitre 4 et présenté dans le présent chapitre pour la classification mono-classe. Les perspectives à long-terme de nos travaux s'inscrivent dans l'espace de représentation optimal et l'apprentissage de caractéristiques pertinentes pour une tâche de classification donnée. La simplicité des méthodes présentées devra ouvrir la voie à des avancées inédites. Le problème de la classification multi-classes demeure ouvert, à la hauteur des défis envisagés.

[Cadre 28] **Taux d'erreur des méthodes de classification multi-classes.**
Le tableau suivant présente les taux d'erreur de classification de diverses méthodes, en utilisant la configuration présentée dans le Cadre 26. Il s'agit de 10 simulations de Monte-Carlo, présentées sous le format $\texttt{moyenne}|^{\texttt{pire}}_{\texttt{meilleur}}$. Les résultats des 6 dernières lignes sont empruntés de [Hsu and Lin, 2002, Huang et al., 2012].

			iris	***wine***	***glass***	***vowel***	***yeast***
Approches proposées dans ce manuscrit	1LSM		$2{,}80\|^{3{,}33}_{2{,}00}$	$0{,}33\|^{1{,}11}_{0{,}00}$	$27{,}39\|^{28{,}7}_{24{,}9}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$38{,}91\|^{39{,}4}_{38{,}1}$
	LSMvect	étiquette ± 1	$2{,}80\|^{3{,}33}_{2{,}66}$	$2{,}06\|^{2{,}77}_{1{,}11}$	$28{,}40\|^{31{,}6}_{25{,}4}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$40{,}78\|^{41{,}9}_{39{,}8}$
		indicatrices	$2{,}80\|^{3{,}33}_{2{,}00}$	$1{,}95\|^{2{,}81}_{1{,}11}$	$27{,}67\|^{30{,}2}_{25{,}5}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$39{,}82\|^{40{,}3}_{39{,}5}$
		alignement	$2{,}66\|^{4{,}00}_{2{,}66}$	$1{,}35\|^{1{,}69}_{1{,}11}$	$26{,}70\|^{29{,}3}_{24{,}1}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$39{,}81\|^{40{,}1}_{39{,}4}$
		consistance	$3{,}00\|^{4{,}00}_{2{,}66}$	$1{,}62\|^{2{,}25}_{1{,}11}$	$26{,}61\|^{29{,}3}_{23{,}6}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$39{,}86\|^{40{,}3}_{39{,}4}$
		min-corr.	$3{,}00\|^{4{,}00}_{2{,}66}$	$1{,}62\|^{2{,}25}_{1{,}11}$	$26{,}61\|^{29{,}3}_{23{,}6}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$39{,}86\|^{40{,}3}_{39{,}4}$
	RLSvect($\boldsymbol{\alpha}$)	étiquette ± 1	$3{,}13\|^{4{,}00}_{2{,}66}$	$1{,}45\|^{2{,}25}_{1{,}11}$	$27{,}65\|^{29{,}0}_{25{,}5}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$38{,}79\|^{39{,}2}_{38{,}4}$
		indicatrices	$3{,}13\|^{4{,}00}_{2{,}66}$	$1{,}68\|^{2{,}25}_{1{,}11}$	$27{,}80\|^{29{,}2}_{25{,}9}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$38{,}94\|^{39{,}1}_{38{,}2}$
		alignement	$3{,}20\|^{4{,}00}_{2{,}66}$	$1{,}28\|^{1{,}69}_{0{,}58}$	$27{,}85\|^{29{,}2}_{25{,}6}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$38{,}59\|^{39{,}1}_{38{,}0}$
		consistance	$3{,}13\|^{4{,}00}_{2{,}66}$	$1{,}34\|^{2{,}25}_{0{,}58}$	$27{,}71\|^{28{,}8}_{25{,}6}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$38{,}62\|^{39{,}0}_{38{,}0}$
		min-corr.	$3{,}20\|^{4{,}00}_{2{,}66}$	$1{,}28\|^{1{,}69}_{0{,}58}$	$27{,}85\|^{29{,}2}_{25{,}6}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$38{,}59\|^{39{,}1}_{38{,}0}$
	RLSvect($\boldsymbol{W}$)	étiquette ± 1	$3{,}40\|^{4{,}00}_{2{,}66}$	$1{,}62\|^{2{,}25}_{1{,}11}$	$27{,}85\|^{29{,}0}_{25{,}9}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$38{,}76\|^{39{,}2}_{38{,}2}$
		indicatrices	$3{,}40\|^{4{,}00}_{2{,}66}$	$1{,}40\|^{2{,}25}_{1{,}11}$	$27{,}93\|^{29{,}2}_{25{,}9}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$38{,}72\|^{39{,}2}_{38{,}3}$
		alignement	$3{,}33\|^{4{,}00}_{2{,}66}$	$1{,}28\|^{1{,}69}_{0{,}58}$	$27{,}84\|^{29{,}2}_{25{,}6}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$38{,}63\|^{39{,}1}_{38{,}1}$
		consistance	$3{,}40\|^{4{,}00}_{2{,}66}$	$1{,}28\|^{2{,}25}_{0{,}58}$	$27{,}83\|^{29{,}2}_{25{,}6}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$38{,}67\|^{39{,}2}_{38{,}2}$
		min-corr.	$3{,}40\|^{4{,}00}_{2{,}66}$	$1{,}28\|^{2{,}25}_{0{,}58}$	$27{,}83\|^{29{,}2}_{25{,}6}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$38{,}67\|^{39{,}2}_{38{,}2}$
	SVMvect	étiquette ± 1	$3{,}13\|^{4{,}00}_{2{,}66}$	$1{,}84\|^{2{,}29}_{1{,}11}$	$27{,}68\|^{29{,}7}_{25{,}0}$	$0{,}62\|^{0{,}93}_{0{,}37}$	$39{,}43\|^{40{,}0}_{38{,}9}$
		indicatrices	$3{,}13\|^{4{,}00}_{2{,}66}$	$1{,}84\|^{2{,}71}_{1{,}11}$	$26{,}20\|^{28{,}2}_{24{,}5}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$39{,}91\|^{40{,}3}_{39{,}6}$
		alignement	$2{,}86\|^{3{,}33}_{2{,}00}$	$1{,}68\|^{2{,}25}_{1{,}11}$	$27{,}08\|^{28{,}9}_{25{,}1}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$39{,}74\|^{40{,}5}_{39{,}3}$
		consistance	$2{,}86\|^{3{,}33}_{2{,}00}$	$1{,}68\|^{2{,}25}_{1{,}11}$	$27{,}08\|^{28{,}9}_{25{,}1}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$39{,}74\|^{40{,}5}_{39{,}3}$
		min-corr.	$2{,}86\|^{3{,}33}_{2{,}00}$	$1{,}68\|^{2{,}25}_{1{,}11}$	$27{,}08\|^{28{,}9}_{25{,}1}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$39{,}74\|^{40{,}5}_{39{,}3}$
	Szedmak *et coll.*	étiquette ± 1	$3{,}73\|^{4{,}66}_{2{,}66}$	$2{,}45\|^{3{,}36}_{1{,}11}$	$34{,}34\|^{35{,}9}_{32{,}0}$	$0{,}83\|^{1{,}32}_{0{,}37}$	$43{,}12\|^{43{,}9}_{42{,}4}$
		indicatrices	$3{,}40\|^{4{,}00}_{2{,}66}$	$2{,}51\|^{2{,}84}_{2{,}19}$	$30{,}09\|^{31{,}2}_{28{,}2}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$42{,}37\|^{42{,}9}_{41{,}7}$
		alignement	$3{,}40\|^{4{,}00}_{2{,}66}$	$1{,}67\|^{2{,}22}_{1{,}11}$	$32{,}23\|^{33{,}4}_{31{,}1}$	$0{,}75\|^{1{,}12}_{0{,}37}$	$45{,}02\|^{45{,}4}_{44{,}4}$
		consistance	$3{,}53\|^{4{,}00}_{2{,}66}$	$1{,}67\|^{2{,}22}_{1{,}11}$	$32{,}18\|^{34{,}8}_{29{,}7}$	$0{,}75\|^{1{,}12}_{0{,}37}$	$45{,}00\|^{45{,}4}_{44{,}4}$
		min-corr.	$3{,}33\|^{4{,}00}_{2{,}66}$	$3{,}19\|^{3{,}92}_{2{,}77}$	$33{,}04\|^{34{,}9}_{31{,}5}$	$0{,}73\|^{1{,}12}_{0{,}37}$	$45{,}00\|^{45{,}7}_{44{,}2}$
	LSM	LSM+OvA	$2{,}80\|^{3{,}33}_{2{,}00}$	$0{,}33\|^{1{,}11}_{0{,}00}$	$27{,}39\|^{28{,}7}_{24{,}9}$	$0{,}60\|^{0{,}93}_{0{,}37}$	$38{,}91\|^{39{,}4}_{38{,}1}$
		LSM+OvO	$2{,}80\|^{3{,}33}_{2{,}00}$	$0{,}33\|^{1{,}11}_{0{,}00}$	$27{,}49\|^{29{,}0}_{24{,}1}$	$0{,}58\|^{0{,}93}_{0{,}37}$	$38{,}92\|^{39{,}3}_{38{,}3}$
	SVM	SVM+OvA	3,33	1,12	28,03	1,51	—
		SVM+OvO	2,66	0,56	28,50	0,94	—
		SVM+DAG	3,33	1,12	26,16	1,32	—
		[Weston and Watkins, 1999]	2,66	1,12	28,97	1,51	—
		[Crammer and Singer, 2002]	2,66	1,12	28,03	1,32	—
	Réseaux de neurones		3,96	1,52	31,59	[a]	—

a. La valeur 41,33 obtenue dans [Huang et al., 2012] résulte d'un jeu de données *vowel* de 528 + 462 instances pour l'apprentissage+test, qui est différent du jeu de données investi dans plusieurs études [Hsu and Lin, 2002, Rifkin and Klautau, 2004], y compris le présent document.

Si j'ai vu si loin, c'est que j'étais monté sur des épaules de géants.
[Isaac Newton]

Un nain a un excellent moyen d'être plus haut qu'un géant, c'est de se jucher sur ses épaules.
[Victor Hugo]

7 Bilan et perspectives

En décembre 2007, j'ai soutenu une thèse de doctorat, sur un travail autour de deux axes de recherche, qui ont été poursuivis dans la suite, et ont permis des avancées importantes.

Le premier concerne la mise en œuvre des méthodes à noyaux pour l'analyse de signaux non stationnaires dans les cadre des représentations temps-fréquence. Ce travail a été poursuivi dans le cadre du projet ANR StaRAC « stationarité relative et approches connexes » (2007-2010), en proposant des tests non paramétriques de détection de non-stationnarité qui s'appuient sur des données générées dites substituts ou *surrogates*. Ces découvertes ouvrent une brèche pour les activités de recherche en apprentissage statistique. Désormais, on peut produire des échantillons de l'ensemble d'apprentissage afin de soutenir la résolution du problème.

Le second axe de recherche concerne l'identification de systèmes non linéaires. J'ai poursuivi ce travail mais selon un autre point de vue. Une rupture remarquable s'est produite avec l'optimisation du dictionnaire, d'une part en proposant des critères qui tiennent en compte de l'erreur de modélisation, et d'autre part en ajustant les éléments du dictionnaire. Le domaine d'application s'est aussi élargi au delà de l'identification des systèmes, avec des domaines d'applications tels que l'estimation en ligne des vecteurs propres ou encore le traitement collaboratif de l'information dans les réseaux de capteurs. Cette dernière application m'a permis de se retrouver au sein des projets de la plateforme CAPSEC sur les réseaux de capteurs.

En parallèle à ces travaux, je me suis lancé dans un défi majeur, à savoir le problème de pré-image. La motivation s'est dressée depuis mes travaux de thèse, sur les représentations temps-fréquence, comme précisé dans les perspectives dans [Honeine, 2007], à savoir

> « ...les méthodes [à noyaux] proposent différentes techniques dans le cadre générique de la *reconstruction de pré-images*. Elles visent à déterminer un échantillon x de $\mathbb{X}$ qui produit à travers l'application ϕ associée au noyau reproduisant considéré, une image $\phi(x)$ proche d'un élément de $\mathbb{H}$ donné. Ces principes pourraient être utilisés pour identifier un signal à partir d'une distribution temps-fréquence donnée, non nécessairement valide car obtenue au terme d'une ACP à noyau temps-fréquence par exemple. »

Il s'est avéré que les avancées acquises, aussi bien sur la définition du problème de pré-image que sur sa résolution, ont trouvé leur place dans les domaines applicatifs divers et étonnants, tels que l'auto-localisation dans les réseaux de capteurs sans fil, le traitement de séries temporelles, ou encore le démélange en imagerie hyperspectrale. L'analyse de la non-stationnarité par représentations temps-fréquence ne devra pas échapper à ce progrès.

Le problème du démélange en imagerie hyperspectrale a été soulevé par Cédric Richard fin 2009, au cours de son départ à l'Université de Nice Sophia-Antipolis. J'ai découvert de nouveaux défis. Mes premiers travaux dans ce domaine ont été sur la géométrie des abondances avec les coordonnées barycentriques, qui décrivent une transversalité impressionnante entre les différentes techniques d'extraction de composants purs. Les coordonnées barycentriques sont souvent oubliées dans la littérature. Paradoxalement, je les ai découvertes par l'intermédiaire des travaux de l'équipe

de José Moura à Carnegie Mellon University sur l'auto-localisation dans les réseaux de capteurs [Khan et al., 2009]. Pourtant, il n'y a, *a priori*, aucun lien entre l'imagerie hyperspectrale et l'auto-localisation des capteurs dans les réseaux sans fil.

La motivation principale des méthodes d'apprentissage est la classification. Il serait dommage que je ne m'y mette pas. C'est le cas avec la classification mono-classe et la classification multi-classes. J'ai traité ses problèmes avec la prétention de simplifier. L'idée de simplifier l'approche mono-classe est inspirée d'un article invité (mais souvent oublié) de Shawe-Taylor et Cristianini paru dans les actes du colloque GRETSI 2003 [Shawe-Taylor and Cristianini, 2003]. Une telle référence devrait rendre nos travaux de recherche plus crédibles. Le problème de classification multi-classes a eu sa part en terme de simplification. L'approche privilégiée n'a pas été de proposer, à notre tour, une méthode plus sophistiquée pour gagner un peu en précision par rapport aux méthodes connues dans la littérature. L'approche défendue n'a pas consisté à proposer « encore » un autre codage des étiquettes. Paradoxalement, il s'agit de montrer que « la simplicité est la sophistication suprême ».

Le problème du choix du noyau reste ouvert. Si les possibilités offertes sont infinies, le choix devrait idéalement être piloté par l'application considérée. Des éléments de solutions ont été proposés :

- nous avons étudié ce problème, depuis les travaux sur les représentations temps-fréquence [Honeine et al., 2006, Honeine and Richard, 2007] ;
- nous avons étudié le choix du noyau en imagerie hyperspectrale, voir la section 5.4.2 ;
- nous avons proposé l'apprentissage à noyaux multiples pour le démélange non linéaire, voir la section 5.4.3 ;
- nous avons développé un noyau adapté dans [Noumir et al., 2012b].

Tous ces résultats demeurent modestes au regard des défis envisagés.

La définition du problème de pré-image et sa résolution ont permis une transversalité remarquable dans mes travaux. Il est clair que ces avancées devront être poursuivies, mais avec une nouvelle vision. En effet, ce que nous apprend la résolution du problème de pré-image n'est autre que l'optimisation dans l'espace des observations. Indépendamment du problème de pré-image, c'est exactement ce qui a été réalisé avec succès avec l'adaptation des échantillons d'un dictionnaire dans [Saidé et al., 2013a, Saidé et al., 2013b]. Actuellement, nous étudions l'optimisation dans l'espace des observations pour résoudre des problèmes dans des domaines très variés, tels que la factorisation en matrices non négatives et la mobilité dans les réseaux de capteurs mobiles [Ghadban et al., 2013b]. Les résultats obtenus sont très encourageants, et ouvrent la voie à des applications imprévues par optimisation des échantillons d'apprentissage.

Perspectives à long terme

Une Recherche au service de l'Être Humain.

Bibliographie — sur les épaules des géants —

[Abdolee et al., 2012] Abdolee, R., Champagne, B., and Sayed, A. H. (2012). Diffusion LMS for source and process estimation in sensor networks. In *Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP)*. Ann Arbor, MI, USA.

[Abrahamsen and Hansen, 2009] Abrahamsen, T. J. and Hansen, L. K. (2009). Input space regularization stabilizes pre-images for kernel PCA de-noising. In *IEEE Workshop on Machine Learning for Signal Processing*, Grenoble, France.

[Abrahamsen and Hansen, 2011] Abrahamsen, T. J. and Hansen, L. K. (2011). Regularized pre-image estimation for kernel pca de-noising : Input space regularization and sparse reconstruction. *Journal of Signal Processing Systems*, 65(3) :403–412.

[Abril et al., 2008] Abril, L. G., Angulo, C., Velasco, F., and Ortega, J. A. (2008). A note on the bias in svms for multiclassification. *IEEE Transactions on Neural Networks*, 19(4) :723–725.

[Aizerman et al., 1964] Aizerman, M., Braverman, E., and Rozonoer, L. (1964). Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25 :821–837.

[Allwein et al., 2001] Allwein, E. L., Schapire, R. E., and Singer, Y. (2001). Reducing multiclass to binary : a unifying approach for margin classifiers. *Journal of Machine Learning Research*, 1 :113–141.

[Alpaydin, 1999] Alpaydin, E. (1999). Combined {cv} f test for comparing supervised classification learning algorithms. *Neural Comput.*, 11(8) :1885–1892.

[Amoud et al., 2009a] Amoud, H., Honeine, P., Richard, C., Borgnat, P., and Flandrin, P. (2009a). Time-frequency learning machines for nonstationarity detection using surrogates. In *Proc. IEEE workshop on Statistical Signal Processing*, Cardiff (Wales), UK.

[Amoud et al., 2009b] Amoud, H., Richard, C., Honeine, P., Flandrin, P., and Borgnat, P. (2009b). Sur la caractérisation de non-stationnarités par la méthode des substituts. In *Actes du 22-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Dijon, France.

[Aronszajn, 1950] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68.

[Bachrach and Taylor, 2005] Bachrach, J. and Taylor, C. (2005). Localization in sensor networks. In Stojmenovic, I., editor, *Handbook of Sensor Networks*.

[Bakir, 2005] Bakir, G. (2005). *Extension to Kernel Dependency Estimation with Applications to Robotics*. PhD thesis, Technische Universitat Berlin.

[Bakir et al., 2004] Bakir, G., Weston, J., and Schölkopf, B. (2004). Learning to find pre-images. In Thrun, S., L. S. and Schölkopf, B., editors, *NIPS 2003*, volume 16, pages 449–456, Cambridge, MA, USA. MIT Press.

[Basseville and Nikiforov, 1993] Basseville, M. and Nikiforov, I. V. (1993). *Detection of abrupt changes : theory and application*. Prentice Hall Information and System Sciences Series. Prentice Hall Inc., Englewood Cliffs, NJ.

[Baudat and Anouar, 2001] Baudat, G. and Anouar, F. (2001). Kernel-based methods and function approximation. In *In International Joint Conference on Neural Networks (IJCNN)*, volume 5, pages 1244–1249, Washington, DC, USA.

[Berman et al., 2004] Berman, M., Kiiveri, H., Lagerstrom, R., Ernst, A., Dunne, R., and Huntington, J. F. (2004). ICE : a statistical approach to identifying endmembers in hyperspectral images. *IEEE Trans. Geoscience and Remote Sensing*, 42(10) :2085–2095.

[Bertsekas, 1999] Bertsekas, D. (1999). *Nonlinear programming*. Athena Scientific, second edition edition.

[Bioucas-Dias and Plaza, 2010] Bioucas-Dias, J. M. and Plaza, A. (2010). Hyperspectral unmixing : Geometrical, statistical, and sparse regression-based approaches. In Bruzzone, L., editor, *Proc. SPIE Image and Signal Processing for Remote Sensing XVI*, volume 7830, page 78300A.

[Bioucas-Dias et al., 2012] Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J. (2012). Hyperspectral unmixing overview : Geometrical, statistical, and sparse regression-based approaches. *IEEE J. Sel. Topics Appl. Earth Observations and Remote Sens.*, 5(2) :354–379.

[Bishop, 1995] Bishop, C. (1995). *Neural networks for pattern recognition*. Oxford University Press.

[Bonnans and Shapiro, 1998] Bonnans, J. F. and Shapiro, A. (1998). Optimization problems with perturbations : A guided tour. *SIAM review*, 40(2) :207–227.

[Borgnat et al., 2010] Borgnat, P., Flandrin, P., Honeine, P., Richard, C., and Xiao, J. (2010). Testing stationarity with surrogates : A time-frequency approach. *IEEE Transactions on Signal Processing*, 58(7) :3459–3470.

[Borgnat et al., 2012] Borgnat, P., Flandrin, P., Richard, C., Ferrari, A., Amoud, H., and Honeine, P. (2012). Time-frequency learning machines for nonstationarity detection using surrogates. In *Advances in Machine Learning and Data Mining for Astronomy, In Eds. M. Way, J. Scargle, K. Ali, and A. Srivastava*, Data Mining and Knowledge Discovery series, chapter 22, pages 487–503. Chapman and Hall / CRC Press (Taylor and Francis).

[Bousquet et al., 2004] Bousquet, O., Boucheron, S., and Lugosi, G. (2004). Introduction to Statistical Learning Theory. In *Advanced Lectures on Machine Learning*, pages 169–207. Springer.

[Boyd et al., 2011] Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1) :1–122.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. University Press, Cambridge.

[Bredensteiner and Bennett, 1999] Bredensteiner, E. J. and Bennett, K. P. (1999). Multicategory classification by support vector machines. *Comput. Optim. Appl.*, 12 :53–79.

[Broadwater and Banerjee, 2009] Broadwater, J. and Banerjee, A. (2009). A comparison of kernel functions for intimate mixture models. In *Proc. IEEE WHISPERS'09*, pages 1–4.

[Broadwater et al., 2008] Broadwater, J., Chellappa, R., Banerjee, A., and Burlina, P. (2008). Kernel fully constrained least squares abundance estimates. In *Proc. IEEE IGARSS'07.*, pages 4041–4044.

[Buchsbaum and Bloch, 2002] Buchsbaum, G. and Bloch, O. (2002). Color categories revealed by non-negative matrix factorization of munsell color spectra. *Vision Research*, 42(5) :559–63.

[Buciu et al., 2008] Buciu, I., Nikolaidis, N., and Pitas, I. (2008). Nonnegative matrix factorization in polynomial feature space. *IEEE Transactions on Neural Networks*, 19(6) :1090–1100.

[Buciu and Pitas, 2004] Buciu, I. and Pitas, I. (2004). Application of non-negative and local non negative matrix factorization to facial expression recognition. In *17th International Conference on Pattern Recognition*, volume 1, pages 288–291, Cambridge, UK.

[Burges, 1999] Burges, C. J. C. (1999). Geometry and invariance in kernel based methods. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in kernel methods*, pages 89–116, Cambridge, MA, USA. MIT Press.

[Cadima and Jolliffe, 2009] Cadima, J. and Jolliffe, I. (2009). On relationships between uncentred and column-centred principal component analysis,. *Pakistan Journal of Statistics*, 25(4) :473–503.

[Candès et al., 2006] Candès, E., Romberg, J., and Tao, T. (2006). Robust uncertainty principles : exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Info. Theory*, 52(2) :489–509.

[Candès and Wakin, 2008] Candès, E. J. and Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2) :21–30.

[Cantú-Paz and Kamath, 2005] Cantú-Paz, E. and Kamath, C. (2005). An empirical comparison of combinations of evolutionary algorithms and neural networks for classification problems. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, pages 915–927.

[Cattivelli et al., 2008] Cattivelli, F., Lopes, C., and Sayed, A. (2008). Diffusion recursive least-squares for distributed estimation over adaptive networks. *IEEE Transactions on Signal Processing*, 56(5) :1865–1877.

[Cauwenberghs and Poggio, 2001] Cauwenberghs, G. and Poggio, T. (2001). Incremental and decremental support vector machine learning. In *Advances in Neural Information Processing Systems*, volume 13.

[Censor et al., 2009] Censor, Y., Herman, G. T., and Jiang, M. (2009). A note on the behavior of the randomized Kaczmarz algorithm of Strohmer and Vershynin. *J. Fourier Anal. Appl.*, 15(4) :431–436.

[Chang and Du, 2004] Chang, C. and Du, Q. (2004). Estimation of number of spectrally distinct signal sources in hyperspectral imagery. *IEEE Trans. Geoscience and Remote Sensing*, 42(3) :608–619.

[Chang et al., 2006] Chang, C., Wuand, C.-C., Liu, W.-M., and Quyang, Y.-C. (2006). A new growing method for simplex-based endmember extraction algorithm. *IEEE Trans. Geoscience and Remote Sensing*, 44(10) :2804–2819.

[Chang, 2005] Chang, C.-I. (2005). Orthogonal subspace projection (OSP) revisited : a comprehensive study and analysis. *IEEE Trans. Geoscience and Remote Sensing*, 43(3) :502–518.

[Chen et al., 2011a] Chen, J., Richard, C., Bermudez, J. C. M., and Honeine, P. (2011a). A modified non-negative lms algorithm and its stochastic behavior analysis. In *Proc. 45th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove (CA), USA.

[Chen et al., 2011b] Chen, J., Richard, C., Bermudez, J. C. M., and Honeine, P. (2011b). Non-negative least-mean-square algorithm. *IEEE Transactions on Signal Processing*, 59(11) :5225–5235.

[Chen et al., 2013a] Chen, J., Richard, C., Bermudez, J. C. M., and Honeine, P. (2013a). Identification en ligne avec régularisation l1. algorithme et analyse de convergence en environnement non-stationnaire. In *Actes du 24-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Brest, France.

[Chen et al., 2013b] Chen, J., Richard, C., Ferrari, A., and Honeine, P. (2013b). Nonlinear unmixing of hyperspectral data with partially linear least-squares support vector regression. In *Proc. 38th IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada.

[Chen et al., 2011c] Chen, J., Richard, C., and Honeine, P. (2011c). A novel kernel-based nonlinear unmixing scheme of hyperspectral images. In *Proc. 45th Asilomar Conference on Signals, Systems, and Computers*, pages 1898–1902, Pacific Grove (CA), USA. IEEE.

[Chen et al., 2011d] Chen, J., Richard, C., and Honeine, P. (2011d). Un nouveau paradigme pour le démélange non-linéaire des images hyperspectrales. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France.

[Chen et al., 2012a] Chen, J., Richard, C., and Honeine, P. (2012a). Nonlinear unmixing of hyperspectral images based on multi-kernel learning. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Shanghai, China.

[Chen et al., 2013c] Chen, J., Richard, C., and Honeine, P. (2013c). Estimating abundance fractions of materials in hyperspectral images by fitting a post-nonlinear mixing model. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Gainesville, Florida, USA.

[Chen et al., 2013d] Chen, J., Richard, C., and Honeine, P. (2013d). Nonlinear estimation of material abundances of hyperspectral images with ℓ_1-norm spatial regularization. *IEEE Transactions on Geoscience and Remote Sensing*.

[Chen et al., 2013e] Chen, J., Richard, C., and Honeine, P. (2013e). Nonlinear unmixing of hyperspectral data based on a linear-mixture/nonlinear-fluctuation model. *IEEE Transactions on Signal Processing*, 61(2) :480–492.

[Chen et al., 2010a] Chen, J., Richard, C., Honeine, P., and Bermudez, J. C. M. (2010a). Non-negative distributed regression for data inference in wireless sensor networks. In *Proc. 44th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove (CA), USA.

[Chen et al., 2010b] Chen, J., Richard, C., Honeine, P., Lantéri, H., and Theys, C. (2010b). System identification under non-negativity constraints. In *Proc. 18th European Conference on Signal Processing*, Aalborg, Denmark.

[Chen et al., 2010c] Chen, J., Richard, C., Honeine, P., Snoussi, H., Lantéri, H., and Theys, C. (2010c). Techniques d'apprentissage non-linéaires en ligne avec contraintes de positivité. In *Actes de la VI-ème Conférence Internationale Francophone d'Automatique*, Nancy, France.

[Chen et al., 2012b] Chen, J., Richard, C., Honeine, P., and Tourneret, J.-Y. (2012b). Prediction of rain attenuation series with discretized spectral model. In *Proc. IEEE International Geoscience and Remote Sensing Symposium*, Munich, Germany.

[Chen et al., 2011e] Chen, J., Richard, C., Lantéri, H., Theys, C., and Honeine, P. (2011e). A gradient based method for fully constrained least-squares unmixing of hyperspectral images. In *Proc. IEEE workshop on Statistical Signal Processing*, Nice, France.

[Chen et al., 2011f] Chen, J., Richard, C., Lantéri, H., Theys, C., and Honeine, P. (2011f). Online system identification under non-negativity and ℓ_1-norm constraints algorithm and weight behavior analysis. In *Proc. 19th European Conference on Signal Processing*, Barcelona, Spain.

[Chen and Chang, 1995] Chen, L.-H. and Chang, S. (1995). An adaptive learning algorithm for principal component analysis. *IEEE Trans. Neural Networks*, 6(5) :1255 –1263.

[Choi, 2009] Choi, Y.-S. (2009). Least squares one-class support vector machine. *Pattern Recogn. Lett.*, 30 :1236–1240.

[Cimmino, 1938] Cimmino, G. (1938). Calcolo approssimato per le soluzioni dei sistemi di equazioni lineari. *La Ricerca Scientifica*, II, 9 :326–333.

[Clark and Geological Survey (U.S.), 2007] Clark, R. N. and Geological Survey (U.S.) (2007). *USGS digital spectral library splib06a [electronic resource]*. U.S. Geological Survey, Denver, CO, US.

[Comon and Jutten, 2010] Comon, P. and Jutten, C., editors (2010). *Handbook of Blind Source Separation : Independent Component Analysis and Applications*. Academic Press.

[Cortes et al., 2005] Cortes, C., Mohri, M., and Weston, J. (2005). A general regression technique for learning transductions. In *Proc. of the 22nd international conference on machine learning (ICML)*, pages 153–160, New York, NY, USA. ACM.

[Costa et al., 2006] Costa, J. A., Patwari, N., and Hero, A. O. (2006). Distributed weighted-multidimensional scaling for node localization in sensor networks. *ACM Trans. Sen. Netw.*, 2(1) :39–64.

[Cox and Cox, 2000] Cox, T. F. and Cox, M. A. A. (2000). *Multidimensional Scaling*. Monographs on Statistics and Applied Probability. Chapman and Hall / CRC, London, 2nd edition edition.

[Crammer and Singer, 2002] Crammer, K. and Singer, Y. (2002). On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2 :265–292.

[Cristianini et al., 2002] Cristianini, N., Elisseeff, A., Shawe-Taylor, J., and Kandola, J. (2002). On kernel target alignment. *Proc. of the Neural Information Processing Systems (NIPS)*, pages 367–373.

[Csató and Opper, 2001] Csató, L. and Opper, M. (2001). Sparse representation for gaussian process models. In *Advances in Neural Information Processing Systems 13*, pages 444–450. MIT Press.

[Cucker and Smale, 2002] Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39 :1–49.

[Darken et al., 1992] Darken, C., Chang, J., and Moody, J. (1992). Learning rate schedules for faster stochastic gradient search. In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Piscataway, NJ. IEEE.

[Davy et al., 2006] Davy, M., Desobry, F., Gretton, A., and Doncarli, C. (2006). An Online Support Vector Machine for Abnormal Events Detection. *Signal Processing*, 86(8) :2009–2025.

[Davy and Godsill, 2002] Davy, M. and Godsill, S. (2002). Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation. In *Proc IEEE ICASSP*, pages 1313–1316.

[Devarajan, 2008] Devarajan, K. (2008). Nonnegative Matrix Factorization : An Analytical and Interpretive Tool in Computational Biology. *PLoS Comput Biol*, 4(7).

[Deveughèle et al., 2012] Deveughèle, S., Yin, H., Fillatre, L., Honeine, P., Nikiforov, I., Richard, C., Snoussi, H., Azzaoui, N., Guépié, B. K., and Noumir, Z. (2012). Vigires'eau. In *Proc. 10th International Conference on Hydroinformatics*, Hamburg, Germany.

[Dietterich, 1998] Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10 :1895–1923.

[Dietterich and Bakiri, 1995] Dietterich, T. G. and Bakiri, G. (1995). Solving multiclass learning problems via error-correcting output codes. *Journal of Artificial Intelligence Research*, 2 :263–286.

[Ding et al., 2010] Ding, C., Li, T., and Jordan, M. I. (2010). Convex and Semi-Nonnegative Matrix Factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1) :45–55.

[Dobigeon and Tourneret, 2011] Dobigeon, N. and Tourneret, J.-Y. (2011). Enhancing hyperspectral image unmixing with spatial correlations. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11) :4239–4247.

[Dodd et al., 2003] Dodd, T., Kadirkamanathan, V., and Harrison, R. (2003). Function estimation in Hilbert space using sequential projections. In *Proceedings of the IFAC Conference on Intelligent Control Systems and Signal Processing (ICONS 2003)*, pages 113–118, University of Algarve, Portugal.

[Dodd and Harris, 2002] Dodd, T. J. and Harris, C. J. (2002). Identification of nonlinear time series via kernels. *International Journal of System Science*, 33(9) :737–750.

[Doherty et al., 2001] Doherty, L., Pister, K. S. J., and Ghaoui, L. E. (2001). Convex position estimation in wireless sensor networks. In *Proc. IEEE Infocom*, Anchorage, USA.

[Du et al., 2008] Du, Q., Raksuntorn, N., Younan, N. H., and King, R. L. (2008). Variants of n-findr algorithm for endmember extraction. In *Proc. SPIE - Image and Signal Processing for Remote Sensing XIV*.

[Eckstein and Bertsekas, 1992] Eckstein, J. and Bertsekas, D. (1992). On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1) :293–318.

[Engel et al., 2004] Engel, Y., Mannor, S., and Meir, R. (2004). The kernel recursive least squares algorithm. *IEEE Trans. Signal Processing*, 52(8) :2275–2285.

[Ertekin et al., 2011] Ertekin, S., Bottou, L., and Giles, C. L. (2011). Nonconvex online support vector machines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33 :368–381.

[Essoloh et al., 2009] Essoloh, M., Honeine, P., Richard, C., and Snoussi, H. (2009). Apprentissage non-linéaire en ligne dans les réseaux de capteurs sans fil. In *Actes du 22-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Dijon, France.

[Essoloh et al., 2008] Essoloh, M., Richard, C., Snoussi, H., and Honeine, P. (2008). Distributed localization in wireless sensor networks as a pre-image problem in a reproducing kernel hilbert space. In *Proc. 16th European Conference on Signal Processing*, Lausanne, Switzerland.

[Etyngier et al., 2007] Etyngier, P., Ségonne, F., and Keriven, R. (2007). Shape priors using manifold learning techniques. In *Proc. 11th IEEE International Conference on Computer Vision*, Rio de Janeiro, Brazil.

[Evgeniou et al., 2005] Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6 :615–637.

[Fillatre et al., 2010] Fillatre, L., Honeine, P., Nikiforov, I., Richard, C., Snoussi, H., and Azzaoui, N. (2010). Vigires'eau : Surveillance en temps réel de la qualité de l'eau potable d'un réseau de distribution en vue de la détection d'intrusions. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'10), (ANR - CSOSG)*, pages 1–7, Troyes, France.

[Fillatre et al., 2011] Fillatre, L., Honeine, P., Nikiforov, I., Richard, C., Snoussi, H., Azzaoui, N., Guépié, B. K., Noumir, Z., Deveughèle, S., and Yin, H. (2011). Vigires'eau : Surveillance en temps réel de la qualité de l'eau potable d'un réseau de distribution en vue de la détection d'intrusions. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'11), (ANR - CSOSG)*, pages 1–7, Troyes, France.

[Flandrin and Borgnat, 2008] Flandrin, P. and Borgnat, P. (2008). Revisiting and testing stationarity. *Journal of Physics : Conference Series*, 139(1) :012004.

[Fürnkranz, 2002] Fürnkranz, J. (2002). Round robin classification. *Journal of Machine Learning Research*, 2 :721–747.

[Ghadban et al., 2013a] Ghadban, N., Honeine, P., Francis, C., Mourad-Chehade, F., Farah, J., and Kallas, M. (2013a). Estimation locale d'un champ de diffusion par modèles à noyaux. In *Actes de la 14-ème conférence ROADEF de la Société Française de Recherche Opérationnele et Aide à la Décision*, Troyes, France.

[Ghadban et al., 2013b] Ghadban, N., Honeine, P., Francis, C., Mourad-Chehade, F., Farah, J., and Kallas, M. (2013b). Mobilité d'un réseau de capteurs sans fil basée sur les méthodes à noyau. In *Actes du 24-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Brest, France.

[Gilbert et al., 2003] Gilbert, A. C., Muthukrishnan, S., and Strauss, M. J. (2003). Approximation of functions over redundant dictionaries using coherence. In *Proc. 14-th annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 243–252, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[Girolami, 2002] Girolami, M. (2002). Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14 :669–688.

[Girosi et al., 1993] Girosi, F., Jones, M., and Poggio, T. (1993). Priors stabilizers and basis functions : From regularization to radial, tensor and additive splines. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA.

[Goldstein and Osher, 2009] Goldstein, T. and Osher, S. (2009). The split Bregman method for L1 regularized problems. *SIAM Journal on Imaging Sciences*, 2(2) :323–343.

[Gómez-Verdejo et al., 2011] Gómez-Verdejo, V., Arenas-García, J., Lázaro-Gredilla, M., and Navia-Vázquez, A. (2011). Adaptive one-class support vector machine. *IEEE Transactions on Signal Processing*, 59(6) :2975–2981.

[Gretton and Désobry, 2003] Gretton, A. and Désobry, F. (2003). On-line one-class support vector machines. an application to signal segmentation. In *Proc. IEEE ICASSP, HongKong*, pages II–709–12 vol.2.

[Guermeur, 2008] Guermeur, Y. (2008). *SVM Multiclasses, Théorie et Applications*. Habilitation à diriger des recherches, Université Nancy I.

[Guilfoyle et al., 2001] Guilfoyle, K. J., Althouse, M. L., and Chang, C.-I. (2001). A quantitative and comparative analysis of linear and nonlinear spectral mixture models using radial basis function neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 39(8) :2314–2318.

[Halimi et al., 2011] Halimi, A., Altman, Y., Dobigeon, N., and Tourneret, J.-Y. (2011). Nonlinear unmixing of hyperspectral images using a generalized bilinear model. *IEEE Transactions on Geoscience and Remote Sensing*, 49(11) :4153–4162.

[Hapke, 1981] Hapke, B. (1981). Bidirectional reflectance spectroscopy, 1, Theory. *J. Geophys. Res*, 86 :3039–3054.

[Harsanyi and Chang, 1994] Harsanyi, J. C. and Chang, C.-I. (1994). Hyperspectral image classification and dimensionality reduction : an orthogonal subspace projection approach. *IEEE Trans. Geoscience and Remote Sensing*, 32(4) :779–785.

[Haussler, 1999] Haussler, D. (1999). Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, Department of Computer Science, University of California at Santa Cruz.

[Haykin, 1999] Haykin, S. (1999). *Neural networks : a comprehensive foundation*. Prentice Hall, Englewood Cliffs, NJ.

[Heinz and Chang, 2001] Heinz, D. and Chang, C. (2001). Fully constrained least squares linear spectral mixture analysis method for material quantification in hyperspectral imagery. *IEEE Trans. Geoscience and Remote Sensing*, 39(3) :529–545.

[Hempstalk and Frank, 2008] Hempstalk, K. and Frank, E. (2008). Discriminating against new classes : One-class versus multi-class classification. In *Proc. 21st Australasian Joint Conference on Artificial Intelligence : Advances in Artificial Intelligence*, pages 325–336, Berlin, Heidelberg. Springer-Verlag.

[Herbrich, 2002] Herbrich, R. (2002). *Learning kernel classifiers. Theory and algorithms*. The MIT Press, Cambridge, MA, USA.

[Heylen et al., ress] Heylen, R., Akhter, M. A., and Scheunders, P. (2014 (In Press)). On using projection onto convex sets for solving the hyperspectral unmixing problem. *IEEE Geoscience and Remote Sensing Letters*.

[Heylen et al., 2011] Heylen, R., Burazerovic, D., and Scheunders, P. (2011). Nonlinear spectral unmixing by geodesic simplex volume maximization. *IEEE Journal of Selected Topics in Signal Processing*, 5(3) :534–542.

[Heylen and Scheunders, 2012] Heylen, R. and Scheunders, P. (2012). Calculation of geodesic distances in non-linear mixing models : demonstration on the generalized bilinear model. *IEEE Geoscience and Remote Sensing letters*, 9(4) :644–648.

[Honeine, 2007] Honeine, P. (2007). *Méthodes à noyau pour l'analyse et la décision en environnement non-stationnaire*. PhD thesis, mémoire de thèse de doctorat en Optimisation et Sûreté des Systèmes, Ecole doctoral SSTO - UTT, Troyes, France.

[Honeine, 2012] Honeine, P. (2012). Online kernel principal component analysis : a reduced-order model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9) :1814–1826.

[Honeine et al., 2008a] Honeine, P., Essoloh, M., Richard, C., and Snoussi, H. (2008a). Distributed regression in sensor networks with a reduced-order kernel model. In *Proc. 51st IEEE GLOBECOM Global Communications Conference*, New Orleans, LA, USA.

[Honeine and Lantéri, 2013] Honeine, P. and Lantéri, H. (2013). Constrained reflect-then-combine methods for unmixing hyperspectral data. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Gainesville, Florida, USA.

[Honeine et al., 2013a] Honeine, P., Lantéri, H., and Richard, C. (2013a). Constrained kaczmarz's cyclic projections for unmixing hyperspectral data. In *Proc. 21th European Conference on Signal Processing*, Marrakech, Morocco.

[Honeine et al., 2011] Honeine, P., Mourad, F., Kallas, M., Snoussi, H., Amoud, H., and Francis, C. (2011). Wireless sensor networks in biomedical : body area networks. In *Proc. 7th International Workshop on Systems, Signal Processing and their Applications*, Algeria.

[Honeine et al., 2013b] Honeine, P., Noumir, Z., and Richard, C. (2013b). Multiclass classification machines with the complexity of a single binary classifier. *Signal Processing*, 93(5) :1013–1026.

[Honeine and Richard, 2007] Honeine, P. and Richard, C. (2007). Signal-dependent time-frequency representations for classification using a radially gaussian kernel and the alignment criterion. In *Proc. IEEE workshop on Statistical Signal Processing*, Madison, WI, USA.

[Honeine and Richard, 2009] Honeine, P. and Richard, C. (2009). Solving the pre-image problem in kernel machines : a direct method. In *Proc. 19th IEEE workshop on Machine Learning for Signal Processing*, Grenoble, France. - best paper award -.

[Honeine and Richard, 2010a] Honeine, P. and Richard, C. (2010a). The angular kernel in machine learning for hyperspectral data classification. In *Proc. IEEE Workshop on Hyperspectral Image and Signal Processing : Evolution in Remote Sensing*, Reykjavik, Iceland.

[Honeine and Richard, 2010b] Honeine, P. and Richard, C. (2010b). A simple scheme for unmixing hyperspectral data based on the geometry of the n-dimensional simplex. In *Proc. IEEE International Geoscience and Remote Sensing Symposium*, Honolulu (Hawaii), USA.

[Honeine and Richard, 2011a] Honeine, P. and Richard, C. (2011a). Approches géométriques pour l'estimation des fractions d'abondance en traitement de données hyperspectales. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France.

[Honeine and Richard, 2011b] Honeine, P. and Richard, C. (2011b). A closed-form solution for the pre-image problem in kernel-based machines. *Journal of Signal Processing Systems*, 65(3) :289–299.

[Honeine and Richard, 2011c] Honeine, P. and Richard, C. (2011c). Preimage problem in kernel-based machine learning. *IEEE Signal Processing Magazine*, 28(2) :77–88.

[Honeine and Richard, 2012] Honeine, P. and Richard, C. (2012). Geometric unmixing of large hyperspectral images : a barycentric coordinate approach. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6) :2185–2195.

[Honeine et al., 2007a] Honeine, P., Richard, C., and Bermudez, J. C. M. (2007a). Modélisation parcimonieuse non linéaire en ligne par une méthode à noyau reproduisant et un critère de cohérence. In *Actes du XXI-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Troyes, France.

[Honeine et al., 2007b] Honeine, P., Richard, C., and Bermudez, J. C. M. (2007b). On-line nonlinear sparse approximation of functions. In *Proc. IEEE International Symposium on Information Theory*, pages 956–960, Nice, France.

[Honeine et al., 2008b] Honeine, P., Richard, C., Bermudez, J. C. M., and Snoussi, H. (2008b). Distributed prediction of time series data with kernels and adaptive filtering techniques in sensor networks. In *Proc. 42nd Annual ASILOMAR Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA. invited paper.

[Honeine et al., 2009a] Honeine, P., Richard, C., Bermudez, J. C. M., Snoussi, H., Essoloh, M., and Vincent, F. (2009a). Functional estimation in hilbert space for distributed learning in wireless sensor networks. In *Proc. 34th IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan.

[Honeine et al., 2008c] Honeine, P., Richard, C., Essoloh, M., and Snoussi, H. (2008c). Localization in sensor networks - a matrix regression approach. In *Proc. 5th IEEE Sensor Array and Multichannel Signal Processing Workshop*, Darmstadt, Germany.

[Honeine et al., 2005] Honeine, P., Richard, C., and Flandrin, P. (2005). Reconnaissance des formes par méthodes à noyau dans le domaine temps-fréquence. In *Actes du XX-ème Colloque GRETSI sur le Traitement du Signal et des Images*, pages 969–972, Louvain-la-Neuve, Belgium.

[Honeine et al., 2006] Honeine, P., Richard, C., Flandrin, P., and Pothin, J.-B. (2006). Optimal selection of time-frequency representations for signal classification : A kernel-target alignment approach. In *Proc. 31st IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France.

[Honeine et al., 2013c] Honeine, P., Richard, C., and Nguyen, N. H. (2013c). Approches géométriques pour l'estimation des fractions d'abondance en traitement de données hyperspectales. extensions aux modèles de mélange non-linéaires. *Traitement du signal*, 30(1-2) :61–86.

[Honeine et al., 2009b] Honeine, P., Richard, C., and Snoussi, H. (2009b). Auto-localisation dans les réseaux de capteurs sans fil par régression de matrices de gram. In *Actes du 22-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Dijon, France.

[Honeine et al., 2010] Honeine, P., Richard, C., Snoussi, H., Bermudez, J. C. M., and Chen, J. (2010). A decentralized approach for non-linear prediction of time series data in sensor networks. *Journal on Wireless Communications and Networking*, Special issue on theoretical and algorithmic foundations of wireless ad hoc and sensor networks :12 :1–12 :12.

[Hsu and Lin, 2002] Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13 :415–425.

[Hsu and hua K. Burke, 2003] Hsu, S. M. and hua K. Burke, H. (2003). Multisensor fusion with hyperspectral imaging data : Detection and classification. *Lincoln Laboratory Journal*, 14(1) :145–159.

[Huang et al., 2012] Huang, G.-B., Zhou, H., Ding, X., and Zhang, R. (2012). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics - Part B : Cybernetics*. in press.

[Hull, 1994] Hull, J. J. (1994). A database for handwritten text recognition research. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16 :550–554.

[Hyvärinen et al., 2001] Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis*. Wiley.

[Hyvärinen and Oja, 2000] Hyvärinen, A. and Oja, E. (2000). Independent component analysis : algorithms and applications. *Neural Netw.*, 13(4-5) :411–430.

[Iordache et al., 2012] Iordache, M.-D., Bioucas-Dias, J., and Plaza, A. (2012). Total variation spatial regularization for sparse hyperspectral unmixing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing (to appear)*.

[Jolliffe, 1986] Jolliffe, I. (1986). *Principal Component Analysis*. Springer-Verlag, New York, NY, USA.

[Kaczmarz, 1937] Kaczmarz, S. (1937). Angenäherte Auflösung von Systemen linearer Gleichungen. *Bulletin International de l'Académie Polonaise des Sciences et des Lettres*, 35 :355–357.

[Kallas et al., 2012a] Kallas, M., Francis, C., Honeine, P., Amoud, H., and Richard, C. (2012a). Modeling electrocardiogram using yule-walker equations and kernel machines. In *Proc. 19th International Conference on Telecommunications*, Jounieh, Lebanon.

[Kallas et al., 2012b] Kallas, M., Francis, C., Kanaan, L., Merheb, D., Honeine, P., and Amoud, H. (2012b). Multi-class svm classification combined with kernel pca feature extraction of ecg signals. In *Proc. 19th International Conference on Telecommunications*, Jounieh, Lebanon.

[Kallas et al., 2011a] Kallas, M., Honeine, P., Amoud, H., and Francis, C. (2011a). Sur le problème de la pré-image en reconnaissance des formes avec contraintes de non-négativité. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France.

[Kallas et al., 2010a] Kallas, M., Honeine, P., Amoud, H., Francis, C., and Richard, C. (2010a). Constrained pattern recognition with nonlinear principal component analysis. In *Journées Scientifiques à l'Ecole Doctorale de Sciences et Technologie*, Liban.

[Kallas et al., 2011b] Kallas, M., Honeine, P., Francis, C., and Amoud, H. (2011b). A comparative study of pre-image techniques : The kernel autoregressive case. In *Proc. IEEE workshop on Signal Processing Systems*, pages 379–384, Beirut, Lebanon.

[Kallas et al., 2013a] Kallas, M., Honeine, P., Francis, C., and Amoud, H. (2013a). Kernel autoregressive models using yule-walker equations. *Signal Processing*, 93(11) :3053–3061.

[Kallas et al., 2010b] Kallas, M., Honeine, P., Richard, C., Amoud, H., and Francis, C. (2010b). Nonlinear feature extraction using kernel principal component analysis with non-negative pre-image. In *Proc. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Buenos Aires, Argentina.

[Kallas et al., 2011c] Kallas, M., Honeine, P., Richard, C., Francis, C., and Amoud, H. (2011c). Kernel-based autoregressive modeling with a pre-image technique. In *Proc. IEEE workshop on Statistical Signal Processing*, Nice, France.

[Kallas et al., 2011d] Kallas, M., Honeine, P., Richard, C., Francis, C., and Amoud, H. (2011d). Modèle autorégressif non-linéaire à noyau. une première approche. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France.

[Kallas et al., 2011e] Kallas, M., Honeine, P., Richard, C., Francis, C., and Amoud, H. (2011e). Non-negative pre-image in machine learning for pattern recognition. In *Proc. 19th European Conference on Signal Processing*, Barcelona, Spain.

[Kallas et al., 2012c] Kallas, M., Honeine, P., Richard, C., Francis, C., and Amoud, H. (2012c). Prediction of time series using yule-walker equations with kernels. In *Proc. 37th IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan.

[Kallas et al., 2013b] Kallas, M., Honeine, P., Richard, C., Francis, C., and Amoud, H. (2013b). Non-negativity constraints on the pre-image for pattern recognition with kernel machines. *Pattern Recognition*, 46(11) :3066–3080.

[Kanaan et al., 2011] Kanaan, L., Merheb, D., Kallas, M., Francis, C., Amoud, H., and Honeine, P. (2011). Pca and kpca of ecg signals with binary svm classification. In *Proc. IEEE workshop on Signal Processing Systems*, pages 344–348, Beirut, Lebanon.

[Kemmler et al., 2010] Kemmler, M., Rodner, E., and Denzler, J. (2010). One-class classification with gaussian processes. In *Proc. Asian Conference on Computer Vision*, pages 489–500.

[Keshava and Mustard, 2002] Keshava, N. and Mustard, J. F. (2002). Spectral unmixing. *IEEE Signal Processing Magazine*, 19(1) :44—57.

[Khan et al., 2009] Khan, U. A., Kar, S., and Moura, J. M. F. (2009). Distributed sensor localization in random environments using minimal number of anchor nodes. *Trans. Sig. Proc.*, 57(5) :2000–2016.

[Khodor et al., 2011] Khodor, N., Amoud, H., Kallas, M., Honeine, P., and Francis, C. (2011). Le problème de pré-image dans la reconnaissance des formes. In *Proc. 1st International Conference on Advances in Biomedical Engineering*, pages 1–2, Tripoli, Lebanon.

[Kim et al., 2005] Kim, K., Franz, M., and Schölkopf, B. (2005). Iterative kernel principal component analysis for image modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(9) :1351–1366.

[Kim and Tidor, 2003] Kim, P. M. and Tidor, B. (2003). Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.*, 13(7) :1706–1718.

[Kimeldorf and Wahba, 1971] Kimeldorf, G. and Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33 :82–95.

[Kivinen et al., 2004] Kivinen, J., Smola, A. J., and Williamson, R. C. (2004). Online learning with kernels. *IEEE Transactions on Signal Processing*, 52(8).

[Kruskal, 1964a] Kruskal, J. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1) :1–27.

[Kruskal, 1964b] Kruskal, J. (1964b). Nonmetric multidimensional scaling : A numerical method. *Psychometrika*, 29(2) :115–129.

[Kurková, 2004] Kurková, V. (2004). Learning from data as an inverse problem. In Antoch, J., editor, *Proc. Computational Statistics (CompStat'04)*, pages 1377–1384, Heidelberg,Germany. Physica-Verlag/SpringerAcademic Press, Inc.

[Kwok and Tsang, 2003] Kwok, J. T. and Tsang, I. W. (2003). The pre-image problem in kernel methods. In Fawcett, T. and Mishra, N., editors, *Proc. of the 20th International Conference on Machine Learning (ICML 2003)*, pages 408–415. AAAI Press.

[Lantéri et al., 2001] Lantéri, H., Roche, M., Cuevas, O., and Aime, C. (2001). A general method to devise maximum-likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Processing*, 81 :945–974.

[Lantéri et al., 2011] Lantéri, H., Theys, C., Richard, C., and Mary, D. (2011). Regularized split gradient method for nonnegative matrix factorization. In *ICASSP*, pages 1133–1136.

[Lawson and Hanson, 1987] Lawson, C. L. and Hanson, R. J. (1987). *Solving Least Squares Problems (Classics in Applied Mathematics)*. Society for Industrial Mathematics.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755) :788–791.

[Lee et al., 2004] Lee, Y., Lin, Y., and Wahba, G. (2004). Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99 :67–81.

[Li and Ding, 2006] Li, T. and Ding, C. (2006). The relationships among various nonnegative matrix factorization methods for clustering. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 362–371, Washington, DC, USA. IEEE Computer Society.

[Li and Ngom, 2012] Li, Y. and Ngom, A. (2012). A new kernel non-negative matrix factorization and its application in microarray data analysis. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB*, pages 371–378, San Diego, CA, USA.

[Liu and Zheng, 2004] Liu, W. and Zheng, N. (2004). Non-negative matrix factorization based methods for object recognition. *Pattern Recognition Letters*, 25(8) :893–897.

[Liu et al., 2010] Liu, Y.-H., Liu, Y.-C., and Chen, Y.-J. (2010). Fast support vector data descriptions for novelty detection. *IEEE Trans. on Neural Networks*, 21(8) :1296–1313.

[Lopes and Sayed, 2007] Lopes, C. and Sayed, A. (2007). Incremental adaptive strategies over distributed networks. *IEEE Transactions on Signal Processing*, 55(8) :4064–4077.

[Lopes and Sayed, 2008] Lopes, C. and Sayed, A. (2008). Diffusion least-mean squares over adaptive networks : Formulation and performance analysis. *IEEE Transactions on Signal Processing*, 56(7) :3122–3136.

[Luenberger and Ye, 2008] Luenberger, D. and Ye, Y. (2008). *Linear and Nonlinear Programming*. Springer Verlag.

[Luo et al., 2008] Luo, W., Zhong, L., and Zhang, B. (2008). Null subspace analysis for spectral unmixing in hyperspectral remote sensing. In *Proc. Congress on Image and Signal Processing*, volume 4, pages 763–767, Washington, DC, USA. IEEE Computer Society.

[Mahfouz et al., 2013a] Mahfouz, S., Mourad-Chehade, F., Honeine, P., Farah, J., and Snoussi, H. (2013a). Kernel-based localization using fingerprinting in wireless sensor networks. In *Proc. 14th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Darmstadt, Germany.

[Mahfouz et al., 2013b] Mahfouz, S., Mourad-Chehade, F., Honeine, P., Farah, J., and Snoussi, H. (2013b). Localisation par fingerprinting et méthodes à noyaux dans les réseaux de capteurs sans fil. In *Actes de la 14-ème conférence ROADEF de la Société Française de Recherche Opérationnelle et Aide à la Décision*, Troyes, France.

[Mahfouz et al., 2013c] Mahfouz, S., Mourad-Chehade, F., Paul, H., Farah, J., and Snoussi, H. (2013c). Decentralized localization using fingerprinting and kernel methods in wireless sensor networks. In *Proc. 21th European Conference on Signal Processing*, Marrakech, Morocco.

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

[Micchelli, 1986] Micchelli, C. A. (1986). Interpolation of scattered data : Distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2(1) :11–22.

[Mika et al., 1999] Mika, S., Schölkopf, B., Smola, A., Müller, K.-R., Scholz, M., and Rätsch, G. (1999). Kernel PCA and de-noising in feature spaces. In *Proc. of the 1998 conference on advances in neural information processing systems II*, pages 536–542, Cambridge, MA, USA. MIT Press.

[Mourad et al., 2012] Mourad, F., Honeine, P., and Snoussi, H. (2012). Indoor localization using polar intervals in wireless sensor networks. In *Proc. 19th International Conference on Telecommunications*, pages 1–6, Jounieh, Lebanon.

[Mourad-Chehade et al., 2013] Mourad-Chehade, F., Honeine, P., and Snoussi, H. (2013). Polar interval-based localization in mobile sensor networks. *IEEE Transactions on Aerospace and Electronic Systems*, 49(4) :2310–2322.

[Muñoz and de Diego, 2006] Muñoz, A. and de Diego, I. (2006). From indefinite to positive semidefinite matrices. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 764–772. Springer-Verlag.

[Nader et al., 2013] Nader, P., Honeine, P., and Beauseroy, P. (2013). Intrusion detection in scada systems using one-class classification. In *Proc. 21th European Conference on Signal Processing*, Marrakech, Morocco.

[Nascimento and Bioucas-Dias, 2009] Nascimento, J. M. P. and Bioucas-Dias, J. M. (2009). Nonlinear mixture model for hyperspectral unmixing. In *Proc. SPIE*, volume 7477.

[Nascimento and Dias, 2004] Nascimento, J. M. P. and Dias, J. M. B. (2004). Vertex component analysis : A fast algorithm to unmix hyperspectral data. *IEEE Trans. Geoscience and Remote Sensing*, 43(4) :898–910.

[Needell, 2010] Needell, D. (2010). Randomized Kaczmarz solver for noisy linear systems. *BIT Numerical Mathematics*, 50(2) :395–403.

[Nguyen et al., 2013] Nguyen, N. H., Chen, J., Richard, C., Honeine, P., and Theys, C. (2013). Supervised nonlinear unmixing of hyperspectral images using a pre-image method. In *New Concepts in Imaging : Optical and Statistical Models, In Eds. D. Mary, C. Theys, and C. Aime*, volume 59 of *EAS Publications Series*, pages 417–437. EDP Sciences.

[Nguyen et al., 2012] Nguyen, N. H., Richard, C., Honeine, P., and Theys, C. (2012). Hyperspectral image unmixing using manifold learning : methods derivations and comparative tests. In *Proc. IEEE International Geoscience and Remote Sensing Symposium*, pages 3086–3089, Munich, Germany. IEEE.

[Nguyen et al., 2005a] Nguyen, X., Jordan, M. I., and Sinopoli, B. (2005a). A kernel-based learning approach to ad hoc sensor network localization. *ACM Trans. Sen. Netw.*, 1(1) :134–152.

[Nguyen et al., 2005b] Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2005b). Nonparametric decentralized detection using kernel methods. *IEEE Trans. Signal Processing*, 53 :4053–4066.

[Noumir et al., 2012a] Noumir, Z., Guépié, B. K., Fillatre, L., Honeine, P., Nikiforov, I., Snoussi, H., Richard, C., Jarrige, P.-A., and Campan, F. (2012a). Detection of contamination in water distribution network. In *2nd International Conference SimHydro : New trends in simulation hydroinformatics and 3D modeling*, Nice, France.

[Noumir et al., 2013] Noumir, Z., Guépié, B. K., Fillatre, L., Honeine, P., Nikiforov, I., Snoussi, H., Richard, C., Jarrige, P.-A., and Campan, F. (2013). Detection of contamination in water distribution network. In *Advances in Hydroinformatics : SIMHYDRO - New Frontiers of Simulation, In Eds. Philippe Gourbesville, Jean Cunge, and Guy Caignaert*, Springer Hydrogeology, chapter 12, pages 1–11. Springer Science.

[Noumir et al., 2011a] Noumir, Z., Honeine, P., and Richard, C. (2011a). Classification multi-classes au prix d'un classifieur binaire. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France.

[Noumir et al., 2011b] Noumir, Z., Honeine, P., and Richard, C. (2011b). Multi-class least squares classification at binary-classification complexity. In *Proc. IEEE workshop on Statistical Signal Processing*, Nice, France.

[Noumir et al., 2012b] Noumir, Z., Honeine, P., and Richard, C. (2012b). Kernels for time series of exponential decay/growth processes. In *Proc. 22nd IEEE workshop on Machine Learning for Signal Processing*, Santander, Spain.

[Noumir et al., 2012c] Noumir, Z., Honeine, P., and Richard, C. (2012c). On simple one-class classification methods. In *Proc. IEEE International Symposium on Information Theory*, MIT, Cambridge (MA), USA.

[Noumir et al., 2012d] Noumir, Z., Honeine, P., and Richard, C. (2012d). One-class machines based on the coherence criterion. In *Proc. IEEE workshop on Statistical Signal Processing*, Ann Arbor, Michigan, USA.

[Noumir et al., 2012e] Noumir, Z., Honeine, P., and Richard, C. (2012e). Online one-class machines based on the coherence criterion. In *Proc. 20th European Conference on Signal Processing*, Bucharest, Romania.

[Oja, 1982] Oja, E. (1982). A simplified neuron model as a principal component analyzer. *J. Math. Biology*, 15 :267–273.

[Ou and Murphey, 2007] Ou, G. and Murphey, Y. (2007). Multi-class pattern classification using neural networks. *Pattern Recognition*, 40(1) :4–18.

[Paatero and Tapper, 1994] Paatero, P. and Tapper, U. (1994). Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2) :111–126.

[Pan et al., 2011] Pan, B., Lai, J., and Chen, W.-S. (2011). Nonlinear nonnegative matrix factorization based on Mercer kernel construction. *Pattern Recognition*, 44(10-11) :2800 – 2810.

[Patwari and Hero, 2004] Patwari, N. and Hero, A. (2004). Manifold learning algorithms for localization in wireless sensor networks. *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3 :857–860.

[Patwari et al., 2003] Patwari, N., Hero, A. O., Perkins, M., Correal, N. S., and O'Dea, R. J. (2003). Relative location estimation in wireless sensor networks. *IEEE Trans. on Signal Processing*, 51(8) :2137–2148.

[Platt et al., 2000] Platt, J. C., Cristianini, N., and Shawe-Taylor, J. (2000). Large margin DAGs for multiclass classification. In *Proc. of Neural Information Processing Systems, NIPS'99*, pages 547–553. MIT Press.

[Plaza and Chang, 2005] Plaza, A. and Chang, C.-I. (2005). An improved n-findr algorithm in implementation. In Shen, Sylvia S. ; Lewis, P. E., editor, *Proc. SPIE Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XI*, volume 5806, pages 298–306.

[Poggio et al., 2002] Poggio, T., Mukherjee, S., Rifkin, R., Rakhlin, A., and Verri., A. (2002). b. *Uncertainty in Geometric Computations*, pages 131–141.

[Predd et al., 2006] Predd, J. B., Kulkarni, S. R., and Poor, H. V. (2006). Distributed learning in wireless sensor networks. *IEEE Signal Processing Magazine*, 23(4) :56–69.

[Rabbat and Nowak, 2004] Rabbat, M. and Nowak, R. (2004). Distributed optimization in sensor networks. In *Proc. third international symposium on Information Processing in Sensor Networks (IPSN)*, pages 20–27, New York, USA. ACM.

[Rakotomamonjy et al., 2008] Rakotomamonjy, A., Bach, F., Canu, S., and Granvalet, Y. (2008). SimpleMKL. *Journal of Machine Learning Research*, 9 :2491–2521.

[Raksuntorn and Du, 2010] Raksuntorn, N. and Du, Q. (2010). Nonlinear spectral mixture analysis for hyperspectral imagery in an unknown environment. *IEEE Geoscience and Remote Sensing Letters*, 7(99) :836–840.

[Ralaivola and d'Alché-Buc, 2005] Ralaivola, L. and d'Alché-Buc, F. (2005). Time series filtering, smoothing and learning using the kernel Kalman filter. In *Proc. International Joint Conference on Neural Networks*, volume 3, pages 1449–1454.

[Rasmussen and Williams, 2006] Rasmussen, C. E. and Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

[Ratle et al., 2007] Ratle, F., Kanevski, M., Terrettaz-Zufferey, A. L., Esseiva, P., and Ribaux, O. (2007). A comparison of one-class classifiers for novelty detection in forensic case data. In *Proc. 8th international conference on Intelligent data engineering and automated learning*, pages 67–76, Berlin, Heidelberg. Springer-Verlag.

[Ray and Murray, 1996] Ray, T. W. and Murray, B. C. (1996). Nonlinear spectral mixing in desert vegetation. *Remote Sensing of Environment*, 55(1) :59–64.

[Ren and Chang, 2003] Ren, H. and Chang, C.-I. (2003). Automatic spectral target recognition in hyperspectral imagery. *IEEE Transactions on Aerospace and Electronic Systems*, 39(4) :1232–1249.

[Richard et al., 2009] Richard, C., Bermudez, J. C. M., and Honeine, P. (2009). Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3) :1058–1067.

[Richard et al., 2011] Richard, C., Chen, J., Honeine, P., and Bermudez, J. C. M. (2011). Filtrage adaptatif avec contrainte de non-négativité. principes de l'algorithme nn-lms et modèle de convergence. In *Actes du 23-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Bordeaux, France.

[Richard et al., 2010a] Richard, C., Ferrari, A., Amoud, H., Honeine, P., Flandrin, P., and Borgnat, P. (2010a). Statistical hypothesis testing with time-frequency surrogates to check signal stationarity. In *Proc. 35th IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas.

[Richard et al., 2008] Richard, C., Honeine, P., Snoussi, H., Essoloh, M., and Bermudez, J. C. M. (2008). Distributed learning in wireless sensor networks. In *5th Workshop on Sensor Networks (CNRS RECAP Sensor and Self-Organized Networks)*.

[Richard et al., 2010b] Richard, C., Honeine, P., Snoussi, H., Ferrari, A., and Theys, C. (2010b). Distributed learning with kernels in wireless sensor networks for physical phenomena modeling and tracking. In *Proc. IEEE International Geoscience and Remote Sensing Symposium*, Honolulu (Hawaii), USA.

[Rifkin, 2002] Rifkin, R. (2002). *Everything Old Is New Again : A Fresh Look at Historical Approaches in Machine Learning*. Phd thesis, Sloan School of Management Science : Massachusetts Institute of Technology.

[Rifkin and Klautau, 2004] Rifkin, R. and Klautau, A. (2004). In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5 :101–141.

[Rifkin et al., 2003] Rifkin, R., Yeo, G., and Poggio, T. (2003). Regularized least squares classification. In Suykens, J., Horvath, G., Basu, S., Micchelli, C., and Vandewalle, J., editors, *Advances in Learning Theory : Methods, Model and Applications*, volume 190 of *NATO Science Series III : Computer and Systems Sciences*, pages 131–154, Amsterdam, Netherlands. VIOS Press.

[Rogge et al., 2007] Rogge, D. M., Rivard, B., Zhang, J., Sanchez, A., Harris, J., and Feng, J. (2007). Integration of spatial-spectral information for the improved extraction of endmembers. *Remote Sensing of Environment*, 110(3) :287–303.

[Rosenblatt, 1958] Rosenblatt, F. (1958). The perceptron : a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6) :386.

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290 :2323–2326.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(Oct) :533–536.

[Saidé et al., 2013a] Saidé, C., Honeine, P., Lengellé, R., Richard, C., and Achkar, R. (2013a). Adaptation en ligne d'un dictionnaire pour les méthodes à noyau. In *Actes du 24-ème Colloque GRETSI sur le Traitement du Signal et des Images*, Brest, France.

[Saidé et al., 2013b] Saidé, C., Lengellé, R., Honeine, P., and Achkar, R. (2013b). Online kernel adaptive algorithms with dictionary adaptation for mimo models. *IEEE Signal Processing Letters*, 20(5) :535–538.

[Saidé et al., 2012] Saidé, C., Lengellé, R., Honeine, P., Richard, C., and Achkar, R. (2012). Dictionary adaptation for online prediction of time series data with kernels. In *Proc. IEEE workshop on Statistical Signal Processing*, Ann Arbor, Michigan, USA.

[Sanger, 1989a] Sanger, T. D. (1989a). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural Networks*, 2 :459–473.

[Sanger, 1989b] Sanger, T. D. (1989b). Optimal unsupervised learning in feedforward neural networks. Technical report, MIT, Cambridge, MA, USA.

[Schagen, 1979] Schagen, I. P. (1979). Interpolation in two dimensions—a new technique. *Journal of the Institute of Mathematics and its Applications*, 23(1) :53–59.

[Schölkopf et al., 2000] Schölkopf, B., Herbrich, R., and Williamson, R. (2000). A generalized representer theorem. Technical Report NC2-TR-2000-81, Royal Holloway College, Univ. of London, UK.

[Schölkopf et al., 2001] Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7) :1443–1471.

[Schölkopf et al., 1998] Schölkopf, B., Smola, A., and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5) :1299–1319.

[Schölkopf and Smola, 2001] Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

[Schraudolph et al., 2007] Schraudolph, N. N., Günter, S., and Vishwanathan, S. V. N. (2007). Fast iterative kernel pca. In *Advances in Neural Information Processing Systems*. MIT Press.

[Shang et al., 2003] Shang, Y., Ruml, W., Zhang, Y., and Fromherz, M. (2003). Localization from mere connectivity. In *Proc. MobiHoc*, Annapolis, USA.

[Shawe-Taylor and Cristianini, 2003] Shawe-Taylor, J. and Cristianini, N. (2003). Estimating the moments of a random vector with applications. In *Actes du 19ème Colloque GRETSI sur le Traitement du Signal et des Images*, volume I, pages 47–52, Paris. Invited Talk.

[Shawe-Taylor and Cristianini, 2004] Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.

[Sommerville, 1958] Sommerville, D. M. Y. (1958). *An Introduction to the Geometry of N Dimensions*. Dover, New York.

[Sonnenburg et al., 2008] Sonnenburg, S., Zien, A., Philips, P., and Ratsch, G. (2008). Poims : positional oligomer importance matrices–understanding support vector machine-based signal detectors. *Bioinformatics*, 24(13) :i6–14.

[Strang, 2003] Strang, G. (2003). *Introduction to Linear Algebra*. Wellesly-Cambridge Press, 3rd edition.

[Strohmer and Vershynin, 2009] Strohmer, T. and Vershynin, R. (2009). A randomized Kaczmarz algorithm with exponential convergence. *J. Fourier Anal. Appl.*, 15(2) :262–278.

[Suykens and Vandewalle, 1999a] Suykens, J. and Vandewalle, J. (1999a). Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3) :293–300.

[Suykens and Vandewalle, 1999b] Suykens, J. and Vandewalle, J. (1999b). Multiclass least squares support vector machines. In *Proc. International Joint Conference on Neural Networks*. World Scientific.

[Suykens et al., 2002] Suykens, J. A. K., Gestel, T. V., Brabanter, J. D., Moor, B. D., and Vandewalle, J. (2002). *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore.

[Szedmak and Shawe-Taylor, 2005] Szedmak, S. and Shawe-Taylor, J. (2005). Multiclass learning at one-class complexity. Technical Report, ISIS Group, Electronics and Computer Science.(Unpublished).

[Szedmak et al., 2006] Szedmak, S., Shawe-Taylor, J., and Parado-Hernandez, E. (2006). Learning via linear operators : Maximum margin regression ; multiclass and multiview learning at one-class complexity. Technical report, University of Southampton.

[Tax, 2001] Tax, D. (2001). *One-class classification ; Concept-learning in the absence of counter-examples*. Phd thesis, Advanced School for Computing and Imaging – Delft University of Technology.

[Tax and Duin, 2004] Tax, D. M. and Duin, R. P. W. (2004). Support vector data description. *Mach. Learn.*, 54 :45–66.

[Tax and Juszczak, 2003] Tax, D. M. J. and Juszczak, P. (2003). Kernel whitening for one-class classification. *International Journal of Pattern Recognition and Artificial Intelligence*, 17 :333–347.

[Tenenbaum et al., 2000] Tenenbaum, J. B., Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500) :2319–2323.

[Theiler et al., 1992] Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Doyne Farmer, J. (1992). Testing for nonlinearity in time series : the method of surrogate data. *Physica D : Nonlinear Phenomena*, 58(1-4) :77–94.

[Theodoridis et al., 2011] Theodoridis, S., Slavakis, K., and Yamada, I. (2011). Adaptive learning in a world of projections. *IEEE Signal Processing Magazine*, 28(1) :97–123.

[Theys et al., 2009] Theys, C., Dobigeon, N., Tourneret, J.-Y., and Lanteri, H. (2009). Linear unmixing of hyperspectral images using a scaled gradient method. In *IEEE 15th Workshop on Statistical Signal Processing*, pages 729–732.

[Tikhonov and Arsenin, 1977] Tikhonov, A. and Arsenin, V. (1977). *Solutions of ill-posed problems*. John Wiley, New York.

[Tohmé and Lengellé, 2011] Tohmé, M. and Lengellé, R. (2011). Maximum Margin One Class Support Vector Machines for multiclass problems. *Pattern Recognition Letters*, 32 :1652–1658.

[Torgerson, 1952] Torgerson, W. S. (1952). Multidimensional scaling : I. theory and method. *Psychometrika*, 17 :401–419.

[Tropp, 2004] Tropp, J. A. (2004). Greed is good : algorithmic results for sparse approximation. *IEEE Trans. Information Theory*, 50 :2231–2242.

[Tu and Sayed, 2021] Tu, S.-Y. and Sayed, A. H. (2021). Diffusion networks outperform consensus networks. In *Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP)*. Ann Arbor, MI, USA.

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA.

[Vapnik, 1998] Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York, NY, USA.

[Voloshynovskiy et al., 2011] Voloshynovskiy, S., Koval, O., Beekhof, F., and Holotyak, T. (2011). Information—theoretic multiclass classification based on binary classifiers. *J. Signal Process. Syst.*, 65(3) :413–430.

[Wahba, 1990] Wahba, G. (1990). *Spline Models for Observational Data*. Society for Industrial and Applied Math (SIAM), Philadelphia.

[Wang et al., 2006] Wang, D., Yeung, D. S., and Tsang, E. C. C. (2006). Structured one class classification. *IEEE Trans. on systems, Man, and Cybernetics, Part B*, 36 :1283–1295.

[Weston and Watkins, 1999] Weston, J. and Watkins, C. (1999). Support vector machines for multiclass pattern recognition. In *Proc Seventh European Symposium On Artificial Neural Networks*.

[Wild et al., 2004] Wild, S., Curry, J., and Dougherty, A. (2004). Improving non-negative matrix factorizations through structured initialization. *Pattern Recognition*, 37(11) :2217–2232.

[Williams, 2002] Williams, C. (2002). On a connection between kernel PCA and metric multidimensional scaling. *Mach. Learn.*, 46(1-3) :11–19.

[Winter, 1999] Winter, M. (1999). N-FINDR : an algorithm for fast autonomous spectral endmember determination in hyperspectral data : an algorithm for fast autonomous spectral endmember determination in hyperspectral data. *Proc. of SPIE : Imaging Spectrometry V*, 3753(10).

[Xiao et al., 2007] Xiao, J., Borgnat, P., and Flandrin, P. (2007). Testing stationarity with time-frequency surrogates. In *Proc. Eusipco*, pages 2020–2024, Poznan, Poland.

[Yamanishi and Vert, 2007] Yamanishi, Y. and Vert, J.-P. (2007). Kernel matrix regression. Technical report, http ://arxiv.org/abs/q-bio/0702054v1. (shorter version in Proc. 12th International Conference on Applied Stochastic Models and Data Analysis, 2007).

[Yin et al., 2012] Yin, H., Campan, F., Guépié, B. K., Noumir, Z., Fillatre, L., Honeine, P., Nikiforov, I., Richard, C., Snoussi, H., Jarrige, P.-A., and Morio, C. (2012). Vigires'eau : Surveiller un réseau de distribution d'eau potable. In *Workshop Interdisciplinaire sur la Sécurité Globale (WISG'12), (ANR - CSOSG)*, pages 1–8, Troyes, France.

[Youla and Webb, 1982] Youla, D. and Webb, H. (1982). Image restoration by the method of convex projections : Part 1 - theory. *IEEE Transactions on Medical Imaging*, 1(2) :81–94.

[Yukawa, 2010] Yukawa, M. (December 6–8, 2010). Adaptive filtering based on projection method. Lecture Notes.

[Zhang et al., 2006] Zhang, D., Zhou, Z., and Chen, S. (2006). Non-negative matrix factorization on kernels. In *Lecture Notes in Computer Science*, volume 4099, pages 404–412. Springer.

[Zhang et al., 2009] Zhang, Y., Meratnia, N., and Havinga, P. (2009). Adaptive and online one-class support vector machine-based outlier detection techniques for wireless sensor networks. In *Proc. of the IEEE 23rd International Conference on Advanced Information Networking and Applications Workshops/Symposia*, pages 990–995, Bradford, United Kingdom.

[Zheng et al., 2010a] Zheng, W., Lai, J., and Yuen, P. C. (2010a). Penalized preimage learning in kernel principal component analysis. *IEEE Transaction Neural Networks*, 21 :551–570.

[Zheng and Lai, 2006] Zheng, W.-S. and Lai, J.-H. (2006). Regularized locality preserving learning of pre-image problem in kernel principal component analysis. In *Proc. of the 18th International Conference on Pattern Recognition (ICPR)*, pages 456–459, Washington, DC, USA. IEEE Computer Society.

[Zheng et al., 2010b] Zheng, W.-S., Lai, J. H., and Yuen, P. C. (2010b). Penalized preimage learning in kernel principal component analysis. *IEEE Trans. on Neural Networks*.

[Zortea and Plaza, 2009] Zortea, M. and Plaza, A. (2009). Spatial preprocessing for endmember extraction. *IEEE Transactions on Geoscience and Remote Sensing*, 47(8) :2679–2693.

[Zymnis et al., 2007] Zymnis, A., Kim, S. J., Skaf, J., Parente, M., and Boyd, S. (2007). Hyperspectral image unmixing via alternating projected subgradients. In *Proc. of Asilomar*.