Méthodes à noyau pour l'analyse et la décision en environnement non-stationnaire

Paul Honeine

École Doctorale *Sciences des Systèmes Technologiques et Organisationnels* (SSTO) – Troyes

Méthodes à noyau pour l'analyse et la décision en environnement non-stationnaire

THÈSE

présentée et soutenue publiquement le 13 décembre 2007

pour l'obtention du

Doctorat de l'Université de technologie de Troyes

(spécialité Optimisation et Sûreté des Systèmes)

par

Paul HONEINE

Composition du jury

Président :	Patrick Flandrin	Directeur de Recherches au CNRS, ENS de Lyon
Rapporteurs :	Stéphane Canu	Professeur des Universités, INSA de Rouen
	Bruno Torrésani	Professeur des Universités, Université de Provence
Examinateurs :	Manuel Davy	Chargé de Recherches au CNRS (HdR), LAGIS Lille
	Rémy Gribonval	Chargé de Recherches INRIA (HdR), IRISA Rennes
	Cédric Richard	Professeur des Universités, UTT (Directeur de thèse)
	Hichem Snoussi	Maître de Conférences, UTT

Institut Charles Delaunay — FRE CNRS 2848
Université de technologie de Troyes

Mis en page avec la classe thloria.

Remerciements

Le chemin sinueux est devenu une belle aventure, merci.

Mes premiers mots ne peuvent qu'aller à la personne qui a le plus compté, pour moi et sur moi, au cours de cette thèse, celle qui a su m'insuffler son enthousiasme, sa curiosité et son engouement scientifique, le directeur de ma thèse, **Cédric Richard**, Professeur à l'Université de technologie de Troyes et Directeur du Laboratoire de Modélisation et Sûreté des Systèmes. Cette thèse n'aurait sans doute pas eu lieu s'il n'avait pas été à mes côtés, toujours prêt à m'aider, à me guider en éclairant le chemin sinueux de ce travail. C'est avec lui que la pierre angulaire de cet édifice a été posée, et c'est grâce à lui que j'ai pu poursuivre. MERCI.

Je suis reconnaissant à **Patrick Flandrin**, Directeur de Recherches au CNRS, et à **José Carlos M. Bermudez**, Professeur au Federal University of Santa Catarina au Brésil, pour l'intérêt qu'ils ont porté à mon travail. Je leur adresse mes plus vifs remerciements.

Je tiens également à remercier **Bernard Durr**, PDG de la société Sonalyse, pour m'avoir ouvert les portes de son entreprise dans le cadre d'une convention CIFRE avec l'UTT, et pour m'avoir transmis son enthousiasme incessant malgré les épreuves, jusqu'au dépôt de bilan de la société. J'adresse également une pensée aux sonalysiens **Thomas** Lebesnerais, **Emmanuel** Belotti, **Sébastien** Courtin, et **Nadia** Hammachi, qui m'ont accueilli avec tant de gentillesse et de soutien.

J'exprime mes remerciements à **Patrick Flandrin**, Directeur de Recherches au CNRS, pour l'honneur qu'il m'a fait en acceptant de présider le jury de cette soutenance de thèse. Je tiens également à exprimer mes sincères remerciements à **Stéphane Canu**, Professeur à l'INSA de Rouen, et à **Bruno Torrésani**, Professeur à l'Université de Provence, qui ont accepté la responsabilité de juger ce travail en qualité de rapporteurs. Je voudrais également remercier **Manuel Davy**, Chargé de Recherches au CNRS, **Rémy Gribonval**, Chargé de Recherches INRIA, et **Hichem Snoussi**, Maître de Conférences à l'UTT, qui ont participé à ce jury de thèse.

L'Université de technologie de Troyes et l'Institut Charles Delaunay (ICD) offrent un cadre privilégié pour l'épanouissement de la recherche. Pour cela, je tiens à remercier **Jacques Duchêne**, Directeur de l'ICD et **Antoine Grall**, Directeur du Pôle Recherche Opérationnelle, Statistiques Appliquées et Simulation (ROSAS), de m'avoir accueilli au sein de l'équipe.

Je remercie également **Régis Lengellé**, Directeur de l'Ecole Doctorale SSTO à l'UTT. Sans lui, cette expérience tellement enrichissante n'aurait pas pu voir le jour.

Mes remerciements vont également à **Pascale Denis** pour sa disponibilité, sa compétence, et pour toute l'aide qu'elle a su apporter depuis mon stage de DEA jusqu'au post-doc.

J'associe à mes remerciements **Marie-José Rousselet** et **Veronique Banse**, secrétaires du Pôle ROSAS, pour leur gentillesse, leur efficacité, leur disponibilité, et le soutien qu'elles m'ont apporté.

Je remercie l'Association Nationale de la Recherche Technique (ANRT) pour le financement CIFRE, et l'Agence Nationale de la Recherche (ANR) pour son soutien financier en fin de thèse.

Enfin une pensée émue à **Bassem Saasouh**, **Ibrahim Khoury**, et **Mireille Tohmé**. Sans même le savoir, ils m'ont épaulé bien plus qu'ils ne le sauront jamais. Merci pour les moments inoubliables. J'adresse également une pensée à mes collègues avec lesquels j'ai eu le plaisir de partager des moments d'amitié : **Georges** Farah, **Yasmine** Hani, **Hassan** Amoud, **Rida** Khatoun, **Hicham** Chehadé, **Nisrine** Jrad, **Farah** Mourad, et **Claire** Deeb. Merci à tous les anciens et nouveaux doctorants à l'UTT, et plus particulièrement à mes collègues de bureau **Mehdi** Essoloh, **Fouzi** Harrou, et **Jing** Teng.

Merci d'avoir consacré du temps à la lecture de ces remerciements, soucieux de rendre à César ce qui appartient à César.

Je dédie cette thèse à
ma mère et mon père

Table des matières

Introduction générale

Nombre d'applications ne permettent plus aujourd'hui de faire abstraction du caractère non-linéaire et non-stationnaire des systèmes dynamiques étudiés, qu'accompagne généralement une pénurie d'information statistique *a priori*. Ce contexte difficile nécessite le développement d'outils sans cesse plus sophistiqués, destinés à pallier les faiblesses des techniques originelles propres à l'analyse des systèmes linéaires et stationnaires. Ainsi l'analyse classique de Fourier échoue-t-elle souvent, bien qu'elle ait été moteur dans le développement de nombreuses disciplines. Rappelons qu'elle permet de représenter un signal temporel ou spatial dans le domaine fréquentiel en le décomposant en une somme de signaux tonals. Cette décomposition nécessite par essence que le signal soit stationnaire au sens où ses propriétés spectrales n'évoluent pas au cours du temps. Afin de remédier à cette lacune, divers prolongements de l'analyse de Fourier vers une configuration non-stationnaire ont été développés. Leur principe est résumé par le paradigme selon lequel il convient de décrire et de manipuler les signaux non-stationnaires dans un plan temps-fréquence. Une grande diversité d'espaces de représentation de ce type s'offre à présent à l'utilisateur, par exemple grâce aux décompositions linéaires telles que les transformées en ondelettes et de Fourier à court-terme, ou aux distributions d'énergie telles que le spectrogramme et la distribution de Wigner. Confrontée aux exigences d'un traitement en-ligne de signaux et systèmes non-stationnaires évoluant dans un contexte statistique inconnu, la discipline du traitement du signal a également connu de multiples développements dans le cadre des méthodes dites adaptatives. Rappelons que celles-ci ont pour vocation de poursuivre les évolutions de systèmes non-stationnaires, caractérisés par l'absence d'une relation analytique simple entre les entrées et sorties d'un système, ou entre les instants successifs d'un signal. La popularité de ces algorithmes tient à une simplicité conceptuelle et une implémentation efficace, principalement inhérente à l'usage de modèles linéaires. Notons cependant qu'il s'agit là d'une limite importante de ces approches à laquelle il conviendrait de remédier afin de pouvoir traiter une plus large classe de problèmes. Il est possible pour cela d'y reconnaître un problème d'apprentissage en-ligne classique, qui consiste en une approximation fonctionnelle à partir d'un flux de réalisations, auquel la littérature correspondante est susceptible de pouvoir répondre.

La théorie des noyaux reproduisants a permis le développement fulgurant d'une classe d'algorithmes d'apprentissage dont la formulation ne dépend pas de la nature des données traitées, ni de l'espace de représentation adopté pour résoudre les problèmes. Au-delà de ce caractère universel, celles que l'on range désormais sous le qualificatif de méthodes à noyau et dont les *Support Vector Machines* (SVM) sont le fer de lance doivent également leur succès à l'essor de la théorie de l'apprentissage statistique, au sein de laquelle la prédiction de leurs performances en généralisation fait aujourd'hui encore l'objet d'études approfondies. Ces techniques reposent sur un principe fondamental, *le coup du noyau*, qui permet de conférer un caractère non-linéaire à nombre de traitements originellement linéaires sans qu'il soit nécessaire de recourir à d'importants développements théoriques. Celui-ci a été appliqué avec une remarquable efficacité à de nombreux algorithmes de reconnaissance des formes sans que leur charge calculatoire s'en ressente. Parmi eux, on trouve l'ACP-à-noyau et l'AFD-à-noyau, des extensions non-linéaires des algorithmes classiques d'analyse en composantes principales (ACP) et d'analyse factorielle

discriminante (AFD). La non-linéarité souvent considérée dans la littérature correspond implicitement à une transformation de l'espace des données vers un espace de représentation de dimension plus élevée, sans aucune exigence quant à l'interprétation de ces représentations. Dans le cas de l'analyse et la classification de signaux non-stationnaires, un choix naturel de représentation est incontestablement le domaine temps-fréquence.

L'objectif de ce manuscrit est de proposer un nouveau cadre pour l'analyse et la décision en environnement non-stationnaire et en situation de pénurie d'information statistique, par une fertilisation croisée des domaines de l'analyse temps-fréquence, du traitement adaptatif du signal et de la reconnaissance des formes par méthodes à noyau.

Dans un premier temps, on propose une mise en œuvre des méthodes à noyau les plus performantes dans le domaine temps-fréquence, grâce à un choix approprié de noyau reproduisant. On profite à cette occasion des plus récentes avancées en reconnaissance des formes et en théorie de l'apprentissage statistique, ainsi que de la diversité des classes de distributions temps-fréquence existantes. On aboutit alors à de nouvelles techniques d'analyse et de décision propres aux signaux non-stationnaires, avec de faibles complexités calculatoires. On s'intéresse en particulier à l'apprentissage non-supervisé par l'adaptation de l'ACP-à-noyau à l'analyse des signaux dans le domaine temps-fréquence. Dans un cadre décisionnel supervisé, on traite des algorithmes d'AFD-à-noyau et de SVM que l'on met en œuvre dans le domaine temps-fréquence.

Le problème du choix de la représentation temps-fréquence optimale pour une application donnée a suscité d'amples études. Récemment, le domaine de la reconnaissance des formes lui a potentiellement apporté des éléments de réponse intéressants dans le cadre de la sélection de noyau reproduisant optimal pour une tâche de classification donnée. Parmi les différentes techniques existantes, le critère d'alignement noyau-cible présente l'avantage de mener à un noyau optimum sans recourir à des apprentissages répétés suivis de procédures de validation croisée. Nous adaptons ce critère au contexte temps-fréquence pour une sélection optimum des distributions. Dans ce contexte, une extension de l'algorithme classique d'optimisation des noyaux temps-fréquence à symétrie radialement Gaussienne dit *RGK* est présentée pour la résolution de problèmes de classification.

Confronté à un environnement non-stationnaire et dynamique, un apprentissage en-ligne peut s'avérer incontournable. Les méthodes à noyau 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. Nous abordons cette question par le biais de critères classiquement utilisés par la communauté ayant trait aux représentations parcimonieuses, que l'on étudie au jour des méthodes à noyau. Notre attention se porte en particulier sur le critère de cohérence d'un dictionnaire, qui nous permet un contrôle de la taille des modèles avec une complexité calculatoire linéaire. Nous établissons alors un lien avec l'entropie quadratique de Rényi notamment, ou encore la procédure d'ACP. Des exemples de mise en œuvre sont présentés dans le cadre de problèmes classiques de traitement du signal, dans le développement de nouveaux algorithmes de filtrage adaptatif non-linéaires en particulier.

Des applications à des signaux réels sont finalement considérées. Nous nous intéressons en particulier au débruitage de signaux magnétoencéphalographiques contaminés par l'activité cardiaque du sujet, à partir d'un enregistrement électrocardiographique. Les méthodes que nous proposons, qui s'inscrivent dans un cadre de filtrage adaptatif non-linéaire, conduisent à une amélioration notable des performances comparée à des techniques antérieures. Nous portons également notre attention sur l'électroencéphalogramme de sommeil, plus particulièrement sur l'analyse de l'un de ses événements transitoires qu'est le complexe K. Les techniques proposées pour l'étude de ce phénomène non-stationnaire tirent parti des méthodes à noyau tout en opérant dans le domaine temps-fréquence.

Plan de lecture du manuscrit

Le manuscrit s'articule en deux grandes parties. La première, du Chapitre 2 au Chapitre 6, concerne la mise en œuvre des méthodes à noyau dans le domaine temps-fréquence pour l'analyse et la classification de signaux non-stationnaires. La seconde partie, du Chapitre 7 au Chapitre 9, présente des méthodes d'apprentissage en-ligne pour des problèmes de modélisation de systèmes non-linéaires et non-stationnaires. Le contenu de chaque chapitre est rappelé en quelques lignes ci-dessous.

Chapitre 1 : Méthodes d'apprentissage statistique à noyau reproduisant

On y introduit brièvement les méthodes à noyau pour l'apprentissage statistique. Après avoir décrit le concept de noyaux reproduisants et les espaces de Hilbert associés, on présente les deux clés de voûte de ces méthodes que sont le coup du noyau et le Théorème de Représentation. On conclut ce chapitre par quelques exemples qui seront réétudiés par la suite.

Chapitre 2 : Distributions temps-fréquence pour l'analyse de signaux non-stationnaires

Ce chapitre a pour but de présenter les représentations temps-fréquence pour l'analyse de signaux non-stationnaires. On s'intéresse plus particulièrement aux distributions temps-fréquences de la classe de Cohen, ainsi qu'aux représentations linéaires telles que la transformée de Fourier à court-terme.

Chapitre 3 : Méthodes à noyau dans le domaine temps-fréquence

On pose les fondements des méthodes à noyau pour l'analyse de signaux non-stationnaires, par un choix approprié du noyau reproduisant. On décrit cette approche pour la distribution de Wigner, avant de l'étendre à d'autres distributions de la classe de Cohen.

Chapitre 4 : Analyse en composantes principales dans le plan temps-fréquence

Ce chapitre décrit l'analyse en composantes principales à noyau, et sa mise en œuvre dans le domaine temps-fréquence grâce à un choix approprié de noyau. La complexité calculatoire de cette approche est alors étudiée, et comparée aux techniques conventionnelles.

Chapitre 5 : Discrimination des signaux dans le domaine temps-fréquence

Dans le cadre de problèmes de discrimination de signaux non-stationnaires, on s'intéresse à l'analyse factorielle discriminante à noyau, que l'on adapte afin qu'elle opère dans le plan temps-fréquence.

Chapitre 6 : Distributions temps-fréquence optimales, une approche par alignement noyau-cible

On traite du problème du choix de l'espace de représentation pour un problème de classification. Après un passage en revue des différentes techniques existantes pour sélectionner un noyau reproduisant, on adopte le critère d'alignement noyau-cible. On modifie celui-ci afin de pouvoir procéder à la sélection de représentations temps-fréquence optimum en ajustant les paramètres, ou encore en combinant plusieurs espaces de représentation.

Chapitre 7 : Contrôle de complexité et critère de cohérence

Les méthodes à noyau conduisant à des modèles dont l'ordre est égal au nombre de données d'apprentissage, on s'intéresse à un critère de parcimonie afin de remédier à ce problème : la cohérence. On étudie ses propriétés et on établit des liens avec d'autres critères existants mais plus exigeants en temps de calcul.

Chapitre 8 : Apprentissage par filtrage adaptatif non-linéaire

Ce chapitre est dédié aux algorithmes d'apprentissage en-ligne, combinés au critère de cohérence. On décrit principalement deux algorithmes, de type moindres carrés récursif à noyau et projection affine à noyau. Un cas particulier de ce dernier est l'algorithme LMS normalisé à noyau, également présenté.

Chapitre 9 : Applications

On expérimente les algorithmes de filtrage adaptatif non-linéaires présentés précédemment. On s'intéresse à des applications biomédicales, le débruitage des signaux magnétoencéphalographiques et l'analyse des complexes K dans l'électroencéphalogramme de sommeil.

Annexe : Classification dans le domaine temps-fréquence par les SVM

On aborde les récentes avancées de la théorie de l'apprentissage statistique, en particulier le principe de minimisation du risque structurel. On présente l'algorithme des Support Vector Machines qui accompagne cette théorie, et son adaptation pour la classification de signaux non-stationnaires.

Produits de la recherche

Cette thèse a fait l'objet de plusieurs publications. Le tableau suivant en donne les références par ordre de leur apparition dans ce manuscrit, ainsi que les chapitres auxquels ils ont trait.

Citation	Référence	Chapitres concernés
[HRF05]	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*, Louvain-la-Neuve, Belgium, 2005.	3, 4, Annexe A
[HRF07]	P. Honeine, C. Richard, and P. Flandrin. Time-frequency learning machines. *IEEE Trans. Signal Processing*, 55 :3930–3936, July 2007.	3, 4, 5, Annexe A
[HRFP06]	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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006.	6
[HR07b]	P. Honeine and C. Richard. Signal-dependent time-frequency representations for classification using a radially gaussian kernel and the alignment criterion. In *Proc. IEEE Statistical Signal Processing (SSP)*, Madison, WI, USA, August 2007. In press.	6
[HR07a]	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. In press.	6
[HRB07b]	P. Honeine, C. Richard, and J. C. M. Bermudez. On-line nonlinear sparse approximation of functions. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, Nice, France, June 2007.	7, 8, 9
[HRB07a]	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. In press.	7, 8, 9
[RBH07]	C. Richard, J. C. M. Bermudez, and P. Honeine. Nonlinear kernel-based adaptive filtering with order controlled by a coherence criterion. *Submitted to IEEE Trans. Signal Processing*, 2007.	7, 8, 9

Chapitre 1

Méthodes d'apprentissage statistique à noyau reproduisant

Sommaire

Ce chapitre a pour objectif de présenter un cadre théorique au problème d'apprentissage statistique, qui consiste à déterminer une fonction à partir de données regroupées dans un ensemble d'apprentissage. Ce problème est mal-posé puisqu'il existe une infinité de fonctions continues qui vérifient les conditions discrètes induites par les données d'apprentissage. Ceci est par exemple le cas des problèmes de régression, où l'on cherche une fonction passant en certains points tandis qu'il en existe une infinité.

La théorie de la régularisation introduite par Tikhonov et Arsenin dans [TA77] propose une solution élégante à ce problème. Une régularisation de type Tikhonov permet de 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 [Aro50]. Ses propriétés sont exploitées par le Théorème de Représentation, initialement proposé pour les problèmes de régression par Kimeldorf et Wahba dans [KW71, Wah90], et récemment généralisé à d'autres problèmes d'apprentissage par Schölkopf *et coll.* dans [SHW00]. La simplicité des méthodes dites à noyau 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, 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 [ABR64] dans le cadre d'un problème de classification, et renforcé par Vapnik dans [Vap95] avec le théorie de l'apprentissage statistique dans un contexte plus général de classification et régression.

Dans ce chapitre, on traite d'une manière concise ces différents concepts, que l'on illustre avec certains exemples de méthodes à noyau que l'on détaillera au cours des chapitres suivants. Dans la Section 1.1, on introduit les méthodes d'apprentissage et la régularisation selon Tikhonov. Après avoir présenté succinctement les concepts de noyau reproduisant et d'espace de Hilbert associé dans la Section 1.2, on introduit les deux clés de voûte des méthodes à noyau dans la Section 1.3 : le coup du noyau et le Théorème de Représentation. On conclut le chapitre par des exemples de méthodes à noyau dans des problèmes d'analyse non-supervisée et supervisée.

1.1 Théorie de l'apprentissage statistique

En théorie de l'apprentissage statistique [Vap95], on cherche à déterminer un modèle qui traduit le mieux possible une relation entre les observations successives recueillies sur un système, ou encore entre ses entrées et sorties, à partir d'un ensemble d'apprentissage. On souhaite que le modèle ainsi élaboré soit généralisable à de nouvelles observations, ce qui nécessite une certaine connaissance *a priori* du comportement du système. Une telle information peut être incorporée grâce à un choix approprié d'espace d'hypothèses, ce dernier étant l'espace fonctionnel dans lequel la solution est recherchée.

1.1.1 Apprentissage statistique

Les méthodes d'apprentissage statistique peuvent être regroupées en deux classes principales : les méthodes supervisées et les méthodes non-supervisées.

Dans le cadre de l'apprentissage supervisé, on cherche la relation entre un compact $\mathcal{X}$ de $\mathbb{C}^l$, dit espace de données ou d'entrée, et un compact $\mathcal{Y}$ de $\mathbb{C}$, dit espace des réponses ou de sortie. Cette relation est décrite par la distribution de probabilité $P(x, y)$ définie pour tout couple $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Ainsi recherche-t-on la fonction ψ^* de $\mathcal{X}$ dans $\mathcal{Y}$ telle que $\psi^*(x)$ soit une estimation appropriée de la sortie y correspondant à la donnée x. L'optimalité de la fonction ψ^* sur toutes les fonctions ψ de $\mathcal{X}$ dans $\mathbb{C}$ est donnée par la minimisation d'une fonctionnelle de risque réelle de la forme

$$\int_{\mathcal{X}\times\mathcal{Y}} V(\psi(x), y)\, \mathrm{d}P(x, y),$$

où $\mathrm{d}P(x, y) = P(x, y)\, \mathrm{d}x\, \mathrm{d}y$, et V une fonction coût qui mesure l'erreur commise entre la sortie désirée y et la sortie estimée $\psi(x)$ pour tout couple $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Dans le cas particulier de la fonction de coût quadratique, $V(\psi(x), y) = (\psi(x) - y)^2$, la fonction minimisant cette expression est donnée par $\psi^*(x) = \int_{\mathcal{Y}} y\, \mathrm{d}P(y|x)$. Puisque $\mathcal{Y}$ est compact, on montre qu'un tel optimum existe. Toutefois, la distribution de probabilité P étant inconnue, l'optimum ne peut pas être obtenu directement. Celle-ci n'est en effet connue qu'à partir d'un ensemble fini de réalisations, appelé ensemble d'apprentissage, que l'on note $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ avec $(x_k, y_k) \in \mathcal{X} \times \mathcal{Y}$. En posant $P_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i)\, \delta(y - y_i)$, le problème d'optimisation se traduit par la minimisation du risque d'apprentissage, appelé aussi risque empirique, selon

$$\psi^* = \arg\min_{\psi} \frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i), y_i). \tag{1.1}$$

Pour un apprentissage non-supervisé, on se contente d'un compact $\mathcal{X}$ de $\mathbb{C}^l$, dit espace de données ou des observations. On cherche alors la relation entre les éléments de cet espace, que décrit la distribution de probabilité $P(x)$ pour tout $x \in \mathcal{X}$. La fonction recherchée ψ^* est alors obtenue en résolvant un problème d'optimisation de la forme

$$\psi^* = \arg\min_{\psi} \int_{\mathcal{X}} V(\psi(x))\, \mathrm{d}P(x),$$

portant sur toutes les fonctions ψ de $\mathcal{X}$ dans $\mathbb{C}$, avec V une fonction coût donnée. N'ayant à disposition qu'un ensemble d'apprentissage fini $\mathcal{A}_n = \{x_1, \ldots, x_n\}$ de réalisations échantillonnées selon P, cette dernière est estimée par $P_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)$. On recherche une fonction optimale en minimisant le risque empirique défini par

$$\psi^* = \arg\min_{\psi} \frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i)). \tag{1.2}$$

Comme cette expression est un cas particulier de (1.1) pour des étiquettes y_i supposées constantes, on considère dans la suite le cas plus général de l'apprentissage supervisé.

Dans les deux cas, il existe une infinité de fonctions ψ^* minimisant le risque empirique, donc vérifiant soit (1.1) pour l'apprentissage supervisé, soit (1.2) pour l'apprentissage non-supervisé. Le problème est dit alors mal-posé, dans le sens où l'ensemble d'apprentissage ne permet pas une reconstruction unique de la fonction recherchée. Pour autant, toutes les fonctions candidates ψ^* n'admettent pas les mêmes capacités en généralisation étant donné de nouvelles observations ne figurant pas dans l'ensemble d'apprentissage. On a alors recours à l'introduction d'hypothèses vis-à-vis de la fonction ψ^* recherchée afin de s'affranchir du caractère mal-posé du problème initial. Une contrainte faible et naturelle, au sens des phénomènes physiques par exemple, consiste à supposer que cette fonction est suffisamment régulière pour que de faibles variations des données produisent de légères fluctuations sur les sorties. Cette contrainte de régularité sur ψ^* permet alors d'interpréter le problème d'apprentissage comme un exercice d'approximation à partir de données bruités. D'autres contraintes, plus fortes, peuvent aussi être considérées préalablement à l'apprentissage, par exemple que le modèle recherché est linéaire ou quadratique.

1.1.2 Méthodes régularisées

Depuis les années 1960, plusieurs techniques de régularisation ont été proposées pour rendre un problème d'optimisation bien-posé, dont les régularisations d'Ivanov [Iva76], de Phillips [Phi62] et de Tikhonov [Tik63][1]. Ces techniques offrent un cadre mathématique général pour résoudre les problèmes d'optimisation (1.1) et (1.2) en restreignant l'espace de recherche des fonctions candidates aux fonctions à faibles oscillations. On considère un espace de Hilbert $\mathcal{H}$ de fonctions de $\mathcal{X}$ dans $\mathbb{C}$, auquel appartient la fonction recherchée ψ^*. Soit $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$ le produit scalaire associé à cet espace, et $\| \cdot \|_{\mathcal{H}}$ sa norme. La pénalisation d'Ivanov consiste à minimiser le risque empirique, avec une contrainte sur la complexité définie par $\|\psi\|_{\mathcal{H}} \leq \tau$ qui vise à pénaliser les solutions oscillantes. Pour mieux comprendre cette pénalisation restreignant l'espace $\mathcal{H}$ aux fonctions à normes réduites, il suffit de considérer par exemple l'espace de Banach $\mathcal{L}_1[a, b]$ des fonctions intégrables en valeur absolue sur l'intervalle $[a, b]$. La norme est alors définie par $\|\psi\|_1 = \int_a^b |\psi(x)| \, \mathrm{d}x$. Un autre type de régularisation concerne l'espace $\mathcal{L}_2[a, b]$ des fonctions d'énergie finie sur $[a, b]$, avec la norme quadratique $\|\psi\|_2^2 = \int_a^b |\psi(x)|^2 \, \mathrm{d}x$. La régularisation de Phillips [Phi62] consiste à minimiser la norme, sous contrainte que le risque empirique reste inférieur à un seuil donné. Ces modes de régularisation d'Ivanov et de Phillips sont équivalents à la régularisation de Tikhonov [Muk04].

Par l'usage de la technique des multiplicateurs de Lagrange, Tikhonov propose dans [TA77] de transformer le problème d'optimisation avec contrainte défini ci-dessus en un problème d'optimisation sans contrainte. La fonctionnelle de risque étant pénalisée, la solution est alors obtenue selon

$$\psi^* = \arg\min_{\psi \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta \, \|\psi\|_{\mathcal{H}}^2,$$

[1] Cette pénalisation est souvent connue sous le nom de Tikhonov–Phillips [MR97].

où η contrôle le compromis entre les deux termes. Le premier terme représente le risque empirique qui mesure l'adéquation entre les sorties estimées et les sorties désirées. Le second terme, de pénalisation, permet d'obtenir des solutions plus régulières. Sans celui-ci, le problème serait mal-posé puisqu'il existerait alors une infinité de fonctions qui minimisent le premier terme. La théorie de la régularisation *à la Tikhonov* des méthodes d'apprentissage statistiques a connu diverses avancées récentes depuis qu'elle a été introduite dans ce contexte par Poggio et Girosi [PG90], et plus récemment avec les techniques de SVM proposées par Vapnik [Vap95]. Une généralisation de cette pénalisation a également été considérée en remplaçant le terme $\|\psi\|^2_{\mathcal{H}}$ par $g(\|\psi\|^2_{\mathcal{H}})$, où $g(\cdot)$ est une fonction monotone croissante sur $\mathbb{R}_+$. La fonctionnelle de risque régularisée s'écrit sous la forme

$$\frac{1}{n}\sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta\, g(\|\psi\|^2_{\mathcal{H}}). \tag{1.3}$$

Le choix du paramètre de régularisation η, appelé souvent hyperparamètre en apprentissage statistique, est crucial pour contrôler le compromis entre erreur d'apprentissage et degré de régularité. Il est donc lié directement à l'erreur de généralisation et conditionne la convergence de l'algorithme. Dans ce manuscrit, nous n'aborderons pas son optimalité en précisant que celle-ci est étudiée par Vapnik [Vap95], ainsi que par Wahba, Lin et Zhang [WLZ00] pour un cas particulier de fonctions coût, celui des SVM. Un exemple plus général a été considéré plus récemment par Cucker et Smale dans [CS02b], ou encore dans [CH02] et [Cap06].

Un nombre important de méthodes d'apprentissage a été développé dans ce cadre. Celles-ci se différencient principalement par deux caractéristiques clés : d'une part la fonctionnelle de coût $V(f(x_i), y_i)$ à minimiser, et d'autre part l'espace fonctionnel $\mathcal{H}$ des fonctions candidates. Cette thèse traite d'une catégorie d'espaces fonctionnels particuliers, les espaces de Hilbert à noyau reproduisant. Les méthodes associées constituent ce qu'on appelle couramment les méthodes à noyau. Avant de continuer, il convient à présent de préciser qu'il existe une interprétation Bayésienne aux méthodes d'apprentissage statistique régularisées.

Sous un angle probabiliste

La communauté traitant de reconnaissance des formes peut être divisée en 2 groupes : les fréquentistes et les probabilistes (ou Bayésiens). Les premiers traitent les observations des phénomènes sans considérer les lois de probabilité les ayant engendrées. Les seconds cherchent à remonter aux distributions de probabilité à partir des observations par des techniques dites d'inférence. Bien que l'on adopte la première philosophie tout au long de ce manuscrit, il est toutefois intéressant de préciser qu'il existe un lien entre les deux approches, ou plus précisément une interprétation probabiliste des méthodes régularisées selon Tikhonov. On désigne par $P(\mathcal{A}_n|\psi)$ la probabilité conditionnelle d'avoir l'ensemble d'apprentissage $\mathcal{A}_n$ ayant la fonction ψ, et par $P(\psi|\mathcal{A}_n)$ la probabilité conditionnelle de ψ sachant $\mathcal{A}_n$. Par la règle de Bayes, la probabilité *a posteriori* $P(\psi|\mathcal{A}_n)$ est alors donnée par

$$P(\psi|\mathcal{A}_n) = P(\mathcal{A}_n|\psi)\, P(\psi), \tag{1.4}$$

où $P(\psi)$ désigne la probabilité *a priori* de la fonction ψ. On suppose dans la suite que les échantillons suivent le modèle $y_i = \psi(x_i) + e_i$, où le bruit e_i est distribué selon une loi normale de variance σ^2. On peut alors écrire la probabilité $P(\mathcal{A}_n|\psi)$ selon

$$P(\mathcal{A}_n|\psi) = \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \psi(x_i))^2\Big).$$

La probabilité *a posteriori* est donnée par (1.4), avec

$$P(\psi|\mathcal{A}_n) = \exp\Big(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \psi(x_i))^2\Big)\,P(\psi).$$

On a recours à la technique du maximum *a posteriori* pour estimer ψ^*, ce qui conduit à la minimisation de l'expression

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \psi(x_i))^2 + \frac{2\sigma^2}{n}\log(1/P(\psi)). \tag{1.5}$$

On retrouve les deux caractéristiques clés de l'apprentissage statistique : le premier terme de l'expression correspond au risque empirique de fonction coût quadratique $V(\psi(x_i), y_i) = (y_i - \psi(x_i))^2$. Le second terme est un terme de pénalisation. En supposant que la probabilité *a priori* puisse s'écrire sous la forme $P(\psi) = \exp(-g(\|\psi\|^2_{\mathcal{H}}))$, on retrouve la régularisation de Tikhonov dans (1.3), avec $2\sigma^2/n$ pour paramètre de régularisation. La fonction ψ^* obtenue par la minimisation du critère (1.5) est optimale sous réserve que le bruit suive une loi normale. Pour une densité de probabilité Laplacienne, $P(e_i) = (1/2)\exp^{-|e_i|}$, la fonction coût correspondante est le coût L_1, à savoir $V(\psi(x_i), y_i) = |y_i - \psi(x_i)|$. L'optimalité de la régression logistique avec $V(\psi(x_i), y_i) = \log(1 + \exp(-y_i\psi(x_i)))$ est étudiée dans [Zha04], avec d'autres fonctions coûts, ou encore dans [SS01] avec la maximisation de la vraisemblance. Plus généralement, l'approche par maximisation de la probabilité *a posteriori* est étudiée par Mackay [Mac03] dans le cadre des processus Gaussiens.

1.2 Noyau reproduisant

Dans le paragraphe précédent, nous avons proposé de restreindre l'espace de recherche de ψ à un espace de Hilbert à noyau reproduisant. Depuis les travaux précurseurs de Aronszajn dans [Aro50] sur les noyaux reproduisants, on a de plus en plus recours à ce type d'espaces, notamment depuis qu'ils ont été retenus pour la résolution de problèmes d'interpolation par Parzen [Par70], Kailath [Kai71] et Wahba [Wah90]. Plus récemment, on peut se référer aux travaux de Saito [SABO99]. On présente ici un aperçu succinct des noyaux reproduisants avant d'exposer à la section suivante les éléments clés qui ont contribué à leurs succès dans le cadre de problèmes d'apprentissage. Dans ce qui suit, on considère un espace mesurable $\mathcal{X}$ auquel on associe le produit scalaire $\langle\cdot\,,\cdot\rangle$ et la norme correspondante $\|\cdot\|^2$.

1.2.1 Espace de Hilbert à noyau reproduisant

Un noyau désigne une fonction $\kappa(\cdot,\cdot)$ de $\mathcal{X}\times\mathcal{X}$ dans $\mathbb{C}$, à symétrie Hermitienne, c'est-à-dire telle que $\kappa(x_i, x_j) = \overline{\kappa}(x_j, x_i)$ pour tout $x_i, x_j \in \mathcal{X}$. On rappelle les deux définitions fondamentales suivantes et les propriétés qui en découlent, en renvoyant le lecteur vers [Aro50, CS02c] pour plus de détails.

Définition 1 (Noyau défini positif)**.** *Un noyau κ est dit défini positif sur $\mathcal{X}$ s'il vérifie*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i\,\overline{a_j}\,\kappa(x_i, x_j) \geq 0 \tag{1.6}$$

pour tout $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ et $a_1, \ldots, a_n \in \mathbb{C}$.

Les noyaux définis positifs sont considérés comme une généralisation du produit scalaire. Ce dernier constitue en effet un noyau défini positif particulier, dont certaines des propriétés sont vérifiées par tout

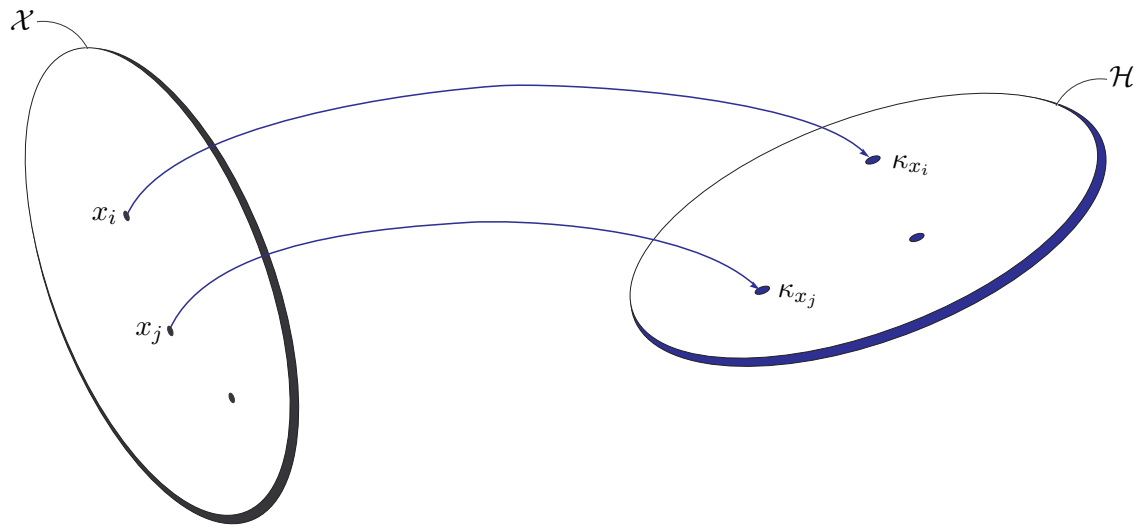


FIG. 1.1 – Espace des données $\mathcal{X}$ et espace $\mathcal{H}$ induit par le noyau reproduisant κ.

noyau défini positif. C'est le cas par exemple de l'inégalité de Cauchy-Schwartz : pour tout noyau $\kappa(\cdot,\cdot)$ défini positif sur $\mathcal{X}$, et pour tout $x_i, x_j \in \mathcal{X}$, on a

$$|\kappa(x_i, x_j)|^2 \leq \kappa(x_i, x_i)\,\kappa(x_j, x_j).$$

Bien que la linéarité ne soit évidemment pas vérifiée par les noyaux définis positifs en général, cette généralisation s'avère très intéressante comme on le montre à la Section 1.3.1 au travers du coup du noyau.

Définition 2 (Noyau reproduisant et RKHS)**.** *Soit $(\mathcal{H}, \langle\cdot\,,\cdot\rangle_{\mathcal{H}})$ un espace de Hilbert constitué de fonctions de $\mathcal{X}$ dans $\mathbb{C}$. La fonction $\kappa(x_i, x_j)$ de $\mathcal{X}\times\mathcal{X}$ dans $\mathbb{C}$ est le noyau reproduisant de $\mathcal{H}$, sous réserve que celui-ci en admette un, si et seulement si*

- *la fonction $\kappa_{x_i} : x_j \mapsto \kappa_{x_i}(x_j) = \kappa(x_i, x_j)$ appartient à $\mathcal{H}$, quel que soit $x_i \in \mathcal{X}$ fixé ;*
- *on a $\psi(x_j) = \langle\psi, \kappa_{x_j}\rangle_{\mathcal{H}}$ pour tout $x_j \in \mathcal{X}$ et $\psi \in \mathcal{H}$.*

On dit que $\mathcal{H}$ est un espace de Hilbert à noyau reproduisant, ou encore un RKHS, acronyme de Reproducing Kernel Hilbert Space.

D'après le premier point de la Définition 2, on désigne par κ_{x_i} et $\kappa(x_i, \cdot)$ la même fonction, qu'on appelle fonction noyau. Le second point de la Définition 2 est connu sous le nom de *propriété reproduisante*. Celle-ci, ainsi que l'existence du noyau reproduisant, sont liées directement au Théorème de Représentation de Riesz. En effet, soit une fonctionnelle $\psi(\cdot) \in \mathcal{H}$ et $\psi(x)$ son évaluation continue pour tout $x \in \mathcal{X}$. Selon ce théorème, il existe une fonction $h_x(\cdot) \in \mathcal{H}$ telle que $\psi(x) = \langle\psi, h_x\rangle_{\mathcal{H}}$. On retrouve la propriété reproduisante et, en posant $\kappa(x_i, x_j) = h_{x_i}(x_j)$ pour tout $x_i, x_j \in \mathcal{X}$, on retrouve le noyau reproduisant. On représente à la Figure 1.1 l'espace des données $\mathcal{X}$ et l'espace $\mathcal{H}$ induit par le noyau reproduisant κ.

Afin de fixer les idées, on présente une analogie noyau reproduisant et RKHS d'une part, et fonction indicatrice et espace $\mathcal{L}_2[a,b]$ d'autre part où toute fonction ψ vérifie $\|\psi\|_2^2 = \int_a^b |\psi(x)|^2\,\mathrm{d}x < \infty$. Pour tout $x \in [a,b]$, la fonction indicatrice $\mathbb{1}_x$ permet l'évaluation de toute fonction ψ en x, selon

$\mathbb{1}_x : \psi \mapsto \mathbb{1}_x\psi = \psi(x)$. On retrouve ainsi la propriété du noyau reproduisant dans un RKHS. L'espace $\mathcal{L}_2[a,b]$ n'est toutefois pas un RKHS car que la fonction indicatrice $\mathbb{1}_x$ n'appartient pas à cet espace, contrairement à la fonction noyau κ_x qui est dans son RKHS.

Corollaire 3 (Coup du noyau)**.** *Tout noyau reproduisant κ d'un espace de Hilbert $\mathcal{H}$ s'écrit comme un produit scalaire dans cet espace, selon*

$$\kappa(x_i, x_j) = \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}}$$

pour tout $x_i, x_j \in \mathcal{X}$.

Ce corollaire constitue une propriété fondamentale des noyaux reproduisants et des méthodes à noyau. Il sera étudié plus en détails dans la Section 1.3.1. Sa démonstration découle directement de la propriété reproduisante $\psi(x_j) = \langle \psi, \kappa_{x_j} \rangle_{\mathcal{H}}$ des noyaux. Il suffit pour cela de remplaçer ψ dans cette expression par la fonction noyau κ_{x_i}.

Théorème 4 (Moore-Aronszajn)**.** *A tout noyau défini positif, il correspond un RKHS unique, et réciproquement.*

Démonstration. On montre tout d'abord que tout noyau reproduisant est défini positif. Pour cela, il suffit de constater que $\sum_i \sum_j a_i \, \overline{a_j} \, \kappa(x_i, x_j) = \| \sum_i a_i \, \kappa_{x_i} \|^2$ ne peut être négatif. Réciproquement, on démontre que tout noyau défini positif κ est le noyau reproduisant d'un espace de Hilbert de fonctions de $\mathcal{X}$ dans $\mathbb{C}$. Pour cela, on considère l'espace vectoriel $\mathcal{H}'$ engendré par l'ensemble des fonctions $\{\kappa_{x_i}\}$ pour $x_i \in \mathcal{X}$. Ceci permet d'exprimer tout élément de $\mathcal{H}'$ comme une combinaison linéaire finie ($n < \infty$) de ces fonctions selon

$$\mathcal{H}' = \Big\{ \psi : \psi(\cdot) = \sum_{i=1}^{n} a_i \, \kappa_{x_i}(\cdot), \; x_i \in \mathcal{X}, \; a_i \in \mathbb{C} \Big\}.$$

A cet espace, on associe le produit scalaire

$$\langle \psi, \phi \rangle_{\mathcal{H}'} = \Big\langle \sum_{i=1}^{n} a_i \, \kappa_{x_i}, \sum_{j=1}^{n} \, b_j \, \kappa_{x_j} \Big\rangle,$$

avec $\psi = \sum_{i=1}^{n} a_i \, \kappa_{x_i}$ et $\phi = \sum_{j=1}^{n} b_j \, \kappa_{x_j}$ appartenant à l'espace $\mathcal{H}'$. En ré-arrangeant les sommations, et par le coup du noyau $\langle \kappa_{x_i}, \kappa_{x_j} \rangle = \kappa(x_i, x_j)$, on peut simplifier l'expression du produit scalaire en

$$\langle \psi, \phi \rangle_{\mathcal{H}'} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \, \overline{b_j} \, \kappa(x_i, x_j).$$

Muni de ce produit scalaire, l'espace ainsi construit est un espace pré-Hilbertien. Pour obtenir un espace de Hilbert, il suffit de le compléter conformément à [Aro50] de sorte que toute suite de Cauchy y converge. ■

Le Théorème de Moore-Aronszajn établit le lien entre noyau défini positif et RKHS. Par abus de langage, on remplace dans la suite la dénomination de noyau défini positif par noyau reproduisant. Avant de préciser leurs propriétés fondamentales dans le cadre de méthodes d'apprentissage régularisées, on présente les noyaux reproduisants les plus connus ainsi que certaines règles permettant de les combiner afin d'en obtenir de nouveaux.

1.2.2 Noyaux reproduisants : construction et exemples

On présente ici des techniques pour construire des noyaux reproduisants, et quelques exemples couramment utilisés. On renvoie le lecteur à [Vap95, Her02, STC04b] pour d'autres propriétés des noyaux reproduisants, ainsi que diverses règles permettant de les combiner.

Théorème 5. *Soit κ_1 et κ_2 deux noyaux reproduisants de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{C}$. La fonction $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ est un noyau reproduisant s'il est défini par une de ces expressions, pour tout $x_i, x_j \in \mathcal{X}$,*

1. *combinaison linéaire :* $\kappa(x_i, x_j) = \beta_1 \kappa_1(x_i, x_j) + \beta_2 \kappa_2(x_i, x_j)$, *pour tout* $\beta_1, \beta_2 \in \mathbb{R}_+$.
2. *décalage :* $\kappa(x_i, x_j) = \kappa_1(x_i, x_j) + c$, *pour tout* $c \in \mathbb{R}_+$.
3. *produit :* $\kappa(x_i, x_j) = \kappa_1(x_i, x_j)\, \kappa_2(x_i, x_j)$.
4. *exposant :* $\kappa(x_i, x_j) = \kappa_1(x_i, x_j)^p$, *pour tout* $p \in \mathbb{N}_+$.
5. *exponentiel :* $\kappa(x_i, x_j) = \exp(\kappa_1(x_i, x_j)/\sigma^2)$, *pour tout* $\sigma \in \mathbb{R}$.
6. *normalisation :* $\kappa(x_i, x_j) = \frac{\kappa_1(x_i, x_j)}{\sqrt{\kappa_1(x_i, x_i)\, \kappa_1(x_j, x_j)}}$.

Éléments de démonstration. A partir du Théorème 4 de Moore-Aronszajn, il suffit de démontrer pour chaque cas que le noyau κ est défini positif, à savoir $\sum_i \sum_j a_i \, \overline{a_j} \, \kappa(x_i, x_j) \geq 0$ pour tout $x_i, x_j \in \mathcal{X}$ et $a_i, a_j \in \mathbb{C}$. On peut facilement montrer cela pour les quatre premiers points. Pour le cinquième point, il suffit de décomposer l'exponentielle en un développement de noyau κ_1 de puissances différentes. Finalement, pour le noyau normalisé, il suffit de remarquer que le dénominateur est toujours positif, et que le numérateur est défini positif. ■

Les noyaux reproduisants classiques sont principalement regroupés en deux catégories : les noyaux radiaux et les noyaux projectifs. Ces derniers dépendent du produit scalaire, comme le noyau linéaire $\kappa(x_i, x_j) = \langle x_i, x_j \rangle$ ou encore les noyaux polynômiaux $\kappa(x_i, x_j) = \langle x_i, x_j \rangle^p$ de degré $p \in \mathbb{N}^*$. La règle du décalage (règle 2) conduit au noyau polynomial complet $\kappa(x_i, x_j) = (\langle x_i, x_j \rangle + c)^p$. Le noyau exponentiel, soit $\kappa(x_i, x_j) = \exp(\langle x_i, x_j \rangle / \sigma_0^2)$ où σ_0 est un paramètre fixe correspondant à la largeur du noyau, résulte de l'application de la règle 5 au noyau linéaire. Le plus connu des noyaux radiaux est sans doute le noyau Gaussien, défini par $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \sigma_0^2)$, que l'on obtient en normalisant (règle 6) le noyau exponentiel. Le noyau de Laplace, $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\| / \sigma_0)$, fait aussi parti de cette catégorie. D'autres noyaux reproduisants sont obtenus en combinant des noyaux classiques selon le Théorème 5. On présente en Annexe B les noyaux reproduisants les plus usités dans le cadre des méthodes de reconnaissance des formes. Toutefois, cette liste n'est pas exhaustive puisqu'en les combinant entre eux, on en obtient de nouveaux sous réserve de conserver le caractère défini positif.

Dans ce manuscrit de thèse, on propose d'utiliser d'autres types de noyaux, par exemple le noyau quadratique $\kappa(x_i, x_j) = |\langle x_i, x_j \rangle|^2$. On précise son intérêt au Chapitre 2 en en proposant une interprétation dans le cadre d'une analyse temps-fréquence. Au-delà de cet exemple, les règles de combinaison présentées permettent de concevoir des noyaux spécifiques à un problème donné. Ce concept est exploité au Chapitre 6, où l'on cherche les coefficients de pondération (β_1, β_2) pour qu'une combinaison linéaire de deux noyaux soit plus performante que les noyaux initiaux dans la résolution d'un problème de classification.

1.3 Au-delà du modèle linéaire : méthodes à noyau

Les méthodes à noyau sont pour la plupart issues d'algorithmes linéaires auxquels on a pu appliquer les deux résultats clés que sont le coup du noyau [ABR64] et le Théorème de Représentation [Wah90, SHW00]. Dans ce qui suit, on note $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un ensemble d'apprentissage donné avec $x_i \in \mathcal{X}$ les données d'entrée et $y_i \in \mathbb{R}$ les sorties désirées.

1.3.1 Coup du noyau

On rappelle la propriété fondamentale introduite au Corollaire 3. Tout noyau reproduisant s'écrit

$$\kappa(x_i, x_j) = \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}}$$

quels que soient $x_i, x_j \in \mathcal{X}$, avec $\mathcal{H}$ l'espace de Hilbert induit par ce noyau. Ainsi, le noyau $\kappa(x_i, x_j)$ fournit le produit scalaire dans $\mathcal{H}$ des images κ_{x_i} et κ_{x_j} de toute paire d'éléments x_i et x_j de $\mathcal{X}$, sans qu'il soit nécessaire d'expliciter ces images. Ce principe, connu sous le nom du *coup du noyau* ou *kernel trick* en anglais, permet de transformer les méthodes linéaires de traitement de données en des méthodes non-linéaires, sous réserve qu'elles puissent s'exprimer uniquement en fonction de produits scalaires des observations. Pour cela, il suffit de remplacer chacun de ces produits scalaires $\langle x_i, x_j \rangle$, qui n'est autre que le noyau linéaire, par un noyau non-linéaire $\kappa(x_i, x_j)$. Ainsi la structure des algorithmes demeure-t-elle inchangée, et le surcoût calculatoire dû à l'évaluation des noyaux négligeable.

On souligne l'importance du noyau pour le calcul implicite d'un produit scalaire dans l'espace $\mathcal{H}$. Cet espace est très souvent de dimension $d_{\mathcal{H}}$ supérieure à celle de l'espace des observations $d_{\mathcal{X}}$. En utilisant par exemple un noyau Gaussien $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2/\sigma_0^2)$, l'espace induit est de dimension infinie. Sans la mise en œuvre du coup du noyau, la détermination du produit scalaire dans de tels espaces serait impossible.

Géométrie associée

L'intérêt du coup du noyau va au-delà d'une simple évaluation de produit scalaire, en incluant l'évaluation de distances et d'angles dans l'espace RKHS. La distance dans $\mathcal{H}$ entre 2 fonctions κ_{x_i} et κ_{x_j} s'exprime par $\|\kappa_{x_i} - \kappa_{x_j}\|^2_{\mathcal{H}} = \langle \kappa_{x_i} - \kappa_{x_j}, \kappa_{x_i} - \kappa_{x_j} \rangle_{\mathcal{H}}$. En développant cette expression, on obtient

$$\|\kappa_{x_i} - \kappa_{x_j}\|^2_{\mathcal{H}} = \kappa(x_i, x_i) - 2\text{Ré}\{\kappa(x_i, x_j)\} + \kappa(x_j, x_j), \tag{1.7}$$

où Ré$(\cdot)$ désigne la partie réelle. Ainsi peut-on calculer la distance entre les éléments de cet espace sans qu'il soit nécessaire de les expliciter. On retrouve l'esprit du coup du noyau, qui permet selon cette expression de transformer toute méthode linéaire qui ne dépendrait uniquement que des distances entre les différents éléments en une méthode non-linéaire basée sur les noyaux. Ceci est le cas de la méthode des k-plus-proches-voisins par exemple. Puisque pour tout $x \in \mathcal{X}$ on a $\kappa(x, x) = \langle \kappa_x, \kappa_x \rangle_{\mathcal{H}} = \|\kappa_x\|^2_{\mathcal{H}}$, on peut aussi réécrire l'équation (1.7) pour un noyau à valeurs réelles selon

$$\kappa(x_i, x_j) = \frac{1}{2}\left(\|\kappa_{x_i}\|^2_{\mathcal{H}} + \|\kappa_{x_j}\|^2_{\mathcal{H}} - \|\kappa_{x_i} - \kappa_{x_j}\|^2_{\mathcal{H}}\right).$$

Cette équation montre que le noyau $\kappa(x_i, x_j)$ est une mesure de similitude entre x_i et x_j, celle-ci étant l'opposé du carré de la distance entre leurs images dans l'espace fonctionnel à deux termes additifs près. Pour le calcul de l'angle entre deux fonctions noyau κ_{x_i} et κ_{x_j}, il suffit de réécrire le cosinus de l'angle en termes de noyau selon

$$\cos(\kappa_{x_i}, \kappa_{x_j}) = \frac{\langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}}}{\|\kappa_{x_i}\|_{\mathcal{H}} \|\kappa_{x_j}\|_{\mathcal{H}}} = \frac{\kappa(x_i, x_j)}{\sqrt{\kappa(x_i, x_i)\kappa(x_j, x_j)}}.$$

Une fois de plus, le coup du noyau permet d'exprimer une mesure dans $\mathcal{H}$ sans qu'il soit nécessaire d'exhiber les éléments de cet espace.

Transformation unitaire

Bien que l'espace de Hilbert $\mathcal{H}$ induit par un noyau reproduisant donné κ soit unique, rien ne s'oppose à l'élaboration d'un tout autre espace muni du même produit scalaire. En effet, pour tout opérateur unitaire $\mathbf{U}$ de $\mathcal{H}$ vers $\mathcal{H}_{\mathbf{U}}$, le produit scalaire dans ce dernier est

$$\langle \mathbf{U}\kappa_{x_i}, \mathbf{U}\kappa_{x_j}\rangle_{\mathcal{H}_{\mathbf{U}}} = \langle \kappa_{x_i}, \kappa_{x_j}\rangle_{\mathcal{H}},$$

qui n'est autre que le noyau reproduisant $\kappa(x_i, x_j) = \langle \kappa_{x_i}, \kappa_{x_j}\rangle_{\mathcal{H}}$. On dit alors que les espaces $\mathcal{H}$ et $\mathcal{H}_{\mathbf{U}}$ sont isomorphes, ou encore que le RKHS $\mathcal{H}$ est unique à un isomorphisme près. Pour plus de détails sur l'équivalence des espaces induits par le même noyau, on renvoie le lecteur intéressé vers [Ste02,MNY06].

1.3.2 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 [KW71, Wah90]. Plus récemment, il a été repris dans le cadre de la résolution de problèmes inverses [Kur04], ainsi qu'en théorie de l'apprentissage [CS02c]. La formulation suivante du Théorème de Représentation et sa démonstration pour différents types de fonctions coûts sont principalement dues à Schölkopf *et coll.* [SHW00].

Théorème 6. *Soient $\mathcal{X}$ un compact, $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un ensemble d'apprentissage donné avec $x_i \in \mathcal{X}$ l'ensemble des données et $y_i \in \mathbb{C}$ l'ensemble des sorties désirées, V une fonction coût arbitraire et $g(\cdot)$ une fonction monotone croissante sur $\mathbb{R}+$. Soit $\mathcal{H}$ un espace de Hilbert induit par le noyau κ défini positif sur $\mathcal{X}$. Toute fonction $\psi^* \in \mathcal{H}$ minimisant la fonctionnelle de risque régularisée*

$$\frac{1}{n}\sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta\, g(\|\psi\|^2_{\mathcal{H}}), \tag{1.8}$$

peut s'écrire sous la forme

$$\psi^*(\cdot) = \sum_{j=1}^{n} \alpha_j^*\, \kappa(x_j, \cdot). \tag{1.9}$$

Démonstration. Soit $\mathcal{H}_n$ le sous-espace de $\mathcal{H}$ engendré par les fonctions $\{\kappa(x_1, \cdot), \ldots, \kappa(x_n, \cdot)\}$, c'est-à-dire

$$\mathcal{H}_n = \Big\{\psi \in \mathcal{H} : \psi(\cdot) = \sum_{j=1}^{n} \alpha_j\, \kappa(x_j, \cdot),\ \alpha_1, \ldots, \alpha_n \in \mathbb{C}\Big\}.$$

Toute fonction ψ de $\mathcal{H}$ admet une et une seule décomposition en deux contributions, l'une appartenant à l'espace $\mathcal{H}_n$ et l'autre qui lui est orthogonale. On peut en effet écrire

$$\psi = \psi^* + \psi^{\perp},$$

avec $\psi^* \in \mathcal{H}_n$ et $\psi^{\perp}$ la composante orthogonale telles que $\langle \psi^{\perp}, \kappa_{x_i}\rangle_{\mathcal{H}} = 0$ pour tout $i = 1, \ldots, n$. La propriété reproduisante permet d'évaluer ψ en x_i selon l'expression

$$\begin{aligned} \psi(x_i) &= \langle \psi, \kappa_{x_i}\rangle_{\mathcal{H}} \\ &= \sum_{j=1}^{n} \alpha_j \langle \kappa_{x_j}, \kappa_{x_i}\rangle_{\mathcal{H}} + \langle \psi^{\perp}, \kappa_{x_i}\rangle_{\mathcal{H}}. \end{aligned}$$

Puisque le second terme s'annule par orthogonalité, on obtient l'expression

$$\psi(x_i) = \sum_{j=1}^{n} \alpha_j \kappa(x_i, x_j).$$

Comme les évaluations de ψ en chaque point de l'ensemble d'apprentissage ne dépendent que des coefficients $\{\alpha_1, \ldots, \alpha_n\}$, le risque empirique (1.8) ne dépend pas de la composante orthogonale $\psi^\perp$. En minimisant ce risque, on obtient la classe des fonctions équivalentes dans $\mathcal{H}$ telle que deux fonctions ψ et ϕ appartiennent à la même classe si et seulement si $\psi(x_i) = \phi(x_i)$ pour tout $i = 1, \ldots, n$. Il reste à déterminer $\psi^\perp$ pour une classe de fonctions équivalentes donnée afin de minimiser le terme régularisant. Par le théorème de Pythagore dans $\mathcal{H}$ appliqué à ψ, soit $\|\psi\|^2_{\mathcal{H}} = \|\psi^*\|^2_{\mathcal{H}} + \|\psi^\perp\|^2_{\mathcal{H}}$, le terme régularisant $g(\|\psi\|^2_{\mathcal{H}})$ dans (1.8) s'écrit

$$g(\|\psi\|^2_{\mathcal{H}}) = g\Big(\big\|\sum_{j=1}^{n} \alpha_j \, \kappa_{x_j}\big\|^2_{\mathcal{H}} + \big\|\psi^\perp\big\|^2_{\mathcal{H}}\Big).$$

Comme $g(\cdot)$ est monotone croissante, la fonction qui minimise l'expression ci-dessus, pour une classe de fonctions équivalentes donnée, doit vérifier $\|\psi^\perp\|^2_{\mathcal{H}} = 0$. ■

L'importance de ce théorème réside dans l'existence d'une solution unique à une fonctionnelle de coût régularisée, celle-ci pouvant s'exprimer comme un développement en série fini de fonctions noyau. La minimisation de cette fonction coût (1.8) se ramène à un problème d'optimisation à n dimensions, celui de la détermination des coefficients optimaux $\alpha_1^*, \ldots, \alpha_n^* \in \mathbb{C}$.

1.4 Exemples de méthodes à noyau

Pour fixer les idées, nous introduisons ici quelques exemples de méthodes d'apprentissage statistiques, d'autres faisant l'objet de l'Annexe C. Ces approches seront étudiées plus en détails dans la suite de ce manuscrit, ainsi que leurs mises en œuvre temps-fréquence pour résoudre des problèmes d'analyse et de classification en environnement non-stationnaire. Pour ces méthodes, la fonctionnelle de risque à minimiser est pénalisée selon le principe de Tikhonov, soit

$$\frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta \, g(\|\psi\|^2_{\mathcal{H}}),$$

où $\mathcal{H}$ désigne un RKHS de noyau reproduisant donné. Comme évoqué précédemment, cette régularisation pénalise les variations importantes de la fonction ψ. Pour s'en convaincre, il suffit de borner l'évaluation de ψ en tout point $x \in \mathcal{X}$ selon

$$|\psi(x)|^2 = |\langle \psi, \kappa_x \rangle|^2 \leq \|\psi\|^2 \, \|\kappa_x\|^2 = \|\psi\|^2 \, \kappa(x, x), \tag{1.10}$$

à partir de la propriété reproduisante et de l'inégalité de Cauchy-Schwartz. En majorant la norme dans un RKHS par $\|\psi\| \leq \tau$, on obtient une borne supérieure pour les valeurs de cette fonction en tout point de $\mathcal{X}$, avec $|\psi(x)| \leq M\tau$ où $M = \sqrt{\kappa(x, x)}$.

Pour une analyse non-supervisée, on cherche à extraire la structure sous-jacente des observations, l'ensemble d'apprentissage étant constitué de données non-étiquetées $\mathcal{A}_n = \{x_1, \ldots, x_n\}$. Parmi les méthodes existantes, l'analyse en composantes principales est incontestablement la plus connue. Celle-ci

détermine un jeu d'axes orthogonaux tel que la variance de l'ensemble d'apprentissage projeté sur celui-ci est maximum. La formulation classique du problème consiste à atténuer la contrainte de régularisation dans l'expression (1.3), en imposant une norme unité aux axes recherchés, similaire à la régularisation de type Ivanov [Iva76]. Le premier axe principal s'obtient alors par maximisation de la fonction coût $|\psi(x_i)|^2$ où $\psi(x_i) = \langle \psi, \kappa_{x_i} \rangle_{\mathcal{H}}$ représente la projection de κ_{x_i} sur l'axe principal défini par ψ dans $\mathcal{H}$. Le risque empirique à minimiser est alors donné par

$$-\frac{1}{n}\sum_{i=1}^{n} |\psi(x_i)|^2,$$

sous la contrainte de normalisation $\|\psi\|_{\mathcal{H}}^2 = 1$. Les axes principaux suivants sont construits à partir de la même fonction coût tout en étant orthogonaux au premier, et entre eux. Dans cette formulation, on a supposé que les fonctions noyau κ_{x_i} sont centrées dans l'espace $\mathcal{H}$. Dans le cas contraire, la fonction coût s'écrit selon $|\psi(x_i) - \frac{1}{n}\sum_{j=1}^{n} \psi(x_j)|^2$, comme on le montre au Chapitre 4.

Dans le cas d'un problème de discrimination, l'ensemble d'apprentissage est formé des données x_i et de leurs étiquettes y_i. La fonction ψ^* recherchée vise à permettre l'identification de l'étiquette à partir d'une donnée. Ceci doit alors être vrai ou presque pour tout couple (x_i, y_i) de l'ensemble d'apprentissage, ce qui correspond à la contrainte $\psi^*(x_i) = y_i$. Pour un problème de classification à 2 classes, les étiquettes peuvent être codées selon $y_i \in \{\pm 1\}$, ce qui se traduit alors par la contrainte $y_i \psi^*(x_i) = 1$. Ainsi cherche-t-on à minimiser la fonctionnelle risque suivante

$$\frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i \psi(x_i)).$$

On retrouve le coût de la marge diffuse, connue par coût charnière ou *soft margin* en anglais, qui a contribué au succès des Support Vector Machines (SVM). Le terme de régularisation s'écrit sous la forme $\eta\, g(\|\psi\|_{\mathcal{H}}^2)$ où le paramètre de pénalisation η contrôle le compromis entre la complexité de la solution et l'erreur d'apprentissage. On aborde au Chapitre 5 et à l'Annexe A les problèmes de discrimination et de classification de signaux non-stationnaires.

Plus généralement, on peut s'intéresser au problème de moindres carrés correspondant à la minimisation du risque empirique

$$\frac{1}{n}\sum_{i=1}^{n} (y_i - \psi(x_i))^2.$$

Par l'usage de la pénalisation quadratique $\|\psi\|_{\mathcal{H}}^2$, on montre au Chapitre 7 que la solution de ce problème est obtenue par la résolution d'un système linéaire. Différentes techniques à complexité linéaire peuvent être considérées pour le résoudre, comme dans [Rif02] où une technique de gradient conjugué est proposée. Contrairement au cas des SVM esquissé ci-dessus, on montre que la solution dépend de tous les éléments de l'ensemble d'apprentissage, les coefficients α_i du développement en noyau issu du Théorème de Représentation n'ayant aucune raison d'être nuls. Afin d'être en mesure de répondre à des problèmes d'apprentissage en-ligne où l'on est confronté à un flux de données, il est nécessaire de se préoccuper de la parcimonie de la solution. Ce problème est traité aux Chapitres 7, 8 et 9.

Première partie

Méthodes à noyau pour l’analyse et la classification de signaux non-stationnaires

Chapitre 2

Distributions temps-fréquence pour l'analyse de signaux non-stationnaires

Sommaire

Du domaine temporel au domaine fréquentiel, en passant par le plan temps-fréquence, il existe une multitude d'espaces de représentation à-même d'offrir un cadre mathématique général pour l'analyse des signaux. Les signaux non-stationnaires nécessitent la mise en œuvre d'outils spécifiques parmi lesquels émergent les distributions temps-fréquence de la classe de Cohen. Celles-ci constituent une alternative intéressante aux techniques d'analyse spectrale classiques en incluant explicitement une dimension supplémentaire d'évolution temporelle. Toutes déclinées par filtrage de la distribution de Wigner, les distributions de la classe de Cohen offrent une variété d'espaces de représentation à la mesure de la diversité des objectifs visés par les utilisateurs. Par exemple, il est possible de privilégier l'intelligibilité de l'information délivrée par une représentation en y limitant les manifestations de bruits et autres termes interférentiels nuisant à sa lisibilité. On peut encore être à la recherche d'une représentation favorisant la résolution d'un problème de décision en maximisant le contraste entre les hypothèses.

Ce chapitre est essentiellement dédié à cette classe de distributions temps-fréquence, sans toutefois négliger les transformations linéaires classiques. Il est divisé en deux parties. Dans un premier temps, on présente des rappels sur les transformations linéaires, en particulier celles de Fourier, de Fourier à court-terme et en ondelettes. Puis on aborde les distributions temps-fréquence covariantes en temps et en fréquence. On commence par la distribution de Wigner avant de s'intéresser aux principales distributions temps-fréquence de la classe de Cohen. Le problème du choix de la distribution optimale est alors posé. On referme ce chapitre avec un rappel sur le problème de discrétisation des représentations.

2.1 Analyse linéaire : Fourier, Fourier à court-terme et ondelettes

2.1.1 Transformée de Fourier

On débute ce chapitre par quelques rappels de définitions et de propriétés élémentaires relatives à l'analyse de Fourier. La définition de la transformée de Fourier d'un signal $x \in \mathcal{L}_2(\mathbb{R})$ est donnée par

$$\widehat{x}(f) = \int x(t)\, \mathrm{e}^{-2j\pi ft}\, \mathrm{d}t.$$

Réciproquement, sa transformée inverse s'exprime par

$$x(t) = \int \widehat{x}(f)\, \mathrm{e}^{+2j\pi ft}\, \mathrm{d}f.$$

La transformée de Fourier et son inverse établissent une relation univoque entre le domaine temporel et le domaine fréquentiel. La transformée de Fourier est une isométrie de $\mathcal{L}_2(\mathbb{R})$ dans lui-même vérifiant les identités de Parseval et Plancherel :

$$\int \widehat{x}_i(f)\, \overline{\widehat{x}_j}(f)\, \mathrm{d}f = \int x_i(t)\, \overline{x_j}(t)\, \mathrm{d}t \quad \text{et} \quad \int |\widehat{x}_i(f)|^2\, \mathrm{d}f = \int |x_i(t)|^2\, \mathrm{d}t.$$

Celles-ci s'expriment sous forme de produits scalaires dans l'espace $\mathcal{L}_2(\mathbb{R})$ selon

$$\langle \widehat{x}_i, \widehat{x}_j \rangle = \langle x_i, x_j \rangle \quad \text{et} \quad \|\widehat{x}_i\|^2 = \|x_i\|^2.$$

En représentant un signal dans le domaine temporel ou fréquentiel, l'information liée à l'autre domaine n'est plus directement accessible. En particulier, la représentation fréquentielle ne permet pas une localisation temporelle des composantes fréquentielles. Réciproquement, une représentation temporelle n'autorise pas d'interprétation fréquentielle explicite. Aucune des deux représentations ne permet non plus d'analyser la durée de composantes fréquentielles données ou d'accéder à leur loi de modulation. Ces informations nécessitent la mise en œuvre d'outils spécifiques parmi lesquels émergent les représentations conjointes temps-fréquence, dont il va être question à présent.

2.1.2 Transformée de Fourier à court-terme

La plus connue des représentations temps-fréquence linéaires est certainement la transformée de Fourier à court-terme. Celle-ci consiste en une succession de transformées de Fourier du signal fenêtré à différents instants, produisant ainsi des spectres localisés temporellement. Étant donné w cette fenêtre d'analyse, la transformée de Fourier à court-terme est définie par

$$F_x(t,f) = \int x(\tau)\, \overline{w}(\tau - t)\, \mathrm{e}^{-j2\pi f\tau}\ \mathrm{d}\tau. \tag{2.1}$$

En posant $w_{t,f}(\tau) = w(\tau - t)\, \mathrm{e}^{-j2\pi f\tau}$ une version de w translatée au temps t et modulée en fréquence de f, l'expression (2.1) devient $F_x(t,f) = \int x(\tau)\, \overline{w_{t,f}}(\tau)\, \mathrm{d}\tau$, soit $F_x(t,f) = \langle x, w_{t,f} \rangle = \langle \widehat{x}, \widehat{w_{t,f}} \rangle$. Afin de reconstruire le signal à partir de cette représentation, on a recours à l'expression

$$x(\tau) = \iint F_x(t,f)\, w(\tau - t)\, \mathrm{e}^{+j2\pi f\tau}\, \mathrm{d}f\, \mathrm{d}t.$$

Le choix et la taille de la fenêtre d'analyse déterminent les propriétés de la représentation, en particulier le compromis entre localisations temporelle et fréquentielle. Ce dernier répond au principe d'incertitude d'Heisenberg-Gabor qui impose une borne inférieure à la concentration d'un atome $w_{t,f}$ dans le

plan temps-fréquence telle $\Delta t\,\Delta f \geq \frac{1}{4\pi}$, où Δt et Δf désignent les largeurs, mesurées par les variances, de l'atome [FS97, Mal98]. Si la fonction Gaussienne définie par $w(t) = C\,\mathrm{e}^{at^2+j2\pi f_0 t}$ avec C, a et f_0 des paramètres est la seule fenêtre qui vérifie l'égalité, le choix de sa taille n'en reste pas moins crucial. Les signaux concentrés temporellement nécessitent des fenêtres d'analyse courtes, tandis que les fenêtres de longue durée sont mieux adaptées aux signaux à composantes sinusoïdales.

Il existe des identités à la Parseval-Plancherel pour la transformée de Fourier à court-terme. On montre aisément que l'énergie d'un signal $x \in \mathcal{L}_2(\mathbb{R})$ peut se mettre sous la forme

$$\int |x(t)|^2\,\mathrm{d}t = \frac{1}{\|w\|^2} \iint |F_x(t,f)|^2\,\mathrm{d}t\,\mathrm{d}f.$$

On peut également montrer l'identité de Parseval pour la transformée de Fourier à court-terme

$$\iint F_{x_i}(t,f)\,\overline{F_{x_j}}(t,f)\,\mathrm{d}t\,\mathrm{d}f = \|w\|^2 \int x_i(t)\,\overline{x_j}(t)\,\mathrm{d}t. \tag{2.2}$$

En utilisant une fenêtre analysante normalisée dans $\mathcal{L}_2(\mathbb{R})$, on obtient une transformation unitaire.

2.1.3 Transformée en ondelettes

La décomposition en ondelettes continue est une représentation atomique appréciée. Elle repose sur une translation d'un temps t et une dilatation/compression en échelle d'un facteur a d'une ondelette w telle que $w_{t,a}(\tau) = \frac{1}{\sqrt{a}}\,w(\frac{\tau-t}{a})$.

Le changement d'échelle s'écrit $a = f_0/f$, où f_0 désigne la fréquence centrale de l'ondelette mère. La transformée en ondelettes du signal $x \in \mathcal{L}_2(\mathbb{C})$ s'exprime selon

$$O_x(t,a) = \int x(\tau)\,\frac{1}{\sqrt{a}}\,\overline{w}\left(\frac{\tau-t}{a}\right)\,\mathrm{d}\tau. \tag{2.3}$$

Pour qu'une telle quantité soit effectivement une représentation, il est souhaitable qu'elle soit inversible. Ceci se traduit par une condition d'admissibilité [Cal64, GM84] peu restrictive sur l'ondelette mère

$$C_w = \int |\widehat{w}(f)|^2 \frac{\mathrm{d}f}{|f|} < +\infty. \tag{2.4}$$

Des exemples classiques d'ondelette sont l'ondelette de Morlet[2] définie par une fenêtre Gaussienne modulée $w(t) = C\ \mathrm{e}^{-t^2/2+j2\pi f_0 t}$, et le chapeau mexicain donné par la dérivée seconde de la Gaussienne, c'est-à-dire $w(t) = C\,(1-t^2)\,\mathrm{e}^{-t^2/2}$ avec C une constante de normalisation. La condition d'admissibilité sur l'ondelette mère permet de définir une identité de type Parseval-Plancherel [SN96], à savoir

$$\iint O_{x_i}(t,a)\,\overline{O_{x_j}}(t,a)\frac{\mathrm{d}t\,\mathrm{d}a}{a^2} = C_w \int x_i(t)\,\overline{x_j}(t)\,\mathrm{d}t, \tag{2.5}$$

où $\mathrm{d}t\,\mathrm{d}a/a^2$ est la mesure naturelle associée au groupe affine des translations et dilatations [Fla98]. Pour démontrer l'équation (2.5), il suffit de remarquer que la transformée de Fourier en t de $O_x(t,a)$ s'exprime

[2] L'ondelette de Morlet n'est pas admissible *stricto sensu*. En effet, la condition d'admissibilité implique que $\widehat{w}(0) = \int_{-\infty}^{+\infty} w(t)\,\mathrm{d}t = 0$, ce qui n'est pas vérifié pour l'ondelette de Morlet. Toutefois, cette intégrale est très petite si f_0 est assez grand.

selon $\widehat{O_x}(f,a) = \widehat{x}(f)\sqrt{a}\,\overline{\widehat{w}}(af)$, ce qui permet d'écrire[3]

$$\begin{aligned}\iint O_{x_i}(t,a)\,\overline{O_{x_j}}(t,a)\frac{\mathrm{d}t\,\mathrm{d}a}{a^2} &= \int \widehat{x_i}(f)\,\overline{\widehat{x_j}}(f)\left[\int |\widehat{w}(af)|^2\frac{\mathrm{d}a}{a}\right]\mathrm{d}f \\ &= \int \widehat{x_i}(f)\,\overline{\widehat{x_j}}(f)\,C_w\,\mathrm{d}f \\ &= C_w\int x_i(t)\,\overline{x_j}(t)\,\mathrm{d}t,\end{aligned}$$

où la dernière égalité résulte de l'identité de Parseval. En utilisant des ondelettes telles que $C_w = 1$, on obtient une transformation unitaire.

2.2 Distributions temps-fréquence quadratiques

Les résolutions temporelle et fréquentielle des décompositions atomiques présentées ci-dessus sont déterminées par la localisation temps-fréquence des atomes d'analyse considérés. Idéalement, on souhaiterait aboutir à une densité d'énergie dans le plan temps-fréquence sans perte de résolution. Lorsqu'elles sont de forme quadratique, les distributions temps-fréquence sont susceptibles d'offrir une interprétation en termes de densité d'énergie. On compte la distribution de Wigner parmi celles-ci qui, avec sa résolution temps-fréquence optimale, sera étudiée dans la section suivante. On s'intéressera ensuite à la classe de Cohen des distributions covariantes par translation dans le plan temps-fréquence. On s'attardera enfin sur un type particulier de distributions temps-fréquence, les distributions adaptées au signal étudié et paramétrées selon un profil radialement Gaussien.

2.2.1 Distribution de Wigner

Étudiée par Wigner en 1932 dans le contexte de la mécanique quantique [Wig32], et introduite dans la communauté du traitement du signal par Ville [Vil48], la distribution de Wigner est un puissant outil pour l'analyse des signaux non-stationnaires. Elle est définie par

$$W_x(t,f) = \int x(t+\tau/2)\,\overline{x}(t-\tau/2)\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}\tau, \tag{2.6}$$

et correspond donc à la transformée de Fourier en τ de la fonction d'autocorrélation instantanée définie par $p(t,\tau) = x(t+\tau/2)\,\overline{x}(t-\tau/2)$. Celle-ci étant à symétrie Hermitienne, soit $p(t,\tau) = \overline{p}(t,-\tau)$, la distribution de Wigner est à valeurs réelles sur tout le domaine de définition. Par l'identité de Parseval, on peut réécrire l'expression (2.6) en fonction de la transformée de Fourier du signal x selon

$$W_x(t,f) = \int \widehat{x}(f+\nu/2)\,\overline{\widehat{x}}(f-\nu/2)\,\mathrm{e}^{+j2\pi t\nu}\,\mathrm{d}\nu.$$

Il est clair que $W_x \in \mathcal{L}_2(\mathbb{R}^2)$ pour tout $x \in \mathcal{L}_2(\mathbb{R})$. A titre d'exemple, la distribution de Wigner d'un signal à modulation fréquentielle linéaire est présenté à la Figure 2.1. Le signal de taille 64 comprend une composante fréquentielle qui croît de 0.1 à 0.4, en fréquence normalisée.

La distribution de Wigner vérifie un nombre important de propriétés mathématiques désirables [Fla98]. Outre le fait qu'elle est à valeurs réelles, elle préserve l'énergie, respecte les marginales temporelles et fréquentielles, et est covariante par translation en temps et en fréquence. De plus, elle permet une localisation parfaite des signaux à modulation de fréquence linéaire. Une propriété importante

[3] On définit la variable d'échelle a dans $\mathbb{R}$.

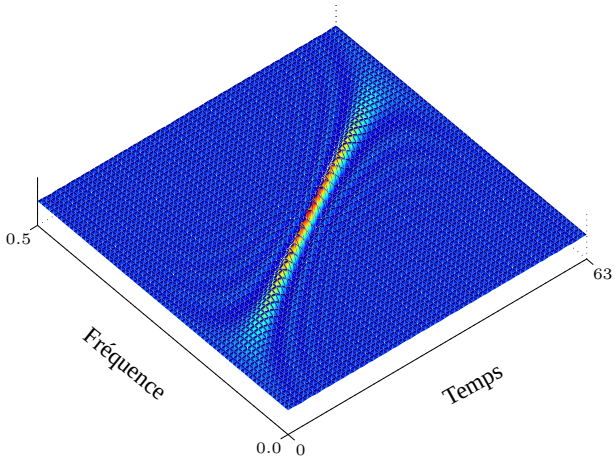


FIG. 2.1 – Représentation de Wigner d'un signal complexe à modulation fréquentielle linéaire. On retrouve la composante fréquentielle qui varie de 0.1 à 0.4 Hz

de la distribution de Wigner est son unitarité, particulièrement intéressante dans le contexte de notre étude comme on le verra par la suite. Cette conservation du produit scalaire entre domaines temporel et temps-fréquence, dite identité de Moyal, s'exprime par [CM80, Moy49]

$$\left| \int x_i(t)\, \overline{x_j}(t)\, \mathrm{d}t \right|^2 = \iint W_{x_i}(t,f)\, W_{x_j}(t,f)\, \mathrm{d}t\, \mathrm{d}f. \tag{2.7}$$

L'identité de Moyal (2.7) s'exprime également dans le domaine Doppler-retard, celui-ci étant le dual au sens de la transformée de Fourier du plan temps-fréquence. On définit pour cela la fonction d'ambiguïté à bande étroite A_x d'un signal x dans le plan Doppler-retard par l'expression

$$A_x(\nu,\tau) = \int x(t+\tau/2)\, \overline{x}(t-\tau/2)\, \mathrm{e}^{-j2\pi\nu t}\, \mathrm{d}t \tag{2.8}$$

et on réécrit la formule de Moyal par

$$\left| \int x_i(t)\, \overline{x_j}(t)\, \mathrm{d}t \right|^2 = \iint A_{x_i}(\nu,\tau)\, \overline{A_{x_j}}(\nu,\tau)\, \mathrm{d}\tau\, \mathrm{d}\nu. \tag{2.9}$$

Par cette dualité, les propriétés de la distribution de Wigner dans le domaine temps-fréquence sont aussi conservées par la fonction d'ambiguïté dans le domaine Doppler-retard.

Bien que la distribution de Wigner vérifie de nombreuses propriétés, sa lisibilité est généralement compromise par la présence de nombreux termes d'interférence. Ceux-ci introduisent de plus des composantes négatives qui ne permettent pas d'interpréter cette distribution comme une densité d'énergie. On illustre ce phénomène dans la Figure 2.2 par la distribution de Wigner et la fonction d'ambiguïté d'un signal comportant deux atomes Gaussiens. On distingue les termes d'interférence à mi-distance entre les deux composantes du signal dans le plan temps-fréquence. Dans le domaine Doppler-retard, ils sont diamétralement opposés et les deux composantes d'intérêt sont localisées à l'origine du plan. Afin de remédier à ce problème de lisibilité, diverses distributions faisant intervenir un lissage ont été proposées. Elles appartiennent à une vaste famille de distributions temps-fréquence : la classe de Cohen.

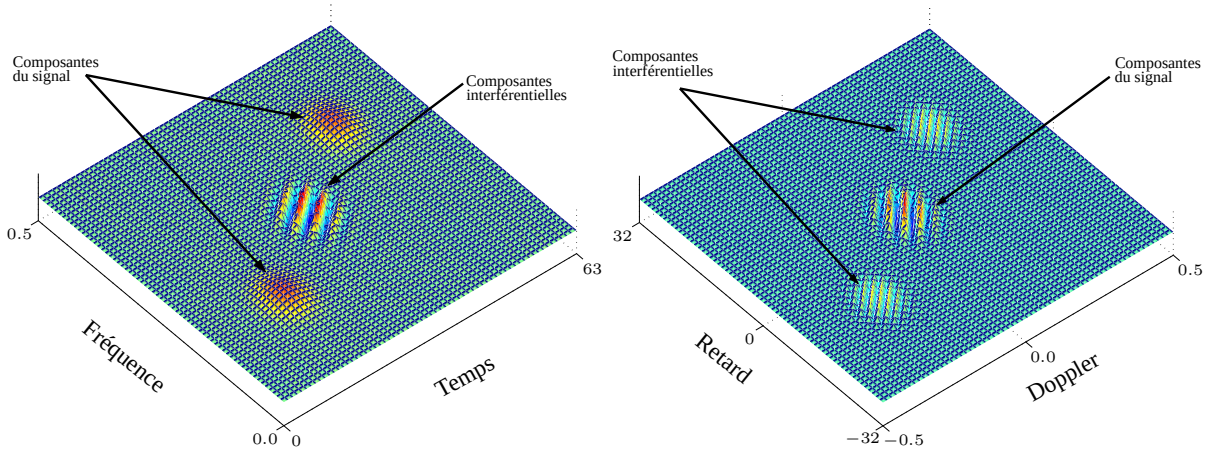


FIG. 2.2 – La distribution de Wigner en (a) et (la partie réelle de) la fonction d'ambiguïté en (b) d'un signal comportant deux atomes Gaussiens. On retrouve les composantes relatives aux atomes, ainsi que les termes d'interférence dans les plans temps-fréquence et Doppler-retard.

2.2.2 Distributions de la classe de Cohen

La classe de Cohen correspond à l'ensemble des distributions temps-fréquence quadratiques et covariantes vis-à-vis des translations temporelles et fréquentielles. Les distributions de cette classe sont définies par

$$C_x^{\Pi}(t,f) = \iint \Pi_{\mathrm{dr}}(\nu,\tau)\; A_x(\nu,\tau)\; \mathrm{e}^{-j2\pi(f\tau - t\nu)}\,\mathrm{d}\nu\,\mathrm{d}\tau, \tag{2.10}$$

où Π_{dr} est une fonction de paramétrisation exprimée dans le plan Doppler-retard, et $\Pi_{\mathrm{dr}}(\nu,\tau)\; A_x(\nu,\tau)$ est appelée fonction caractéristique. Par transformée de Fourier, il existe une définition équivalente qui, dans le plan temps-fréquence, s'exprime par

$$C_x^{\Pi}(t,f) = \iint \Pi_{\mathrm{tf}}(t-t', f-f')\; W_x(t',f')\,\mathrm{d}t'\,\mathrm{d}f', \tag{2.11}$$

avec Π_{tf} la fonction de paramétrisation dans le plan temps-fréquence. Celle-ci se déduit de Π_{dr} grâce à la transformée de Fourier par

$$\Pi_{\mathrm{tf}}(t,f) = \iint \Pi_{\mathrm{dr}}(\nu,\tau)\; \mathrm{e}^{j2\pi(f\tau + t\nu)}\,\mathrm{d}\nu\,\mathrm{d}\tau.$$

Lorsqu'il n'y aura pas d'ambiguïté sur la fonction de paramétrisation, on notera $C_x(t,f) = C_x^{\Pi}(t,f)$.

Chaque distribution de la classe de Cohen est entièrement caractérisée par sa fonction de paramétrisation Π. On retrouve dans le Tableau 2.1 les éléments les plus connus de cette classe ainsi que leur fonction de paramétrisation. La fonction $U(\cdot)$ désigne l'échelon unité de Heaviside. On reconnaît la distribution de Wigner, que l'on obtient à partir de l'expression générale (2.10) en prenant $\Pi_{\mathrm{dr}}(\nu,\tau) = 1$. Ceci revient à noter $\Pi_{\mathrm{tf}}(t,f) = \delta(t)\,\delta(f)$ dans l'expression (2.11), où δ désigne l'impulsion de Dirac. On retrouve certaines distributions temps-fréquence classiques comme la pseudo-Wigner-lissée, construite à partir de deux fenêtres w et v pour un double lissage en temps et en fréquence. On parle alors de paramétrisation à variables séparables. Le spectrogramme quant à lui est obtenu pour $\Pi_{\mathrm{dr}}(\nu,\tau) = \overline{A_w}(\nu,\tau)$. Il correspond au module au carré de la transformée de Fourier à court-terme. Malgré sa positivité et sa simplicité de mise en œuvre, sa faible résolution due à un lissage excessif demeure son plus grand défaut.

	Nom	$\Pi_{\mathrm{dr}}(\nu,\tau)$	$C_x^{\Pi}(t,f)$
Wigner et ses versions lissées	Wigner	1	$\int x(t+\tau/2)\,\overline{x}(t-\tau/2)\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}\tau$
	Pseudo-Wigner	$w(\tau)$	$\int w(\tau)\,x(t+\tau/2)\,\overline{x}(t-\tau/2)\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}\tau$
	Pseudo-Wigner-lissée	$\widehat{w}(\nu)\,v(\tau)$	$\int \overline{v}(\tau)\left[\int w(t'-t)\,x(t'+\tau/2)\,\overline{x}(t'-\tau/2)\,\mathrm{d}t'\right]\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}\tau$
Rihaczek et ses sous-produits	Rihaczek	$\mathrm{e}^{j\pi\nu\tau}$	$x(t)\,\overline{\widehat{x}}(f)\,\mathrm{e}^{-j2\pi ft}$
	Margenau-Hill Ackroyd	$\cos(\pi\nu\tau)$	$\mathrm{R}\{x(t)\,\overline{\widehat{x}}(f)\,\mathrm{e}^{-j2\pi ft}\}$
	Page-Levin	$\mathrm{e}^{\pm j\pi\nu\lvert\tau\rvert}$	$\pm\left\lvert\int x(\tau)\,U(\pm(t-\tau))\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}\tau\right\rvert^2$
Distributions à interférences réduites	Choï-Williams	$\mathrm{e}^{-\frac{1}{2}(\pi\nu\tau/\sigma)^2}$	$\iint \frac{\sigma}{\lvert\tau\rvert}\,\mathrm{e}^{-2\sigma^2(t'-t)^2/\tau^2}x(t'+\tau/2)\,\overline{x}(t'-\tau/2)\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}t'\,\mathrm{d}\tau$
	Butterworth	$\frac{1}{1+(\pi\nu\tau/\sigma)^{2l}}$	$\int \frac{\sqrt{\sigma}}{2\lvert\tau\rvert}\,\mathrm{e}^{-\lvert t'\rvert\sigma 2/\lvert\tau\rvert}x(t+t'+\tau/2)\,\overline{x}(t+t'-\tau/2)\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}t'\,\mathrm{d}\tau$
	Born-Jordan	$\frac{sin\pi\nu\tau}{\pi\nu\tau}$	$\int \left[\frac{1}{\lvert\tau\rvert}\int_{t-\lvert\tau\rvert/2}^{t+\lvert\tau\rvert/2} x(t'+\tau/2)\,\overline{x}(t'-\tau/2)\,\mathrm{d}t'\right]\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}\tau$
	Zhao-Atlas-Marks	$w(\tau)\lvert\tau\rvert\frac{sin\pi\nu\tau}{\pi\nu\tau}$	$\int w(\tau)\int_{t-\lvert\tau\rvert/2}^{t+\lvert\tau\rvert/2} x(t'+\tau/2)\,\overline{x}(t'-\tau/2)\,\mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}t'\,\mathrm{d}\tau$
	Spectrogramme	$\overline{A_w}(\nu,\tau)$	$\left\lvert\int x(\tau)\,\overline{w}(\tau-t)\;\mathrm{e}^{-j2\pi f\tau}\;\mathrm{d}\tau\right\rvert^2$

TAB. 2.1 – Exemples de distributions temps-fréquence de la classe de Cohen, avec leur fonction de paramétrisation.

La fonction de paramétrisation définit les propriétés que vérifie la distribution. On peut rechercher par exemple des représentations qui soient réelles, qui conservent les distributions marginales, ou encore qui vérifient l'égalité de Moyal. Cette dernière contrainte d'unitarité est vérifiée par certaines distributions, dont celles de Wigner, de Page-Levin et de Rihaczek. On montre que cette propriété est vérifiée à condition que $|\Pi_{\mathrm{dr}}(\nu,\tau)| = 1$. Pour le démontrer, il suffit de remarquer que selon l'expression (2.10) et par l'identité de Parseval, on obtient

$$\iint C_{x_i}^{\Pi}(t,f)\,\overline{C_{x_i}^{\Pi}}(t,f)\,\mathrm{d}t\,\mathrm{d}f = \iint |\Pi_{\mathrm{dr}}(\nu,\tau)|^2 A_{x_i}(\nu,\tau)\,\overline{A_{x_j}}(\nu,\tau)\,\mathrm{d}\nu\,\mathrm{d}\tau.$$

La condition est obtenue en comparant cette expression à (2.9). On renvoie le lecteur vers [HBB92,Fla98] pour de plus amples informations sur la relation entre les propriétés de la représentation et les contraintes sur sa paramétrisation.

De nombreuses propriétés peuvent ainsi être envisagées, mais toujours au détriment de quelques autres. Ceci est le cas pour la réduction des termes d'interférence, qui se pratique en renonçant à une bonne localisation des composantes dans le plan temps-fréquence. Pour s'en convaincre, il suffit de considérer une fonction de paramétrisation Π correspondant à un filtrage passe-bas bidimensionnel dans le plan des ambiguïtés et destiné à limiter la présence de termes distants de l'origine. Par l'expression (2.11), on note que ce filtrage correspond à un lissage dans le domaine temps-fréquence Π_{tf}. Parmi ces distributions à interférences réduites, on retrouve celles de paramétrisation de la forme $\Pi_{\mathrm{dr}}(\nu\times\tau)$, comme les distributions de Choï-Williams, de Butterworth et de Born-Jordan. La plupart de ces distributions admettent un ou plusieurs paramètres de réglage, permettant par exemple de définir le degré de lissage. Ceci est le cas du spectrogramme, des distributions de Wigner lissée et pseudo-lissée avec le choix de la taille et du type de fenêtre de lissage.

Si l'on s'est contenté de présenter la classe de Cohen, il convient de préciser qu'il existe moultes autres variantes parmi les distributions temps-fréquence. Par exemple, Davy et Doncarli [DD98, Dav04]

ont proposé l'usage de distributions normalisées telles que

$$C_x^{\Pi,\mathrm{p}}(t,f) = \frac{|C_x^{\Pi}(t,f)|}{\iint |C_x^{\Pi}(t',f')|\,\mathrm{d}t'\,\mathrm{d}f'}. \tag{2.12}$$

où C^{Π} désigne une distribution quelconque de la classe de Cohen. La motivation principale derrière cette normalisation est de conférer un caractère probabiliste aux distributions ainsi définies, permettant par là-même l'usage de divergences telles que celle de Kolmogorov afin de répondre à des problèmes de classification de signaux. On montre toutefois au Chapitre 6 qu'au sens du critère d'alignement, celles-ci s'avèrent moins bien adaptées à la résolution de problèmes décisionnels que les distributions non-normalisées.

2.2.3 Distribution optimale : la paramétrisation à profil radialement Gaussien

Bien que la sélection de représentations vérifiant certaines propriétés théoriques puisse être systématisée par le choix d'une fonction de paramétrisation adéquate, rien ne garantit qu'elles soient pour autant bien adaptées aux signaux étudiés. On peut recourir dans ce cas à une paramétrisation qui s'exprime en fonction des signaux, et l'optimiser au sens d'un critère propre à un problème de classification ou de suppression d'interférences donné. On renvoie le lecteur vers [Suc04] pour un bilan sur les représentations construites à partir du signal. Le premier type de problème sera traité dans le Chapitre 6. On présente ici le second problème qui vise à trouver un compromis entre la résolution temps-fréquence d'une part, et la suppression des termes interférentiels d'autre part, par optimisation d'une fonction de paramétrisation de profil donné. Afin de simplifier la présentation, la même notation sera utilisée pour désigner la fonction en coordonnées rectangulaires et polaires, c'est-à-dire $\Pi_{\mathrm{dr}}(\nu,\tau)$ et $\Pi_{\mathrm{dr}}(r,\theta)$ respectivement. La fonction de paramétrisation optimale constitue un filtre bidimensionnel dans le domaine Doppler-retard qui laisse passer les composantes du signal tout en dégradant les termes d'interférence. Un choix judicieux pour ce filtre passe-bas est la fonction à profil radialement Gaussien, dont l'expression s'avère plus simple en coordonnées polaires. On considère pour cela les variables radiale $r^2 = \nu^2 + \tau^2$ et angulaire $\theta = \arctan(\tau/\nu)$. La fonction de paramétrisation retenue s'exprime alors sous la forme

$$\Pi_{\mathrm{dr}}(r,\theta) = \mathrm{e}^{-\frac{r^2}{2\sigma^2(\theta)}}.$$

La largeur de bande $\sigma(\cdot)$ est fonction de l'angle et détermine la forme de la fonction de paramétrisation, donc le lissage de la distribution temps-fréquence. Le choix du filtre bidimensionnel se réduit alors à un problème d'optimisation à une dimension qui est l'ajustement de $\sigma(\cdot)$.

Étant donné un signal x, les auteurs de [BJ93] proposent de déterminer la largeur de bande $\sigma(\cdot)$ qui maximise le volume de la fonction de caractérisation sous une contrainte de volume sur Π_{dr}, pénalisant ainsi les termes d'interférence distants de l'origine par nature. Le problème d'optimisation s'exprime en coordonnées polaires selon

$$\max_{\sigma} \int_0^{2\pi} \int_0^{\infty} |A_x(r,\theta)|^2 \ \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}}\,\mathrm{d}r\,\mathrm{d}\theta \tag{2.13}$$

sous la contrainte que $\int_0^{2\pi} \sigma^2(\theta)\,\mathrm{d}\theta$ soit égale à une constante donnée. Pour traiter ce problème, une discrétisation des coordonnées polaires dans le plan Doppler-retard est nécessaire. Celle-ci, proposée dans [Bar92], permet de transformer le problème d'optimisation précédent ainsi :

$$\max_{\sigma} \sum_r \sum_{\theta} r|A_x(r,\theta)|^2 \ \mathrm{e}^{-\frac{(r\ \Delta_r)^2}{\sigma^2(\theta)}} \tag{2.14}$$

sous la contrainte

$$\sum_{\theta} \sigma^2(\theta) = V. \tag{2.15}$$

Ici V est le paramètre qui contrôle le compromis lissage/interférences, et $\Delta_r = 2\sqrt{\pi/l}$ où l est la taille du signal échantillonné x. Afin de résoudre ce problème, un algorithme de *projection et descente de gradient* est proposé. Dans un premier temps, l'estimation de la solution à l'itération $k+1$, soit $\sigma_{k+1}(\theta)$, est déterminée par l'évaluation de

$$\sigma_{k+1}(\theta) = \sigma_k(\theta) + \mu_k \, \frac{\partial g}{\partial \sigma_k(\theta)}, \tag{2.16}$$

où μ_k est le pas, et g la fonction objectif à maximiser dans (2.14). Le gradient estimé en $\sigma_k(\theta)$ est défini par le vecteur $\left[\frac{\partial g}{\partial \sigma_k(0)} \; \cdots \; \frac{\partial g}{\partial \sigma_k(l-1)}\right]$, avec

$$\frac{\partial g}{\partial \sigma_k(\theta)} = \frac{2\Delta_r^2}{\sigma_k^3(\theta)} \sum_r |A_x(r,\theta)|^2 \; r^3 \; e^{-(r\;\Delta_r)^2/\sigma_k^2(\theta)}. \tag{2.17}$$

Dans un second temps, la contrainte est prise en compte en normalisant $\sigma_{k+1}(\theta)$ par $||\sigma_{k+1}(\theta)||/V$ à chaque itération, ce qui correspond à une projection dans l'ensemble des fonctions admissibles vérifiant la contrainte (2.15).

2.2.4 Questions de discrétisation

Toute étude considérant des distributions temps-fréquence est incomplète si l'on ne précise pas le mode de discrétisation adopté. Avant de refermer ce chapitre, nous présentons succinctement celui que nous avons utilisé pour les simulations tout au long de ce manuscrit.

En pratique, on dispose d'un espace $\mathcal{X}$ de signaux échantillonnés de taille l. Proposée par Claasen et Mecklenbrauker en 1980, la discrétisation classique de la distribution de Wigner est définie selon

$$W_x^d[t,f) = \sum_{\tau=0}^{l-1} x[t+\tau] \; \overline{x}[t-\tau] \, \mathrm{e}^{-j\frac{2\pi}{l} f\tau}. \tag{2.18}$$

Bien que cette discrétisation soit intuitive et facile à mettre en œuvre, on déplore la perte de la propriété d'unitarité alors qu'elle est essentielle à la bonne résolution de nombreux problèmes. Dans la littérature, plusieurs définitions tentent de remédier à cet inconvénient. Parmi celles-ci, on peut citer les distributions discrètes de Peyrin *et coll.* [PP86], Richman *et coll.* [RPS98], O'Neill *et coll.* [OW99]. Récemment, Chassande-Mottin *et coll.* [CMP05] ont proposé une distribution de Wigner discrète définie par

$$W_x^d[t,f] = \sum_{\tau=-\tau_t}^{\tau_t} x[\lfloor t+\tau/2 \rfloor] \; \overline{x}[\lfloor t-\tau/2 \rfloor] \, \mathrm{e}^{-j\frac{2\pi}{2l} f\tau}, \tag{2.19}$$

où $\tau_t = \min\{2t, 2l-1-2t\}$, et $\lfloor t \rfloor$ désigne le plus grand entier inférieur ou égal à t. Cette dernière définition vérifie la plupart des propriétés auxquelles la forme continue (2.6) satisfait, en particulier l'unitarité. Elle produit de plus des représentations lisibles, sans artifices secondaires comme c'est le cas d'autres définitions discrètes. Enfin, on peut aisément étendre cette définition aux différentes distributions de la classe de Cohen.

L'approche proposée tout au long de ce manuscrit de thèse se prête facilement à une reformulation avec les différentes variantes de distributions discrètes, et plus particulièrement la définition (2.19) pour son unitarité. Toutefois, afin de rendre l'approche plus générique, on se contente dans la suite de développer nos travaux à partir du formalisme en variables continues.

Chapitre 3

Méthodes à noyau dans le domaine temps-fréquence

Sommaire

Au Chapitre 1, nous avons introduit diverses méthodes de reconnaissance des formes à noyau et le formalisme qui leur est propre. Celles-ci offrent un cadre général pour la mise en œuvre de différentes méthodes d'apprentissage statistiques dans des espaces de Hilbert à noyau reproduisant. Un choix naturel pour ces espaces dans un contexte d'analyse et de décision en environnement non-stationnaire est l'espace des distributions temps-fréquence. Introduites au chapitre précédent, celles-ci ont été largement considérées dans la littérature dans le cadre d'applications de reconnaissance des formes.

L'approche que nous proposons dans ce chapitre exploite les avancées dans le domaine des méthodes à noyau pour la mise en œuvre de techniques de reconnaissance des formes dans un contexte temps-fréquence. A chaque distribution temps-fréquence correspond en effet un noyau reproduisant et un RKHS associé, ce qui permet de recourir aux deux clés de voûte que sont le coup du noyau et le Théorème de Représentation.

Ce chapitre débute par quelques rappels sur les méthodes décisionnelles opérant dans le plan temps-fréquence qui ont pu être décrites dans la littérature. On généralise ensuite ces méthodes de reconnaissance des formes en les inscrivant dans le formalisme propre aux noyaux reproduisants. Par soucis de clarté, nous nous intéressons tout d'abord à la distribution de Wigner avant de considérer d'autres distributions temps-fréquence, linéaires et quadratiques.

3.1 Motivations

Dans ce qui suit, nous passons en revue les travaux clés qui ont influencé notre approche et permettent de mieux les comprendre. Nous commençons par ceux ayant trait à l'exploitation d'espaces de représentations temps-fréquence. Puis nous pointons sur des travaux dans lesquels la notion de noyaux reproduisants et de RKHS est sous-jacente, à un stade embryonnaire. A titre pédagogique, nous nous restreignons enfin à des problèmes de détection dans le domaine temps-fréquence, avant de considérer des applications de reconnaissance des formes dans ce même domaine.

3.1.1 Espace de représentation temps-fréquence et noyau reproduisant

Espace de représentation

De nombreux chercheurs se sont intéressés aux espaces engendrés par des familles de distributions temps-fréquence quadratiques. Dans l'un des premiers travaux en ce sens, Saleh et Subotic ont étudié dans [SS85] l'espace des distributions valides, et la projection sur cet espace d'une signature temps-fréquence. La synthèse de signaux à partir de signatures temps-fréquence quadratiques a été longuement traité par Hlawatsch [HK92], à partir d'un espace de représentations temps-fréquence induit par un espace linéaire de signaux. Ces différents travaux sont largement décrits dans son ouvrage [Hla98], dans le cadre de problèmes de filtrage, détection et estimation dans le domaine temps-fréquence. Pour construire ces espaces de représentation, l'auteur a recours à une description de l'espace de signaux à partir d'une base orthonormale dans $\mathcal{L}_2(\mathbb{R})$. Son attention se porte sur les fonctions d'Hermites et Slépiennes. Notre approche se distingue de ces travaux par la construction d'un espace induit par un ensemble de signaux d'apprentissage, sans nécessité de recourir à un espace linéaire de signaux au sens de Hlawatsch. Par cette construction simple et efficace à l'aide de noyaux reproduisants, les différentes méthodes de reconnaissance des formes dans le domaine temps-fréquence sont à portée de main. Plus récemment, l'usage d'espaces de représentation pour un problème de classification de signaux a été abordé dans [Dav00]. De même, Richard a proposé dans [Ric01] de quantifier la redondance informationnelle de la distribution de Wigner discrète vis-à-vis de celle initialement véhiculée par le signal. L'auteur développe d'avantage cette notion dans [Ric04] en exprimant cette redondance sous forme de dépendance linéaire de certaines composantes de la distribution. Il aboutit alors à la dimension de l'espace linéaire engendré par les distributions de Wigner d'un ensemble de signaux donné.

Dans ce manuscrit, nous envisageons une construction de l'espace de représentation similaire à ce dernier, avec la notion de noyau reproduisant et de RKHS. Contrairement à [Ric04], nous considérons avantageusement la redondance des distributions temps-fréquence pour plusieurs raisons. D'une part, le théorème de Balian-Low [Bal81] stipule que les représentations de type Gabor ou Fourier à court-terme ayant de bonnes propriétés de localisation conjointement en temps et en fréquence sont des représentations redondantes[4]. La redondance permet d'avoir une représentation temps-fréquence plus robuste au bruit. D'autre part, si le nombre de paramètres libres à déterminer dans ce type d'espace est quadratique par rapport à la taille des signaux considérés, ce qui peut mener à des problèmes mal-posés, il reste possible de recourir à des techniques de régularisation.

[4] Pour construire des représentations non-redondantes, on dispose de nombreuses bases orthonormées temps-fréquence bien localisées des deux cotés de Fourier ; par exemple, les bases MDCT et les bases de Wilson (lapped orthogonal transforms) [Mal98].

Noyau reproduisant et temps-fréquence

Initiés par les travaux de Kailath sur la détection et l'estimation de signaux dans un RKHS [Kai71], Fowler et Sibul se sont intéressés dans [FS91] à la mise en place d'un RKHS pour la détection par représentations linéaires, temps-fréquence et temps-échelle. Cette approche reste toutefois limitée en raison de la connaissance *a priori* des propriétés statistiques des signaux étudiés qu'elle nécessite d'une part, et d'autre part par la linéarité et l'unitarité requises pour les distributions considérées. Plus récemment, plusieurs travaux ont souligné l'intérêt de combiner des méthodes à noyau et des représentations temps-fréquence. En particulier, un usage direct des SVM pour la discrimination de distributions temps-fréquence est proposé par Davy *et coll.* dans [DGDR02]. La résolution d'un problème de régression non-paramétrique est étudiée par Rakotomamonjy *et coll.* dans [RMC05] avec des noyaux reproduisants d'ondelette, ou encore dans un cadre plus général de structures obliques pour l'apprentissage regularisé selon Tikhonov [RC05].

3.1.2 De la détection dans le plan temps-fréquence ...

La détection de signaux en présence de bruit est un problème ancien auquel de nombreux auteurs ont voulu apporter des réponses dans le domaine temps-fréquence, par exemple pour la détection des ondes gravitationnelles [Ver93, CMF99, Mor02, CM05]. Les différents travaux publiés montrent l'intérêt des représentations linéaires [PF86, Tut89] et quadratiques [Fla88, BO90, SJ95]. Par soucis de clarté, on se limitera à ces dernières selon le cadre adopté par Flandrin dans [Fla86, Fla98]. Dans ce cas, le problème de détection consiste à déterminer, dans un environnement bruité, si un signal est présent (hypothèse H_1) ou absent (hypothèse H_0). Reformulé à l'aide des distributions de Wigner, le problème de détection s'écrit selon

$$\left\{ \begin{array}{ll} H_0 : & W_x(t,f) = W_b(t,f) \\ H_1 : & W_x(t,f) = W_{r+b}(t,f), \end{array} \right.$$

où r est le signal à détecter, noyé dans un bruit additif b, et W_x la distribution de Wigner du signal observé x. Le détecteur optimal au sens du rapport de vraisemblance, lorsque r est à phase initiale aléatoire uniformément distribué sur $[0, 2\pi[$ et b un bruit blanc gaussien,[5] est le filtre adapté temps-fréquence défini par

$$\Lambda(x) = \iint W_x(t,f)\, W_r(t,f)\ \mathrm{d}t\,\mathrm{d}f \underset{H_0}{\overset{H_1}{\gtrless}} \nu_0.$$

Ainsi cette règle de décision correspond-t-elle à une corrélation entre la distribution de Wigner du signal observé x et celle du signal de référence r. L'identité de Moyal permet son évaluation directement à partir des signaux, avec $\Lambda(x) = \left| \int x(t)\, \overline{r}(t)\ \mathrm{d}t \right|^2$, ce qui correspond alors à un filtre adapté classique suivi d'un détecteur d'enveloppe quadratique. Notons que cette formulation n'est autre que la mise en œuvre du *coup du noyau* dans le domaine temps-fréquence, puisque l'on y résout un problème de détection sans expliciter les distributions temps-fréquence dont il est question. Bien que la distribution de Wigner revêt ici un caractère de représentation optimale, renforcé par les nombreuses propriétés intéressantes qu'elle vérifie, la présence de termes interférentiels peut nuire à la robustesse des détecteurs qui lui seraient associés. Dans ces conditions, il est parfois préférable en pratique de considérer une statistique de la forme $\Lambda(x) = \iint C_x^{\Pi}(t,f)\ \overline{C_r^{\Pi}}(t,f)\ \mathrm{d}t\,\mathrm{d}f$. Si la distribution C_x^{Π} n'est pas unitaire, il est en revanche impossible d'exprimer directement celle-ci à partir des signaux.

[5] L'optimalité de différents détecteurs temps-fréquence est étudiée dans [Lem95] au sens de plusieurs critères.

3.1.3 ... à la reconnaissance des formes dans le plan temps-fréquence

Les règles de décision rapidement présentées ci-dessus utilisent une représentation temps-fréquence du signal de référence. Dans une optique dépassant le cadre du filtre adapté temps-fréquence, il est possible de remplacer la représentation de référence par une signature temps-fréquence arbitraire Ψ qui n'est pas nécessairement une distribution valide. Plus généralement, les classiques méthodes paramétriques de reconnaissance des formes appliquées dans le plan temps-fréquence reposent souvent sur la recherche d'une signature temps-fréquence Ψ optimale au le sens d'un critère de performance de la forme (1.3). On dispose pour cela d'un ensemble d'apprentissage formé de n signaux $x_1, \ldots, x_n \in \mathcal{X}$, et éventuellement de leurs étiquettes $y_1, \ldots, y_n \in \mathcal{Y}$. La statistique associée s'écrit alors selon

$$\Lambda(x) = \iint W_x(t,f)\ \Psi(t,f)\ \mathrm{d}t\,\mathrm{d}f, \tag{3.1}$$

la distribution de Wigner figurant dans cette expression pouvant être remplacée par n'importe quelle autre distribution temps-fréquence. Au-delà de contextes décisionnels, ce type de statistique peut être exploité dans le cadre de problèmes de représentation tels que l'analyse en composantes principales. Dans ce cas, on cherche la signature temps-fréquence correspondant à un axe principal. Le critère à optimiser est alors relatif à la maximisation de la variance des données projetées selon cette signature. La statistique (3.1) correspond alors à la composante principale associée à cet axe. Force est de constater qu'une difficulté rencontrée lors de la résolution de tels problèmes est la taille des représentations manipulées. Pourvu que l'on soit en mesure de composer un espace de Hilbert à noyau reproduisant adéquate, le coup du noyau et le Théorème de Représentation pourraient fournir une réponse intéressante et ouvrir une voie vers l'ensemble des méthodes de reconnaissance des formes à noyau.

3.2 Noyau et RKHS associés à la distribution de Wigner

3.2.1 RKHS associé à la distribution de Wigner

Soit Φ une application associant la distribution de Wigner W_{x_i} à tout signal $x_i \in \mathcal{X}$. L'espace image de Φ est alors défini par $\{\Phi\ :\ \Phi(x_i) = W_{x_i}\ ,\ x_i \in \mathcal{X}\}$. Notons que celui-ci n'est pas un espace linéaire puisqu'une combinaison linéaire de distributions de Wigner n'est pas forcément une distribution de Wigner valide. On complète en conséquence cet espace à l'aide des combinaisons linéaires de l'ensemble des distributions, afin d'obtenir[6]

$$\mathcal{H}'_W = \Big\{\Phi : \Phi(x) = \sum_{i=1}^{n} a_i\ W_{x_i}\ ,\ x_i \in \mathcal{X}\ ,\ a_i \in \mathbb{C}\Big\}.$$

A cet espace engendré par les n distributions temps-fréquence $W_{x_1}(t,f), \ldots, W_{x_n}(t,f)$, on associe le produit scalaire canonique dans $\mathcal{L}_2(\mathbb{R}^2)$, selon

$$\langle \psi, \phi \rangle_{\mathcal{H}'_W} = \Big\langle \sum_{i=1}^{n} a_i W_{x_i}, \sum_{j=1}^{n}\ b_j W_{x_j} \Big\rangle, \tag{3.2}$$

pour tous les éléments $\psi = \sum_{i=1}^{n} a_i W_{x_i}$ et $\phi = \sum_{j=1}^{n} b_j W_{x_j}$ de l'espace $\mathcal{H}'_W$. Pour obtenir un espace de Hilbert à partir de l'espace pré-Hilbertien $\mathcal{H}'_W$, il suffit de le compléter conformément à [Aro50] de

[6] Bien que la distribution de Wigner soit à valeurs réelles, on présente ici un espace de fonctions à valeurs complexes par soucis de généralisation, comme on le verra plus loin avec les différentes distributions de la classe de Cohen.

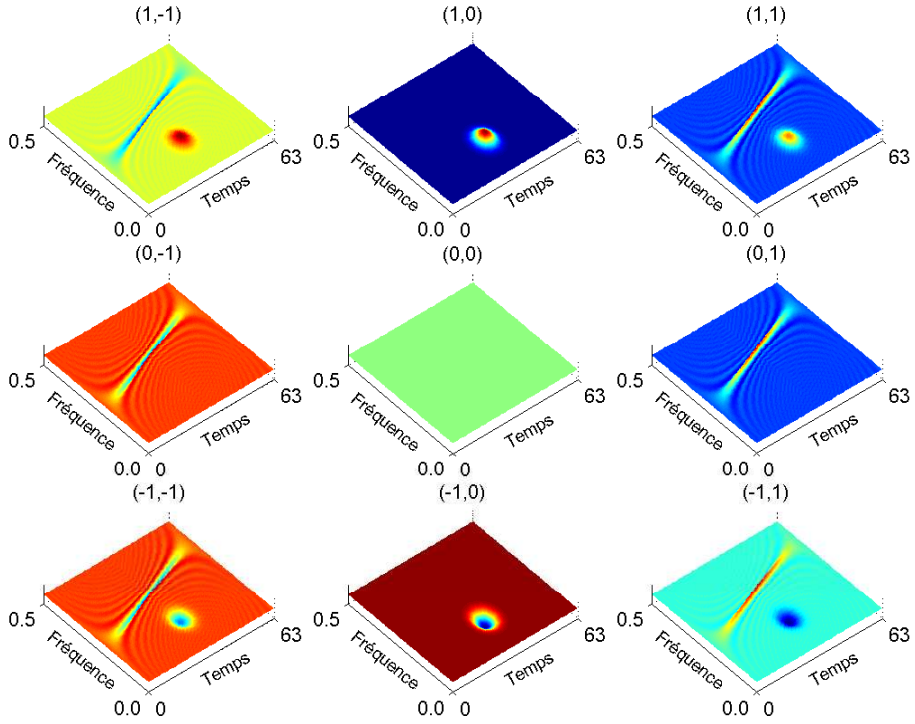


FIG. 3.1 – Certains éléments de l'espace engendré par les distributions $\{W_{x_1}(t,f)$ et $W_{x_2}(t,f)\}$ où x_1 comporte un atome Gaussien et x_2 un chirp. Ces éléments, de la forme $a_1 W_{x_1}(t,f) + a_2 W_{x_2}(t,f)$, correspondent aux coefficients $a_1, a_2 \in \{-1, 0, 1\}$

sorte que toute suite de Cauchy y converge. L'espace ainsi obtenu $\mathcal{H}_W$ est l'espace de Hilbert engendré par $\{W_{x_1}(t,f), \ldots, W_{x_n}(t,f)\}$ muni du produit scalaire $\langle \psi, \phi \rangle_{\mathcal{H}_W}$ défini dans (3.2).

Considérons un exemple simple d'espace engendré par 2 signaux, x_1 comportant un atome temps-fréquence Gaussien à 0.2 Hz, et x_2 une modulation fréquentielle linéaire variant de 0.2 à 0.4 Hz. Les éléments de cet espace s'expriment sous la forme $a_1 W_{x_1}(t,f) + a_2 W_{x_2}(t,f)$ pour $a_1, a_2 \in \mathbb{C}$. Certains de ces éléments sont présentés à la Figure 3.1, pour les couples de coefficients $a_1, a_2 \in \{-1, 0, 1\}$. On retrouve des représentations comportant l'atome Gaussien en contribution positive (première ligne) et négative (troisième ligne), celles comportant la composante à modulation fréquentielle en contribution négative (première colonne) et positive (dernière colonne). Afin de construire une base orthonormée dans cet espace, on peut recourir à une orthonormalisation de type Gram-Schmidt, ou encore à une technique d'ACP-à-noyau comme on verra au Chapitre 4.

3.2.2 Distribution de Wigner et méthodes à noyau

Dans l'expression (3.2) du produit scalaire dans $\mathcal{H}$, on obtient en réarrangeant les sommations

$$\langle \psi, \phi \rangle_{\mathcal{H}_W} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i \, \overline{b_j} \, \langle W_{x_i}, W_{x_j} \rangle,$$

avec $\langle W_{x_i}, W_{x_j} \rangle = \iint W_{x_i}(t,f)\ W_{x_j}(t,f)\,\mathrm{d}t\,\mathrm{d}f$. Le produit scalaire entre les distributions de Wigner se calcule directement à partir des signaux, selon la formule de Moyal (2.7), ce qui permet d'écrire

$$\langle \psi, \phi \rangle_{\mathcal{H}_W} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i\ \overline{b_j}\ |\langle x_i, x_j \rangle|^2.$$

On peut alors déterminer le produit scalaire entre deux éléments de l'espace des représentations, directement à partir des signaux sans qu'il soit nécessaire d'expliciter leurs images dans cet espace. Il s'agit là du coup du noyau en notant

$$\kappa_W(x_i, x_j) = \langle W_{x_i}, W_{x_j} \rangle, \tag{3.3}$$

que l'on reformule ainsi à partir de l'identité de Moyal

$$\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2. \tag{3.4}$$

On peut vérifier facilement que κ_W est un noyau défini positif, puisque la Définition 1 s'écrit dans ce cas $\| \sum_j a_j\ W_{x_j} \|^2 \geq 0$, ce qui est vrai pour tout $x_j \in \mathcal{X}$ et $a_j \in \mathbb{C}$. L'espace $\mathcal{H}_{\mathcal{W}}$ construit est donc l'espace de Hilbert induit par le noyau reproduisant κ_W. Ainsi l'association du noyau (3.3)-(3.4) et du Théorème de Représentation à tout algorithme à noyau décrit dans la littérature permet-elle de mettre en œuvre les méthodes de reconnaissance des formes les plus variées dans le domaine temps-fréquence. Les statistiques considérées s'écrivent alors sous la forme

$$\Lambda(x) = \sum_{j=1}^{n} \alpha_j^*\ \langle W_x, W_{x_j} \rangle = \sum_{j=1}^{n} \alpha_j^*\ \kappa_W(x, x_j),$$

Elles admettent une interprétation temps-fréquence puisque l'on peut écrire $\Lambda(x) = \langle W_x, \Psi_W \rangle$ avec

$$\Psi_W = \sum_{j=1}^{n} \alpha_j^*\ W_{x_j}. \tag{3.5}$$

D'un point de vue calculatoire, on souligne le fait que les n coefficients optimaux, $\alpha_1^*, \ldots, \alpha_n^*$, sont évalués sans calculer aucune distribution de Wigner. Il en est de même pour l'évaluation de $\Lambda(x)$. Si l'on souhaitait toutefois exhiber dans certains cas la signature temps-fréquence Ψ_W, afin de procéder à une analyse classique dans le plan temps-fréquence, il reste possible d'évaluer (3.5) par un calcul itératif sans qu'il soit nécessaire de conserver en mémoire l'ensemble des distributions de Wigner associées à l'ensemble d'apprentissage. On peut également remarquer que l'on peut écrire

$$\Psi_W(t,f) = \int \Big[\sum_{j=1}^{n} \alpha_j\ x_j(t+\tau/2)\ \overline{x_j}(t-\tau/2) \Big]\ \mathrm{e}^{-j2\pi f\tau}\,\mathrm{d}\tau, \tag{3.6}$$

ce qui montre qu'une seule transformée de Fourier suffit à calculer $\Psi_W(t,f)$. Enfin, notons que certaines méthodes à noyau telles que les SVM ont recours à des fonctions de coût menant à des solutions parcimonieuses. Dans ce cas, Ψ_W ne dépend que d'un nombre très réduit de distributions de Wigner, ce qui rend son calcul d'autant plus rapide.

3.2.3 Interprétation en termes de filtre linéaire variant en temps

On propose à présent une interprétation des résultats précédents en termes de filtre linéaire variant dans le temps. On note pour cela $h(t+\tau/2, t-\tau/2)$ l'expression entre crochets dans (3.6), soit

$$h(t,t') = \sum_{j=1}^{n} \alpha_j\ x_j(t)\ \overline{x_j}(t').$$

Puisque la fonction $h(t,t')$ est carré intégrable et Hermitienne, elle peut être vue comme la réponse impulsionnelle d'un système non-stationnaire défini par un opérateur linéaire auto-adjoint L_Ψ. La relation entrée-sortie d'un tel système est alors donnée par

$$\begin{aligned} x'(t) &= (L_\Psi\, x)(t) \\ &= \int h(t,t')\, x(t')\, \mathrm{d}t' \\ &= \int h(t+\tau/2, t-\tau/2)\, x(t+\tau)\, \mathrm{d}\tau \\ &= \int \Psi_W(t,f)\, \widehat{x}(f)\, \mathrm{d}f, \end{aligned}$$

où la dernière égalité est due à l'identité de Parseval. On retrouve la signature temps-fréquence $\Psi_W(t,f)$ donnée par (3.6). Désignée comme étant le symbole de Weyl associé, celle-ci permet de caractériser le filtre linéaire variant en temps outre sa réponse impulsionnelle $h(t,t')$. Le lien entre les deux est donné par l'expression $\Psi_W(t,f) = \int h(t+\tau/2, t-\tau/2)\, \mathrm{e}^{-j2\pi f\tau}\, \mathrm{d}\tau$. Défini ainsi, l'opérateur L_Ψ vérifie alors l'expression suivante

$$\langle L_\Psi x, x\rangle_{\mathcal{X}} = \iint W_x(t,f)\, \Psi_W(t,f)\, \mathrm{d}t\, \mathrm{d}f.$$

On retrouve alors la statistique associée à toute méthode à noyau reproduisant pour la distribution de Wigner, à savoir $\Lambda(x) = \langle W_x, \Psi_W\rangle$. De plus, on peut décomposer L_Ψ en n opérateurs, puisque

$$\begin{aligned} \langle L_\Psi x, x\rangle_{\mathcal{X}} &= \sum_{j=1}^{n} \iint W_x(t,f)\, W_{x_j}(t,f)\, \mathrm{d}t\, \mathrm{d}f \\ &= \sum_{j=1}^{n} \langle L_{W_{x_j}} x, x\rangle_{\mathcal{X}}, \end{aligned}$$

où l'opérateur $L_{W_{x_j}}$ est associé au symbole de Weyl $W_{x_j}(t,f)$. Dans ce manuscrit, on se contentera de cette analogie entre le symbole de Weyl et son opérateur d'une part, et la signature temps-fréquence et la statistique issus d'une méthode à noyau d'autre part. On renvoie le lecteur intéressé par le symbole de Weyl associé au plan temps-fréquence vers l'article précurseur [RT94], ou plus récemment vers les travaux de Hlawatsch sur les filtres linéaires variants en temps [Hla98, HM01].

3.3 Noyaux et RKHS associés aux distributions temps-fréquence

Évidemment, le concept de méthodes temps-fréquence à noyau ne se restreint pas à la distribution de Wigner. Dans ce qui suit, on en propose des mises en œuvre avec d'autres distributions temps-fréquence, linéaires et quadratiques.

3.3.1 Transformations linéaires

On se limite ici aux transformations linéaires les plus couramment utilisées, à savoir la transformée de Fourier à court-terme et la transformée en ondelette.

Pour une fenêtre analysante w donnée, la transformation de Fourier à court-terme définit un opérateur linéaire $F^w : x \mapsto F^w_x$ de $\mathcal{L}_2(\mathbb{R})$ dans $\mathcal{L}_2(\mathbb{R}^2)$. A l'espace image qui est inclus dans $\mathcal{L}_2(\mathbb{R}^2)$, on associe le produit scalaire

$$\langle F^w_{x_i}, F^w_{x_j}\rangle = \iint F^w_{x_i}(t,f)\, \overline{F^w_{x_j}}(t,f)\, \mathrm{d}t\, \mathrm{d}f$$

On définit alors le noyau $\kappa_{F^w}(x_i, x_j) = \langle F^w_{x_i}, F^w_{x_j} \rangle$, qui s'exprime directement à partir des signaux grâce à l'identité de Parseval

$$\kappa_{F^w}(x_i, x_j) = \|w\|^2 \langle x_i, x_j \rangle. \tag{3.7}$$

On peut vérifier que κ_{F^w} est un noyau défini positif puisque la condition (1.6) s'écrit selon $\|w\|^2 \| \sum_i a_i \, x_i\|^2 \geq 0$, ce qui est toujours vérifié. On peut alors parler de RKHS induit par le noyau (3.7), ce dernier pouvant être associé à n'importe quelle méthode à noyau. Comme précédemment, l'expression (1.9) admet une interprétation temps-fréquence $\Lambda(x) = \langle F^w_x, \Psi_{F^w} \rangle$ avec $\Psi_{F^w} = \sum_{j=1}^n \alpha_j^* \, F^w_{x_j}$. Par linéarité de la transformation considérée, cette dernière expression n'est autre que la transformée de Fourier à court-terme du signal $\sum_{j=1}^n \alpha_j^* \, x_j$.

Dans le cas d'une transformation en ondelettes, l'opérateur linéaire est défini selon $O^w : x \mapsto O^w_x$, pour une ondelette mère w donnée. A l'espace image inclus dans $\mathcal{L}_2(\mathbb{R}^2)$, on associe le produit scalaire

$$\langle O^w_{x_i}, O^w_{x_j} \rangle_\Omega = \iint O^w_{x_i}(t,a) \, \overline{O^w_{x_j}}(t,a) \frac{\mathrm{d}t \, \mathrm{d}a}{a^2}.$$

Par l'identité de Parseval, on aboutit à $\langle O^w_{x_i}, O^w_{x_j} \rangle_\Omega = C_w \langle x_i, x_j \rangle$. Le noyau reproduisant associé à cet opérateur, $\kappa_{O^w}(x_i, x_j) = \langle O^w_{x_i}, O^w_{x_j} \rangle$, est un noyau défini positif puisque l'expression $C_w \| \sum_i \alpha_i \, x_i \|^2$ est toujours non-négative. De plus, le produit scalaire dans le RKHS induit par ce noyau s'exprime directement à partir des signaux selon

$$\kappa_{O^w}(x_i, x_j) = C_w \langle x_i, x_j \rangle. \tag{3.8}$$

L'expression (1.9) admet une interprétation temps-échelle $\Lambda(x) = \langle O^w_x, \Psi_{O^w} \rangle$ où $\Psi_{O^w} = \sum_{j=1}^n \alpha_j^* \, O^w_{x_j}$. Comme la transformation en ondelettes est linéaire, on peut réécrire cette dernière expression comme une transformée en ondelettes du signal $\sum_{j=1}^n \alpha_j^* \, x_j$.

3.3.2 Distributions quadratiques

La classe de Cohen, à laquelle appartient la distribution de Wigner, offre une grande diversité de représentations. Ainsi est-il possible de considérer des distributions lissées à la lisibilité améliorée, ou encore des distributions favorisant la résolution de problèmes décisionnels. Les propriétés inhérentes aux espaces de représentation correspondants sont déterminées par le noyau reproduisant considéré.

Dans le cas général des distributions de la classe de Cohen, on considère la forme générale du noyau reproduisant $\kappa_{C^\Pi}(x_i, x_j) = \langle C^\Pi_{x_i}, C^\Pi_{x_j} \rangle$, où Π désigne la fonction de paramétrisation de la représentation considérée. Pour certaines de ces distributions, une simplification du calcul du noyau reproduisant est possible, comme pour le spectrogramme où $\kappa_{\mathrm{S}^w}(x_i, x_j) = \iint |\langle x_i, w_{t,f} \rangle \langle x_j, w_{t,f} \rangle|^2 \, \mathrm{d}t \, \mathrm{d}f$, avec $w_{t,f}(\tau) = w(\tau - t) \, e^{2j\pi f \tau}$. Plus généralement, on évalue le noyau dans le domaine Doppler-retard selon $\kappa_{C^\Pi}(x_i, x_j) = \iint |\Pi_{\mathrm{dr}}(\nu, \tau)|^2 A_{x_i}(\nu, \tau) \, A_{x_j}(\nu, \tau) \, \mathrm{d}\nu \, \mathrm{d}\tau$. Pour la famille des distributions unitaires, on retrouve $\kappa_U(x_i, x_j) = \iint A_{x_i}(\nu, \tau) \, A_{x_j}(\nu, \tau) \, \mathrm{d}\nu \, \mathrm{d}\tau$ car $|\Pi_{\mathrm{dr}}(\nu, \tau)| = 1$, que l'on simplifie grâce à l'identité de Moyal selon $\kappa_U(x_i, x_j) = |\langle x_i, x_j \rangle|^2$. Outre la distribution de Wigner, les distributions unitaires comprennent celles de Page et de Rihaczek. Toutes ces distributions admettent le même noyau reproduisant (3.4) et partagent donc le même espace reproduisant à une transformation unitaire près. Cette propriété a été abordée à la Section 1.3.1. Dans le cas de la distribution à paramétrisation radialement Gaussienne introduite au Paragraphe 2.2.3, notons que le noyau reproduisant s'écrit en coordonnées polaires sous la forme $\kappa_\sigma(x_i, x_j) = \iint r \, A_{x_i}(r, \theta) \, \overline{A_{x_j}}(r, \theta) \, \mathrm{e}^{-r^2/\sigma^2(\theta)} \, \mathrm{d}r \, \mathrm{d}\theta$.

La mise en œuvre des différentes méthodes à noyau dans un RKHS associé à une distribution arbitraire de la classe de Cohen est rendu possible par l'usage du noyau reproduisant κ_{C^Π} correspondant.

On désigne par $\mathcal{H}_{C^\Pi}$ l'espace induit par ce noyau. La statistique de décision correspondante est alors de la forme $\Lambda(x) = \langle C_x^\Pi, \Psi_{C^\Pi} \rangle$, où Ψ_{C^Π} désigne la signature temps-fréquence résultant de l'optimisation d'un critère de performance donné à partir d'un ensemble d'apprentissage formé de n signaux x_j. Cette signature appartient au RKHS engendré par les n distributions $C_{x_j}^\Pi$, et s'écrit $\Psi_{C^\Pi} = \sum_{j=1}^n \beta_j^* \, C_{x_j}^\Pi$. La statistique correspondante est alors donnée par $\Lambda(x) = \sum_{j=1}^n \beta_j^* \, \langle C_x^\Pi, C_{x_j}^\Pi \rangle = \sum_{j=1}^n \beta_j^* \, \kappa_{C^\Pi}(x, x_j)$. Les propriétés de la signature résultante sont inhérentes au type de distribution temps-fréquence choisi. En considérant une distribution à interférences réduites par exemple, l'espace RKHS comprend de telles distributions et la signature résultante connaît une réduction de ses termes interférentiels en comparaison de la distribution de Wigner. Pour s'en convaincre, il suffit d'écrire

$$\begin{aligned} \Psi_{C^\Pi}(t,f) &= \sum_{i=1}^n \beta_i^* \iint \Pi_{\mathrm{tf}}(t-t', f-f') \, W_{x_i}(t',f') \, \mathrm{d}t' \, \mathrm{d}f' \\ &= \iint \Pi_{\mathrm{tf}}(t-t', f-f') \, \Big[\sum_{i=1}^n \beta_i^* W_{x_i}(t',f') \Big] \, \mathrm{d}t' \, \mathrm{d}f'. \end{aligned}$$

On retrouve alors une version lissée par la fonction de paramétrisation Π d'une combinaison linéaire de distributions de Wigner.

3.3.3 Stratégie hybride

Dans le cas général des distributions non-unitaires de la classe de Cohen, le calcul de $\kappa_{C^\Pi}(x_i, x_j)$ pour tout couple (x_i, x_j) de l'ensemble d'apprentissage peut allonger le temps d'apprentissage de la méthode à noyau considérée. Pour les applications où le temps de calcul est crucial, on propose d'utiliser une méthode hybride qui combine les avantages du noyau quadratique de Wigner et des distributions lissées de la classe de Cohen. Une approche similaire a déjà été évoquée dans le cadre de détection dans le plan temps-fréquence. On adapte à présent celle-ci aux méthodes à noyau.

Le principe de l'approche proposée repose sur une phase d'apprentissage effectuée à partir du noyau quadratique $\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2$ associé à la distribution de Wigner. La signature temps-fréquence correspondante s'écrit alors $\Psi_W = \sum_{j=1}^n \alpha_j^* \, W_{x_j}$. Afin de conférer à celle-ci des propriétés complémentaires telles qu'un contenu interférentiels réduit par exemple, on suggère de recourir dans un second temps à un filtrage défini par $\widetilde{\Psi}_{C^\Pi}(t,f) = \iint \Pi_{\mathrm{tf}}(t-t', f-f') \, \Psi_W(t',f') \, \mathrm{d}t' \, \mathrm{d}f'$. En développant cette expression, on obtient

$$\widetilde{\Psi}_{C^\Pi}(t,f) = \iint \Pi_{\mathrm{tf}}(t-t', f-f') \, \Big[\sum_{i=1}^n \alpha_i^* W_{x_i}(t',f') \Big] \, \mathrm{d}t' \, \mathrm{d}f',$$

Il s'agit donc là d'un élément appartenant au RKHS induit par le noyau κ_{C^Π}. La statistique utilisée est alors donnée par $\Lambda(x) = \iint C_x^\Pi(t,f) \; \widetilde{\Psi}_{C^\Pi}(t,f) \; \mathrm{d}t \, \mathrm{d}f$. Notons toutefois que ce principe est sous-optimum au sens du noyau reproduisant κ_{C^Π} dans la mesure où les coefficient α_i^* ont été déterminés à partir du noyau quadratique κ_W. On remarque que l'évaluation de $\Lambda(x)$ nécessite deux filtrages, l'un appliqué à Ψ_W, l'autre à W_x. De manière équivalente, ll est possible de reformuler la statistique afin de faire porter un double filtrage sur la distribution de Wigner de l'observation

$$\begin{aligned} \Lambda(x) &= \iint C_x^\Pi(t,f) \left[\iint \Pi_{\mathrm{tf}}(t-t', f-f') \, \sum_{i=1}^n \alpha_i W_{x_i}(t',f') \, \mathrm{d}t' \, \mathrm{d}f' \right] \mathrm{d}t \, \mathrm{d}f \\ &= \iint \sum_{i=1}^n \alpha_i W_{x_i}(t',f') \left[\iint \Pi_{\mathrm{tf}}(t-t', f-f') \, C_x^\Pi(t,f) \, \mathrm{d}t \, \mathrm{d}f \right] \mathrm{d}t' \, \mathrm{d}f'. \end{aligned}$$

	Nom	Noyau reproduisant associé
linéaire	Fourier	$\kappa_F(x_i, x_j) = \langle x_i, x_j \rangle$
	Fourier à court-terme	$\kappa_{F^w}(x_i, x_j) = \|w\|^2 \langle x_i, x_j \rangle$
	Ondelettes	$\kappa_{O^w}(x_i, x_j) = C_w \langle x_i, x_j \rangle$
quadratique	Wigner	$\kappa_W(x_i, x_j) = \|\langle x_i, x_j \rangle\|^2$
	Spectrogramme	$\kappa_{\mathrm{S}^w}(x_i, x_j) = \iint \|\langle x_i, w_{t,f} \rangle \langle x_j, w_{t,f} \rangle\|^2 \,\mathrm{d}t\,\mathrm{d}f$
	Page-Levin, Rihaczek	$\kappa_U(x_i, x_j) = \|\langle x_i, x_j \rangle\|^2$
	Cohen	$\kappa_{C^\Pi}(x_i, x_j) = \iint \|\Pi_{\mathrm{dr}}(\nu, \tau)\|^2 A_{x_i}(\nu, \tau)\, A_{x_j}(\nu, \tau) \,\mathrm{d}\nu\,\mathrm{d}\tau$
	Profil radialement Gaussien	$\kappa_\sigma(x_i, x_j) = \iint r\; A_{x_i}(r, \theta)\; \overline{A_{x_j}}(r, \theta)\; \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}}\; \mathrm{d}r\;\mathrm{d}\theta$

TAB. 3.1 – Distributions temps-fréquence, linéaires et quadratiques, avec leur noyau reproduisant

On parle alors de filtrage *a posteriori* de l'observation. Par analogie, on peut aussi évaluer la statistique de décision à partir d'un double filtrage de la signature Ψ_W. Il suffit alors d'écrire

$$\begin{aligned} \Lambda(x) &= \iint \left[\iint \Pi_{\mathrm{tf}}(t - t', f - f') W_x(t', f') \,\mathrm{d}t'\,\mathrm{d}f' \right] \Psi_{C^\Pi}(t, f) \,\mathrm{d}t\,\mathrm{d}f \\ &= \iint W_x(t', f') \left[\iint \Pi_{\mathrm{tf}}(t - t', f - f') \Psi_{C^\Pi}(t, f) \,\mathrm{d}t\,\mathrm{d}f \right] \mathrm{d}t'\,\mathrm{d}f'. \end{aligned}$$

On dit alors que l'on opère par filtrage *a priori* de la signature issue de la distribution de Wigner. On note que le double filtrage peut être effectué avec la même fonction de paramétrisation quand cette dernière est paire par rapport à ses variables, soit $\Pi_{\mathrm{tf}}(t, f) = \Pi_{\mathrm{tf}}(-t, -f)$, ce qui constitue une hypothèse peu restrictive.

3.4 Récapitulatif : méthodes à noyau dans le domaine temps-fréquence

Les méthodes à noyau constituent une avancée considérable dans les domaines de la reconnaissance des formes et de la théorie de l'apprentissage statistique. Elles reposent pour la plupart sur la recherche d'une fonctionnelle ψ^* qui minimise un risque régularisé selon Tikhonov, de la forme

$$\psi^* = \arg\min_{\psi \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i), y_i) + \eta \,\|\psi\|^2_{\mathcal{H}},$$

où V est une fonction coût donnée. Les espaces de Hilbert à noyau reproduisant constituent un choix judicieux pour $\mathcal{H}$. L'une de leurs propriétés fondamentales est que $\psi(x) = \langle \psi, \kappa_x \rangle_{\mathcal{H}}$ pour tout $x \in \mathcal{X}$. Il en résulte par le Théorème de Représentation que $\psi^*(\cdot) = \sum_{j=1}^{n} \alpha_j^* \kappa(x_j, \cdot)$, les coefficients α_j^* étant obtenus par la minimisation du risque régularisé évoqué ci-dessus.

Le choix d'un noyau reproduisant associé à une distribution de la classe de Cohen a des effets régularisants sur la solution Ψ_{C^Π}. Ceci est montré à partir de l'expression suivante

$$|\Psi_{C^\Pi}|^2 = |\langle \Pi_{\mathrm{tf}}, W_x \rangle|^2 \leq \|\Pi_{\mathrm{tf}}\|^2 \,\|W_x\|^2 = \|\Pi_{\mathrm{tf}}\|^2 \,\|x\|^4, \tag{3.9}$$

où l'inégalité est due au théorème de Cauchy-Schwartz. Dans certains cas, la statistique peut être évaluée directement à partir des signaux sans qu'il soit nécessaire d'exhiber les représentations temps-fréquence.

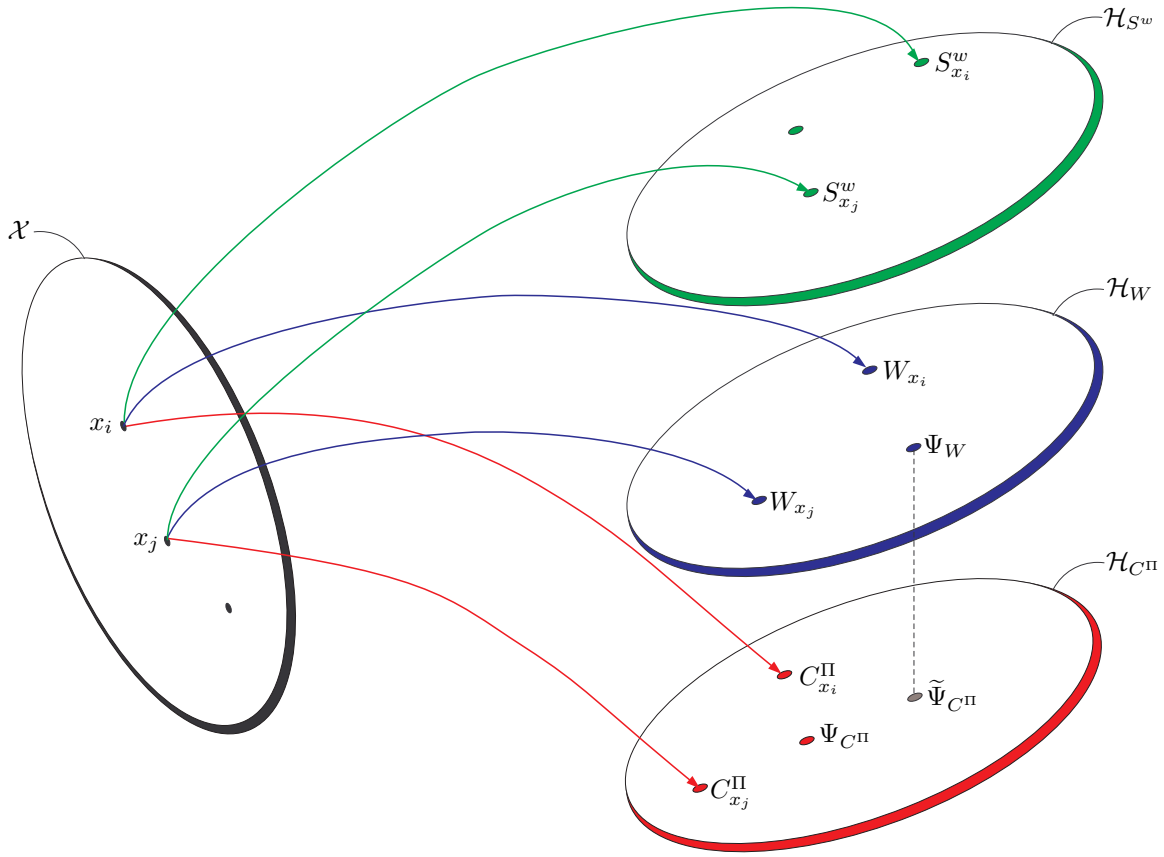


FIG. 3.2 – Illustration des espaces de représentation temps-fréquence associés à la distribution de Wigner, au spectrogramme et à une distribution arbitraire de la classe de Cohen, avec les signatures temps-fréquence optimales correspondant à chaque espace, ainsi que la signature sous-optimale obtenue par la stratégie hybride.

Lorsque tel n'est pas le cas, on suggère de recourir à une stratégie hybride combinant les avantages calculatoires liés à la distribution de Wigner et les propriétés d'autres distributions de la classe de Cohen. Elle consiste à filtrer la signature obtenue à partir du noyau quadratique de Wigner à l'aide de la fonction de paramétrisation de son choix. Cette technique est illustrée sur la Figure 3.4.

Dans la suite, on met en œuvre ces principes dans le cadre de l'ACP-à-noyau et de la discrimination par AFD-à-noyau et SVM. Le choix d'une distribution temps-fréquence adaptée pouvant être considéré comme un problème de sélection de noyau, on traitera cette question à l'aide d'un critère spécifique aux méthodes à noyau.

Chapitre 4

Analyse en composantes principales dans le plan temps-fréquence

Sommaire

L'outil le plus utilisé pour l'analyse et la représentation d'un ensemble de données est incontestablement l'analyse en composantes principales (ACP), étant donné ses propriétés théoriques et sa simplicité de mise en œuvre. Le cadre général des méthodes à noyau offre l'opportunité d'en développer une version non-linéaire. Dans ce chapitre, on adapte cette méthode afin qu'elle puisse opérer dans le domaine temps-fréquence pour l'analyse de signaux non-stationnaires.

Ce chapitre est organisé ainsi. On commence par un court passage en revue de différents travaux sur l'ACP et sur des tentatives d'implémentation dans le domaine temps-fréquence. Puis on présente cet algorithme dans sa version linéaire avant de s'intéresser à son extension non-linéaire. On propose alors sa mise en œuvre dans le domaine temps-fréquence, avec la distribution de Wigner puis d'autres distributions temps-fréquence. Pour montrer l'intérêt de notre approche, on clot ce chapitre par une étude de la complexité calculatoire de l'algorithme que l'on compare à celle de l'ACP classique appliquée directement à des distributions temps-fréquence.

4.1 Introduction

L'analyse en composantes principales (ACP) est une méthode non-supervisée menant au sous-espace optimal qui capte la plus grande part de variance des données. Cette technique linéaire est couramment utilisée à des fins de représentation de données initialement de grande dimension, ou pour en extraire des paramètres destinés à nourrir une règle de décision. L'analyse en composantes principales à noyau (ACP-à-noyau), ou *kernel principal component analysis* en anglais, constitue une extension non-linéaire de l'ACP à des espaces de représentation induits par des noyaux reproduisants. Mieux que l'ACP classique, l'information extraite est liée non-linéairement aux données d'entrée. Comparée à d'autres extensions non-linéaires de l'ACP, par exemple les réseaux de neurones, elle bénéficie d'une stabilité et d'un coût calculatoire réduit. Pour une description de l'ACP et de l'ACP-à-noyau, on renvoie le lecteur aux ouvrages classiques [Jol86, DK96] et aux travaux de Schölkopf *et coll.* [SSM98, Mik98, SSM99].

De nombreux travaux de recherche ont été menés dans le domaine de l'analyse temps-fréquence non-supervisée, conduisant assez naturellement les chercheurs à mettre en œuvre l'ACP sur des données extraites des distributions temps-fréquence. On cite par exemple l'article [SB88], où Stapleton *et coll.* propose d'utiliser l'ACP afin de déterminer une base orthonormée pour une collection de segments musicaux, en rappelant que celle-ci est plus naturelle que celles formées par des tonalités pures à la Fourier, avec l'inconvénient de ne pas opérer dans le domaine temps-fréquence. D'autres travaux tentent de réduire la dimensionnalité d'une représentation en utilisant les premières composantes principales obtenues par ACP. Ceci est le cas dans [BN91, NBD93] où Nawab *et coll.* utilisent une composante principale issue des coefficients de la transformé de Fourier à court-terme. Plus récemment, l'ACP a été mise en œuvre avec des vecteurs spectraux dans [DMCB00, MCD00]. Dans le cadre de la distribution de Wigner, la décomposition de celle-ci sur une base de représentations temps-fréquence a fait l'objet de plusieurs publications, dont [ME85, Mar97, Ami94], ou plus récemment encore avec les ondelettes dans [MCA06].

De manière certes réductrice, les distributions temps-fréquence peuvent être considérées comme des images. On peut alors profiter de la vaste littérature sur l'usage de l'ACP dans le domaine du traitement d'images. Par exemple, Sirovich *et coll.* proposent dans [SK87] de contourner le problème de la taille $l \times l$ élevée des n images traitées en les considérant sous forme de vecteurs de taille l^2 et en y appliquant une ACP. Dans [TP91], Turk *et coll.* ont recours à cette approche pour la classification et la détection de visages. Plus récemment, l'adaptation de cette approche pour l'analyse des distributions temps-fréquence a été proposée dans [EBG05]. Ces démarches ne profitent toutefois pas des avancées théoriques et des coûts calculatoires réduits qu'offre l'ACP-à-noyau.

On propose dans ce chapitre une mise en œuvre temps-fréquence de l'ACP, rendue possible par les deux concepts clés des méthodes à noyau que sont le coup du noyau et le Théorème de Représentation. Par un choix approprié du noyau, l'ACP opère alors implicitement dans le domaine temps-fréquence sans qu'il soit nécessaire de calculer les distributions.

4.2 Analyses en composantes principales, classique et à noyau

4.2.1 Algorithme classique de l'ACP

Considérons un ensemble de n observations $\{x_1, \ldots, x_n\}$ dans un espace donné $\mathcal{X}$ de dimension l. On suppose que ces observations sont centrées dans $\mathcal{X}$. L'ACP vise à rechercher des espaces de projection pertinents pour les données en maximisant leur variance projetée. Les composantes de faible variance sont associées à du bruit, et écartées de fait.

On recherche u maximisant la variance des données projetées $\langle x_i, u\rangle$, pour $i = 1, \ldots, n$. La fonction coût à maximiser est définie par

$$\frac{1}{n}\sum_{i=1}^{n} |\langle x_i, u\rangle|^2,$$

sous la contrainte de normalisation $\langle u, u\rangle = 1$. Le problème d'optimisation avec contrainte est alors donné par le Lagrangien

$$\mathrm{L}(u, \mu) = \frac{1}{n}\sum_{i=1}^{n} |\langle x_i, u\rangle|^2 - \mu(\langle u, u\rangle - 1),$$

où μ est le multiplicateur de Lagrange. La solution à ce problème est obtenue par l'annulation des dérivées par rapport à u et à μ, ce qui conduit au système suivant

$$C\,u = \lambda\,u,$$

où $C = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top$ est la matrice de covariance des données, avec x_i représentant un vecteur colonne de taille l et $x_i^\top$ sa transposée. Les l couples (λ, u) ainsi obtenus forment les valeurs-vecteurs propres de C. Puisque $\lambda\,u = C\,u = \frac{1}{n}\sum_{i=1}^{n}\langle x_i, u\rangle\, x_i$, les vecteurs propres appartiennent à l'espace engendré par les n données $x_1, \ldots, x_n$. On parle alors d'une transformation de base, où la nouvelle base formée par $u_1, \ldots, u_n$ est orthonormée. Les nouvelles coordonnées d'un x donné, appelées composantes principales, sont déterminées par projection sur cette base, $\langle x, u_1\rangle, \ldots, \langle x, u_n\rangle$. Les valeurs propres de cette matrice correspondent à la variance dans chacune des directions définies par les vecteurs propres. Par cette construction, le sous-espace engendré par les k premiers vecteurs propres, ordonnés dans le sens décroissant de leurs valeurs propres, minimise l'erreur quadratique de reconstruction. Soit U_k la matrice formée en colonnes par les k premiers vecteurs propres. La matrice $P_{U_k} = U_k U_k^\top$ correspond alors à la projection sur le sous-espace engendré par les colonnes de U_k, soit $u_1, \ldots, u_k$. Pour une dimension k donnée, l'ACP minimise alors $\sum_{i=1}^{n} \|P_{U_k} x_i - x_i\|^2$.

4.2.2 ACP dans un espace transformé

L'un des inconvénients de l'ACP réside dans sa linéarité. Des vecteurs principaux sont obtenus indépendamment du fait que des lois non-linéaires puissent régir le comportement du système étudié. Afin de pallier cet inconvénient, on s'intéresse aux composantes principales dans un espace transformé lié par une relation non-linéaire à l'espace des données.

On considère donc la transformation $\phi : x_i \mapsto \phi(x_i)$ de $\mathcal{X}$ vers l'espace transformé $\mathcal{H}$. On suppose que les images $\phi(x_1), \ldots, \phi(x_n)$ sont centrées à l'origine de $\mathcal{H}$, c'est-à-dire $\sum_{i=1}^{n}\phi(x_i) = 0$. L'ACP pratiquée dans l'espace $\mathcal{H}$ consiste à diagonaliser la matrice de covariance

$$C_\phi = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\,\phi(x_i)^\top. \tag{4.1}$$

On cherche alors un vecteur propre Ψ et la valeur propre λ non-nulle associée qui vérifient l'équation

$$\lambda\,\Psi = C_\phi\,\Psi. \tag{4.2}$$

En substituant l'expression (4.1) dans (4.2), on obtient

$$\lambda\, n\,\Psi = \sum_{i=1}^{n} \langle \phi(x_i)\,, \Psi\rangle\,\phi(x_i), \tag{4.3}$$

où $\langle \cdot , \cdot \rangle$ désigne le produit scalaire dans l'espace $\mathcal{H}$, et $\langle \phi(x_i) , \Psi \rangle$ est obtenue par la projection de $\phi(x_i)$ sur le vecteur propre Ψ. En examinant cette expression, on remarque que toute solution Ψ_k appartient à l'espace engendré par $\{\phi(x_1), \ldots, \phi(x_n)\}$, permettant ainsi de l'écrire sous forme de combinaison linéaire

$$\Psi_k = \sum_{i=1}^{n} \alpha_{i,k}\, \phi(x_i). \tag{4.4}$$

En plaçant cette expression dans l'expression (4.3), on obtient

$$\lambda\, n \sum_{i=1}^{n} \alpha_{i,k}\, \phi(x_i) = \sum_{i=1}^{n} \alpha_{i,k} \sum_{j=1}^{n} \phi(x_j) \, \langle \phi(x_j) \, , \phi(x_i) \rangle$$

En calculant le produit scalaire de chacun des deux membres avec $\phi(x_{i'})$ pour tout $i' = 1, \ldots, n$, on obtient des expression ne faisant intervenir les données qu'à travers des produits scalaires de leurs transformées. On définit alors la matrice de Gram K, de dimensions $n \times n$, dont l'élément (i, j) est

$$(K)_{i,j} = \langle \phi(x_i), \phi(x_j) \rangle,$$

ou sous forme vectorielle $(K)_{i,j} = \phi(x_i)^\top \phi(x_j)$. L'expression résultante est alors

$$n\lambda K \boldsymbol{\alpha}_k = K^2 \boldsymbol{\alpha}_k, \tag{4.5}$$

$\boldsymbol{\alpha}_k$ désignant un vecteur colonne formé par $\alpha_{1,k}, \ldots, \alpha_{n,k}$. Pour déterminer la solution à ce problème, on cherche les vecteurs $\boldsymbol{\alpha}$ solutions de

$$n\lambda\, \boldsymbol{\alpha} = K\boldsymbol{\alpha}. \tag{4.6}$$

Soient $\mu_n \geq \ldots \geq \mu_1$ les valeurs propres et $\boldsymbol{\alpha}_n, \ldots, \boldsymbol{\alpha}_1$ les vecteurs propres correspondants. Afin d'avoir des vecteurs propres Ψ_k normalisés dans $\mathcal{H}$, selon $\langle \Psi_k \, , \Psi_k \rangle = 1$ pour tout $k = p, \ldots, n$ avec μ_p la première valeur propre non-nul, on écrit

$$\begin{aligned}
\langle \Psi_k \, , \Psi_k \rangle &= \sum_{i,j=1}^{n} \alpha_{i,k}\, \alpha_{j,k} \langle \phi(x_i) \, , \phi(x_j) \rangle \\
&= \sum_{i,j=1}^{n} \alpha_{i,k}\, \alpha_{j,k} K_{i,j} \\
&= \langle \boldsymbol{\alpha}_k \, , K\boldsymbol{\alpha}_k \rangle \\
&= \mu_k \langle \boldsymbol{\alpha}_k \, , \boldsymbol{\alpha}_k \rangle,
\end{aligned}$$

pour tout $k = p, \ldots, n$. La condition de normalisation de Ψ_k se traduit alors en une normalisation des vecteurs $\boldsymbol{\alpha}_k$, selon

$$\|\boldsymbol{\alpha}_k\|^2 = \frac{1}{\mu_k}. \tag{4.7}$$

La $k^{\text{ème}}$ composante principale d'une donnée test x est obtenue par projection de son image $\phi(x)$ sur Ψ_k selon l'expression

$$\Lambda_k(x) = \langle \phi(x) \, , \Psi_k \rangle = \sum_{i=1}^{n} \alpha_{i,k} \langle \phi(x) \, , \phi(x_i) \rangle \tag{4.8}$$

On remarque que ni l'équation à résoudre (4.6), ni l'expression de la projection (4.8), ne nécessitent d'expliciter les images $\phi(x_i)$ puisqu'elles reposent uniquement sur leurs produits scalaires. On peut alors recourir à des fonctions qui permettent de déterminer les produits scalaires dans l'espace transformé, sans calculer les images. Il s'agit là de l'esprit des méthodes à noyau. L'algorithme de l'ACP-à-noyau est introduit ci-dessous, après la présentation du cas des données non-centrées.

4.2.3 Centrage des données dans l'espace transformé

Pour simplifier, nous avons supposé que les données sont centrées dans l'espace $\mathcal{H}$. Si cette opération est aisée dans $\mathcal{X}$, ceci est moins immédiat dans l'espace transformé $\mathcal{H}$ où il faut expliciter les images afin de les centrer selon $\phi^c(x_i) = \phi(x_i) - \frac{1}{n}\sum_k \phi(x_k)$. Sachant que l'algorithme final de l'ACP dans l'espace $\mathcal{H}$ ne nécessite que le calcul de la matrice de Gram K, on souhaite adopter celle-ci pour centrer les données dans $\mathcal{H}$ sans expliciter leurs images.

Chaque élément de la matrice de Gram des données centrées dans $\mathcal{H}$ peut s'écrire

$$\begin{aligned}
(K^c)_{i,j} &= \langle \phi^c(x_i)\,,\phi^c(x_j)\rangle \\
&= \Big\langle \phi(x_i) - \frac{1}{n}\sum_k \phi(x_k)\,,\phi(x_j) - \frac{1}{n}\sum_{k'}\phi(x_{k'})\Big\rangle \\
&= \langle \phi(x_i)\,,\phi(x_j)\rangle - \frac{1}{n}\sum_k \langle \phi(x_k)\,,\phi(x_j)\rangle \\
&\quad -\frac{1}{n}\sum_{k'}\langle \phi(x_i)\,,\phi(x_{k'})\rangle + \frac{1}{n^2}\sum_{k,k'}\langle \phi(x_k)\,,\phi(x_{k'})\rangle.
\end{aligned}$$

Ceci permet d'écrire sous forme matricielle la matrice de Gram des données centrées K^c à partir de la matrice des données non-centrées K, selon

$$K^c = K - \tfrac{1}{n}\,\mathbf{1}_n K - \tfrac{1}{n}\,K\mathbf{1}_n + \tfrac{1}{n^2}\,\mathbf{1}_n K \mathbf{1}_n, \tag{4.9}$$

avec $\mathbf{1}_n$ désignant la matrice unité de dimensions $n \times n$ telle que $(\mathbf{1}_n)_{ij} = 1$. En factorisant cette expression, on trouve l'expression $K^c = \left(\mathbf{1} - \frac{1}{n}\mathbf{1}_n\right) K \left(\mathbf{1} - \frac{1}{n}\mathbf{1}_n\right)$, où $\mathbf{1}$ est la matrice identité telle que $(\mathbf{1})_{ij} = \delta_{ij}$. On peut alors déterminer la matrice centrée directement à partir de celle non-centrée, sans exhiber aucune des images, pour tout l'ensemble d'apprentissage.

Il nous faut évaluer la matrice de Gram $K^{t,c}$ des données test $x_1^t, \ldots, x_{n_t}^t$ centrées. On détermine cette matrice à partir de la matrice non-centrée K^t formée par les éléments $(K^t)_{i,j} = \langle \phi(x_i^t)\,,\phi(x_j)\rangle$. Pour cela, on utilise la même démarche que précédemment, et on obtient une expression similaire à (4.9) pour $K^{t,c}$, avec

$$K^{t,c} = K^t - \tfrac{1}{n}\,K^t\mathbf{1}_n - \tfrac{1}{n}\,\mathbf{1}_{t,n}K + \tfrac{1}{n^2}\,\mathbf{1}_{t,n}K\mathbf{1}_n, \tag{4.10}$$

où $\mathbf{1}_{t,n}$ désigne la matrice unité de dimensions $t \times n$. On remarque que les deux premiers termes de cette expression dépendent de K^t, donc du produit scalaire $\langle \phi(x^t)\,,\cdot\rangle$, alors que les deux derniers ne contiennent que la matrice K. On peut alors considérer que la $k^{\text{ème}}$ coordonnée principale donnée en (4.8) de tout x_j^t donné s'écrit $\Lambda_k(x_j^t) = \sum_{i=1}^n \alpha_{i,k}\kappa'(x_j^t, x_i) + b$, où $\kappa'(x_j^t, x_i) = (K^t - \frac{1}{n}\,K^t\mathbf{1}_n)_{j,i}$ et le biais b est déterminé à partir de l'expression $\sum_{j=1}^n \alpha_{j,k}\,(\frac{1}{n^2}\mathbf{1}_{t,n}K\mathbf{1}_n - \frac{1}{n}\,\mathbf{1}_{t,n}K)_{j,i}$. On peut alors évaluer les composantes principales pour tout point test sans même avoir à recourir au calcul de son image.

4.2.4 Algorithme de l'ACP-à-noyau

Comme on vient de voir, le calcul d'une ACP dans un espace transformé $\mathcal{H}$ s'exprime sous forme de produits scalaires des observations. D'après le résultat obtenu, tout vecteur propre Ψ_k appartient à l'espace engendré par les images des observations. Ceci permet de l'exprimer comme une combinaison linéaire de ces dernières conformément à l'expression (4.4). Ces deux observations ne font que traduire le coup du noyau et le Théorème de Représentation, qui sont les fondements des méthodes à noyau présentés au Chapitre 1.

Données	$\mathcal{A} = \{x_1, \ldots, x_n\}$	Base d'apprentissage
	$\kappa(\cdot\,,\cdot)\ ,\ x^t$	noyau reproduisant, donnée test
Algorithme	$(K)_{i,j} = \kappa(x_i, x_j)\ ,\ i,j = 1, \ldots, n$	Matrice de Gram des données non-centrées
	$K = K - \frac{1}{n}\,\mathbf{1}_n K - \frac{1}{n}\,K\mathbf{1}_n + \frac{1}{n^2}\,\mathbf{1}_n K \mathbf{1}_n$	Matrice de Gram des données centrées
	$[V, \Lambda] = \mathrm{eig}(K)$	Vecteurs et valeurs propres de K
	$\alpha_{\cdot,k} = \frac{1}{\sqrt{\lambda_k}} v_k$	Vecteurs propres de la matrice de covariance C
Sorties	$\Psi_k = \sum_i \alpha_{i,k} \Psi(x_i)$	$k^{\text{ème}}$ axe principal
	$[x^t]_k = \sum_i \alpha_{i,k} \kappa(x^t, x_i)$	$k^{\text{ème}}$ composante principale de x^t

TAB. 4.1 – L'algorithme ACP-à-noyau

Un résumé de l'algorithme ACP-à-noyau est donné au Tableau 4.1. En utilisant un noyau défini positif $\kappa(\cdot\,,\cdot)$ sur des observations dans $\mathcal{X}$, cet algorithme correspond à une ACP dans un autre espace $\mathcal{H}$. Cette observation nous permet de profiter de la vaste littérature dans le domaine puisque les différentes propriétés de l'ACP se trouvent transférées à l'ACP-à-noyau. Les axes principaux sont orthonormaux dans l'espace $\mathcal{H}$. En supposant qu'ils sont ordonnés par valeurs décroissantes de leurs valeurs propres, les k premiers axes principaux, $\Psi_1, \ldots, \Psi_k$, expliquent plus de variance (de l'ensemble d'apprentissage) que n'importe quelles k autres directions. En projetant les données sur ce sous-espace U_k, on obtient les k premières composantes principales (non-linéaires), minimisant ainsi l'erreur quadratique de reconstruction. On posant $P_{U_k}\phi(x) = \sum_{j=1}^{k} \beta_j \Psi_j$ la projection de $\phi(x)$ sur ce sous-espace, l'ACP-à-noyau minimise $\sum_{i=1}^{n} \|P_{U_k}\phi(x_i) - \phi(x_i)\|^2$ pour une dimension k donnée du sous-espace. Comme $\sum_{i=1}^{k}\lambda_i$ correspond à la variance des données dans le sous-espace U_k, la dimension k est souvent choisie de sorte que $\sum_{i=1}^{k}\lambda_i / \sum_{i=1}^{n}\lambda_i$ correspond à la variance qu'on cherche à restituer.

Tout comme l'ACP classique, l'ACP-à-noyau peut être mise en œuvre dans différents problèmes, en réduction de dimensionnalité et sélection de variables [WB05], détection de nouveauté [Hof06], compression, débruitage et reconstruction [MSS+99, TK02]. Toutes ces références correspondent à l'ACP-à-noyau pour différentes applications, et s'appliquent aisément au domaine temps-fréquence comme on le montre au paragraphe suivant.

4.3 Mise en œuvre de l'ACP dans le domaine temps-fréquence

Comme évoqué au Chapitre 3, on peut adapter le concept des méthodes à noyau dans le domaine temps-fréquence grâce à l'usage d'un noyau approprié. Il s'agit en particulier du cas de l'ACP-à-noyau comme présenté ci-dessous dans le contexte de la distribution de Wigner, que l'on étend ensuite aux représentations linéaires et quadratiques. On conclut par une analyse de la complexité de notre algorithme, que l'on compare à un algorithme classique de l'ACP appliqué aux distributions temps-fréquence des données.

4.3.1 Distribution de Wigner

On considère un ensemble de n signaux $x_1, \ldots, x_n \in \mathcal{X}$, chacun de taille l. L'ACP-à-noyau peut être configurée de sorte à opérer directement sur leurs distributions de Wigner avec le noyau (3.3)-(3.4), soit $\kappa_W(x_i, x_j) = |\langle x_i, x_j\rangle|^2$. Soit $\mathcal{H}_W$ l'espace de Hilbert induit par ce noyau. L'étape principale de l'algorithme est la diagonalisation de la matrice formée par $K_W - \frac{1}{n}\,\mathbf{1}_n K_W - \frac{1}{n}\,K_W\mathbf{1}_n + \frac{1}{n^2}\,\mathbf{1}_n K_W \mathbf{1}_n$, où $(K_W)_{i,j} = \kappa_W(x_i, x_j)$ et $\mathbf{1}_n$ est la matrice unité de taille $n \times n$. Les vecteurs propres $\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_k$ ainsi obtenus, ordonnés par ordre décroissant de leurs valeurs propres, sont normalisés selon $\mu_k \|\boldsymbol{\alpha}_k\|^2 = 1$ comme expliqué précédemment.

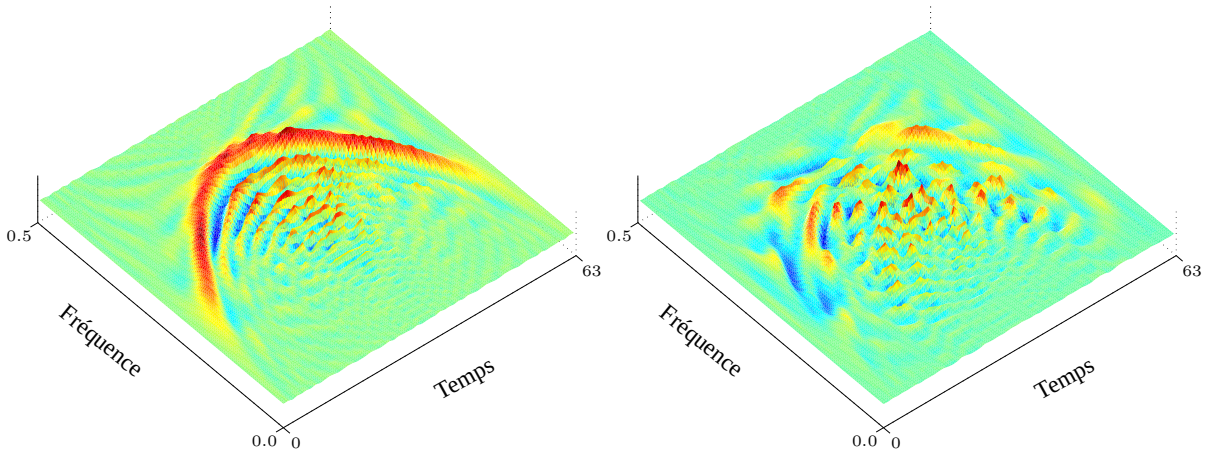


FIG. 4.1 – La première (a) et la seconde (b) signatures principales obtenues par la mise en œuvre de l'ACP-à-noyau avec la distribution de Wigner, pour une famille de signaux à modulation fréquentielle parabolique.

Dans ce contexte, les axes principaux sont définis dans un espace de représentation temps-fréquence, plus particulièrement dans l'espace $\mathcal{H}_W$ engendré par $\{W_{x_1}, \ldots, W_{x_n}\}$. On propose de parler de signature temps-fréquence principale pour désigner un axe principal, bien qu'elle ne soit pas forcément une distribution temps-fréquence valide. La $k^{\text{ème}}$ signature principale s'écrit alors

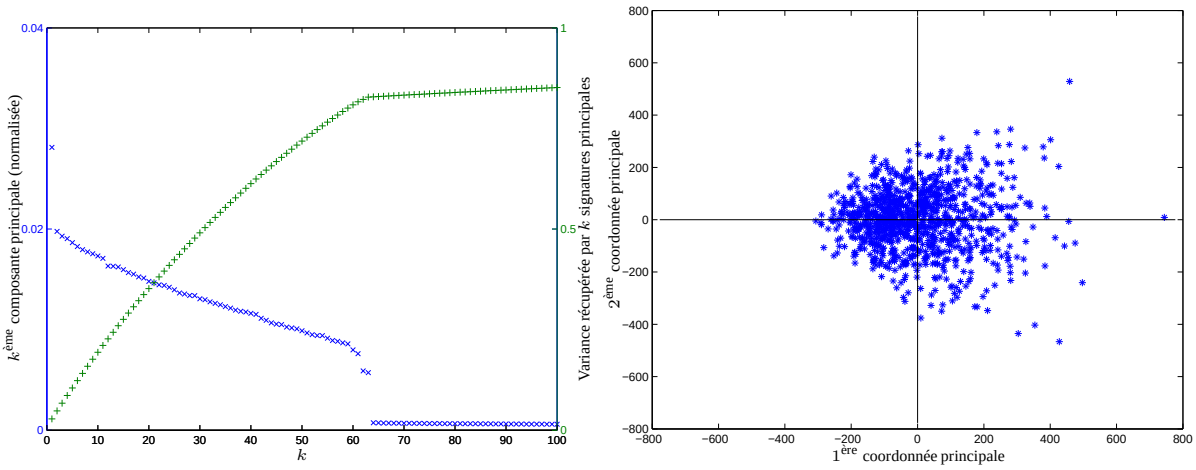
$$\Psi_k = \sum_{i=1}^{n} \alpha_{i,k}\, W_{x_i}, \tag{4.11}$$

avec $\alpha_{i,k}$ la $i^{\text{ème}}$ composante de $\boldsymbol{\alpha}_k$. Les signatures principales sont orthonormales, vérifiant l'expression $\langle \Psi_i, \Psi_j \rangle = \delta_{ij}$ où δ désigne le symbole de Kronecker. Cette propriété se traduit dans le plan temps-fréquence par l'expression d'orthonormalité $\iint \Psi_i(t,f)\, \overline{\Psi_j}(t,f)\, \mathrm{d}t\, \mathrm{d}f = \delta_{ij}$, pour tout $i, j = 1, \ldots, n$. Pour un signal x donné, la composante principale extraite de la $k^{\text{ème}}$ signature principale est obtenue par la projection de W_x sur Ψ_k. Ceci se traduit par l'expression

$$\Lambda_k(x) = \langle W_x, \Psi_k \rangle = \sum_{i=1}^{n} \alpha_{i,k}\, \kappa_W(x, x_i).$$

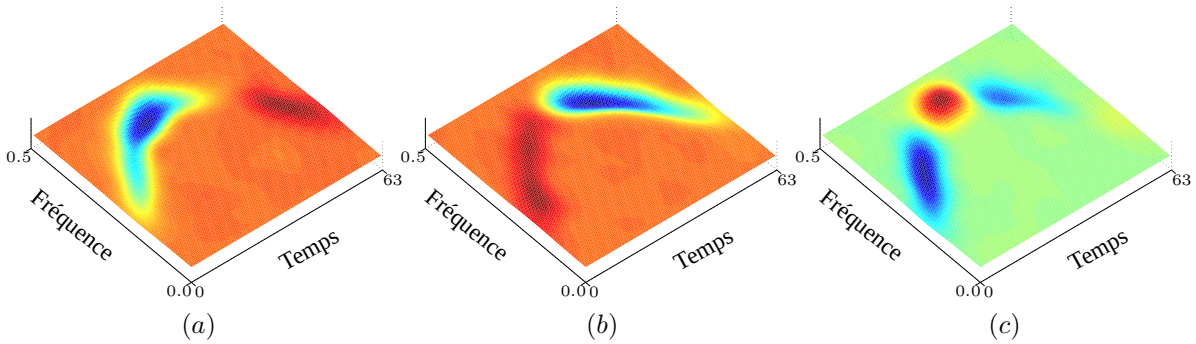
Cette expression correspond à la $k^{\text{ème}}$ coordonnée dite non-centrée de x. En modifiant $\kappa_W(x, x_i)$ selon (4.10), on centre implicitement W_x dans l'espace engendré par les Ψ_k. On considère l'espace engendré par les k premières distributions propres, $U_k = \{\Psi_1, \ldots, \Psi_k\}$. Il est possible de projeter les représentations dans cet espace, obtenant ainsi un ensemble de k composantes principales $\Lambda_1(x), \ldots, \Lambda_k(x)$ pour la distribution de Wigner d'un signal x. On peut alors représenter le signal par ces k coordonnées. En considérant $k < n$, on parle de réduction de dimensionnalité. On rappelle que l'espace U_k obtenu par l'ACP minimise l'erreur quadratique moyenne de reconstruction $\sum_{i=1}^{n} \|P_{U_k} W_{x_i} - W_{x_i}\|^2$.

Pour illustrer la mise en œuvre de l'ACP-à-noyau avec la distribution de Wigner, on considère l'exemple suivant. Soit un ensemble de $1\,000$ signaux de taille 64. Chaque signal contient une modulation fréquentielle parabolique variant de 0.1 à 0.4 Hz, noyé dans un bruit blanc Gaussien additif, avec un rapport signal-sur-bruit de 0 dB. L'ACP-à-noyau de Wigner est alors utilisée pour déterminer les signatures principales Ψ_k et les composantes principales $\Lambda_k(\cdot)$ de l'ensemble d'apprentissage. La première



(a) (b)

FIG. 4.2 – On illustre en (a) les valeurs propres normalisées λ_k par $\times$ et leurs cumulées $\sum_{i=1}^{k} \lambda_i$ par $+$ associées aux 100 premières distributions principales. En (b), les signaux sont représentés dans l'espace défini par les deux premières signatures principales.



(a) (b) (c)

FIG. 4.3 – Les 3 premières signatures principales obtenues à partir du spectrogramme pour la même base de données qu'à la Figure 4.1.

signature principale Ψ_1 représentée à la Figure 4.1 (a) montre qu'elle est en mesure à elle seule d'extraire l'information temps-fréquence des données bruitées. Cette caractéristique des signaux est absente de la seconde signature principale, illustrée à la Figure 4.1 (b). Cette observation peut être quantifiée à partir des valeurs propres λ_i. On considère la valeur λ_k qui correspond à la variance extraite par la signature principale Ψ_k. Cette valeur est souvent normalisée par la variance totale $\sum_{i=1}^{n} \lambda_i$ afin d'évaluer le pourcentage de variance expliquée. A la Figure 4.2 (a), on représente en $\times$ ces valeurs normalisées pour les 100 premières signatures principales. On retrouve le fait que la première signature principale est beaucoup plus intéressante que les autres. D'autre part, on représente par $+$ sur la même figure les valeurs de $\sum_{i=1}^{k} \lambda_i$, pour $k = 1, \ldots, 100$. Ces valeurs correspondent à la variance expliquée par les sous-espaces optimaux de dimensions k, normalisées par la variance totale dans $\mathcal{H}_W$.

4.3.2 Autres distributions temps-fréquence

Dans le cas d'une distribution de la classe de Cohen, un algorithme similaire au cas précédent est mis en place, la seule différence majeure étant l'usage du noyau reproduisant $\kappa_{C^\Pi}(x_i, x_j) = \langle C^\Pi_{x_i}, C^\Pi_{x_j} \rangle$, où Π désigne la fonction de paramétrisation de la distribution.

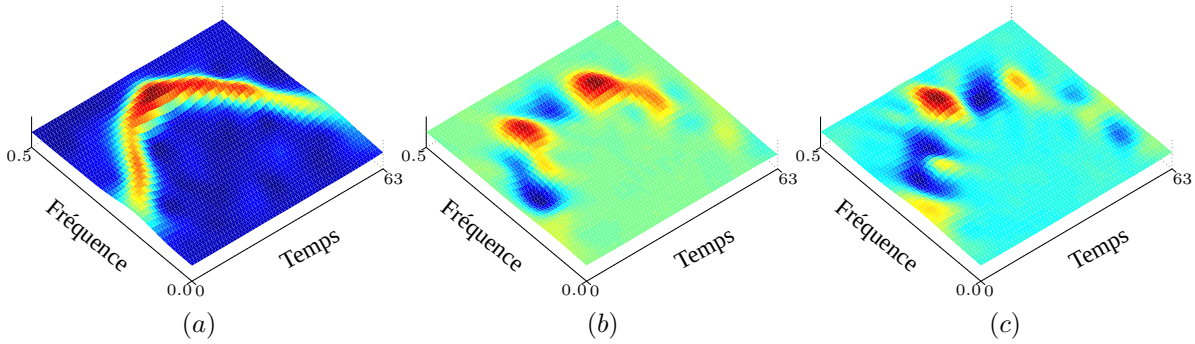


FIG. 4.4 – Les 3 premières signatures principales obtenues par la stratégie hybride combinant noyau quadratique et spectrogramme.

Pour le spectrogramme, on considère le noyau $\kappa_{S^w}(x_i, x_j) = \iint |\langle x_i, w_{t,f}\rangle\langle x_j, w_{t,f}\rangle|^2 \,\mathrm{d}t\,\mathrm{d}f$, où $w_{t,f}(\tau) = w(\tau - t)\; e^{2j\pi f\tau}$. On adopte une fenêtre de lissage de Hamming de taille 16. Comme le montre la Figure 4.3, les trois premières signatures principales ont hérité du caractère lissé du spectrogramme et de l'absence d'interférence. Cette propriété produit des distributions quasi-dépourvues de bruit. Le Figure 4.5 (a) présente la variance expliquée par chacune des distributions principales. En projetant les signaux d'apprentissage sur les deux premières, on obtient leurs coordonnées principales, présentées à la Figure 4.5 (b). Pour comparer les résultats obtenus avec la distribution de Wigner et le spectrogramme, on propose de projeter les données sur leur première signature principale, comme présenté à la Figure 4.5 (c). On retrouve l'effet de lissage du spectrogramme qui produit des représentations à faible variance. On s'intéresse maintenant à la stratégie hybride introduite à la Section 3.3.3. En conjuguant le noyau reproduisant associé à la distribution de Wigner et l'espace de représentation associé au spectrogramme, on obtient les trois premières signatures principales illustrées à la Figure 4.4. Cette approche s'étend facilement à d'autres distributions de la classe de Cohen. On peut alors aborder l'étude des premières signatures principales pour chaque type de distributions temps-fréquence, ou encore les comparer entre elles, comme on l'a illustré à la Figure 4.5 (c) avec le spectrogramme et la distribution de Wigner. On illustre en Figure 4.6 les 200 premières valeurs propres normalisées associées aux différentes distributions temps-fréquence de la classe de Cohen. Enfin, on s'intéresse en Figure 4.7 à la transformée de Fourier à court-terme. On y représente le module des deux premières signatures principales en (a) et (b), ainsi que les 100 premières valeurs propres normalisées en (c).

4.3.3 Complexité calculatoire

Distribution de Wigner

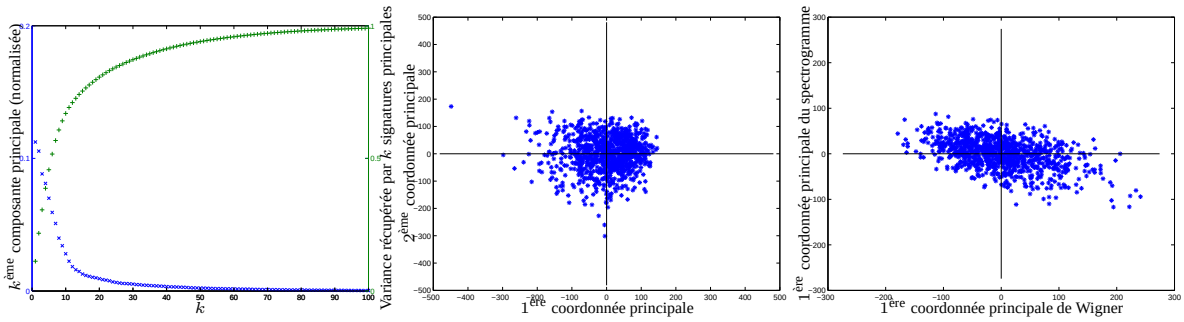
L'utilisation de l'ACP classique directement sur les distributions de Wigner, $W_{x_1}, \ldots, W_{x_n}$, produit le même résultat que l'approche préconisée dans ce chapitre, à savoir l'usage du noyau de Wigner κ_W avec l'ACP-à-noyau. Toutefois, l'approche classique est souvent très lourde en calcul. Ceci est principalement dû à la génération de la matrice de covariance, et à sa diagonalisation. Puisque chaque distribution de Wigner est de taille l^2, la matrice de covariance des n distributions, $C = \frac{1}{n}\sum_{i=1}^{n} W_{x_i} W_{x_i}^{\top}$, est de taille $l^2 \times l^2$, nécessitant ainsi $\mathcal{O}(l^4 n)$ opérations. La détermination des vecteurs propres et valeurs propres d'une matrice de cette taille requiert $\mathcal{O}(l^6)$ opérations.

L'ACP-à-noyau conduit à un algorithme beaucoup plus efficace. Le Tableau 4.2 en présente les différentes étapes de l'algorithme dans le cas de la distribution de Wigner, ainsi que leur complexité calculatoire. Pour comparer cet algorithme à l'ACP classique, on remplace dans celle-ci les deux étapes

Instructions	Expressions	Complexité
Algorithme de base		
1. Calculer la matrice de Gram	$(K)_{i,j} = \kappa_W(x_i, x_j) = \lvert\langle x_i, x_j\rangle\rvert^2$	$\mathcal{O}(n^2 l)$
2. Centrer implicitement les données	$K = K - \frac{1}{n}\mathbf{1}_n K - \frac{1}{n} K\mathbf{1}_n + \frac{1}{n^2}\mathbf{1}_n K\mathbf{1}_n$	$\mathcal{O}(n^2)$
3. Décomposition de K et normalisation	$[V, \Lambda] = \mathrm{eig}(K)$	$\mathcal{O}(n^3)$
Sortie : **Signature principale**		
4. Calculer n distributions de Wigner	W_{x_i}	$\mathcal{O}(nl^2 \log l)$
5. La $k^{\text{ème}}$ signature principale	$\Psi_k = \sum_{i=1}^{n} \alpha_{i,k} W_{x_i}$	$\mathcal{O}(nl^2)$
Sortie : **Composante principale**		
6. Calculer la matrice de Gram test	$(K^t)_{i,j} = \kappa_W(x_i^t, x_j) = \lvert\langle x_i^t, x_j\rangle\rvert^2$	$\mathcal{O}(ntl)$
7. Centrer implicitement les données test	$K^t = K^t - \frac{1}{n}\mathbf{1}_{t,n} K - \frac{1}{n} K^t\mathbf{1}_n + \frac{1}{n^2}\mathbf{1}_{t,n} K\mathbf{1}_n$	$\mathcal{O}(nt)$
8. Les $k^{\text{ème}}$ composantes principales	$\Lambda_k(x_i^t) = \sum_{j=1}^{n} \alpha_{j,k} (K^t)_{i,j}$	$\mathcal{O}(tn^2)$

TAB. 4.2 – L'algorithme ACP-à-noyau pour la distribution de Wigner et le coût calculatoire de chacune de ses étapes, l étant la taille des signaux, n et n^t les tailles respectives de l'ensemble d'apprentissage $\{x_i\}$ et de l'ensemble de test $\{x_i^t\}$.

évoquées ci-dessus et de complexité $\mathcal{O}(l^4 n)$ et $\mathcal{O}(l^6)$, par les étapes 1., 2. et 3. du tableau qui sont de complexités $\mathcal{O}(n^2 l)$, $\mathcal{O}(n^2)$ et $\mathcal{O}(n^3)$ respectivement compte tenu du fait que la matrice à diagonaliser est de taille $n \times n$. Cette efficacité calculatoire s'étend aussi à la détermination des signatures principales et des composantes principales. Pour les résultats, on traite séparément dans le tableau le cas de la signature principale et celui de la composante principale. Dans le premier cas, on suppose que le calcul de la distribution de Wigner nécessite $\mathcal{O}(l^2 \log l)$ opérations, sans tenir compte de la nature Hermitienne de la fonction d'ambiguïté, ni du type de discrétisation. Un calcul itératif permet la mise à jour de Ψ_k pour chaque distribution de Wigner W_{x_i}, selon $\Psi_k \leftarrow \Psi_k + \alpha_{i,k} W_{x_i}$, sans la nécessité de mettre en mémoire toutes les distributions temps-fréquence nécessaires. D'autre part, la détermination de la composante principale ne nécessite aucun calcul de distribution.



(a) (b) (c)

FIG. 4.5 – Résultats obtenus dans le cas du spectrogramme : en (a) on présente les valeurs propres normalisées et leur somme cumulée, et en (b) on représente les signaux dans l'espace défini par les deux premières signatures principales. On compare les deux approches, Wigner et spectrogramme, en représentant les données dans leur espace en (c).

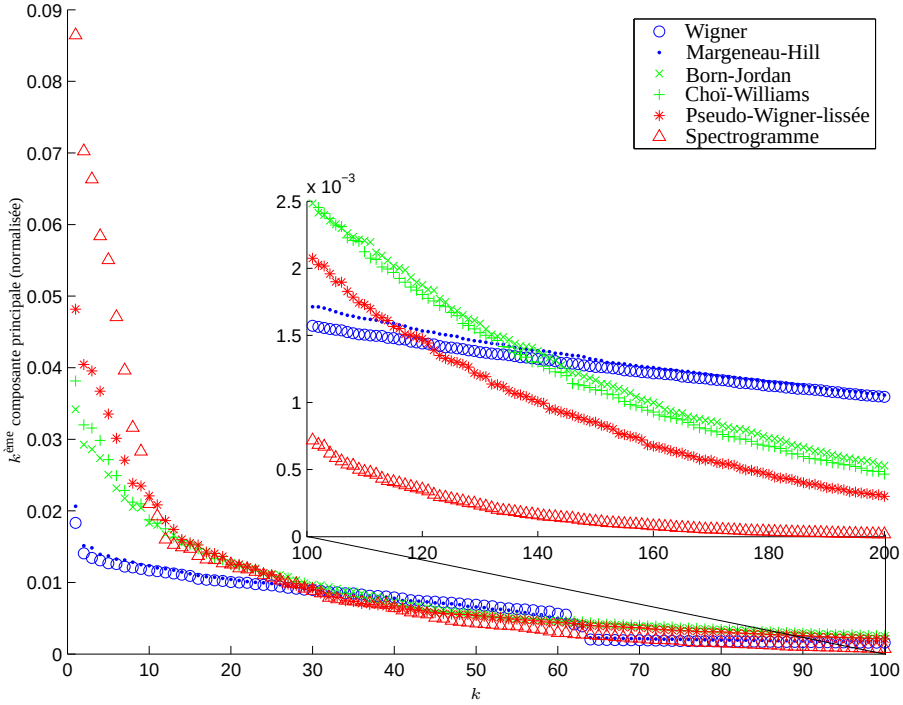
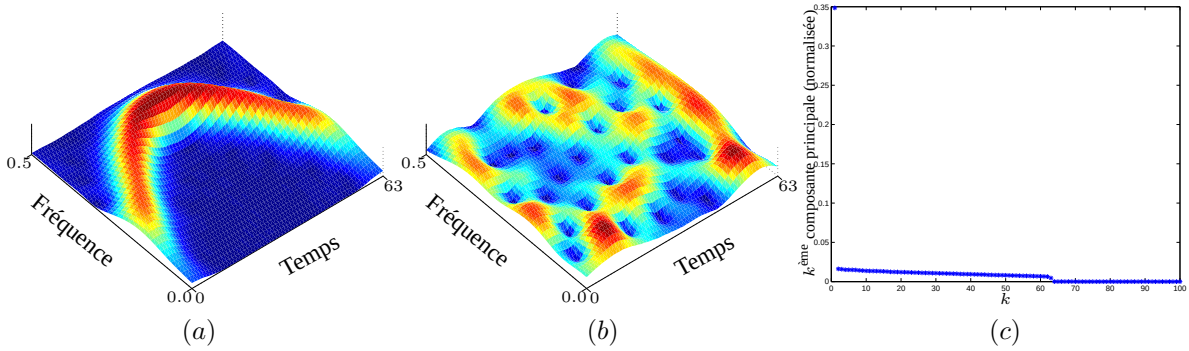


FIG. 4.6 – Comparaison de la répartition des 200 premières valeurs propres pour différentes distributions temps-fréquence de la classe de Cohen

Classe de Cohen

Pour les distributions de la classe de Cohen, on obtient un algorithme similaire dans le cas des distributions unitaires, donc d'une complexité équivalente à celle obtenue par Wigner. Ceci n'est pas le cas de la plupart des distributions quadratiques, où la différence principale est alors la détermination de la matrice de Gram qui nécessite en général le calcul des n distributions. On considère que toutes les distributions de la classe de Cohen nécessitent $\mathcal{O}(l^2 \log l)$ opérations. L'étape 1. dans le Tableau 4.2 est alors remplacée par deux étapes, le calcul des n distributions C_{x_i} avec une complexité de $\mathcal{O}(nl^2 \log l)$, et la détermination de la matrice de Gram $(K)_{i,j} = \langle C_{x_i}, C_{x_j} \rangle$, avec $\mathcal{O}(n^2 l^2)$ opérations. Dans le cas de l'approche hybride, on détermine les coefficients de pondérations $\alpha_{\cdot,k}$ par l'algorithme de Wigner, puis



(a) (b) (c)

FIG. 4.7 – On présente en (a) et (b) le module des deux premières signatures principales obtenues par la transformée de Fourier à court-terme, et en (c) les 100 premières valeurs propres.

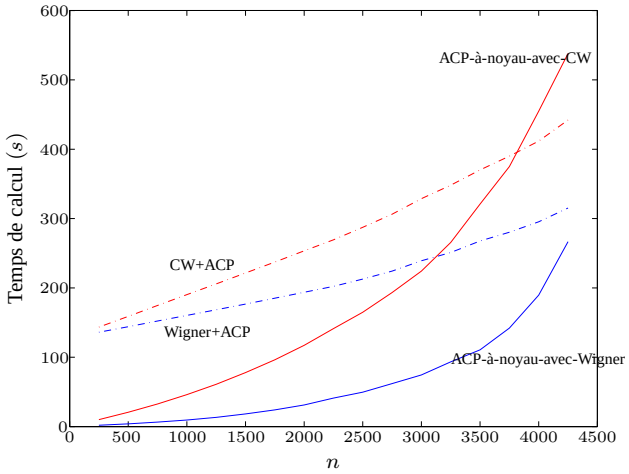


FIG. 4.8 – Temps de calcul pour l'ACP-à-noyau et l'ACP appliquée aux distributions temps-fréquence, en fonction du nombre n de signaux de taille 64. On présente les résultats obtenus à partir de deux distributions, Wigner et Choï-Williams (CW).

les signatures principales sont déterminées en calculant les distributions temps-fréquence en question. On obtient alors une complexité similaire à celle de la distribution de Wigner.

La Figure 4.8 présente un comparatif du temps de calcul, en fonction de la taille n de l'ensemble d'apprentissage de signaux de taille 64, de l'ACP-à-noyau et de l'ACP classique appliquée aux distributions temps-fréquence. On considère en particulier la distribution de Wigner et la distribution de Choï-Williams. La première étant unitaire, alors que la seconde ne l'est pas. L'implémentation est faite sur Matlab avec la Toolbox Temps-Fréquence [AFGL05], et les expérimentations reposent sur un ordinateur portable de processeur Pentium M 1.60 GHz, et de mémoire vive de 1 GB. Comme prévu, le temps calculatoire est linéaire en n pour l'ACP appliquée directement sur les différentes distributions temps-fréquence. Ceci est dû à la complexité de calcul de la matrice de covariance, qui est de l'ordre de $\mathcal{O}(l^4 n)$, donc linéaire en n. D'autre part, celui-ci est polynômial pour l'ACP-à-noyau, comme dans le cas de la distribution de Wigner, dû aux étapes 1., 2. et 3. qui sont en $\mathcal{O}(n^2 l)$, $\mathcal{O}(n^2)$ et $\mathcal{O}(n^3)$ respectivement. On peut vérifier que l'algorithme de l'ACP-à-noyau est plus efficace que celui de l'ACP, tant que n est inférieur à l^2, vu que les deux approches reposent sur la diagonalisation de matrices de dimensions $n \times n$ et $l^2 \times l^2$ respectivement. En pratique, cette condition est souvent remplie.

Transformation linéaire

Dans le cas des transformations linéaires, le calcul est plus simple puisque le noyau utilisé est un noyau linéaire, et l'axe principal est obtenu par une seule transformation d'un signal formé par combinaison linéaire des signaux de l'ensemble d'apprentissage. Le Tableau 4.3 présente la complexité des différentes étapes de l'algorithme de l'ACP pour une transformation linéaire, en particulier le cas de la transformée de Fourier à court-terme. Comme centrer implicitement les données revient à les centrer dans l'espace d'entrée, on considère la seconde approche, ne nécessitant aucun calcul matriciel.

Optimisation de la complexité calculatoire

Pour déterminer les vecteurs propres et les valeurs propres de la matrice de Gram de taille $n \times n$, on a besoin de $\mathcal{O}(n^3)$ opérations. Un grand nombre d'applications ne nécessite pas le calcul de tous les

Instructions	Expressions	Complexité
Algorithme de base		
1. Centrer les données	$x_i^c = x_i - \frac{1}{n}\sum_{j=1}^n x_j$	$\mathcal{O}(n)$
2. Calculer la matrice de Gram	$(K)_{i,j} = \kappa_{F^w}(x_i^c, x_j^c) = \|w\|^2 \langle x_i^c, x_j^c \rangle$	$\mathcal{O}(n^2 l)$
3. Décomposition de K et normalisation	$[V, \Lambda] = \mathrm{eig}(K)$	$\mathcal{O}(n^3)$
Sortie : **Signature principale**		
4. Le $k^{\text{ème}}$ signal principal	$z_k = \sum_{i=1}^n \alpha_{i,k}\, x_i$	$\mathcal{O}(n)$
5. La $k^{\text{ème}}$ signature principale	$\Psi_k = F^w_{z_k}$	$\mathcal{O}(l^2)$
Sortie : **Composante principale**		
6. Centrer les données test	$x_i^{t,c} = x_i^t - \frac{1}{n^t}\sum_{j=1}^n x_j^t$	$\mathcal{O}(n^t)$
8. Les $k^{\text{ème}}$ composantes principales	$\Lambda_k(x_i^t) = \kappa_{F^w}(x_i^{t,c}, z_k) = \|w\|^2 \langle x_i^{t,c}, z_k \rangle$	$\mathcal{O}(n^t l)$

TAB. 4.3 – L'algorithme ACP-à-noyau pour la transformée de Fourier à court-terme, ainsi que sa complexité calculatoire, l étant la taille des signaux, n et n^t les tailles respectives de l'ensemble d'apprentissage $\{x_i\}$ et de l'ensemble de test $\{x_i^t\}$.

vecteurs propres. Dans ce cas, on peut avoir recours à des méthodes itératives pour calculer les vecteurs propres correspondants aux valeurs propres les plus pertinentes. Dans [KFS03, KFS05] par exemple, les auteurs proposent une technique d'ACP-à-noyau itérative basée sur l'algorithme Hebbien généralisé, obtenant non seulement une implémentation efficace mais aussi une capacité de traitement en-ligne de l'ACP. Les auteurs de [RA03] proposent une méthode itérative pour réduire la complexité calculatoire de calcul d'une composante principale.

Au-delà de ces approches destinées à proposer des algorithmes séquentiels pour l'ACP-à-noyau, on propose dans la Partie II de ce document une approche originale pour déterminer un sous-espace optimal au sens de l'ACP-à-noyau. Celle-ci est soutenue par des résultats théoriques, dont la Proposition 7.7 au Chapitre 7. Au Chapitre 8, on propose un algorithme séquentiel d'ACP-à-noyau, avec une complexité calculatoire linéaire par rapport à l'ordre du modèle réduit considéré. Le dernier chapitre est dédié à la mise en œuvre de ces algorithmes sur des signaux biomédicaux. Mais avant, on poursuit l'étude de méthodes de reconnaissance des formes dans le domaine temps-fréquence dans un cadre d'apprentissage supervisé.

Chapitre 5

Discrimination de signaux dans le domaine temps-fréquence : l'analyse factorielle discriminante

Sommaire

Ce chapitre est consacré au problème d'apprentissage supervisé, plus précisément à la discrimination de signaux non-stationnaires. On s'intéresse à l'algorithme classique de l'analyse factorielle discriminante qui permet de déterminer une direction maximisant la séparabilité des classes tout en minimisant la variance au sens de celles-ci, conformément au critère de Fisher. Par l'usage du coup du noyau et du Théorème de Représentation, il est possible de reconsidérer cette méthode dans un espace transformé plus pertinent pour représenter les données. Pour la discrimination de signaux non-stationnaires, un choix naturel est donné par les distributions temps-fréquence.

Ce chapitre est organisé ainsi. Tout d'abord, on rappelle le contexte de la discrimination en reconnaissance des formes. On décrit alors l'analyse factorielle discriminante, ainsi que sa version non-linéaire obtenue au moyen d'un noyau reproduisant et du RKHS associé. On étudie alors cette approche dans le domaine temps-fréquence. On conclut ce chapitre par des applications dans le cadre de la discrimination à deux classes et plus.

5.1 Introduction

En reconnaissance des formes, l'analyse discriminante consiste principalement à déterminer un espace de représentation qui maximise la séparabilité de données appartenant à différentes classes, permet-

tant ainsi de définir une statistique de décision qui détermine l'appartenance de nouvelles données à l'une des classes en question. Dans le cas particulier de données appartenant à deux classes, on cherche une direction telle que leurs projections selon cette direction présentent le plus fort contraste. Cette idée a été initialement développée par Fisher [Fis36] avec l'analyse factorielle discriminante (AFD), qui conduit à une direction qui maximise l'écart des moyennes des deux classes dans cette direction, tout en minimisant leurs variances respectives.

Si la portée de l'AFD sous sa forme standard se limite à la détermination de statistiques linéaires, il est possible d'élaborer des règles de discrimination non-linéaires par l'application préalable d'une transformation non-linéaire sur les données. Or, il s'avère que ce problème peut être résolu en ne considérant que les produits scalaires des données transformées de l'ensemble d'apprentissage. Ce résultat, initialement présenté par Mika *et coll.* dans [MRW$^+$99], fait usage du coup du noyau sans qu'il y ait nécessité d'exhiber la transformation non-linéaire associée. L'algorithme AFD-à-noyau [MRW$^+$99, Mik02] profite de ce concept pour résoudre un problème de discrimination non-linéaire à deux classes. Une généralisation de cette méthode aux problèmes multi-classes est proposé dans [BA00], connue sous le nom d'analyse discriminante généralisée à noyau (ADG-à-noyau).

Pour la classification de signaux non-stationnaires, il peut être naturel d'adopter le plan temps-fréquence comme espace de représentation préalable à la phase de discrimination. Dans ce chapitre, on considère la mise en œuvre de l'AFD-à-noyau dans le domaine temps-fréquence par un choix approprié du noyau reproduisant. L'usage de critères de séparabilité inter-classes pour classer des signaux à partir de leurs représentations temps-fréquence a fait l'objet de plusieurs travaux. Parmi ceux-ci, voir par exemple [Dav04] pour un résumé de ces approches, on retrouve les travaux de Heitz *et coll.* [Hei95, HT97] sur la maximisation de la distance entre les centres des classes des distributions. Outre la séparabilité des centres, l'importance de prendre en considération la variance de chaque classe est soulignée dans les travaux de Atlas *et coll.*, et en particulier dans [GA01]. Dans la plupart de ces articles, la question du choix de la distribution temps-fréquence est posée. On laisse ici ce problème ouvert en attendant de le traiter au Chapitre 6 à l'aide du critère d'alignement noyau-cible.

De l'ACP à l'AFD : le quotient de Rayleigh

Lors du Chapitre 4, nous avons introduit l'ACP et son extension à noyau par maximisation de la variance projetée, qui en est certainement la formulation la plus connue. Il est toutefois possible de traiter ce problème en considérant d'autres critères, par exemple la minimisation de l'erreur de reconstruction ou encore la maximisation d'un quotient de Rayleigh. Dans ce qui suit, on adopte ce dernier point de vue car il offre un cadre général pour l'extraction de caractéristiques et s'applique à l'AFD.

Pour un problème d'ACP, on recherche la direction u qui maximise la variance de la projection des données $\phi(x_i)$, pour tout $i = 1, \ldots, n$. En exprimant ces quantités sous forme vectorielle, on peut écrire le risque empirique selon

$$\frac{1}{n}\sum_{i=1}^{n} |\langle \phi(x_i), u\rangle|^2 = \frac{1}{n}\sum_{i=1}^{n} u^\top \phi(x_i)\ \phi(x_i)^\top u = u^\top \left[\frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\ \phi(x_i)^\top\right] u = u^\top C\, u,$$

où $C = \frac{1}{n}\sum_{i=1}^{n} \phi(x_i)\ \phi(x_i)^\top$ est la matrice de covariance des données transformées. Pour avoir une solution unique, on inclut une contrainte sur la norme de u, avec $u^\top u = 1$. Ceci revient à maximiser

$$Q(u) = \frac{u^\top C\, u}{u^\top u}, \tag{5.1}$$

une expression connue dans le littérature sous le nom de quotient de Rayleigh. Le principe de Rayleigh stipule que la plus grande valeur propre de la matrice de covariance coïncide avec le maximum de ce quotient. Pour montrer cela, il suffit d'en annuler la dérivée par rapport à u, soit

$$\frac{\partial Q}{\partial u} = \frac{(u^\top u) C\, u - (u^\top C\, u)\, u}{(u^\top u)^2} = 0.$$

Ceci implique $(u^\top u) C\, u = (u^\top C\, u)\, u$, ou encore

$$C\, u = \left(\frac{u^\top C\, u}{u^\top u} \right) u.$$

Le vecteur u est alors un vecteur propre de la matrice de covariance C correspondant à la valeur propre $\frac{u^\top C\, u}{u^\top u}$. La direction optimale correspond à la plus grande valeur de $Q(u)$, c'est-à-dire à la plus grande valeur propre. Si l'ACP est un puissant outil pour la représentation des données, rien ne garantit son intérêt pour la résolution de problèmes de discrimination du fait que cette technique non-supervisée ne tient pas compte des classes d'appartenance des données. Les techniques d'AFD permettent de traiter ce problème de discrimination, par maximisation d'un quotient de Rayleigh spécifique.

5.2 Analyses factorielles discriminantes, classique et à noyau

L'idée principale de Fisher consiste à déterminer la direction qui maximise la séparabilité inter-classes, tout en minimisant les variances intra-classes. Dans ce qui suit, on considère un ensemble d'apprentissage $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ formé de n signaux $x_i \in \mathcal{X}$, et de leurs étiquettes $y_i \in \mathcal{Y}$, selon l'appartenance de x_i à l'une des J classes. On désigne par $\mathcal{C}_j$ l'ensemble des n_j signaux x_i appartenant à la $j^{\text{ème}}$ classe, et par $\mathcal{C}_0 = \bigcup_{j=1}^{J} \mathcal{C}_j$ l'ensemble de tous les signaux de l'ensemble d'apprentissage sans tenir compte de leur appartenance aux différentes classes, avec $n_0 = n$. Soit m_j la moyenne (empirique) des données appartenant à $\mathcal{C}_j$, soit

$$m_j = \frac{1}{n_j} \sum_{x_i \in \mathcal{C}_j} x_i.$$

La moyenne (empirique) totale est définie par $m_0 = \frac{1}{n} \sum_{i=1}^{n} x_i$.

5.2.1 Analyse factorielle discriminante linéaire

Le problème traité peut être décomposé en deux parties. D'une part, on souhaite déterminer la direction u qui maximise la séparabilité des données par le biais des moyennes projetées de chaque classe en considérant le critère suivant

$$\frac{1}{n} \sum_{j=1}^{J} n_j \big(u^\top (m_j - m_0)\big)^2 = u^\top S\, u,$$

avec $S = \frac{1}{n} \sum_{j=1}^{J} (m_j - m_0)(m_j - m_0)^\top$ la matrice de dispersion des moyennes. D'autre part, on souhaite minimiser conjointement la dispersion intra-classe des données projetées par le critère suivant

$$\sum_{j=1}^{J} \sum_{x_i \in \mathcal{C}_j} \big(u^\top (x_i - m_j)\big)^2 = u^\top D\, u.$$

En combinant ces deux critères, on aboutit au contraste de Fisher qui consiste à maximiser le quotient

$$Q(u) = \frac{u^\top S\, u}{u^\top D\, u}. \tag{5.2}$$

Cette expression correspond à un quotient de Rayleigh. Une solution peut être obtenue en annulant la dérivée de Q par rapport à u, c'est-à-dire

$$\frac{\partial Q}{\partial u} = \frac{(u^\top Du)\, S\, u - (u^\top S\, u)\, D\, u}{(u^\top Du)^2} = 0.$$

La direction optimale correspond alors à l'un des vecteurs propres du problème généralisé

$$S\, u = \lambda\, D\, u, \tag{5.3}$$

avec $\lambda = \frac{u^\top S\, u}{u^\top Du}$ la valeur propre correspondante. Le vecteur u qui maximise la séparabilité des classes est alors donné par le vecteur propre de (5.3) associé à la plus grande valeur propre, cette dernière correspondant à la valeur du quotient de Rayleigh.

Une étude plus détaillée du problème [DHS00] montre que le nombre de valeurs propres non-nulles de la matrice S est inférieur ou égal à $J-1$, avec J le nombre de classes dans le problème considéré. Ceci est dû au fait que S est obtenue par la sommation de J matrices, $(m_j - m_0)(m_j - m_0)^\top$ pour $j = 1, \ldots, J$, chacune de rang égal au plus à 1, et que parmi celles-ci seules $J-1$ sont linéairement indépendantes. Comme chacune de ces valeurs propres correspond à la variance des données dans la direction correspondante, il est alors souhaitable de maximiser le produit de ses valeurs propres non-nulles, ce qui correspond au déterminant de la matrice en question. Le problème est alors donné par la maximisation d'un quotient de Rayleigh de la forme

$$Q(U) = \frac{|U^\top S\, U|}{|U^\top D\, U|}, \tag{5.4}$$

où $|\cdot|$ désigne le déterminant d'une matrice, et U la matrice formée en colonnes par les $J-1$ vecteurs propres $u_1, \ldots, u_{J-1}$. En récapitulant, l'approche consiste à résoudre le problème généralisé (5.3) et à déterminer les $J-1$ vecteurs propres correspondant aux valeurs propres les plus importantes.

5.2.2 Analyse factorielle discriminante à noyau

On considère une transformation non-linéaire ϕ de $\mathcal{X}$ vers un autre espace $\mathcal{H}$, et κ le noyau reproduisant associé à cette transformation. Pour utiliser l'approche AFD linéaire dans le cadre d'un problème de discrimination non-linéaire, on considère la mise en œuvre de l'AFD dans l'espace $\mathcal{H}$, sur les données transformées $\phi(x_i)$. Recourir au coup du noyau nécessite la reformulation du problème sous forme de produits scalaires des différents données transformées.

On désigne par $m_0^\phi, m_1^\phi, \ldots, m_J^\phi$ les moyennes des données dans l'espace transformé, soit

$$m_j^\phi = \frac{1}{n_j} \sum_{x_i \in \mathcal{C}_j} \phi(x_i).$$

On cherche la direction, représentée par le vecteur Ψ, qui maximise le quotient de Rayleigh

$$Q_\phi(\Psi) = \frac{\Psi^\top S_\phi\, \Psi}{\Psi^\top D_\phi\, \Psi}, \tag{5.5}$$

où S_ϕ et D_ϕ désignent respectivement la matrice de dispersion des moyennes des classes et la matrice de covariance des données transformées. On écrit alors

$$S_\phi = \frac{1}{n}\sum_{j=1}^{J}(m_j^\phi - m_0^\phi)(m_j^\phi - m_0^\phi)^\top$$

$$D_\phi = \frac{1}{n}\sum_{j=1}^{J}\sum_{x_i\in\mathcal{C}_j}(\phi(x_i) - m_j^\phi)(\phi(x_i) - m_j^\phi)^\top.$$

Comme présenté précédemment avec l'équation (5.3) dans le cas linéaire, le vecteur Ψ est alors donné par un vecteur propre de

$$S_\phi\,\Psi = \lambda\, D_\phi\,\Psi$$

qui correspond à la plus grande valeur propre. Il est connu que ces vecteurs propres appartiennent à l'espace engendré par les n vecteurs transformés $\phi(x_1),\ldots,\phi(x_n)$, soit

$$\Psi = \sum_{k=1}^{n}\alpha_k\phi(x_k), \tag{5.6}$$

Ce résultat n'est autre que le Théorème de Représentation, qui s'applique dans ce cadre particulier comme on sera amené à le revoir dans la Section 5.2.4. La projection sur ce vecteur Ψ d'une des J moyennes $m_1^\phi,\ldots,m_J^\phi$, est donnée par l'expression

$$\langle\Psi, m_j^\phi\rangle = \frac{1}{n_j}\sum_{x_i\in\mathcal{C}_j}\sum_{k=1}^{n}\alpha_k\,\langle\phi(x_k),\phi(x_i)\rangle = \frac{1}{n_j}\sum_{x_i\in\mathcal{C}_j}\sum_{k=1}^{n}\alpha_k\,\kappa(x_k,x_i) = \boldsymbol{\alpha}^\top\boldsymbol{\kappa}_j,$$

où $\boldsymbol{\alpha}$ est un vecteur colonne dont le $k^{\text{ème}}$ terme est α_k, et $\boldsymbol{\kappa}_j$ un vecteur colonne dont le $k^{\text{ème}}$ terme est $\frac{1}{n_j}\sum_{x_i\in\mathcal{C}_j}\kappa(x_k,x_i)$. On dispose alors d'une expression, en termes de noyau reproduisant, de la projection de chacune des moyennes sur la direction optimale. On peut utiliser cette expression pour écrire le quotient de Rayleigh (5.5) sous une forme ne faisant intervenir que des noyaux appliqués aux données d'apprentissage. Pour cela, on considère d'une part son numérateur, et on l'écrit

$$\Psi^\top S_\phi\,\Psi = \Psi^\top\left[\frac{1}{n}\sum_{j=1}^{J}(m_j^\phi - m_0^\phi)(m_j^\phi - m_0^\phi)^\top\right]\Psi = \boldsymbol{\alpha}^\top\left[\frac{1}{n}\sum_{j=1}^{J}(\boldsymbol{\kappa}_j - \boldsymbol{\kappa}_0)(\boldsymbol{\kappa}_j - \boldsymbol{\kappa}_0)^\top\right]\boldsymbol{\alpha}.$$

D'autre part, on développe son dénominateur suivant l'expression

$$\Psi^\top D_\phi\,\Psi = \Psi^\top\left[(\phi(x_i) - m_j^\phi)(\phi(x_i) - m_j^\phi)^\top\right]\Psi = \boldsymbol{\alpha}^\top\left[\frac{1}{n}\sum_{j=1}^{J}\sum_{x_i\in\mathcal{C}_j}(\boldsymbol{\kappa}(x_i) - \boldsymbol{\kappa}_j)(\boldsymbol{\kappa}(x_i) - \boldsymbol{\kappa}_j)^\top\right]\boldsymbol{\alpha},$$

où $\boldsymbol{\kappa}(x) = [\kappa(x,x_1)\cdots\kappa(x,x_n)]^\top$. En replaçant ces deux expressions dans (5.5), on retrouve le quotient de Rayleigh, que l'on écrit

$$Q_\kappa(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^\top S_\kappa\,\boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top D_\kappa\,\boldsymbol{\alpha}}, \tag{5.7}$$

où $S_\kappa = \frac{1}{n}\sum_{j=1}^{J}(\boldsymbol{\kappa}_j - \boldsymbol{\kappa}_0)(\boldsymbol{\kappa}_j - \boldsymbol{\kappa}_0)^\top$ et $D_\kappa = \frac{1}{n}\sum_{j=1}^{J}\sum_{x_i\in\mathcal{C}_j}(\boldsymbol{\kappa}(x_i) - \boldsymbol{\kappa}_j)(\boldsymbol{\kappa}(x_i) - \boldsymbol{\kappa}_j)^\top$. La résolution de ce problème est alors similaire au cas linéaire, avec une solution vérifiant un problème de vecteur propre généralisé de la forme (5.3). Contrairement au quotient de Rayleigh défini dans $\mathcal{H}$ par

(5.5), dont la solution appartient à un espace de dimension élevée, voire infinie pour certains noyaux, le vecteur $\boldsymbol{\alpha}$ à déterminer dans (5.7) est de taille finie égale au cardinal de l'ensemble d'apprentissage.

Une fois les coefficients optimaux de pondération obtenus, la direction optimale de discrimination dans l'espace transformé $\mathcal{H}$ est donnée par $\Psi = \sum_{k=1}^{n} \alpha_k \phi(x_k)$. La représentation $\phi(x)$ de tout élément x de $\mathcal{X}$ admet comme projection selon cette direction

$$\Psi^\top \phi(x) = \Big(\sum_{k=1}^{n} \alpha_k \phi(x_k) \Big)^\top \phi(x) = \sum_{k=1}^{n} \alpha_k \big(\phi(x_k)^\top \phi(x) \big) = \sum_{k=1}^{n} \alpha_k \kappa(x, x_k) = \boldsymbol{\alpha}^\top \boldsymbol{\kappa}(x).$$

On peut alors évaluer cette quantité sans qu'il soit nécessaire d'exhiber l'application $\phi(\cdot)$. Dans le cas plus général d'une discrimination à J classes, on a $J-1$ directions optimales, donc $J-1$ projections de $\phi(x)$ selon ces directions. La résolution du problème de discrimination multi-classes est étudié plus en détails dans [BA00]. On se contente dans la suite d'étudier le cas particulier d'un problème de discrimination à deux classes.

5.2.3 Cas particulier : discrimination entre deux classes

Pour un problème de discrimination à deux classes, on considère un ensemble d'apprentissage $\mathcal{A}_n$, formé de n signaux x_i munis de leurs étiquettes $y_i = \pm 1$. On désigne par n_1 et n_2 le nombre d'observations disponibles dans la classe représentée par les étiquettes -1 et $+1$, respectivement. Soient $m_1 = \frac{1}{n_1} \sum_{y_i=-1} \phi(x_i)$, $m_2 = \frac{1}{n_2} \sum_{y_i=+1} \phi(x_i)$, et $m = \sum_{i=1}^{n} \phi(x_i)$ les moyennes des données sans tenir compte de leurs étiquettes.

Dans le cas d'une AFD à noyau, on cherche à maximiser le quotient de Rayleigh (5.2), à savoir $Q_\phi(\Psi) = \Psi^\top S_\phi \, \Psi / \Psi^\top D_\phi \, \Psi$. La matrice de séparation des moyennes est donnée par

$$S_\phi = \frac{1}{n} \sum_{j=1,2} n_j (m_j^\phi - m^\phi)(m_j^\phi - m^\phi)^\top = \frac{n_1 \, n_2}{n} \, (m_1^\phi - m_2^\phi)(m_1^\phi - m_2^\phi)^\top.$$

Ceci permet d'écrire

$$S_\phi \Psi = \frac{n_1 \, n_2}{n} \, (m_1^\phi - m_2^\phi) \Big[(m_1^\phi - m_2^\phi)^\top \Psi \Big],$$

où le terme entre crochets est un scalaire. On trouve alors que $S_\phi \Psi$ est colinéaire à $(m_1^\phi - m_2^\phi)$. En utilisant l'expression $S_\phi \, \Psi = \lambda \, D_\phi \, \Psi$, on peut alors écrire

$$D_\phi \, \Psi = m_1^\phi - m_2^\phi,$$

où les différents termes multiplicatifs (scalaires) sont écartés puisqu'ils n'influent pas sur la direction recherchée. Par l'usage du Théorème de Représentation, on a $\Psi = \sum_{k=1}^{n} \alpha_k \phi(x_k)$ et, en remplaçant dans (5.2), on obtient le quotient de Rayleigh

$$Q_\kappa(\boldsymbol{\alpha}) = \frac{\boldsymbol{\alpha}^\top S_\kappa \, \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top D_\kappa \, \boldsymbol{\alpha}}, \tag{5.8}$$

avec

$$S_\kappa = (\boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2)(\boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2)^\top \tag{5.9}$$

et

$$D_\kappa = \sum_{j=1,2} K_j K_j^\top - n_j \boldsymbol{\kappa}_j \boldsymbol{\kappa}_j^\top. \tag{5.10}$$

Sous une forme équivalente, on a également $D_\kappa = \sum_{j=1,2} K_j \Big(\boldsymbol{I}_{n_j} - \frac{1}{n_j} \mathbf{1}_{n_j} \Big) K_j^\top$. Dans ces expressions, K_1 et K_2 désignent les matrices de Gram associées à chacune des deux classes. La matrice $\boldsymbol{I}_{n_j}$ est la matrice identité de taille $n_j \times n_j$, et $\mathbf{1}_{n_j}$ désigne la matrice unité de taille $n_j \times n_j$.

La résolution passe par celle d'un problème aux valeurs propres de la forme $S_\kappa \, \Psi = \lambda \, D_\kappa \, \Psi$. Or, par analogie avec les résultats ci-dessus, on écrit $S_\kappa \boldsymbol{\alpha} = (\boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2) \left[(\boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2)^\top \boldsymbol{\alpha} \right]$ pour finalement obtenir

$$D_\kappa \boldsymbol{\alpha} = \boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2.$$

Les n coefficients α_k qui déterminent l'axe principal de discrimination, $\Psi = \sum_{k=1}^n \alpha_k \phi(x_k)$, sont alors obtenus par la résolution de ce système linéaire. En supposant que l'inverse de D_κ existe, la solution est alors donnée par $\boldsymbol{\alpha} = D_\kappa^{-1}(\boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2)$. Toutefois, cette supposition n'est pas toujours valide puisque cette matrice est généralement mal-conditionnée, voire singulière. Pour pallier ceci, on a recours à une régularisation du problème comme on le montre à la section suivante.

Pour procéder à une discrimination non-linéaire, on considère finalement la statistique de décision suivante

$$\Lambda(x) = \Psi^\top \phi(x) = \sum_{k=1}^n \alpha_k \big(\phi(x_k)^\top \phi(x) \big) = \sum_{k=1}^n \alpha_k \kappa(x, x_k) = \boldsymbol{\alpha}^\top \boldsymbol{\kappa}(x),$$

ce qui correspond à la projection de $\phi(x)$ sur Ψ. On détermine l'appartenance de x à l'une des deux classes en comparant cette statistique à un seuil. On écrit alors

$$\Lambda(x) = \sum_{k=1}^n \alpha_k \kappa(x, x_k) \underset{y_k=-1}{\overset{y_k=+1}{\gtrless}} \nu_0.$$

La maximisation du critère de contraste de Fisher ne permet pas de déterminer le seuil optimal ν_0. Cette quantité est perdue par du fait que l'on ne considère que la distance entre les moyennes. On écrit alors

$$\nu_0 = \Psi^\top \Big(\frac{n_1 \, m_1^\phi + n_2 \, m_2^\phi}{n} \Big) = \frac{1}{n} \, \boldsymbol{\alpha}^\top (n_1 \boldsymbol{\kappa}_1 + n_2 \boldsymbol{\kappa}_2), \tag{5.11}$$

avec $\boldsymbol{\kappa}_1$ et $\boldsymbol{\kappa}_2$ des vecteurs colonnes formés respectivement par les termes $\frac{1}{n_1} \sum_{y_i=-1} \kappa(x_k, x_i)$ et $\frac{1}{n_2} \sum_{y_i=+1} \kappa(x_k, x_i)$, pour $k = 1, \ldots, n$. La solution obtenue avec ce seuil est optimale au sens de la minimisation de l'erreur quadratique [DHS00]. Une règle plus classique pour le choix du seuil consiste à le mettre à mi-distance des moyennes des deux classes, après leur projection sur la direction de discrimination optimale. On a alors $\nu_0 = \Psi^\top (m_1^\phi + m_2^\phi)/2 = \frac{1}{2} \, \boldsymbol{\alpha}^\top (\boldsymbol{\kappa}_1 + \boldsymbol{\kappa}_2)$. On présente au Tableau 5.1 un résumé des différentes étapes de l'algorithme AFD-à-noyau, dans le cas particulier de la discrimination à deux classes.

5.2.4 Paramètre de régularisation : interprétation selon Tikhonov

Les méthodes régularisées selon Tikhonov forment un cadre général pour traiter des problèmes de reconnaissance des formes. La discrimination n'échappe pas à cette règle, comme on le montre dans la suite pour le cas particulier d'une discrimination à deux classes.

Le lien entre l'AFD classique et la minimisation de l'erreur quadratique est connu depuis un certains temps, [DHS00]. Sa généralisation à l'AFD-à-noyau a fait l'objet de plusieurs travaux, dont la thèse de Mika [Mik02], [XZL01] pour une équivalence avec la régression ridge à noyau, ou plus récemment dans [ZR05] où les auteurs décrivent l'équivalence entre l'AFD-à-noyau avec terme régularisant et la

Données	$\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, $y_i = \pm 1$	Base d'apprentissage
	$\kappa(\cdot\,,\cdot)$, x^t	noyau reproduisant, donnée test
Algorithme	$(\boldsymbol{\kappa}_1)_k = \frac{1}{n_1} \sum_{y_i=-1} \kappa(x_k, x_i)$, $k = 1, \ldots, n$	Moyennes des noyaux
	$(\boldsymbol{\kappa}_2)_k = \frac{1}{n_2} \sum_{y_i=+1} \kappa(x_k, x_i)$, $k = 1, \ldots, n$	de chaque classe
	$(K_1)_{i,k} = \kappa(x_i, x_k)$, $y_k = -1$, $i = 1, \ldots, n$	Matrices de Gram
	$(K_2)_{i,k} = \kappa(x_i, x_k)$, $y_k = +1$, $i = 1, \ldots, n$	associées à chaque classe
	$D_\kappa = \sum_{j=1,2} K_j K_j^\top - n_j \boldsymbol{\kappa}_j \boldsymbol{\kappa}_j^\top$	Matrice des variances
	$D_\kappa \boldsymbol{\alpha} = \boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2$	Résolution du système linéaire
Sorties	$\Psi = \sum_k \alpha_k \phi(x_k)$	Axe principal de discrimination
	$\sum_{k=1}^n \alpha_k \kappa(x^t, x_k)$	Projection de $\phi(x^t)$ sur Ψ
	$\Lambda(x^t) = \sum_{k=1}^n \alpha_k \kappa(x^t, x_k) \underset{y=-1}{\overset{y=+1}{\gtrless}} \nu_0$	Statistique de décision
	avec $\nu_0 = \boldsymbol{\alpha}^\top (n_1 \boldsymbol{\kappa}_1 + n_2 \boldsymbol{\kappa}_2)/n$	et seuil

TAB. 5.1 – L'algorithme AFD-à-noyau pour une discrimination à deux classes

minimisation de l'erreur quadratique avec régularisation de type Tikhonov. On rappelle que ce problème consiste à résoudre

$$\psi^* = \arg\min_{\psi \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (y_i - \psi(x_i))^2 + \eta \, \|\psi\|_{\mathcal{H}}^2,$$

où η est le paramètre de régularisation. Ce paramètre permet de contrôler le compromis entre erreur d'apprentissage et degré de régularité de la solution retenue. Afin d'aboutir au même résultat que celui donné par l'algorithme AFD-à-noyau, il faut adopter les réponses désirées suivantes

$$y_i = \begin{cases} +n/n_1 & \text{, si } \boldsymbol{x}_i \in \mathcal{C}_1; \\ -n/n_2 & \text{, si } \boldsymbol{x}_i \in \mathcal{C}_2, \end{cases}$$

où $\mathcal{C}_1$ et $\mathcal{C}_2$ désignent chacune des deux classes, et n_1 et n_2 leur cardinalité. De plus, et par analogie avec l'AFD-à-noyau, il convient d'ajouter à cette fonction coût une contrainte sur la norme du vecteur des coefficients $\boldsymbol{\alpha}$, soit $\|\boldsymbol{\alpha}\| = 1$.

L'importance de la régularisation est souvent mise en évidence dans la littérature, depuis l'article de référence [MRW+99]. Plus récemment, différentes approches ont été adoptées afin de déterminer le paramètre de régularisation optimal, voir par exemple [STC04a]. Cette régularisation est nécessaire pour la mise en œuvre de ces algorithmes dans le domaine temps-fréquence, où l'on passe de signaux de taille l à des représentations de tailles $l \times l$. Différents travaux ont montré l'inefficacité d'une approche AFD brute, sans régularisation, dans ce domaine. Dans [BD95], les auteurs considèrent une technique dite *neo-classique* pour remédier à ce problème, alors que dans [Hei95], et plus récemment dans [GA01], une réduction du nombre de degrés de liberté de la fonction de paramétrisation est adoptée. Toutefois, ces approches ne sont pas soutenues par des résultats théoriques sur la généralisation, comme c'est le cas de l'AFD et de l'AFD-à-noyau avec la régularisation de type Tikhonov considérée dans ce document.

5.3 Analyse discriminante dans le domaine temps-fréquence

Au Chapitre 3, on a introduit le concept de méthodes à noyau dans le domaine temps-fréquence, par l'usage d'un noyau associé à une distribution temps-fréquence. On étudie à présent ce concept pour la discrimination de signaux à partir de leur distribution temps-fréquence. Pour cela, on considère un ensemble d'apprentissage formé par n signaux $x_i \in \mathcal{X}$ de taille l et de leurs étiquettes y_i.

Pour une distribution temps-fréquence quelconque, désignée par $\phi(\cdot)$, on considère le noyau reproduisant κ qui correspond au produit scalaire $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$ entre les distributions, et on désigne par $\mathcal{H}$ le RKHS induit par ce noyau. Le Théorème de Représentation permet d'exhiber l'axe de discrimination comme une combinaison linéaire des distributions temps-fréquence de l'ensemble d'apprentissage. La signature obtenue, exprimée sous la forme $\psi(\cdot) = \sum_{k=1}^{n} \alpha_k \kappa(x_k, \cdot)$ peut être qualifiée de *signature (temps-fréquence) discriminante*. Il faut toutefois souligner que celle-ci n'est pas nécessairement une distribution temps-fréquence valide. La statistique de décision est alors donnée par

$$\Lambda(x) = \langle \psi(\cdot), \phi(x) \rangle_{\mathcal{H}} = \Big\langle \sum_{k=1}^{n} \alpha_k \kappa(x_k, \cdot), \phi(x) \Big\rangle_{\mathcal{H}} = \sum_{k=1}^{n} \alpha_k \kappa(x, x_k) = \boldsymbol{\alpha}^{\top} \boldsymbol{\kappa}(x).$$

Dans ce qui suit, on étudie ce problème de façon plus détaillée en considérant la distribution de Wigner dans un premier temps, puis d'autres distributions de la classe de Cohen.

5.3.1 Discrimination par la distribution de Wigner

On se contente dès à présent d'étudier le cas d'une discrimination à deux classes, le contexte multi-classes s'en déduisant directement. Pour la distribution de Wigner, on considère la transformation du signal suivante $\phi : x \mapsto W_x$. Le noyau reproduisant associé à cette distribution est $\kappa_W = \langle W_{x_i}, W_{x_j} \rangle$, et l'espace RKHS induit par ce noyau est désigné par $\mathcal{H}_W$. Dans le cadre de discrimination par AFD, on cherche la signature discriminante qui maximise la séparation des données dans l'espace $\mathcal{H}_W$. La projection selon cette direction doit maximiser la distance entre les moyennes des distributions temps-fréquence de chaque classe, tout en minimisant les variances intra-classes.

Le coup du noyau autorise la mise en œuvre de l'AFD-à-noyau sans nécessité d'évaluer les représentations temps-fréquence de l'ensemble d'apprentissage, le noyau reproduisant étant donné par $\kappa_W = |\langle x_i, x_j \rangle|^2$. L'expression (5.6) permet une interprétation temps-fréquence de la caractéristique retenue, celle-ci correspondant à une signature de la forme $\Psi = \sum_{i=1}^{n} \alpha_i W_{x_i}$. Pour un signal $x \in \mathcal{X}$ donné, on considère la projection de sa distribution de Wigner sur la signature discriminante, soit la statistique $\Lambda(x) = \sum_{k=1}^{n} \alpha_k \kappa_W(x, x_k)$, ou encore en tenant compte de coup du noyau on a

$$\Lambda(x) = \sum_{k=1}^{n} \alpha_k |\langle x, x_k \rangle|^2. \tag{5.12}$$

Ceci n'est autre que la coordonnée de W_x dans cette direction. En considérant un seuil sur cette quantité, on définit une statistique de décision sur l'appartenance de x à l'une des deux classes, selon

$$\Lambda(x) \underset{y=-1}{\overset{y=+1}{\gtrless}} \nu_0.$$

D'un point de vue complexité algorithmique, l'usage d'un algorithme d'AFD classique sur les distributions de Wigner de l'ensemble d'apprentissage est plus exigeant que la mise en œuvre de l'AFD-à-noyau dans le domaine de la distribution de Wigner, comme proposé dans ce chapitre. La plus importante contribution est due au coup du noyau qui permet une réduction du coût calculatoire de la matrice de Gram associée à tout l'ensemble d'apprentissage, de $\mathcal{O}(n^3 l^2 \log l)$ pour l'AFD classique à $\mathcal{O}(n^3 l)$ pour l'AFD-à-noyau. Si l'on s'intéresse à la signature temps-fréquence, le Théorème de Représentation permet son évaluation d'une manière itérative, sans qu'il soit nécessaire de stocker les distributions de Wigner de tout l'ensemble d'apprentissage. Le Tableau 5.2 résume les différentes étapes de l'algorithme résultant de la mise en œuvre de l'AFD-à-noyau dans le domaine de la distribution de Wigner, ainsi que la complexité calculatoire de chaque étape.

Instructions	Expressions	Complexité
Algorithme de base		
1. Calculer des matrices de Gram	$(K_1)_{i,k} = \kappa(x_i, x_k) = \lvert\langle x_i, x_k\rangle\rvert^2\ ,\ y_k = -1$	$\mathcal{O}(n_1 nl)$
	$(K_2)_{i,k} = \kappa(x_i, x_k) = \lvert\langle x_i, x_k\rangle\rvert^2\ ,\ y_k = +1$	$\mathcal{O}(n_2 nl)$
2. Moyennes des matrices de Gram	$(\boldsymbol{\kappa}_1)_i = \frac{1}{n_1}\sum_{y_i=-1}\kappa(x_i, x_k) = \frac{1}{n_1}\sum_k (K_1)_{i,k}$	$\mathcal{O}(n_1 n)$
	$(\boldsymbol{\kappa}_2)_i = \frac{1}{n_2}\sum_{y_i=+1}\kappa(x_i, x_k) = \frac{1}{n_2}\sum_k (K_2)_{i,k}$	$\mathcal{O}(n_2 n)$
3. Calcul de la matrice des variances	$D_\kappa = \sum_{j=1,2} K_j K_j^\top - n_j \boldsymbol{\kappa}_j \boldsymbol{\kappa}_j^\top$	$\mathcal{O}(n^2)$
4. Résolution du système linéaire	$D_\kappa \boldsymbol{\alpha} = \boldsymbol{\kappa}_1 - \boldsymbol{\kappa}_2$	$\mathcal{O}(n^3)$
Sortie : **Signature discriminante**		
5. Calculer n distributions de Wigner	W_{x_i}	$\mathcal{O}(nl^2 \log l)$
6. La signature discriminante	$\Psi = \sum_{i=1}^n \alpha_i W_{x_i}$	$\mathcal{O}(nl^2)$
Sortie : **Règle de discrimination**		
8. Vecteur d'evaluation en x	$(\boldsymbol{\kappa}(x))_i = \kappa(x, x_i) = \lvert\langle x, x_i\rangle\rvert^2$	$\mathcal{O}(nl)$
7. Seuil de discrimination	$\nu_0 = \boldsymbol{\alpha}^\top (n_1 \boldsymbol{\kappa}_1 + n_2 \boldsymbol{\kappa}_2)/n$	$\mathcal{O}(n)$
8. Statistique de décision	$\Lambda(x) = \boldsymbol{\alpha}^\top \boldsymbol{\kappa}(x) \underset{y=-1}{\overset{y=+1}{\gtrless}} \nu_0$	$\mathcal{O}(n)$

TAB. 5.2 – Les différentes étapes de l'algorithme AFD-à-noyau pour la distribution de Wigner, et leur coût calculatoire, dans le cas d'une discrimination à deux classes, l étant la taille des signaux, et n_1, n_2 et $n = n_1 + n_2$ désignent les tailles respectives de chacune des deux classes ainsi que la taille totale de l'ensemble d'apprentissage.

5.3.2 Au-delà de la distribution de Wigner, la classe de Cohen

La mise en œuvre de l'AFD-à-noyau pour une distribution quelconque de la classe de Cohen est donnée par un choix approprié du noyau reproduisant. En désignant par C_x^Π une distribution temps-fréquence d'un signal x, le noyau reproduisant est donné par

$$\kappa_{C^\Pi} = \langle C_{x_i}^\Pi, C_{x_j}^\Pi \rangle.$$

Une fois cette quantité évaluée pour tout couple de signaux de l'ensemble d'apprentissage, l'algorithme détermine les coefficients de pondération β_i définissant la signature temps-fréquence discriminante sous la forme $\Psi_{C^\Pi} = \sum_{i=1}^n \beta_i C_{x_i}^\Pi$. La règle de décision est donnée par

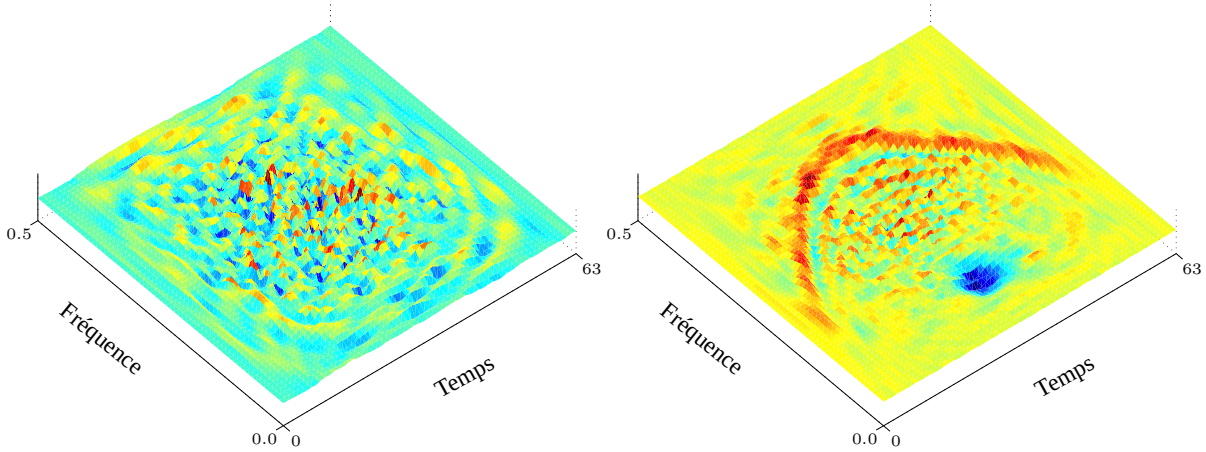
$$\Lambda(x) = \sum_{k=1}^n \beta_k \kappa_{C^\Pi}(x, x_k) \underset{y=-1}{\overset{y=+1}{\gtrless}} \nu_0.$$

Toutefois, le noyau reproduisant défini ci-dessus admet souvent une complexité calculatoire élevée, comparé au noyau quadratique de la distribution de Wigner. Pour contourner cet inconvénient, on peut recourir à la stratégie hybride proposé dans le Paragraphe 3.3.3. On rappelle que cette approche, sous-optimale, consiste à calculer les coefficients α_i à partir du noyau quadratique $\kappa_W = \lvert\langle x_i, x_j\rangle\rvert^2$, la signature résultante étant donnée par $\widetilde{\Psi}_{C^\Pi} = \sum_{i=1}^n \alpha_i C_{x_i}^\Pi$.

5.3.3 Applications

Discrimination à deux classes

La première application concerne un problème de discrimination entre deux familles de $1\,000$ signaux de taille 64 noyés dans un bruit blanc Gaussien de moyenne nulle et de variance 1.25. La première famille



(a) (b)

FIG. 5.1 – Problème de discrimination concernant deux familles de 1 000 signaux de taille 64, l'une à modulation de fréquence parabolique et l'autre à atome gaussien, noyés dans un bruit blanc Gaussien. Signatures discriminantes issus d'une AFD-à-noyau avec le noyau quadratique et la distribution de Wigner, sans régularisation en (a) et avec régularisation en (b).

est formée de 500 signaux à modulation parabolique de fréquence variant de 0.1 à 0.4 en fréquence normalisée, la seconde contient 500 signaux à atome Gaussien de fréquence normalisée 0.1 et à l'instant central.

On commence par l'AFD-à-noyau, sur la base de la distribution de Wigner et de son noyau quadratique. La Figure 5.1 (a) présente la signature Ψ_0 obtenue par l'application directe de la méthode d'AFD-à-noyau. Cette représentation ne met pas en évidence la configuration temps-fréquence des signaux étudiés, en raison de problèmes d'instabilité numérique identifiés dans [MRW+99]. On peut contenir ces effets à l'aide d'une technique de régularisation de type Tikhonov comme le montre la signature obtenue Ψ, illustrée à la Figure 5.1 (b). On retrouve les deux composantes clés pour la discrimination, que sont la modulation parabolique représentée avec une orientation positive (couleur rouge) et l'atome Gaussien représenté avec une orientation négative (couleur bleue). La régularisation de type Tikhonov offre un compromis entre le sur-apprentissage et la régularité de la solution obtenue. Pour comprendre ce mécanisme, on considère la projection des données dans le domaine temps-fréquence sur les signatures Ψ_0 et Ψ, obtenues respectivement sans et avec régularisation. Un signal x est alors représenté par ses coordonnées dans Ψ_0 et Ψ selon $\langle W_x, \Psi_0 \rangle$ et $\langle W_x, \Psi \rangle$ respectivement. On compare les résultats obtenus à partir de deux ensembles de données, en les représentant dans le plan $\Psi_0\Psi$. D'une part, on considère l'ensemble d'apprentissage que l'on représente dans ce plan, comme illustré à la Figure 5.2 (a). Bien que les deux familles semblent être mieux séparées avec la signature Ψ_0, obtenue sans terme régularisant, elle admet un pouvoir de généralisation faible. Pour montrer cela, on considère un nouvel ensemble de 1 000 signaux, appartenant aux deux familles en question mais n'ayant pas servi à l'apprentissage. En représentant ces signaux dans le plan $\Psi_0\Psi$ comme illustré à la Figure 5.2 (b), on trouve que le résultat donné avec régularisation a un meilleur pouvoir de généralisation.

Pour le même problème, on considère d'autres distributions de la classe de Cohen, et en particulier la distribution de Choï-Williams et le spectrogramme. Les signatures temps-fréquence obtenues avec ces distributions sont illustrées aux Figures 5.3 (a) et 5.3 (b). On retrouve les deux composantes clés pour la discrimination, que sont la modulation de fréquence parabolique et l'atome Gaussien, chacune ayant une orientation opposée. Contrairement à celle obtenue par la distribution de Wigner, ces signatures

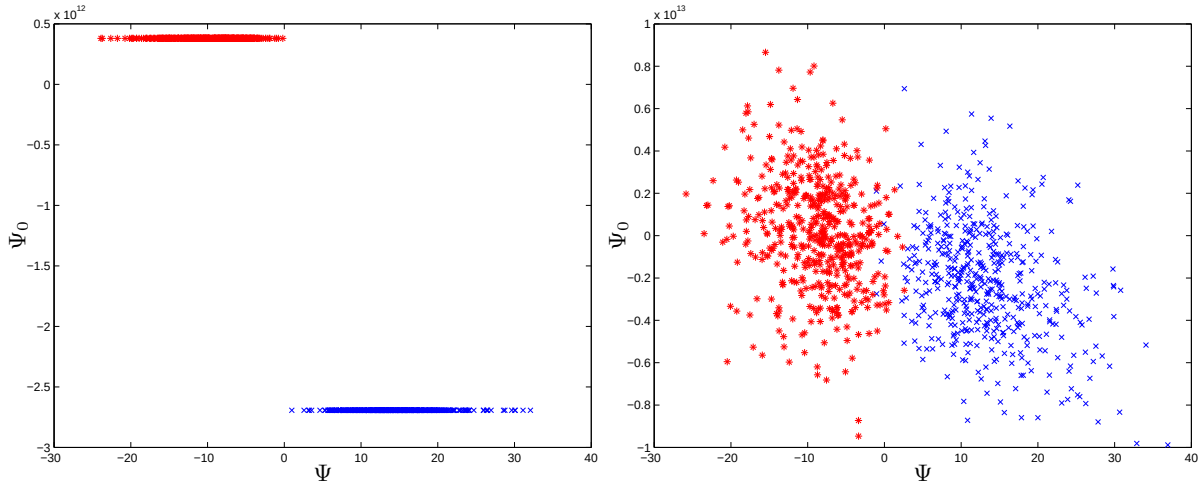


FIG. 5.2 – L'effet de régularisation sur la séparabilité des données, en les représentants dans le plan $\Psi_0\Psi$ des signatures Ψ_0 et Ψ obtenues sans et avec régularisation, respectivement. Sans terme de régularisation, on est dans un schéma de sur-apprentissage, avec une bonne séparation sur l'ensemble d'apprentissage (a) et une mauvaise généralisation sur de nouvelles données (b). La régularisation de type Tikhonov permet de pallier cet inconvénient, comme on l'observe avec la séparabilité des deux ensembles de données considérés selon l'axe Ψ.

sont à interférences réduites, une propriété inhérente des distributions considérées. Le prix à payer étant une complexité algorithmique pour le calcul du noyau reproduisant en question, le stratégie hybride permet de la réduire en proposant une solution sous-optimale. Les Figures 5.3 (c) et 5.3 (d) illustrent les signatures hybrides ainsi obtenues avec le noyau quadratique, pour la distribution de Choï-Williams et le spectrogramme, respectivement.

Discrimination multi-classe

Ayant un problème de discrimination entre J familles de signaux, l'ADG-à-noyau détermine $J-1$ statistiques de décision de la forme (5.12) dans le cas de la distribution de Wigner. Pour illustrer cette approche, on considère un exemple de 3 familles de 1500 signaux, chacun de taille 64. Chaque signal est formé soit d'un atome Gaussien à la fréquence normalisée 0.25, soit d'une fréquence sinusoïdale pure de 0.1, soit d'une modulation fréquentielle parabolique variant entre 0.15 et 0.45. Ces signaux sont noyés

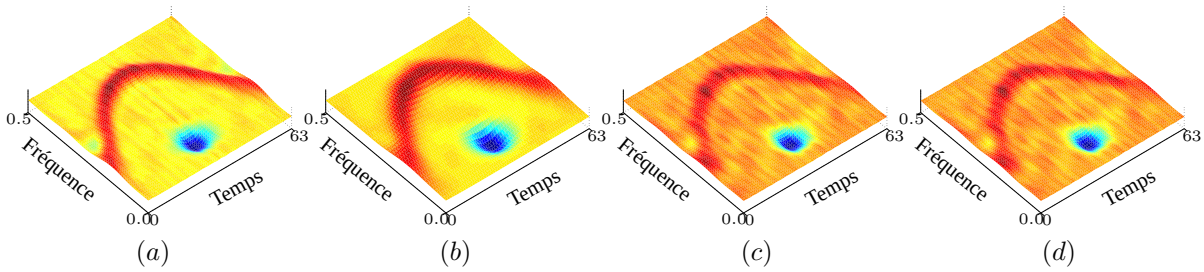
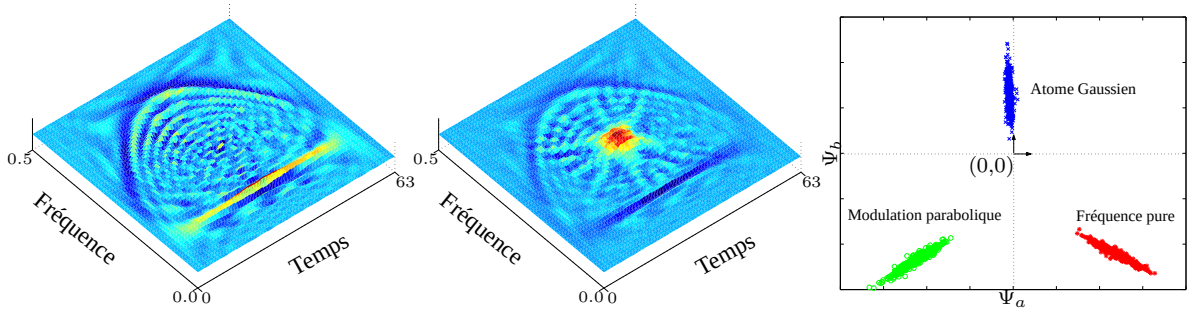


FIG. 5.3 – Signatures temps-fréquence discriminantes obtenues à partir de la distribution de Choï-Williams (a) et du spectrogramme (b). L'usage d'une stratégie hybride produit des résultats sous-optimaux, comme illustré dans (c) pour la première distribution, et dans (d) pour la seconde.



(a) (b) (c)

FIG. 5.4 – Les signature temps-fréquences Ψ_a et Ψ_b, illustrées dans (a) et (b) respectivement, obtenues par l'usage d'un algorithme de ADG-à-noyau avec la distribution de Wigner dans le cas d'une discrimination à 3 classes. Les signaux appartenant à chaque classe sont formés d'une composante noyée dans un bruit blanc Gaussien : une modulation fréquentielle parabolique, une fréquence sinusoïdale pure, et un atome Gaussien. Dans (c), les signaux de l'ensemble d'apprentissage sont représentés dans le plan $\Psi_a\Psi_b$.

dans un bruit Gaussien additif de moyenne nulle et de variance 0.4. Les Figures 5.4 (a) et 5.4 (b) représentent les signatures temps-fréquence obtenues par l'ADG-à-noyau basée sur la distribution de Wigner. Ces signatures discriminantes, désignées par Ψ_a et Ψ_b, recouvrent toute l'information nécessaire pour discerner les distributions des différentes familles. En particulier, la signature Ψ_a permet de différencier la modulation parabolique de la fréquence pure, qui sont respectivement orientées négativement et positivement, tandis que l'atome Gaussien est absent. La signature Ψ_b permet de discerner ce dernier, orienté positivement, des deux autres composantes orientées négativement. Cette analyse est validée par la projection des signaux x_i de l'ensemble d'apprentissage sur ses deux signatures, comme illustré à la Figure 5.4 (c) où on représente les données par leurs coordonnées $\langle W_{x_i}, \Psi_a \rangle$ et $\langle W_{x_i}, \Psi_b \rangle$ dans le plan $\Psi_a\Psi_b$. Selon l'axe Ψ_a, les signaux à modulation parabolique, à atome Gaussien et à fréquence pure admettent respectivement des coordonnées négatives, presque nulles et positives. Avec des coordonnées presque nulles, les signaux à atome Gaussien ne pourront pas être distingués de signaux formés simplement de bruit blanc. Ceci est résolu par l'usage de l'axe Ψ_b, produisant des coordonnées positives pour ceux-ci, alors que les deux autres composantes ont des coordonnées négatives.

Pour terminer, on considère le même problème de discrimination multi-classes pour d'autres distributions de la classe de Cohen. Pour cela, on utilise les paramètres de lissage considérés par défaut

Nom	Rang du problème	1ère valeur propre	2ème valeur propre
Spectrogramme	41	0.979275	0.978300
Pseudo-Wigner-Ville lissée	77	0.980557	0.979304
Choï-Williams	103	0.980757	0.979481
Born-Jordan	111	0.980899	0.979615
Margeneau-Hill	183	0.982456	0.980965
Wigner	183	0.982526	0.981071

TAB. 5.3 – Comparaison des résultats obtenus avec différentes distributions temps-fréquence. Les distributions temps-fréquence sont affichées par ordre croissant du rang du problème. La variance expliquée par les deux premiers axes discriminants est donnée par la première et la seconde valeurs propres du problème, respectivement.

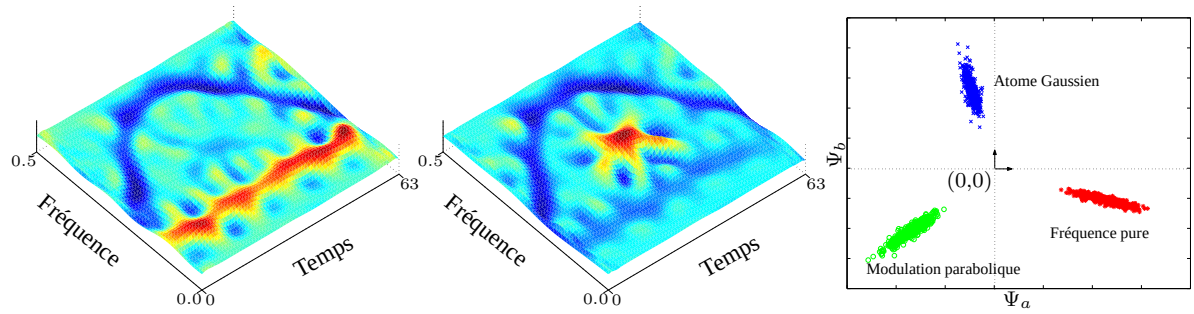


FIG. 5.5 – Les signatures temps-fréquence Ψ_a et Ψ_b, et la représentation des données d'apprentissage sur le plan $\Psi_a\Psi_b$, illustrées dans (a), (b) et (c) respectivement, comme présenté à la Figure 5.4 mais avec la distribution Pseudo-Wigner-Ville lissée.

dans la Toolbox Temps-Fréquence [AFGL05]. Le Tableau 5.3 regroupe des résultats obtenus à partir de différentes distributions temps-fréquence. Les distributions temps-fréquence à faibles termes interférentiels produisent de faible rang, la cause principale étant le lissage qui engendre un espace de dimension plus petite. De plus, dans le cas du spectrogramme, on a considéré l'algorithme proposé dans la Toolbox Temps-Fréquence [AFGL05], et qui produit des représentations de taille 64×32, d'où un rang encore moindre. Dans le même tableau, on représente les deux plus importantes valeurs propres du problème généralisé. Celles-ci correspondent à la variance des données expliquée par les deux signatures obtenues pour chaque distribution temps-fréquence. On trouve que la distribution de Wigner maximise ces quantités, alors qu'elles sont de plus en plus faibles lors de l'application du lissage. Du point de vue signature temps-fréquence et à titre indicatif, on illustre à la Figure 5.5 les signatures obtenues à partir de la distribution Pseudo-Wigner-Ville lissée, pour le même problème traité à la Figure 5.4.

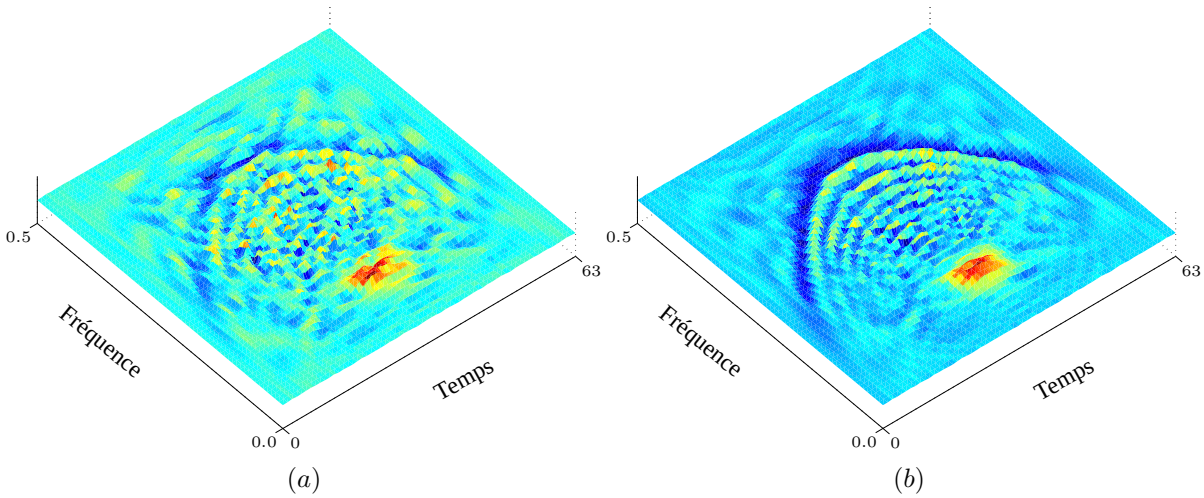


FIG. 5.6 – Problème de classification concernant deux familles de 1 000 signaux de taille 64, l'une à modulation de fréquence parabolique et l'autre à atome gaussien, noyés dans un bruit blanc Gaussien. Signatures obtenues de la distribution de Wigner avec son noyau quadratique, (a) par l'AFD-à-noyau et (b) par les SVM.

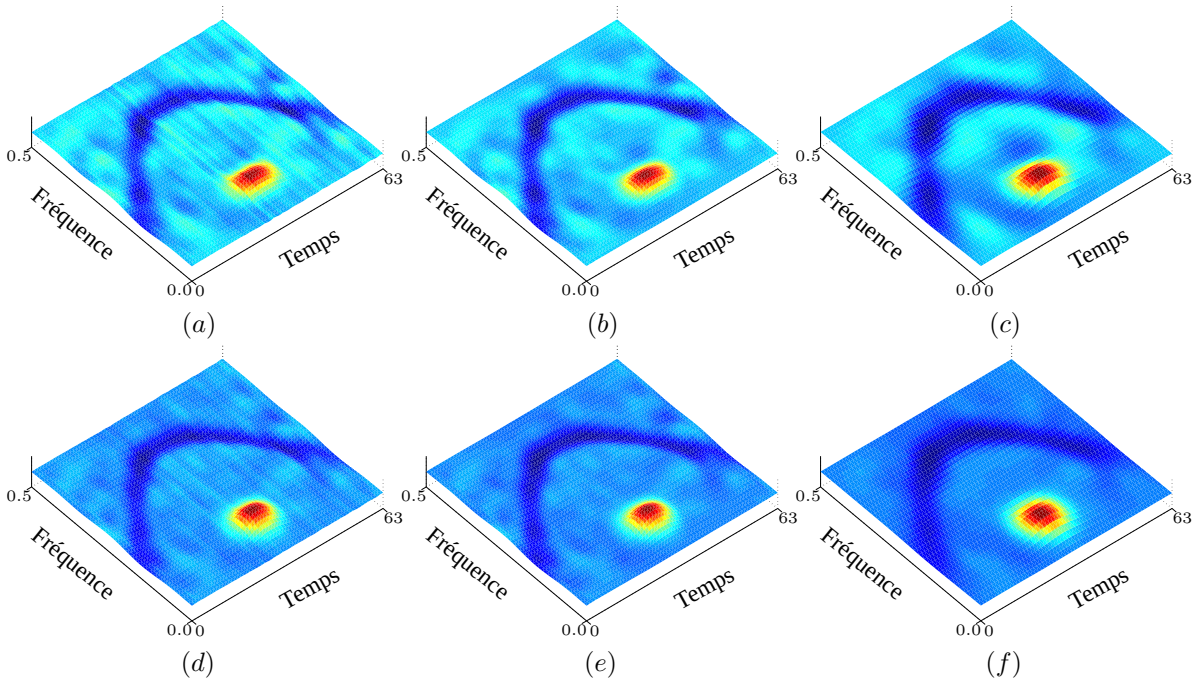


FIG. 5.7 – Signatures discriminantes obtenues $(a) - (c)$ par l'AFD-à-noyau et $(d) - (f)$ par les SVM, pour les distributions $(a), (d)$ de Choï-Williams, $(b), (e)$ de pseudo-Wigner-lissée et $(c), (f)$ pour le spectrogramme.

Classification de signaux dans le domaine temps-fréquence

On propose d'étudier les performances en classification pour différentes distributions temps-fréquence, d'une part avec l'AFD-à-noyau et d'autre part avec les SVM. Ces dernières déterminent l'hyperplan séparateur à marge maximale entre les échantillons d'un ensemble d'apprentissage. L'Annexe A est consacrée à l'étude des SVM et de la théorie qui les accompagne. On présente sa mise en œuvre dans le domaine temps-fréquence par un choix approprié de noyau reproduisant, ce qui revient à déterminer la signature temps-fréquence qui maximise la distance entre elle et les représentations temps-fréquence des signaux d'apprentissage.

On reprend la même configuration que dans le cadre de la discrimination à deux classes vue précédemment, avec une famille de signaux à modulation parabolique et une à atome Gaussien. Afin de comparer les erreurs estimées, on considère des signaux avec un rapport signal-sur-bruit d'environ -10 dB. On illustre à la Figure 5.6 (a) la signature discriminante obtenue par l'AFD-à-noyau, à partir du noyau quadratique et de la distribution de Wigner, et en 5.6 (b) la signature séparatrice résultant des SVM. On trouve que la signature obtenue par les SVM offre un meilleur contraste pour les deux composantes clés que sont la modulation parabolique et l'atome Gaussien. On illustre à la Figure 5.7 les signatures discriminantes obtenues avec l'AFD-à-noyau et les SVM à partir de différentes distributions temps-fréquence de la classe de Cohen. On rappelle que l'évaluation du noyau reproduisant associé à chacune de ces distributions occasionne un coût calculatoire supplémentaire, comparé au noyau quadratique de Wigner. Pour remédier à cela, on propose d'utiliser la technique hybride présentée dans la Section 3.3.3. Les signatures ainsi obtenues avec les SVM sont illustrées à la Figure 5.8.

Afin de comparer les performances des différentes distributions temps-fréquence utilisées, on considère un ensemble test de $10\,000$ signaux appartenant aux deux classes afin d'estimer l'erreur de générali-

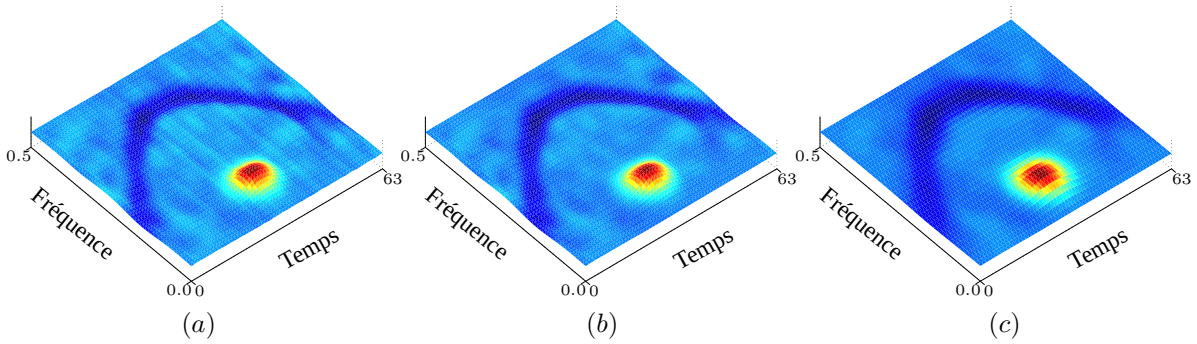


FIG. 5.8 – Signatures séparatrices obtenues par les SVM, avec la technique hybride, en combinant noyau quadratique et (a) distribution de Choï-Williams, (b) pseudo-Wigner-lissée et (c) spectrogramme.

sation par une technique de validation croisée. On présente au Tableau 5.4 les erreurs estimées obtenues, tout en rappelant qu'on n'a pas cherché à déterminer le paramètre de régularisation optimal pour chacun des deux algorithmes. La distribution de Wigner est donc la meilleure pour ce problème de classification. Les distributions les plus lissées, dont le spectrogramme, sont moins adaptées à ce type de problème. Ce résultat ouvre la voie à plusieurs questions. Y a-t-il une technique pour déterminer l'optimalité d'une représentation temps-fréquence pour un problème de classification donné ? Comme le lissage arbitraire nuit à la représentation, quelle approche permet d'ajuster la représentation au problème considéré ? Le chapitre suivant traite de ces problèmes, et propose une solution originale avec le critère d'alignement noyau-cible. Celui-ci ne nécessite aucun apprentissage de la statistique de décision, ou de techniques de validation croisée pour estimer l'erreur de généralisation.

Distribution temps-fréquence	SVM (%)	AFD-à-noyau (%)	SVM hybride (%)	AFD-à-noyau hybride (%)
Spectrogramme	19.86	28.04	19.5	23.57
Pseudo-Wigner-Ville lissée	16.36	26.55	17.07	19.72
Choï-Williams	16.83	26.47	16.08	19.07
Wigner	13.45	19.35		

TAB. 5.4 – Comparaison des erreurs estimées pour différentes distributions de la classe de Cohen avec l'AFD-à-noyau et les SVM, ainsi que pour la technique hybride.

Chapitre 6

Distributions temps-fréquence optimales par alignement noyau-cible

Sommaire

Tout au long des chapitres précédents, nous avons montré que les méthodes à noyau les plus performantes et les plus diverses peuvent être mises en œuvre dans le plan temps-fréquence grâce à un choix approprié de noyau reproduisant. La sélection d'une représentation temps-fréquence adaptée à la résolution d'un problème de classification de signaux demeure toutefois une question récurrente. Nous nous proposons donc de montrer que le domaine de la reconnaissance des formes lui a récemment apporté des éléments de réponse intéressants dans le cadre de la sélection de noyau reproduisant en général. Parmi ces approches, l'alignement noyau-cible présente en particulier l'avantage de pouvoir sélectionner *a priori* une représentation temps-fréquence sans recourir à des apprentissages répétés de règles de décision suivis de validation croisée. Cette notion d'alignement pour mesurer une similitude entre deux noyaux, permet de quantifier la qualité d'un noyau reproduisant vis-à-vis d'une tâche de discrimination.

Ce chapitre est organisé ainsi. Dans un premier temps, on décrit brièvement les méthodes existantes pour l'optimisation de la distribution temps-fréquence d'une part, et pour la sélection de noyaux reproduisants d'autre part. La suite du chapitre est alors consacrée à un critère de sélection particulier, le critère d'alignement noyau-cible. On adapte ce critère pour la sélection et l'élaboration de distributions temps-fréquence optimales pour un problème de classification donné.

6.1 Introduction

Le choix de la distribution temps-fréquence optimale pour une application donnée a suscité d'amples études. Par le lien que l'on a établi entre distribution temps-fréquence et noyau reproduisant, on traduit ce problème par la recherche d'un noyau reproduisant optimal au sens d'un critère à définir.

6.1.1 Distribution temps-fréquence optimale : un aperçu

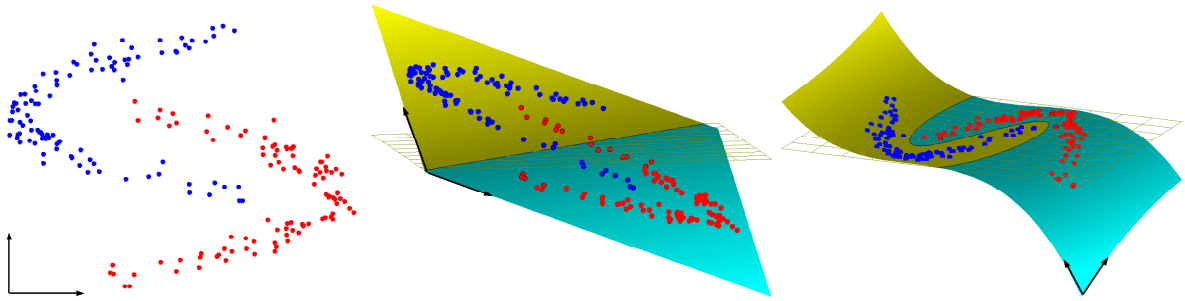
Toutes déclinées de la distribution de Wigner, les distributions de la classe de Cohen offrent une variété d'espaces de représentation à la mesure des objectifs visés par les utilisateurs. Toutefois, ceci passe nécessairement par un choix soigneux de la fonction de paramétrisation en fonction du type de signal analysé, des propriétés voulues et de l'application visée.

Il est possible de privilégier l'intelligibilité de l'information délivrée par une représentation en y limitant les manifestations de bruits et autres termes interférentiels qui pourraient nuire à sa lisibilité. Ceci est le cas par exemple de la paramétrisation à profil radialement Gaussien [BJ93, JB95, RB93], de la modélisation par mélange de fonctions Gaussiennes [CF99], de la réallocation [AF95, CMAF05], ou encore de la diffusion adaptative [GRG05]. On peut encore être à la recherche d'un espace de représentation favorisant la résolution d'un problème de décision. Dans [McL97, MDA97], la distance Euclidienne entre les (représentations) moyennes de chacune des classes de distributions est maximisée. Ne prenant pas en compte la variance intra-classe des représentations temps-fréquence, ce critère est souvent remplacé par un critère du second ordre, le contraste de Fisher. Ceci est le cas de [GA01], avec l'usage de la distance Euclidienne, et l'optimisation du critère par un ensemble réduit de coordonnées du plan des ambiguïtés afin de régulariser la solution. Ce critère est aussi considéré dans [Hei95, HT97], mais avec deux modifications majeures. D'une part, le critère du second ordre considéré est construit avec une mesure de dissimilarité entre deux représentations obtenue à partir de leur produit scalaire[7]. D'autre part, la régularisation est pratiquée grâce à un choix d'une sous-classe de distributions telles que celle de pseudo Wigner-Ville lissées, voir Chapitre 2, et normalisée selon (2.12). Cette normalisation permet de considérer les distributions temps-fréquence comme des densités de probabilité, ouvrant ainsi la voie à d'autres types de distances, ou plus précisément de divergences entre les lois de probabilités [MBF94]. Ceci est utilisé dans [DD98], où les auteurs considèrent une paramétrisation à profil radialement Gaussien qu'ils optimisent grâce à une distance Euclidienne entre représentations normalisées. Dans [DD99], les auteurs étendent cette approche à d'autres types de distances et de classes de distributions temps-fréquence. Dans cette perspective, diverses techniques pour optimiser les distributions temps-fréquence ont été élaborées, comme par exemple dans [DDBB01], ou encore plus récemment dans [Dav04]. Mis à part [Dav00], peu de travaux ont abordé l'aspect théorique de ces méthodes de détermination de représentations temps-fréquence optimales, et plus précisément leur capacité en généralisation. Par une reformulation du problème en un problème de méthode à noyau, on a accès aux différentes avancées dans ce domaine.

6.1.2 Un point de vue noyau reproduisant

Les méthodes à noyau sont des algorithmes linéaires appliquées implicitement dans un espace de Hilbert transformé, par le coup du noyau. Confronté à un problème décisionnel, on choisit généralement en premier lieu l'algorithme de classification à utiliser, avant de se focaliser sur le choix du noyau reproduisant et l'estimation de ses paramètres. Les aspects lié à la sélection du noyau sont fondamentaux car ils définissent l'espace de représentation dans lequel la règle de décision va opérer. Afin de s'en convaincre,

[7] On retrouve le principe des méthodes à noyau et en particulier le coup du noyau, puisque l'algorithme résultant ne dépend que des produits scalaires des représentations temps-fréquence.



(a) (b) (c)

FIG. 6.1 – Problème de classification d'un jeu de données en dimension 2 (a). Séparatrice obtenue par un algorithme de SVM, à partir d'un noyau linéaire (b) et polynômial de degré 3 (c).

il suffit de considérer un algorithme SVM appliqué à un problème de classification tel que celui présenté à la Figure 6.1 (a). Les deux classes ne sont évidemment pas linéairement séparables, comme le montre le séparateur illustré à la Figure 6.1 (b). Pour remédier à cela, on peut ici recourir par exemple à un noyau polynômial de degré 3 comme le montre la Figure 6.1 (c). Différentes approches pour systématiser la recherche du meilleur noyau reproduisant ont été proposées. Après une brève présentation de ces approches, on considère le critère d'alignement noyau-cible qui, par ses propriétés statistiques et géométriques, offre une complexité algorithmique réduite par rapport à des méthodes plus classiques.

6.1.3 Critères de sélection de modèle : un aperçu

Plusieurs résultats de la théorie d'apprentissage montrent l'importance des connaissances a priori pour un problème étudié. Ceci se traduit par les théorèmes *Ugly Duckling* [Wat85] et *no free lunch* [WM97, Wol01], qui affirment qu'il n'existe aucun algorithme optimal ou représentation optimale des données valable quel que soit le problème. Pour les méthodes à noyau, le théorème *no free kernel* [CKEST06] stipule qu'aucun noyau n'est optimal pour toutes les applications, et qu'une connaissance a priori se manifestant par un choix approprié du noyau est nécessaire. On peut alors recourir à des algorithmes spécifiques mettant en évidence une relation entre les données d'entrée et leurs cibles, appelées aussi étiquettes en classification. Dans ce chapitre, on s'intéresse au critère d'alignement noyau-cible, après cette brève présentation de différentes approches existantes pour la sélection de noyau.

Outre les techniques de transformation conforme [AW99], on compte les approches de validation croisée et autres *leave-one-out*. Pour chaque noyau étudié, l'erreur de généralisation est estimée à partir de plusieurs classifieurs élaborés pour différents sous-ensembles de la base d'apprentissage et validés sur les données restantes [MLH03]. Le noyau optimal, celui qui correspond à la plus faible erreur, est souvent obtenu par une recherche sur une grille de valeurs. Cette approche s'avère très coûteuse puisqu'on a recours à plusieurs phases d'apprentissage et de validation. On parle alors d'une complexité calculatoire en $\mathcal{O}(n^4)$ ou, dans le meilleur des cas en $\mathcal{O}(n^3)$ pour une implémentation efficace avec certains algorithmes [CT03]. Afin d'accélérer cette recherche, des méthodes d'estimation d'une borne de l'erreur de généralisation ont été récemment proposées, nécessitant une seule construction de classifieur, pour chaque noyau étudié. Parmi celles-ci, on a la borne de la VC dimension [Bur98], celle de rayon-marge [CKS$^+$03], ou encore *generalized approximate cross validation* [WLZ00]. D'autres possibilités sont également mentionnées dans [CVBM02, DKP03]. De nouvelles méthodes permettent de déterminer une combinaison optimal d'un ensemble de noyaux, ceux-ci pouvant être différents par nature, ou de même nature mais à paramétrisation distincte. L'idée de base, empruntée du *boosting* [CSS02], consiste à combiner plusieurs classifieurs aux performances modestes afin d'aboutir à un classifieur

plus compétitif. Ce principe, reformulé dans le contexte de l'ingénierie des noyaux, vise à combiner des noyaux génériques afin d'améliorer les performances, sous réserve que le noyau résultant soit lui même un noyau défini positif, donc reproduisant. Parmi ces méthodes on retrouve le critère de séparation des classes [KMB06] à partir du quotient de Fisher [WC02], le critère d'alignement noyau-cible [CSTEK01, CKEST06], la programmation semi-définie [LCB$^+$04], ainsi que le principe d'apprentissage du noyau proposé dans [Ong05]. L'apprentissage du noyau consiste ici à partir d'un ensemble de plusieurs noyaux disponibles, le noyau optimal étant obtenu par combinaison linéaire de ces derniers. Bien que le cadre défini par cette méthode soit intéressant, sa complexité calculatoire peut s'avérer être un facteur bloquant.

Il est important de noter que les différentes approches présentées ci-dessus se transposent facilement au domaine temps-fréquence. Il en est de même pour leurs propriétés statistiques et théoriques. On se contente dans ce chapitre de considérer le critère d'alignement noyau-cible, et de montrer sa mise en œuvre dans le plan temps-fréquence. On présente aussi le lien avec les différentes approches, en montrant par exemple que le noyau optimal ainsi obtenu maximise la séparabilité des classes.

6.2 Critère d'alignement noyau-cible

6.2.1 Notions d'alignement noyau-cible

Les performances des méthodes de reconnaissance des formes à noyau sont principalement déterminées par le noyau reproduisant considéré. Pour un algorithme donné, par exemple les SVM, deux noyaux produiront des résultats d'autant plus proches qu'ils sont similaires. L'alignement, introduit par Cristianini *et coll.* [CSTEK01], est une mesure de similarité entre deux noyaux reproduisants, ou entre un noyau reproduisant et une fonction cible. Étant donné une base d'apprentissage $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, l'alignement (empirique) des deux noyaux κ_1 et κ_2 est défini par

$$A(\kappa_1, \kappa_2; \mathcal{A}_n) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}},$$

où $\langle \cdot, \cdot \rangle_F$ est le produit scalaire de Frobénius, et K_1 et K_2 les matrices de Gram de termes respectifs $\kappa_1(x_i, x_j)$ et $\kappa_2(x_i, x_j)$. Puisque les matrices de Gram contiennent toutes les informations nécessaires à l'apprentissage, nous noterons l'alignement par $A(K_1, K_2)$. Ainsi cette mesure correspond-t-elle plus concrètement au cosinus de l'angle entre K_1 et K_2. Par cette analogie, on a $-1 \leq A(K_1, K_2) \leq 1$ en général et $A(K_1, K_2) \geq 0$ pour les matrices définies positives [STC04b]. On peut écrire l'alignement sous forme de distance entre les deux matrices de Gram normalisées, selon

$$A(K_1, K_2) = 2 - \left\| \frac{K_1}{\|K_1\|_F} - \frac{K_2}{\|K_2\|_F} \right\|_F \tag{6.1}$$

où $\|K\|_F = \sqrt{\langle K, K \rangle_F}$ est la norme de Frobénius de la matrice K.

Le concept d'alignement peut aussi quantifier la similarité entre un noyau et une fonction cible. Dans le cadre d'un problème de classification, on cherche à apprendre la statistique de décision ψ qui vise à respecter $\psi(x_i) = y_i$, où y_i est l'étiquette de l'observation x_i. Pour un problème de classification bi-classes, on cherche alors à acquérir la fonction cible $\psi : x \mapsto \{-1, +1\}$ selon la classe d'appartenance de x. Le noyau reproduisant correspondant à la transformation idéale ϕ^* telle que $\phi^*(x_i) = y_i$ est $\kappa^*(x_i, x_j) = y_i\, y_j$. La fonction cible $\psi^*(\cdot) = \sum_{i=1}^n y_i \kappa^*(x_i, \cdot)$ ainsi obtenue permet de retrouver les étiquettes des données de l'ensemble d'apprentissage. La matrice de Gram idéale associée à κ^* est alors

$K^* = \boldsymbol{y}\,\boldsymbol{y}^\top$, où $\boldsymbol{y}$ est un vecteur colonne cible dont le $j^{\text{ème}}$ élément est y_j, ce qui correspond à la matrice de Gram

$$K^*(i,j) = \begin{cases} 1 & \text{si } y_i = y_j \\ -1 & \text{si } y_i \neq y_j. \end{cases}$$

Dans ce qui suit, on se contente d'étudier un problème de classification bi-classes avec la matrice K^* définie ci-dessus comme matrice cible. On rappelle que l'on peut généraliser facilement les développements à venir au cas multi-classes. Pour le cas de c classes, il suffit de considérer comme cibles les c vecteurs de norme unité et équiangulaire [Bur98], c'est-à-dire : $\kappa^*(x_i, x_j) = -1/(c-1)$ si les classes de x_i et x_j sont différentes, sinon $\kappa^*(x_i, x_j) = 1$.

6.2.2 Critère d'alignement

Dans [CSTEK01, CKEST06], Cristianini *et coll.* suggèrent d'utiliser cette matrice cible $K^* = \boldsymbol{y}\,\boldsymbol{y}^\top$ et le critère d'alignement afin de rechercher le noyau reproduisant le mieux adapté à la résolution d'un problème de classification bi-classes donné. L'expression de l'alignement noyau-cible, entre un noyau κ donné et le noyau optimal κ^*, s'exprime alors selon

$$A(K, K^*) = \frac{\langle K, \boldsymbol{y}\,\boldsymbol{y}^\top\rangle_F}{\sqrt{\langle K, K\rangle_F \langle \boldsymbol{y}\,\boldsymbol{y}^\top, \boldsymbol{y}\,\boldsymbol{y}^\top\rangle_F}} = \frac{\boldsymbol{y}^\top K \boldsymbol{y}}{n\|K\|_F}.$$

où K est la matrice de Gram de termes $\kappa(x_i, x_j)$. Le noyau optimal correspond donc à celui maximisant $\langle K, \boldsymbol{y}\,\boldsymbol{y}^\top\rangle_F$ sous la contrainte de normalisation $\|K\|_F = 1$.

Afin de mieux comprendre le critère d'alignement noyau-cible, on établit un lien direct avec la séparabilité des classes. L'alignement noyau-cible d'un noyau normalisé selon $\|K\| = 1$ s'exprime selon $A(K, K^*) = \langle K, \boldsymbol{y}\,\boldsymbol{y}^\top\rangle_F / n$ où encore, en développant

$$A(K, K^*) = \frac{1}{n}\sum_{i,j=1}^{n} y_i y_j \kappa(x_i, x_j) = \frac{1}{n}\Big\|\sum_{i=1}^{n} y_i \kappa(x_i, \cdot)\Big\|^2 = \frac{1}{n}\Big\|\sum_{y_i=+1} \kappa(x_i, \cdot) - \sum_{y_i=-1} \kappa(x_i, \cdot)\Big\|^2,$$

Ceci correspond donc à une mesure de distance interclasse. En supposant qu'on ait le même nombre de données dans les deux classes, on peut écrire l'expression ci-dessus selon

$$A(K, K^*) = \frac{n}{4}\|\mu_+ - \mu_-\|^2,$$

où μ_+ et μ_- désignent les centres d'inertie des deux classes. Le noyau reproduisant qui maximise le critère d'alignement noyau-cible correspond alors, à une isométrie près, à celui qui induit un RKHS où les classes des données engendrées sont le mieux séparables. D'où une similitude entre le critère d'alignement et le critère de contraste de Fisher étudié au Chapitre 5. D'autres interprétations du critère d'alignement peuvent être établies dans la même esprit, comme par exemple un lien avec la maximisation de l'aire sous la courbe ROC [Rak04].

Un biais nul et une variance à décroissance exponentielle en fonction de la taille de l'ensemble d'apprentissage garantissent la qualité et la constance de l'estimation. Il s'agit du principe de concentration au sens de McDiarmid. La pertinence du critère d'alignement est assurée par un lien direct avec l'erreur de généralisation, en démontrant qu'un estimateur de Parzen de la forme $h(\cdot) = \frac{1}{n}\sum_{i=1}^{n} y_i \kappa(x_i, \cdot)$ admet de bonnes performances en généralisation quand l'alignement est élevé. La concentration de l'alignement et son lien avec l'erreur en généralisation sont établies dans [CSTEK01]. Depuis les premiers travaux de Cristianini *et coll.*, le principe de maximisation de l'alignement noyau-cible a été étendu à

d'autres problèmes, par exemple celui de la régression [KSTC02a]. Il a également suscité plusieurs travaux sur l'apprentissage de métriques [WCP05, LCB+02], la décomposition en valeurs propres de la matrice de Gram [KSTC02a, CKEST06], ou encore la combinaison de noyaux [KSTC02b, PR05, PR06] en vue d'une amélioration des performances de l'ensemble.

Cette démarche présente l'intérêt de ne nécessiter aucun apprentissage de la statistique de décision, la sélection du noyau étant pratiquée *a priori*. Trois applications potentielles sont alors proposées dans ce cadre. Dans une première application, on s'intéresse à la sélection d'un noyau optimal parmi un ensemble de noyaux candidats, où l'optimalité est déterminée par le plus grand alignement avec la cible. Une seconde application consiste à ajuster les paramètres d'un noyau donné de sorte à maximiser son alignement. Un exemple est la paramétrisation du noyau Gaussien, où on cherche à déterminer la largeur de bande optimale. Une troisième application consiste à combiner linéairement des noyaux, souvent de natures différentes, afin d'obtenir un noyau mieux aligné avec la cible et donc plus performant.

6.2.3 Ajustement optimal des paramètres

La plus simple stratégie pour déterminer les paramètres d'un noyau consiste à évaluer l'alignement sur une grille de valeurs possibles. Bien que cette approche puisse être envisagée pour un paramètre unique, elle s'avère coûteuse lorsqu'il s'agit de considérer un vecteur de paramétrisation. Pour remédier à cela, on considère une approche de montée de gradient pour optimiser le paramètre du noyau en question. Cette approche ne peut évidemment être envisagée que sous réserve que l'alignement soit differentiable par rapport au paramètre. Il en est ainsi pour les noyaux Gaussien et de Laplace par rapport à leur largeur de bande. Ceci n'est en revanche pas le cas du noyau polynômial, $\kappa_q(x_i, x_j) = (1+\langle x_i, x_j\rangle)^q$ par rapport à son paramètre q, tandis que $\kappa_p(x_i, x_j) = (p + \langle x_i, x_j\rangle)^q$ est differentiable par rapport à p.

Soit κ_σ un noyau reproduisant de paramètre σ. Le paramètre optimal est alors donné par la maximisation de l'alignement noyau-cible, selon

$$\sigma^* = \arg\max_\sigma A(K_\sigma, K^*) = \arg\max_\sigma \frac{\langle K_\sigma, K^*\rangle_F}{n\|K_\sigma\|_F}, \tag{6.2}$$

où K_σ désigne la matrice de Gram associée au noyau et K^* la matrice cible. Pour mettre en œuvre une approche d'optimisation par montée de gradient, on recherche une expression du gradient de l'alignement par rapport au paramètre σ à partir du gradient du noyau étudié. L'expression du gradient du numérateur dans (6.2) par rapport à σ s'écrit

$$\nabla_\sigma \langle K_\sigma, K^*\rangle_F = \sum_{i,j=1}^{n} y_i y_j \nabla_\sigma \kappa_\sigma(x_i, x_j).$$

Le gradient de $\|K_\sigma\|_F$ par rapport à σ s'exprime selon

$$\begin{aligned}
\nabla_\sigma \|K_\sigma\|_F &= \nabla_\sigma \Big(\sum_{i,j=1}^{n} \kappa_\sigma^2(x_i, x_j)\Big)^{-\frac{1}{2}} \\
&= \Big(\sum_{i,j=1}^{n} \kappa_\sigma^2(x_i, x_j)\Big)^{-\frac{1}{2}} \sum_{i,j=1}^{n} \kappa_\sigma(x_i, x_j)\ \nabla_\sigma \kappa_\sigma(x_i, x_j) \\
&= \frac{1}{\|K_\sigma\|_F} \sum_{i,j=1}^{n} \kappa_\sigma^2(x_i, x_j)\ \nabla_\sigma \kappa_\sigma(x_i, x_j).
\end{aligned}$$

En combinant les deux expressions, on peut alors écrire le gradient de l'alignement $A(K_\sigma, K^*)$ par rapport à σ ainsi

$$\nabla_\sigma A(K_\sigma, K^*) = \frac{1}{n\|K_\sigma\|_F} \sum_{i,j=1}^{n} y_i y_j \nabla_\sigma \kappa_\sigma(x_i, x_j) - \frac{\langle K_\sigma, K^* \rangle_F}{n\|K_\sigma\|_F^3} \sum_{i,j=1}^{n} \kappa_\sigma(x_i, x_j)\, \nabla_\sigma \kappa_\sigma(x_i, x_j).$$

La méthode de montée de gradient pour la maximisation de l'alignement est constituée principalement de l'itération

$$\sigma_k \longleftarrow \sigma_k + \rho\, \nabla_\sigma A(K_\sigma, K^*).$$

D'autres étapes peuvent être aussi considérée à chaque itération, afin d'imposer certaines contraintes au noyau. On parle alors d'optimisation alternée. Cette démarche est étudiée en détails dans la suite pour l'optimisation de la distribution temps-fréquence à paramétrisation radialement Gaussienne, où le noyau reproduisant associé est differentiable par rapport à la largeur de bande.

6.2.4 Combinaison linéaire

Grâce aux propriétés algébriques des fonctions définies positives, on peut construire un noyau reproduisant à partir d'une combinaison linéaire de noyaux, présenté dans la Section 1.2.2, à condition que leurs coefficients de pondération soient non-négatifs. La détermination de la pondération optimale, dans le cadre du critère d'alignement, fait l'objet de ce paragraphe. Des solutions plus ou moins coûteuses en temps de calcul ont été proposées dans le contexte général des méthodes à noyau pour résoudre ce problème d'optimisation. Dans [KSTC02b], on recherche les coefficients α_k tels que le noyau composite suivant maximise l'alignement noyau-cible :

$$\kappa(x_i, x_j) = \sum_{k=1}^{m} \alpha_k \kappa_k(x_i, x_j), \tag{6.3}$$

où m désigne le nombre de noyaux candidats. On impose la positivité des m coefficients α_k afin d'assurer le caractère défini-positif du noyau résultant κ. Plus précisément, on a le problème d'optimisation avec contrainte suivant :

$$\begin{array}{rl} \max\limits_{\boldsymbol{\alpha}} & A(\sum_{k=1}^{m} \alpha_k K_k, K^*) \\ \text{sous contrainte} & \alpha_1, \ldots, \alpha_m \geq 0, \end{array}$$

où K_k est la matrice de Gram du noyau κ_k. L'optimisation de l'alignement relativement au modèle linéaire (6.3) se ramène selon [KSTC02b] au problème d'optimisation convexe suivant :

$$\max_{\boldsymbol{\alpha}} \sum_i \alpha_i \langle K_i, K^* \rangle_F - \sum_{i,j} \alpha_i \alpha_j \langle K_i, K_j \rangle_F - \eta \sum_i \alpha_i^2$$

sous les m contraintes $\alpha_1, \ldots, \alpha_m \geq 0$. Dans cette expression, le terme de régularisation $\eta \sum_i \alpha_i^2$ est nécessaire pour éviter un phénomène de sur-apprentissage par rapport au critère d'alignement. Ce problème d'optimisation est similaire à celui obtenu dans [LCB+02] en maximisant la marge par l'usage de la programmation semi-définie. Pour résoudre se problème d'optimisation, les auteurs ont recours à des algorithmes classiques de programmation quadratique, les mêmes outils utilisés dans le cadre des SVM et nécessitant un coût calculatoire équivalent. Afin d'éviter un tel surcoût, on a recours à une stratégie de division du problème en sous-problèmes. Selon [PR05], il existe une solution analytique simple à

ce problème pour le cas d'une combinaison de deux noyaux. Le problème d'optimisation se réduit à la recherche des coefficients non-négatifs α_1 et α_2 qui maximisent

$$\alpha_1 \langle K_1, K^* \rangle_F + \alpha_2 \langle K_2, K^* \rangle_F - \alpha_1^2(\|K_1\|_F^2 + \eta) - \alpha_2^2(\|K_2\|_F^2 + \eta) - 2\alpha_1\alpha_2\langle K_1, K_2 \rangle_F.$$

On rappelle que la contrainte de non-négativité des coefficients est une condition suffisante pour avoir un noyau défini positif. L'optimalité est obtenue par l'annulation du gradient par rapport à α_1 et α_2. La solution à ce problème est donnée par

$$(\alpha_1^*, \alpha_2^*) = \begin{cases} (\alpha_1, \alpha_2) & \text{si} \quad \alpha_1, \alpha_2 > 0 \\ (1, 0) & \text{si} \quad \alpha_2 \leq 0 \\ (0, 1) & \text{si} \quad \alpha_1 \leq 0, \end{cases} \tag{6.4}$$

où

$$\alpha_1 = \frac{1}{2} \frac{\langle K_1, K^* \rangle_F - 2\langle K_1, K_2 \rangle_F \alpha_2}{\|K_1\|_F^2 + \eta}$$

$$\alpha_2 = \frac{1}{2} \frac{(\|K_1\|_F^2 + \eta)\langle K_2, K^* \rangle_F - \langle K_1, K_2 \rangle_F \langle K_1, K^* \rangle_F}{(\|K_1\|_F^2 + \eta)(\|K_2\|_F^2 + \eta) - \langle K_1, K_2 \rangle_F^2},$$

avec $\eta \geq 0$ le terme de régularisation fixé par l'utilisateur. Le noyau optimal s'exprime alors selon

$$\kappa(x_i, x_j) = \alpha_1^* \kappa_1(x_i, x_j) + \alpha_2^* \kappa_2(x_i, x_j). \tag{6.5}$$

Pour traiter le cas d'une combinaison linéaire à plus de deux noyaux, une stratégie de type *Branch and Bound* peut être adoptée. Elle consiste à retenir le meilleur noyau à disposition au sens de l'alignement, puis à sélectionner parmi les noyaux restants celui qui maximise l'alignement par combinaison linéaire. Ce principe est réitéré tant que l'accroissement de l'alignement est jugé suffisant.

6.3 Distributions temps-fréquence optimales par le critère d'alignement

Dans un cadre décisionnel, les méthodes de sélection de représentations temps-fréquence existantes consistent essentiellement en une recherche de celle conduisant à une erreur de classification minimum par validation croisée. Nous avons essayé de montrer, tout au long de ce manuscrit, que les méthodes à noyau reproduisant constitueront sans nul doute une nouvelle source de progrès en analyse temps-fréquence par la diversité des traitements envisageables, leurs performances et leur coût calculatoire généralement réduit. La mise en œuvre de différentes méthodes de choix de noyau, et plus précisément du critère d'alignement noyau-cible, pour la sélection de représentations temps-fréquence s'inscrit dans cette démarche. On considère dans la suite un ensemble d'apprentissage $\mathcal{A}_n$ formé de n signaux x_i d'énergie finie et de leurs étiquettes y_i. On désigne par κ^* le noyau cible et par $K^* = \boldsymbol{y}^\top \boldsymbol{y}$ la matrice cible correspondante.

6.3.1 Sélection de distribution optimale

On propose dans cette application de sélectionner une distribution temps-fréquence optimale au sein d'un ensemble de distributions candidates de la classe de Cohen. On rappelle pour cela qu'une distribution appartenant à cette classe s'exprime sous la forme $C_x^\Pi(t,f) = \iint \Pi_{\text{dr}}(\nu,\tau)\ A_x(\nu,\tau)\ \mathrm{e}^{-j2\pi(f\tau - t\nu)}\,\mathrm{d}\nu\,\mathrm{d}\tau$, où Π_{dr} est la fonction de paramétrisation exprimée dans le plan Doppler-retard et A_x la fonction d'ambiguïté du signal x. Le noyau reproduisant associé à cette distribution s'écrit sous forme $\kappa_{C^\Pi}(x_i, x_j) = \iint |\Pi_{\text{dr}}(\nu,\tau)|^2 A_{x_i}(\nu,\tau)\, A_{x_j}(\nu,\tau)\,\mathrm{d}\nu\,\mathrm{d}\tau$. On désigne par

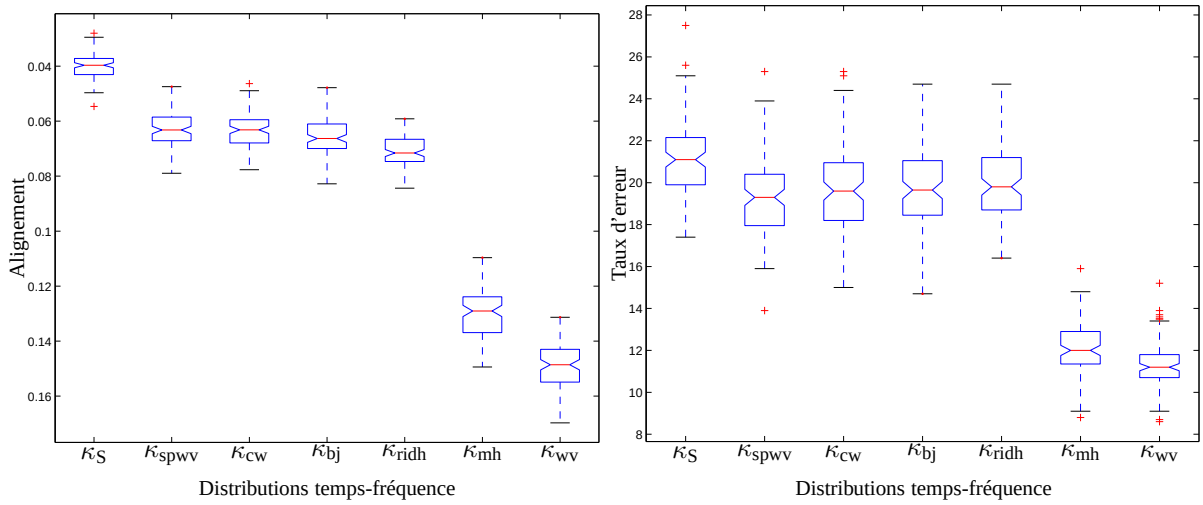


FIG. 6.2 – Boites à moustaches représentant les variations de l'alignement et du taux d'erreur pour différentes distributions temps-fréquence

K_{C^Π} la matrice de Gram de termes $\kappa_{C^\Pi}(x_i, x_j)$ pour $i, j \in \{1, \ldots, n\}$. L'alignement entre le noyau κ_{C^Π} et le noyau cible κ^* est donné par

$$A(K_{C^\Pi}, K^*) = \frac{\langle K_{C^\Pi}, K^* \rangle_F}{n \, \|K_{C^\Pi}\|_F}. \tag{6.6}$$

Disposant d'un ensemble de distributions temps-fréquence candidates, l'approche consiste à les comparer à partir de leur alignement avec la cible et à sélectionner celle qui présente le meilleur. On propose d'illustrer cela à partir de l'application suivante.

Le problème traité concerne la détection du signal $\exp^{2j\pi[\phi(t)+\phi_0]}$ noyé dans un bruit blanc gaussien de variance 2.25, où $\phi(t)$ définit une loi de fréquence instantanée parabolique, et ϕ_0 est une phase initiale aléatoire uniformément distribuée sur $[0, 2\pi[$. On dispose pour cela de 200 réalisations de 64 échantillons pour chacune des hypothèses, et on note par $\mathcal{A}_{200}$ cet ensemble d'apprentissage. On souhaite comparer les distributions[8] de Wigner (κ_{wv}) et pseudo-Wigner lissée (κ_{spwv}), de Margenau-Hill (κ_{mh}), de Choï-Williams (κ_{cw}), de Born-Jordan (κ_{bj}), à interférences réduites par fenêtrage de Hamming (κ_{ridh}), ainsi que le spectrogramme (κ_S).

On a choisi dans un premier temps de mettre en avant la concentration de l'alignement avec la cible pour les différentes distributions temps-fréquence considérées. Pour chaque distribution, on estime la distribution de l'alignement sur une base de 100 ensembles d'apprentissage $\mathcal{A}_{200}$ déterminés comme ci-dessus. On représente à la Figure 6.2 (a) les réalisations de l'alignement pour chacune des distributions. On compare l'alignement avec le taux d'erreur dans un second temps. Le taux d'erreur est obtenu en considérant un classifieur SVM reposant sur le même noyau, élaboré à partir des mêmes données d'apprentissage et testé sur un ensemble de 500 nouvelles réalisations pour chacune des classes. On représente à la Figure 6.2 (b) le taux d'erreur pour chacune des distributions temps-fréquence. Comme illustré aux Figures 6.2, les distributions de Choï-Williams, Born-Jordan, pseudo-Wigner lissée, et à fenêtrage de Hamming, présentent approximativement les mêmes valeurs d'alignement, ainsi que des taux d'erreur

[8] Les configurations retenues pour ces distributions sont celles proposées par défaut par la boite à outils temps-fréquence sous Matlab.

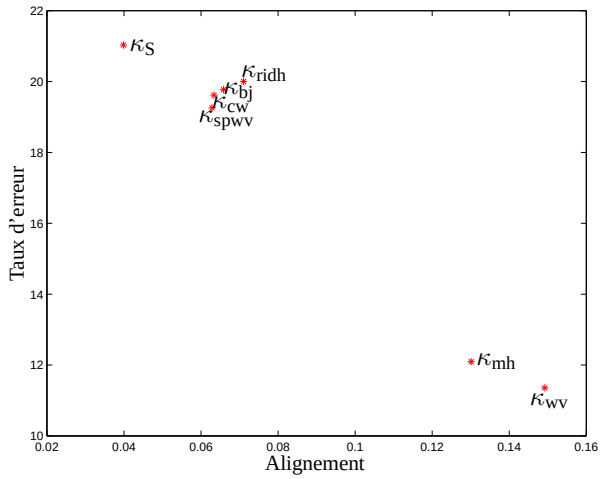


FIG. 6.3 – Alignement et taux d'erreur pour différents noyaux reproduisants de distribution temps-fréquence

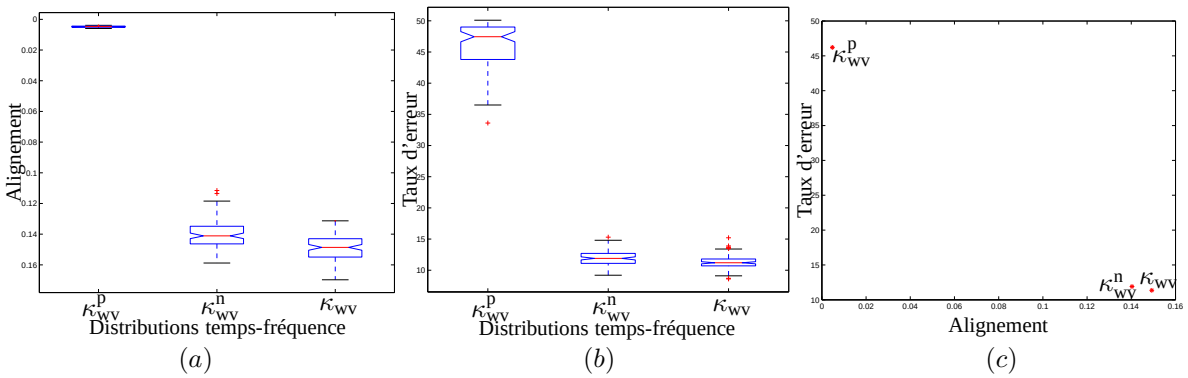


FIG. 6.4 – Comparaison des distributions normalisées par rapport à la distribution de Wigner. Boites à moustaches représentant les variations de l'alignement (a), du taux d'erreur (b) et la corrélation entre les deux (c).

très proches. Du fait de leur lissage, ces distributions sont souvent considérées comme des distributions à interférences réduites. Le spectrogramme se manifeste par un alignement plus faible et un taux d'erreur plus élevé. On distingue d'autre part les distributions de Wigner et de Margeneau-Hill, qui sont toutes deux des distributions non-lissée. En combinant les deux résultats, la Figure 6.3 fournit l'alignement et le taux d'erreur moyens sur les 100 ensembles d'apprentissage, pour chaque distribution temps-fréquence. La corrélation entre taux d'erreur et alignement confirme la pertinence des informations fournies par ce dernier, ici exploité pour le choix d'un noyau parmi d'autres. Le noyau associé à la distribution de Wigner se détache nettement. Ce résultat est conforme à celui fourni par la théorie statistique de la décision, le filtre adapté avec détecteur d'enveloppe désignant la distribution de Wigner comme mode de représentation optimum pour le problème traité en raison de son unitarité [Fla98].

Nous proposons d'étudier la normalisation d'une représentation temps-fréquence et ses effets sur l'alignement et le taux d'erreur. Pour cela, on se contente d'étudier la distribution de Wigner, avec deux normalisations particulières. La première, définie par l'expression (2.12), a été proposée par Davy et *al.* dans le cadre de classification de signaux non-stationnaires. Pour la distribution de Wigner, elle s'exprime

ainsi

$$W_x^{\mathrm{p}}(t,f) = \frac{|W_x(t,f)|}{\iint |W_x(t,f)|\,\mathrm{d}t\,\mathrm{d}f}.$$

La seconde consiste à considérer la normalisation selon

$$W_x^{\mathrm{n}}(t,f) = \frac{W_x(t,f)}{\sqrt{\iint (W_x(t,f))^2\,\mathrm{d}t\,\mathrm{d}f}}.$$

Cette normalisation est naturelle dans le domaine des méthodes à noyau. Elle peut être évaluée à partir du noyau reproduisant κ_W associé à la distribution de Wigner, sans la nécessité de l'évaluer. Pour cela, il suffit de considérer le noyau reproduisant normalisé

$$\kappa_{W^{\mathrm{n}}}(x_i,x_j) = \langle W_{x_i}^{\mathrm{n}}, W_{x_j}^{\mathrm{n}} \rangle = \frac{\iint W_{x_i}(t,f) W_{x_j}(t,f)\,\mathrm{d}t\,\mathrm{d}f}{\sqrt{\iint (W_{x_i}(t,f))^2\,\mathrm{d}t\,\mathrm{d}f}\sqrt{\iint (W_{x_j}(t,f))^2\,\mathrm{d}t\,\mathrm{d}f}} = \frac{\kappa_W(x_i,x_j)}{\sqrt{\kappa_W(x_i,x_i)\,\kappa_W(x_j,x_j)}}.$$

La Figure 6.4 illustre le taux d'erreur et l'alignement obtenus pour chacune des distributions, avec les mêmes signaux et la même configuration que ci-dessus. La première normalisation étudiée, avec son très faible alignement, ne conduit pas à de bonnes performances en classification. La seconde normalisation produit des performances très proches de la distribution de Wigner, comme le critère d'alignement le laissait le prévoir.

6.3.2 Estimation des paramètres du spectrogramme

Cette section concerne le choix optimum des paramètres d'une représentation temps-fréquence, au sens du critère d'alignement noyau-cible. Dans le cas général d'une distribution de la classe de Cohen, la maximisation du critère d'alignement consiste à déterminer la fonction de paramétrisation Π^* telle que

$$\Pi^* = \arg \max_{\Pi \in \mathbb{R}^{l\times l}} \frac{\langle K_{C^\Pi}, K^* \rangle_F}{n \|K_{C^\Pi}\|_F}, \tag{6.7}$$

avec K_{C^Π} la matrice de Gram associé à la distribution temps-fréquence considérée. La résolution de ce problème nécessite la détermination d'une fonction bi-dimensionnelle de taille $l \times l$. On peut diminuer la dimensionnalité du problème en se restreignant à une classe de distribution donnée. Sans que cela nuise au caractère général de la démarche, on considère dans ce qui suit l'ajustement de la fenêtre de lissage du spectrogramme. Dans la section suivante, nous traiterons le problème d'optimisation de la paramétrisation temps-fréquence à profil radialement Gaussien avec un algorithme de descente de gradient.

On considère l'ajustement de la largeur de fenêtre w du spectrogramme S_x^w dans le but final de minimiser les erreurs de classification. Pour cela, on a recours au critère d'alignement, évalué sur une grille de valeurs possibles. Le noyau reproduisant du spectrogramme est défini par $\kappa_{\mathrm{S}^w}(x_i,x_j) = \langle S_{x_i}, S_{x_j} \rangle$. La largeur de fenêtre optimale w^* est alors donnée par

$$w^* = \arg\max_{w\in\mathbb{R}} \frac{\langle K_{\mathrm{S}^w}, K^* \rangle_F}{n \|K_{\mathrm{S}^w}\|_F},$$

avec K_{S^w} la matrice de Gram associée au noyau reproduisant κ_{S^w}. Le problème traité concerne la détection du signal $\mathrm{e}^{2j\pi[\phi(t)+\phi_0]}$ noyé dans un bruit blanc gaussien de variance 2.25, où $\phi(t)$ définit une loi de fréquence instantanée parabolique, et ϕ_0 est une phase initiale aléatoire uniformément distribuée sur $[0, 2\pi[$. On dispose pour cela de 100 réalisations de 64 échantillons pour chacune des hypothèses. La Figure 6.5 présente les valeurs d'alignement du noyau κ_{S} en fonction de la largeur de la fenêtre de Hamming utilisée pour le spectrogramme. On note que l'alignement maximum est obtenu pour une fenêtre

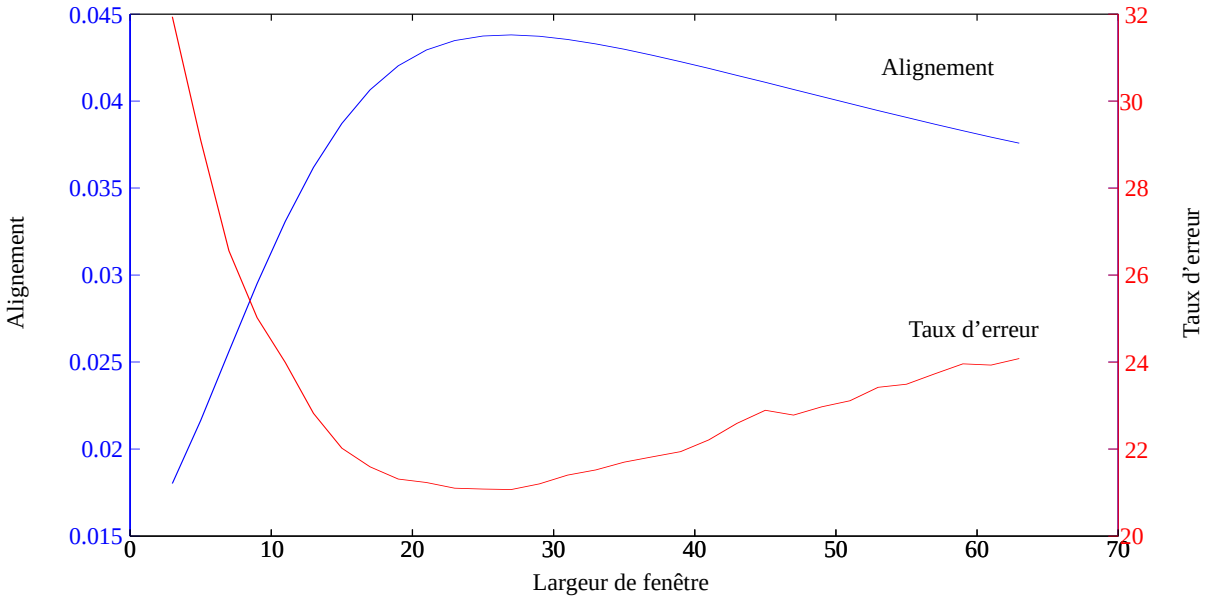


FIG. 6.5 – Ajustement de la largeur de la fenêtre de lissage pour le spectrogramme, avec le critère d'alignement noyau-cible. Comparaison avec le taux d'erreur obtenu par un classifieur SVM.

de largeur 23. Ces résultats sont confrontés aux performances de classifieurs SVM reposant sur le même noyau, élaborés à partir des mêmes données d'apprentissage, et testés sur 500 nouvelles réalisations pour chacune des classes. Il apparaît que le taux d'erreur le plus faible coincide avec l'alignement maximum. Notons toutefois que ce dernier ne nécessite aucunement la mise en œuvre d'un coûteux algorithme d'apprentissage.

6.3.3 Elaboration d'une paramétrisation optimale à profil radialement Gaussien

En considérant la classe des distributions temps-fréquence à paramétrisation radialement Gaussienne, on cherche la largeur de bande optimale σ^* maximisant l'alignement noyau-cible selon l'expression

$$\sigma^* = \arg\max_{\sigma} \frac{\langle K_\sigma, K^* \rangle_F}{n \|K_\sigma\|_F}.$$

Ce problème d'optimisation est équivalent à un problème d'optimisation avec contrainte, correspondant à la maximisation du numérateur de l'expression de l'alignement, sous contrainte que son dénominateur soit constant. On peut alors résoudre le problème d'optimisation sous contrainte suivant :

$$\max_{\sigma} \sum_{i,j=1}^{n} y_i \, y_j \, \kappa_\sigma(x_i, x_j),$$

sous la contrainte

$$\sum_{i,j=1}^{n} (\kappa_\sigma(x_i, x_j))^2 = V_0, \tag{6.8}$$

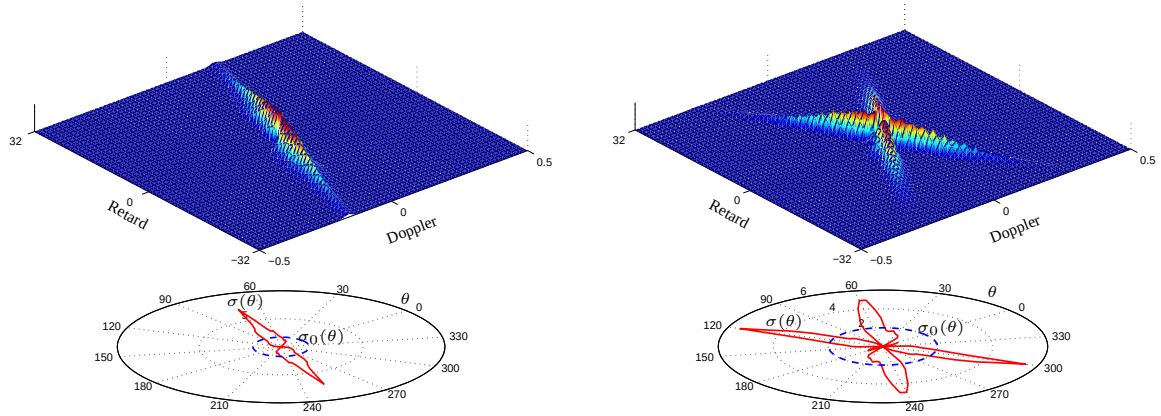


FIG. 6.6 – Résultats obtenus pour la 1ère application (a) et la 2ème application (b). En haut : fonction de paramétrisation optimale résultante. En bas : son contour en rouge, et le contour initial en bleu, en coordonnées polaires.

où V_0 est un paramètre de normalisation. En développant la fonction objectif à maximiser, on aboutit à

$$\begin{aligned}\sum_{i,j=1}^{n} y_i y_j\, \kappa_\sigma(x_i,x_j) &= \sum_{i,j=1}^{n} y_i y_j \iint r\; A_{x_i}(r,\theta)\, \overline{A_{x_j}}(r,\theta)\; \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}}\; \mathrm{d}r\; \mathrm{d}\theta \\ &= \iint r \Big[\sum_{i,j=1}^{n} y_i y_j\; A_{x_i}(r,\theta)\, \overline{A_{x_j}}(r,\theta) \Big]\; \mathrm{e}^{-\frac{r^2}{\sigma^2(\theta)}}\; \mathrm{d}r\; \mathrm{d}\theta\end{aligned}$$

On retrouve la fonction objectif à maximiser (2.13) proposée dans [BJ93], à savoir $\iint r|A_x(r,\theta)|^2\; \mathrm{e}^{-r^2/\sigma^2(\theta)}\; \mathrm{d}r\; \mathrm{d}\theta$, où la partie dépendante du signal, $|A_x(r,\theta)|^2$, est remplacée par la représentation équivalente $\sum_{i,j} y_i y_j\; A_{x_i}(r,\theta)\; \overline{A_{x_j}}(r,\theta)$ qui ne dépend que de l'ensemble d'apprentissage, signaux et étiquettes. Il est à noter que cette dernière peut être évaluée préalablement à toute optimisation. On peut alors avoir recours à l'algorithme d'optimisation de descente de gradient proposé dans [BJ93], avec la même complexité calculatoire une fois la représentation équivalente évaluée. Pour cela, on relâche la contrainte (6.8) coûteuse en temps de calcul en la remplaçant par la contrainte sur le volume de la fonction de paramétrisation, selon $\int \sigma^2(\theta)\, \mathrm{d}\theta = V_0'$ comme préconisé dans [BJ93]. Dans ce qui suit, on présente la mise en œuvre de l'algorithme proposé.

L'algorithme nécessite une discrétisation du plan Doppler-retard, que l'on opère comme proposé dans [BJ93]. Le noyau reproduisant est alors donné par

$$\kappa_\sigma(x_i,x_j) = \sum_{r,\theta} r\; A_{x_i}(r,\theta)\, \overline{A_{x_j}}(r,\theta)\; \mathrm{e}^{-(r\Delta_r)^2/\sigma^2(\theta)}.$$

pour la distribution temps-fréquence à noyau radialement Gaussien, avec $\Delta_r = 2\sqrt{\pi/l}$, l étant la largeur des signaux échantillonnés. En reprenant la fonction objectif dans (2.14), le problème d'optimisation s'écrit en coordonnées polaires selon

$$\max_{\sigma} \sum_{r,\theta} r \Big[\sum_{i,j=1}^{n} y_i y_j\; A_{x_i}(r,\theta)\, \overline{A_{x_j}}(r,\theta) \Big]\; \mathrm{e}^{-(r\Delta_r)^2/\sigma^2(\theta)}, \tag{6.9}$$

	1ère application		**2ème application**	
	Taux d'erreur (%)	Nombre de *SV*	Taux d'erreur (%)	Nombre de *SV*
Distribution de Wigner	19.10 ± 1.23	161.9 ± 4.97	19.41 ± 1.34	164.3 ± 5.62
Distribution optimale	16.89 ± 2.01	65.2 ± 4.46	17.81 ± 1.72	83.85 ± 5.65

TAB. 6.1 – Comparaison du taux d'erreur (%) et du nombre de vecteurs support (*SV*) pour un classifieur SVM associé d'une part à la distribution de Wigner, et d'autre part à la distribution optimale, pour chacune des deux applications.

sous la contrainte

$$\sum_{\theta} \sigma^2(\theta) = V_0'.$$

Pour résoudre ce problème d'optimisation avec contrainte, on considère l'algorithme de descente de gradient alternée suivant. A l'itération $k+1$, on opère dans un premier temps une mise à jour de la solution selon

$$\sigma_{k+1}(\theta) = \sigma_k(\theta) + \mu_k \frac{\partial f}{\partial \sigma_k(\theta)},$$

où μ_k est un paramètre contrôlant la vitesse de convergence, et f la fonction objectif à maximiser dans (6.9), et dont le gradient évalué en $\sigma_k(\theta)$ est défini par le vecteur $\left[\frac{\partial f}{\partial \sigma_k(0)} \ \ldots \ \frac{\partial f}{\partial \sigma_k(l-1)}\right]$, avec

$$\frac{\partial f}{\partial \sigma_k(\theta)} = \frac{2\Delta_r^2}{\sigma_k^3(\theta)} \sum_r r^3 \ \Psi(r,\theta) \ \mathrm{e}^{-(r\,\Delta_r)^2/\sigma^2(\theta)}.$$

Dans cette expression, la représentation équivalente donnée par l'expression

$$\Psi(r,\theta) = \sum_{i,j} y_i y_j A_{x_i}(r,\theta)\overline{A_{x_j}}(r,\theta), \tag{6.10}$$

est évaluée préalablement à la phase d'optimisation. Dans une seconde étape, on prend en compte la contrainte en projetant la solution sur l'ensemble des fonctions admissibles, ce qui revient à normaliser $\sigma_{k+1}(\theta)$ à chaque itération selon $\|\sigma_{k+1}(\theta)\|/V_0'$.

On insiste sur le fait que la représentation $\Psi(r,\theta)$ peut être calculée dans une étape d'initialisation. De plus, l'expression (6.10) se prête à un calcul itératif ne nécessitant pas de conserver en mémoire l'ensemble des fonctions d'ambiguïté des signaux de l'ensemble d'apprentissage. Une fois la représentation $\Psi(r,\theta)$ obtenue, la technique d'optimisation devient indépendante de la taille de l'ensemble d'apprentissage. Ceci n'est pas le cas pour les approches classiques, telle que celle proposée dans [DD98], qui nécessitent l'évaluation des représentations temps-fréquence de chaque signal de l'ensemble d'apprentissage, à chaque itération de l'algorithme d'optimisation. Pour cette raison certainement, les auteurs de cet article se restreignent à un ensemble d'apprentissage de 15 signaux pour chaque classe.

On considère successivement deux problèmes de classification de deux familles de 200 signaux de taille 64, à modulation fréquentielle linéaire, noyés dans un bruit blanc Gaussien de variance 4. Ceci correspond à un rapport signal-sur-bruit de l'ordre de -8 dB. La première application concerne des signaux à modulation fréquentielle croissante, de 0.1 à 0.25 pour la première classe, et de 0.25 à 0.4 pour la seconde, en échelle fréquentielle normalisée. La Figure 6.6 (a) en haut présente la fonction de paramétrisation à profil radialement Gaussien ainsi obtenue, ce qui montre la pertinence de cette

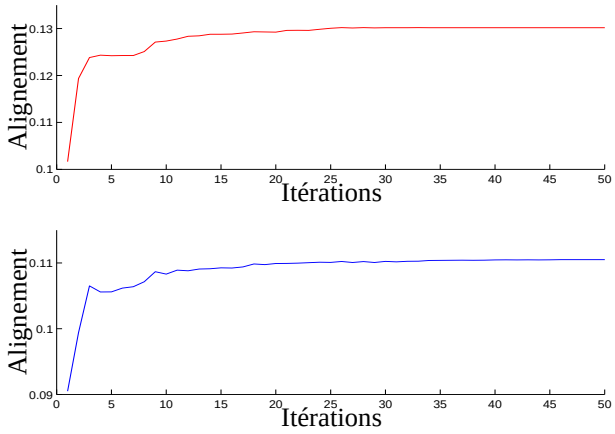


FIG. 6.7 – Evolution de l'alignement, moyenné sur 20 réalisations, pour la 1ère (en haut) et la 2ème (en bas) application.

région dans le plan Doppler-retard pour la classification. La Figure 6.6 (a) en bas illustre son contour en rouge $\sigma(\theta)$, ainsi que le contour initial $\sigma_0(\theta)$ en bleu avant optimisation. Ce dernier est déterminé par la contrainte de volume, que l'on a fixé à $V_0' = 2$. Dans une seconde application, on propose d'étudier le cas où les régions des signaux des deux classes sont distinctes dans le plan Doppler-retard. On considère pour cela des signaux comportant une modulation fréquentielle linéairement croissante pour la première classe, de 0.1 à 0.4, alors qu'elle décroît de 0.4 à 0.1 pour la seconde classe. Dans la Figure 6.6 (b), on représente la fonction de paramétrisation ainsi obtenue. Elle correspond à un filtrage de l'information pertinente des deux régions d'intérêt pour la classification. Pour illustrer la maximisation de l'alignement, on représente sur la Figure 6.7 l'évolution moyenne sur 20 réalisations de ce paramètre au cours des itérations, pour chacune des deux applications. Ceci met en évidence la convergence de l'algorithme vers une valeur maximale de l'alignement, malgré le remplacement de la contrainte $\|K_\sigma\| = V_0$ par $\int \sigma^2(\theta)\ \mathrm{d}\theta = V_0'$.

Afin d'illustrer la pertinence de cette stratégie, on propose d'estimer l'erreur de classification obtenue à partir d'un classifieur de type Support Vector Machines (SVM), associée à chacune des distributions temps-fréquence : la distribution de Wigner et la distribution optimale. Le Tableau 6.1 présente, en les moyennant sur 20 réalisations, le taux d'erreur obtenu sur un ensemble de test de $2\,000$ signaux, et le nombre de vecteurs support correspondant. Non seulement la distribution optimale minimise l'erreur de classification, mais elle conduit aussi à une division par deux environ du nombre de vecteurs support. Ceci est principalement dû, d'une part au caractère optimal de la distribution ainsi obtenue, et d'autre part au caractère régularisant de ce traitement.

6.3.4 Combinaison de représentations

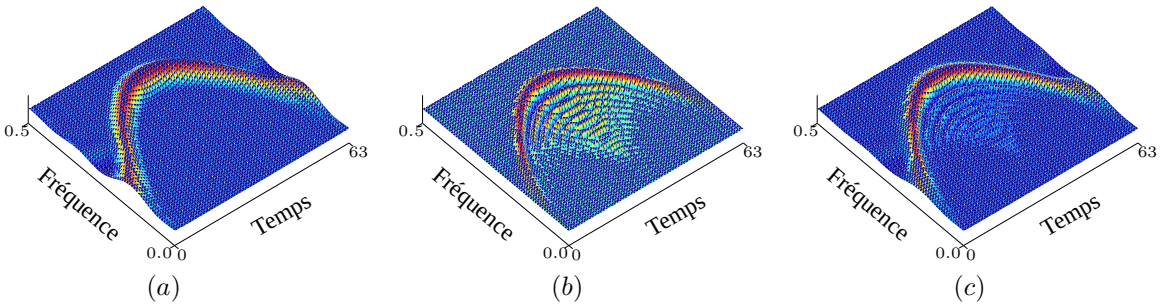
Le formalisme à noyau reproduisant adopté jusqu'ici pour les distributions temps-fréquence autorise la mise en œuvre des différents algorithmes de combinaison de noyau reproduisant, et en particulier l'optimisation de la combinaison linéaire de ceux-ci comme étudié au Paragraphe 6.2.4. On suggère ici de combiner les espaces de représentations temps-fréquence par ce moyen. L'illustration proposée concerne un problème de détection. L'hypothèse H_1 (signal + bruit) concerne le signal $\mathrm{e}^{2j\pi[\varphi(t)+\varphi_0]}$ noyé dans un bruit blanc gaussien de variance 4, où $\varphi(t)$ définit une loi de fréquence instantanée parabolique, et φ_0 est une phase initiale aléatoire uniformément distribuée sur $[0, 2\pi[$. L'hypothèse H_0 (bruit seul) concerne un bruit blanc gaussien de variance 9.

Dans un premier temps, nous avons repris l'ensemble des distributions candidates considérées dans la Section 6.3.1 dédiée à la sélection de distributions. L'algorithme proposé a conduit à la sélection du noyau κ_{spwv}, d'alignement 0.1039, puis du noyau κ_{wv} pour un alignement du noyau composite de 0.1076. La combinaison linéaire optimale de ces deux noyaux est alors donnée par les coefficients de pondération selon les expressions (6.4), ce qui défini le noyau composite optimal selon

$$\kappa_{\text{comp}} = \kappa_{\text{spwv}} + 0.208\kappa_{\text{wv}}.$$

Pour illustrer les qualités de ce noyau, on considère la mise en œuvre de l'algorithme des SVM avec les différents noyaux, κ_{spwv}, κ_{wv}, et le noyau composite κ_{comp}. En estimant le taux d'erreur *a posteriori*, on remarque qu'il est passé de 4.7% avec les noyaux simples à 3.2% avec le noyau composite. La Figure 6.8 présente la distribution temps-fréquence composite correspondante obtenue, appliquée ici au signal à détecter.

Une seconde expérimentation a été menée en ajoutant le noyau linéaire (3.7) de la transformée de Fourier à court-terme au groupe de noyaux quadratiques considérés jusqu'ici. L'algorithme a mené à la combinaison de celui-ci avec le noyau κ_{spwv}, pour un alignement final de 0.1698 et un taux d'erreur de 2.7%. On obtient donc un taux d'erreur plus faible qu'à la première expérimentation. La combinaison des ces deux noyaux, est conforme à la théorie puisque la statistique optimale au sens du rapport de vraisemblance pour le problème traité, où la variance du bruit est différente sous les 2 hypothèses, combine composantes linéaires et quadratiques.



(a) (b) (c)

FIG. 6.8 – Distribution pseudo-Wigner lissée (gauche), Wigner (centre) et composite selon la pondération (κ_{spwv} + 0.208 κ_{wv}) déterminée par les expressions (6.4). Ces distributions sont ici mises en œuvre sur le signal à détecter.

Deuxième partie

Apprentissage en-ligne et filtrage adaptatif non-linéaire

Chapitre 7

Contrôle de complexité et critère de cohérence

Sommaire

Les méthodes à noyau permettent la résolution de problèmes de reconnaissance des formes dans des espaces de dimension élevée, voire infinie. La dimensionnalité du problème, qui correspondant au nombre de paramètres à estimer, est alors réduite au nombre de données d'apprentissage disponibles. Si cette propriété est particulièrement appréciable tant que la taille de la base d'apprentissage reste dans les limites de la capacité du calculateur utilisé, elle s'avère rédhibitoire dans le cadre d'applications en-ligne. Il en est ainsi par exemple lorsqu'il s'agit d'identifier un système dynamique non-stationnaire à l'aide de méthodes adaptatives. Afin de combler cette lacune, il convient de contrôler l'ordre du modèle à noyau à chaque nouvelle arrivée de donnée, et adopter des méthodes de résolution itératives. Comme nous le verrons dans ce chapitre, l'usage de méthodes de représentation parcimonieuse est indiqué pour traiter le premier point ayant trait à l'élaboration d'un modèle à noyau le plus creux possible. On complète le dispositif à l'aide de méthodes d'apprentissage en-ligne visant à mettre les paramètres du modèle à jour.

Dans ce chapitre, on s'intéresse essentiellement à un critère de parcimonie, originale dans le cadre des méthodes à noyau : le critère de cohérence. Issu de la littérature sur l'approximation parcimonieuse de fonctions, ce critère nous offre un moyen de contrôle en-ligne de l'ordre du modèle avec une complexité calculatoire linéaire. Ceci nous permettra de développer des méthodes à noyau en-ligne à faible coût calculatoire dans le chapitre suivant, que l'on appliquera à la problématique du filtrage adaptatif non-linéaire.

7.1 Introduction

Les méthodes à noyau offrent des algorithme ne dépendant pas de la nature des données traitées, ni de l'espace de représentation adopté pour résoudre les problèmes. Au cours des chapitres précédents, nous avons profité de cette propriété pour proposer de nouveaux outils pour le traitement des signaux non-stationnaires dans un espace de représentation particulier : le plan temps-fréquence. Nous avons cependant noté que ces techniques, en vertu du Théorème de Représentation, engendrent des modèles dont l'ordre correspond au nombre de données disponibles. Cette caractéristique réduit à néant toute perspective de traitement de vastes ensembles de données, de façon batch ou en-ligne.

Sur la parcimonie dans le cadre des méthodes à noyau

De nombreux travaux se sont intéressés à la mise en œuvre des SVM sur des bases d'apprentissage de grande taille. Plusieurs méthodes séquentielles ont été développées en décomposant le problème d'optimisation en sous-problèmes, comme par exemple avec la technique *chunking* [Vap82], la décomposition [OFG97, OG99], l'approche *shrinking* ou sélection de sous-ensembles optimaux [Joa99]. L'actualisation d'un ou de deux coefficients du modèle à chaque itération est proposée respectivement par l'adatron à noyau basé sur un algorithme de descente de gradient stochastique [FCC98], et l'approche SMO ou optimisation par minimisation séquentielle [Pla99]. Toutes ces méthodes ont été présentées dans le cadre des SVM, dont la fonction coût favorise la parcimonie de la solution.

Dans le cadre plus général des méthodes à noyau, on peut imposer la parcimonie en introduisant un terme de pénalisation supplémentaire sur la fonctionnelle de risque régularisée. Afin d'aboutir à un modèle d'ordre réduit, il conviendrait d'inclure une pénalisation de type ℓ_0 sur le vecteur $\boldsymbol{\alpha}$, ce qui revient à minimiser le nombre de coefficients α_i non-nuls. Toutefois, la résolution d'un tel problème d'optimisation s'avère difficile, voire impossible, puisque la solution nécessite la résolution de C_n^m problèmes, où m désigne le nombre d'éléments non-nuls de $\boldsymbol{\alpha}$. Pour remédier à cet inconvénient, on lui substitue la norme ℓ_1, soit $\|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^{n} |\alpha_i|$. Cette pénalisation est proposée dans [CDS98] avec la technique *Basis Pursuit De-Noising*, et dont le lien avec les SVM est présenté dans [Gir98]. Dans [MRM00], les auteurs proposent de l'utiliser pour l'AFD-à-noyau afin d'obtenir un modèle d'ordre réduit. Des techniques d'élagage ont été également proposées, consistant à supprimer les fonctions noyau à faibles contributions dans le modèle. Dans le même esprit, Burges introduit dans [Bur96] une technique plus avancée pour construire un modèle de faible ordre, ou plus précisément une règle de décision simplifiée pour les SVM. Ces diverses approches nécessitent toutefois la résolution d'un problème d'optimisation impliquant la manipulation et l'inversion de matrices de la taille de la base d'apprentissage. Pour pallier ce défaut, différentes techniques de type glouton ont été suggérées pour l'approximation de la matrice de Gram par une matrice de rang inférieur, en considérant aléatoirement un sous-ensemble de la base d'apprentissage à partir duquel on estime cette matrice. Cette approche est mise en avant par les travaux de Smola *et coll.* dans [SS00] dans le cadre général des méthodes à noyau, et pour les processus Gaussiens dans [SB00], ainsi que dans [MSS01] pour son adaptation à l'AFD-à-noyau. D'autres travaux considèrent la méthode de Nyström [WS01], ou encore la factorisation incomplète de Cholesky [FS02]. Ces schémas ne permettent pas un apprentissage en-ligne où il s'agit de traiter un flux de données.

Différentes techniques d'apprentissage itératives ont été développées pour les méthodes à noyau les plus courantes, voir [SY06] et les références citées. Dans [KFS03, KFS05, GSV07] par exemple, les auteurs proposent un algorithme itératif pour l'ACP-à-noyau, en se basant sur l'apprentissage Hebbien dans un espace RKHS. Toutefois, l'algorithme résultant n'est pas un algorithme en-ligne puisqu'il repose sur la connaissance de la totalité de l'ensemble d'apprentissage, qui doit être de taille finie. Cette faiblesse a été abordée par les auteurs, avec une approche dite semi-en-ligne destinée à y remédier. Dif-

férents algorithmes itératifs ont également été proposés pour la classification. En particulier, un schéma itératif pour l'AFD-à-noyau a été développé dans [FDBR04]. Dans [MM01] les auteurs présentent différents schémas itératifs sous le nom de *Lagrangian Support Vector Machines* en modifiant l'algorithme usuel des SVM. D'autres algorithmes d'apprentissage séquentiel pour les SVM ont été aussi proposés dans [KSW04, VSS06], où les auteurs s'appuie sur la parcimonie induite par la fonction coût. Ce type d'approche est toutefois spécifique aux SVM et ne peut profiter aux méthodes à noyau ne reposant pas sur ce même critère.

Problèmes des moindres carrés dans un RKHS

On s'intéresse à la fonctionnelle de coût quadratique telle qu'elle est utilisée pour l'ACP-à-noyau, l'AFD-à-noyau, certains schémas de classification voisins des SVM, ou encore en régression. Il s'agit là de rechercher une fonction $\psi^*(\cdot)$ minimisant l'erreur quadratique entre la sortie du modèle $\psi(\boldsymbol{x}_i)$ et la réponse désirée d_i assortie d'un terme de régularisation

$$\frac{1}{n}\sum_{i=1}^{n}(d_i - \psi(\boldsymbol{x}_i))^2 + \eta \left\|\psi\right\|^2_{\mathcal{H}}. \tag{7.1}$$

Pour une analyse non-supervisée, on considérerait la même réponse désirée pour tous les $\boldsymbol{x}_i$ de l'ensemble d'apprentissage, soit $d_i = 0$. Le premier terme de l'expression (7.1) correspondrait alors à la variance de $\psi(\boldsymbol{x}_i)$, dont la maximisation mènerait au premier axe principal d'inertie. Ceci constitue un résultat équivalent, à une normalisation près, à celui fournit par une ACP-à-noyau telle qu'elle a été présentée au Chapitre 4. Dans le cadre d'un apprentissage supervisé, on a montré à la Section 5.2.4 le lien entre l'AFD-à-noyau et la minimisation de (7.1) lorsque les réponses désirées sont définies ainsi :

$$d_i = \begin{cases} +n/n_+ \text{ , si } \boldsymbol{x}_i \in \mathcal{X}_+; \\ -n/n_- \text{ , si } \boldsymbol{x}_i \in \mathcal{X}_-, \end{cases} \tag{7.2}$$

où $\mathcal{X}_+$ et $\mathcal{X}_-$ désignent respectivement l'ensemble des données d'apprentissage de chacune des deux classes, et n_+ et n_- leur taille. Ce lien est également traité dans [Mik02], ou encore dans [XZL01] où les auteurs s'intéressent de plus à la régression ridge à noyau. Les LS-SVM pour *least squares support vector machines* [SV99, SGB$^+$02], les réseaux régularisés [EPP99, RYP03] ou encore les *proximal SVM* [FM01] reposent également sur la minimisation de la fonctionnelle de risque (7.1), avec

$$d_i = \begin{cases} +1 \text{ , si } \boldsymbol{x}_i \in \mathcal{X}_+; \\ -1 \text{ , si } \boldsymbol{x}_i \in \mathcal{X}_-. \end{cases}$$

pour la résolution de problèmes de classification. Pour les problèmes d'estimation fonctionnelle, on considère une sortie désirée $d_i \in \mathbb{R}$ et l'on cherche le modèle ψ^* apte à reproduire celle-ci à partir de $\boldsymbol{x}_i$. Ce cas de figure englobe les cas précédents, puisqu'il regroupe les cas particuliers $d_i = 0$ pour l'ACP-à-noyau, $d_i = +n/n_+$ ou $d_i = -n/n_-$ pour l'AFD-à-noyau, et $d_i = \pm 1$ pour les LS-SVM.

Afin de simplifier la présentation, on adopte dans le reste de ce manuscrit le schéma de la régression avec $d_i \in \mathbb{R}$. Aussi la fonctionnelle de coût quadratique constituera-t-elle pour nous un terrain commun d'investigation pour l'analyse et la classification de données. Le prix à payer réside dans la non-parcimonie de la solution, avec une représentation finale dépendant de tout l'ensemble d'apprentissage contrairement aux SVM. Dans la section suivante, on s'intéresse à la résolution du problème d'apprentissage. Puis on propose de nouvelles techniques de contrôle de l'ordre du modèle, permettant son utilisation dans le cadre d'applications en-ligne.

Contrôle de complexité de modèle dans un RKHS

Soient $\mathcal{X}$ un compact de $\mathbb{R}^p$, $\kappa\colon \mathcal{X}\times\mathcal{X}\to\mathbb{R}$ un noyau reproduisant associé à l'espace de Hilbert $\mathcal{H}$ de produit scalaire $\langle\cdot,\cdot\rangle_{\mathcal{H}}$). Par la propriété reproduisante, toute fonction $\psi(\cdot)$ de $\mathcal{H}$ peut être évaluée en tout $\boldsymbol{x}_i\in\mathcal{X}$ selon $\psi(\boldsymbol{x}_i)=\langle\psi(\cdot),\kappa(\boldsymbol{x}_i,\cdot)\rangle_{\mathcal{H}}$, la fonction noyau $\kappa(\boldsymbol{x}_i,\cdot)$ étant telle que pour tout $\boldsymbol{x}_k\in\mathcal{X}$ on a $\kappa(\boldsymbol{x}_k,\boldsymbol{x}_i)$.

On s'intéresse à la recherche d'une fonction $\psi(\cdot)$ minimisant l'erreur quadratique, entre la sortie du modèle $\psi(\boldsymbol{x}_i)$ et la réponse désirée d_i, soit

$$\frac{1}{n}\sum_{i=1}^{n}(d_i-\psi(\boldsymbol{x}_i))^2+\eta\,\|\psi\|_{\mathcal{H}}^2, \tag{7.3}$$

où η désigne le paramètre de régularisation. Comme le Théorème de Représentation le précise, la fonction optimale ψ^* recherchée appartient à l'espace engendré par les fonctions noyau des données d'apprentissage, soit

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

En injectant ce développement dans (7.3), on obtient la forme duale du problème d'optimisation, avec la recherche du vecteur colonne des coefficients $\boldsymbol{\alpha}^*=[\alpha_1\cdots\alpha_n]^\top$ tel que

$$\boldsymbol{\alpha}^*=\arg\min_{\boldsymbol{\alpha}}\|\boldsymbol{d}-\boldsymbol{K}\boldsymbol{\alpha}\|^2+\eta\,\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha}, \tag{7.4}$$

où $\boldsymbol{K}$ désigne la matrice de Gram dont le $(i,j)^{\text{ème}}$ élément est $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$, et $\boldsymbol{d}=[d_1\cdots d_n]^\top$ le vecteur colonne des réponses désirées. La solution est alors donnée par la résolution du système linéaire à n équations et n inconnues suivant :

$$(\boldsymbol{K}+\eta\boldsymbol{I}_n)\,\boldsymbol{\alpha}=\boldsymbol{d},$$

où $\boldsymbol{I}_n$ est la matrice identité de taille $n\times n$. Les coefficients du modèle sont alors déterminés par l'inversion de la matrice $\boldsymbol{K}+\eta\boldsymbol{I}_n$ de taille $n\times n$.

Le modèle optimal $\psi^*(\cdot)=\sum_{j=1}^{n}\alpha_j\,\kappa(\boldsymbol{x}_j,\cdot)$ n'est pas adapté au contexte des applications en-ligne pour plusieurs raisons. D'une part, la matrice de Gram nécessite $\mathcal{O}(n^2)$ espace mémoire, et la résolution du système linéaire admet une complexité calculatoire en $\mathcal{O}(n^3)$. D'autre part, l'ordre du modèle correspond au nombre de couples de données disponibles, toute prédiction à partir de ce modèle nécessitant $\mathcal{O}(n)$ opérations. Pour remédier à ce problème, on considère un modèle réduit d'ordre m consistant en une combinaison linéaire d'un nombre plus restreint de fonctions noyau que celles disponibles. Le modèle réduit est de la forme

$$\psi_n^*(\cdot)=\sum_{j=1}^{m}\alpha_j\,\kappa(\boldsymbol{x}_{\omega_j},\cdot), \tag{7.5}$$

où les indices $\omega_1,\ldots,\omega_m$ appartiennent à l'ensemble $\{1,\ldots,n\}$ et sont ordonnés par ordre croissant. L'ensemble de ces m fonctions noyau est désigné par $\mathcal{D}_m=\{\kappa(\boldsymbol{x}_{\omega_1},\cdot),\ldots,\kappa(\boldsymbol{x}_{\omega_m},\cdot)\}$, que l'on nomme dictionnaire. La stratégie de contrôle de l'ordre du modèle consiste à sélectionner les m fonctions noyau les plus pertinentes parmi les n disponibles. La solution obtenue est alors sous-optimale puisqu'elle appartient au sous-espace engendré par les m fonctions noyau.

Dans le cadre d'un apprentissage en-ligne, l'arrivée d'une nouvelle observation $\boldsymbol{x}_n$ à l'instant n pose la question de l'ajout ou non de la fonction noyau candidate $\kappa(\boldsymbol{x}_n,\cdot)$ au dictionnaire. Pour cela, on a recours à un critère dit de parcimonie pour déterminer s'il convient de mettre à jour le dictionnaire ou non. On est alors confronté à l'une des deux options suivantes :

- Si $\kappa(\boldsymbol{x}_n, \cdot)$ ne satisfait pas la règle de parcimonie, le dictionnaire n'est pas modifié.
- Si $\kappa(\boldsymbol{x}_n, \cdot)$ vérifie la règle de parcimonie, on modifie le dictionnaire en l'y ajoutant.

Pour la modification du dictionnaire, deux approches sont envisageables. Certaines techniques ont recours à une approche de substitution afin que la taille du dictionnaire reste constante, fixée préalablement à l'étape d'apprentissage. Pour cela, l'ajout d'un élément au dictionnaire s'accompagne du retrait d'un autre. D'autres techniques ne procèdent à aucune suppression. Il en résulte un dictionnaire de taille indéterminée à l'avance. La suite de la procédure d'apprentissage en-ligne vise à mettre les coefficients du modèle à jour, en prenant en compte l'éventuel ajout du nouvel élément dans le dictionnaire qui se traduirait par l'accroissement de l'ordre du modèle. La suite de ce chapitre est consacrée à la première étape, avec la définition d'un critère de parcimonie et l'étude de ses propriétés.

Critère de dépendance linéaire

Différentes approches ont été proposées pour sélectionner les fonctions noyau les plus significatives en vue de constituer un dictionnaire. Parmi celles-ci, on retrouve la technique de sélection de sous-espace par l'ACP-à-noyau, le critère de maximisation d'entropie, ainsi que le critère de dépendance linéaire, voir [Hoe05]. Nous allons à présent nous concentrer sur ce dernier.

A l'instant n, la nouvelle fonction noyau $\kappa(\boldsymbol{x}_n, \cdot)$ est ajoutée dans $\mathcal{D}_m$ si la règle d'approximation linéaire est satisfaite

$$\min_{\boldsymbol{\gamma}} \left\| \kappa(\boldsymbol{x}_n, \cdot) + \sum_{j=1}^{m} \gamma_j \, \kappa(\boldsymbol{x}_{\omega_j}, \cdot) \right\|_{\mathcal{H}}^2 > \eta_0^2, \tag{7.6}$$

où η_0 est un seuil donné déterminant le niveau de parcimonie du modèle. Dans cette expression, le noyau reproduisant est supposé de norme unité, soit $\kappa(\boldsymbol{x}_k, \boldsymbol{x}_k) = 1$ pour tout $\boldsymbol{x}_k \in \mathcal{X}$; dans le cas contraire, on remplace $\kappa(\boldsymbol{x}_k, \cdot)$ par $\kappa(\boldsymbol{x}_k, \cdot)/\sqrt{\kappa(\boldsymbol{x}_k, \boldsymbol{x}_k)}$ dans l'expression ci-dessus. Ce critère a été initialement proposé dans [BA01] pour les schémas classiques de classification et régression par méthodes à noyau, avant d'être adapté à l'approximation en-ligne. Depuis, il a fait l'objet de plusieurs travaux sur l'apprentissage en-ligne avec Csató *et coll.* [CO01] dans le cadre des processus Gaussiens, Dodd *et coll.* [DKH03] avec un schéma séquentiel de projection, ou encore Engel *et coll.* [EMM04].

Ce critère admet une interprétation géométrique simple. En effet, la règle (7.6) consiste à évaluer la distance de la fonction noyau candidate à l'espace engendré par les éléments de $\mathcal{D}_m$, et à la comparer à un seuil donné. On peut écrire alors cette règle selon $\|(I - P_{\mathcal{D}_m})\kappa(\boldsymbol{x}_n, \cdot)\|_{\mathcal{H}}^2 > \eta_0^2$, où $P_{\mathcal{D}_m}$ désigne l'opérateur de projection sur l'espace engendré par les fonctions noyau de $\mathcal{D}_m$. En exprimant cette règle sous forme de produit scalaire, comme on l'a montré à la Section 1.3.1, on obtient

$$\kappa(\boldsymbol{x}_n, \boldsymbol{x}_n) - \boldsymbol{\kappa}(\boldsymbol{x}_n)^\top \boldsymbol{\gamma} > \eta_0^2,$$

où $\boldsymbol{\kappa}(\boldsymbol{x}_n)$ est le vecteur colonne de termes $\kappa(\boldsymbol{x}_n, \boldsymbol{x}_{\omega_j})$, pour $j = 1, \ldots, m$, et le vecteur colonne des coefficients optimaux $\boldsymbol{\gamma}$ est donné par

$$\boldsymbol{\gamma} = \boldsymbol{K}_n^{-1} \boldsymbol{\kappa}(\boldsymbol{x}_n),$$

avec $\boldsymbol{K}_n$ la matrice de Gram du dictionnaire à l'instant n, dont le $(i,j)^{\text{ème}}$ élément correspond à $\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$. L'évaluation de la condition d'indépendance linéaire (7.6) est alors donnée par

$$\kappa(\boldsymbol{x}_n, \boldsymbol{x}_n) - \boldsymbol{\kappa}(\boldsymbol{x}_n)^\top \boldsymbol{K}_n^{-1} \boldsymbol{\kappa}(\boldsymbol{x}_n) > \eta_0^2. \tag{7.7}$$

Bien qu'une telle approche produit un modèle réduit avec une faible erreur d'approximation, l'évaluation du critère nécessite l'inversion de la matrice de Gram du dictionnaire. La complexité calculatoire est d'ordre $\mathcal{O}(m^3)$, que l'on peut réduire à $\mathcal{O}(m^2)$ si l'inversion est effectuée récursivement lorsqu'un

élément est ajouté au dictionnaire. Dans ce chapitre, on introduit un nouveau critère de parcimonie avec une complexité linéaire par rapport à la taille du dictionnaire. Il repose sur un paramètre proposé pour la caractérisation de dictionnaires dans le cadre de l'approximation parcimonieuse de fonctions : la cohérence.

7.2 Cohérence d'un dictionnaire de fonctions noyau

La cohérence est une quantité fondamentale considérée pour la caractérisation de dictionnaires dans le cadre de techniques d'approximation parcimonieuse. Elle désigne la plus grande corrélation entre les éléments d'un dictionnaire, ou mutuellement entre les éléments de deux dictionnaires. Ce concept a été initialement proposé au début des années 1990 par Mallat et Zhang dans le contexte du *matching pursuit* [MZ93]. Il a pris son essor au début des années $2\,000$ avec les premières études formelles par Donoho *et coll.* dans [DH01] pour l'union de deux bases orthonormées, puis étendue à des dictionnaires arbitraires pour la technique *basis pursuit* dans [DE03]. La qualité de la représentation d'un signal à l'aide d'un dictionnaire de cohérence donnée a été étudiée dans [GMS03, GMST03]. Plus récemment, Gribonval *et coll.* ont proposé dans [GV06] l'usage du paramètre de cohérence pour le traitement de signal en transmissions multi-canaux et pour les problèmes de séparation de source. On adapte ce paramètre à un dictionnaire de fonctions noyau [HRB07b, RBH07].

Soit $\mathcal{D}_m$ l'ensemble de m fonctions noyau de norme unité, $\kappa(\boldsymbol{x}_{\omega_1}, \cdot), \kappa(\boldsymbol{x}_{\omega_2}, \cdot), \ldots, \kappa(\boldsymbol{x}_{\omega_m}, \cdot)$. La cohérence du dictionnaire est définie par

$$\mu = \max_{i \neq j} |\langle \kappa(\boldsymbol{x}_{\omega_i}, \cdot), \kappa(\boldsymbol{x}_{\omega_j}, \cdot) \rangle_{\mathcal{H}}| = \max_{i \neq j} |\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})|,$$

pour tout $i, j = 1, \ldots, m$, ce qui correspond au plus grand terme, en valeur absolue, non-diagonal de la matrice de Gram du dictionnaire. La cohérence correspond donc à la plus grande corrélation entre deux éléments du dictionnaire et, par conséquent, est nulle pour toute base orthonormale. En revanche, elle vaut 1 quand le dictionnaire contient au moins deux éléments identiques. On dit que le dictionnaire est incohérent quand μ est faible. Le paramètre de cohérence admet comme interprétation géométrique d'être le plus faible angle entre les fonctions noyau dans $\mathcal{H}$. Toutefois, ceci conduit à une contrainte, puisque l'on peut pas avoir d'angles arbitrairement grands entre les m fonctions noyau dans un espace de dimension d_0, comme le montre l'inégalité $\mu^2 \geq \frac{m-d_0}{d_0(m-1)}$ connue dans la littérature sous le nom de borne inférieure de Welch.

Ne dépendant que des deux éléments les plus corrélés du dictionnaire, le paramètre de cohérence souffre d'un manque de description plus approfondie. Pour cette raison, Tropp introduit dans [TGMS03] la fonction de Babel d'un dictionnaire, aussi connue sous le nom de cohérence cumulée [Tro04]. La fonction de Babel correspond à la maximum corrélation cumulée entre un élément et un sous-ensemble des autres éléments du dictionnaire. Pour le dictionnaire $\mathcal{D}_m$, elle est définie pour $k = 1, \ldots, m-1$ par

$$\mu_1(k) = \max_i \max_{J_k} \sum_{j \in J_k} |\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})|,$$

où J_k est un ensemble de k indices, excluant le $i^{\text{ème}}$ terme. On peut facilement vérifier que $\mu_1(1) = \mu$, la cohérence du dictionnaire.

L'indice 1 dans $\mu_1(\cdot)$ renvoie vers la norme matricielle ℓ_1. Cette dernière correspond à la plus grande somme, en valeur absolue, des éléments de chaque colonne. Pour la matrice de Gram $\boldsymbol{K}$, elle s'exprime sous la forme $\|\boldsymbol{K}\|_1 = \max_{\|x\|_1=1} \|\boldsymbol{K}x\|_1 = \max_k \sum_j |\kappa(\boldsymbol{x}_{\omega_j}, \boldsymbol{x}_{\omega_k})|$. Par cette analogie, le paramètre

de cohérence correspond alors à la norme matricielle infinie. Cette analogie ouvre la voie à d'autres mesures de corrélation entre les éléments d'un dictionnaire. On définit[9] alors la fonction $\mu_p(\cdot)$ par analogie avec la norme ℓ_p avec

$$\mu_p(k) = \max_i \max_{J_k} \Big(\sum_{j \in J_k} |\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})|^p \Big)^{1/p},$$

pour $k = 1, \ldots, m-1$ et J_k un ensemble de k indices, excluant le $i^{\text{ème}}$ terme. Une telle généralisation n'est pas soutenue par une interprétation géométrique, comme c'est le cas avec le paramètre de cohérence. Cependant, un résultat intéressant est obtenu à la Proposition 7.4 en considérant $\mu_2(\cdot)$, définie par l'expression

$$\mu_2(k) = \max_i \max_{J_k} \sqrt{\sum_{j \in J_k} |\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})|^2}.$$

Le paramètre de cohérence et ses extensions, en particulier la fonction de Babel, vérifient certaines relations d'inégalités. Les relations suivantes sont utilisées tout au long de ce chapitre. Soit $\mathcal{D}_m$ un dictionnaire de cohérence μ, et soit $\mu_1(\cdot)$ sa fonction de Babel. Les propriétés suivantes sont satisfaites pour tout $k = 1, \ldots, m-1$:

Propriété 1 : $\mu_1(k) \leq k\,\mu$

Cette propriété résulte des définitions de la cohérence et de la fonction de Babel.

Propriété 2 : $\mu_1(\cdot)$ est monotone croissante

On peut facilement vérifier que $\mu_1(k) \geq \mu_1(k-1)$ pour tout $k = 2, \ldots, m-1$. C'est pour cette propriété que la fonction de Babel est souvent appelée la cohérence cumulée.

Propriété 3 : $\mu_2^2(k) \leq k\,\mu^2$

Cette propriété résulte de l'expression $\sum_{j \in J_k} |\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})|^2 \leq \sum_{j \in J_k} \mu^2 = k\mu^2$.

Propriété 4 : $\mu_2(k) < \mu_1(k)$

Pour montrer cela, on écrit

$$\Big[\sum_{j \in J_k} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)| \Big]^2 = \sum_{j \in J_k} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)|^2 + \sum_{j \neq j'} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)||\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{j'})| > \sum_{j \in J_k} |\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)|^2,$$

et en les combinant à leur définition, on obtient l'inégalité recherchée.

Propriété 5 : $\mu_2(k) < \mu_1(k) \leq k\,\mu$

Cette propriété découle directement des deux Propriétés 1 et 4 ci-dessus. Toutefois, on précise que l'inégalité $\mu_2(k) < k\,\mu$ est faible, comparée à la borne plus restrictive $\mu_2^2(k) < k\,\mu^2$, ce qui produit $\mu_2(k) \leq \sqrt{k}\,\mu$.

7.2.1 Méthodes à noyau avec un dictionnaire de cohérence μ

On considère un problème d'apprentissage avec un ensemble d'apprentissage $\mathcal{A}_m$ induisant un dictionnaire $\mathcal{D}_m$ de cohérence μ et de fonction de Babel $\mu_1(\cdot)$, soit

$$\mathcal{A}_m = \{(\boldsymbol{x}_{\omega_1}, d_{\omega_1}), (\boldsymbol{x}_{\omega_2}, d_{\omega_2}), \ldots, (\boldsymbol{x}_{\omega_n}, d_{\omega_n})\}$$

$$\mathcal{D}_m = \{\kappa(\boldsymbol{x}_{\omega_1}, \cdot), \kappa(\boldsymbol{x}_{\omega_2}, \cdot), \ldots, \kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}.$$

En désignant par $\boldsymbol{K}$ une matrice de Gram associée à ce dictionnaire, on élabore des bornes supérieures et inférieures sur les valeurs propres de cette matrice, à partir de la fonction de Babel du dictionnaire. La

[9] Récemment, Gribonval *et coll.* ont proposé dans [GRSV07] une approche similaire dite *p-thresholding*.

proposition suivante ainsi que sa démonstration sont due principalement à Tropp [Tro04], et établies ici dans un cadre plus large de fonctions noyau. Ce résultat fournit de nouvelles perspectives dans le cadre général des méthodes à noyau.

Proposition 7.1. *La plus petite et la plus grande valeurs propres de la matrice de Gram d'un dictionnaire de m fonctions noyau de fonction de Babel $\mu_1(\cdot)$, désignées respectivement par $\lambda_{\min}$ et $\lambda_{\max}$, vérifient les bornes suivantes*

$$\begin{aligned} 1 - \mu_1(m-1) \leq \quad & \lambda_{\min} \quad \leq 1 \\ 1 \leq \quad & \lambda_{\max} \quad \leq 1 + \mu_1(m-1) \end{aligned}$$

De plus, puisque $\mu_1(m-1) \leq (m-1)\,\mu$, ses inégalités fournissent d'autres bornes en fonction du paramètre de cohérence du dictionnaire, soit $1 - (m-1)\,\mu \leq \lambda_{\min} \leq 1 \leq \lambda_{\max} \leq 1 + (m-1)\,\mu$.

Démonstration. La démonstration se déroule en deux temps. D'abord, on montre que $\lambda_{\min} \leq 1 \leq \lambda_{\max}$. Pour cela, soit $\lambda_1, \ldots, \lambda_m$ les valeurs propres de la matrice de Gram $\boldsymbol{K}$, ce qui permet d'écrire $\sum_{k=1}^{m} \lambda_k = \mathrm{Trace}(\boldsymbol{K})$, où $\mathrm{Trace}(\cdot)$ désigne la trace de la matrice. D'une part, on a $\mathrm{Trace}(\boldsymbol{K}) = \sum_{k=1}^{m} \kappa(\boldsymbol{x}_{\omega_k}, \boldsymbol{x}_{\omega_k}) = m$ pour les fonctions noyau de norme unité. D'autre part, on a par définition $\lambda_{\min} \leq \lambda_k \leq \lambda_{\max}$, et par conséquent les inégalités $m\lambda_{\min} \leq \sum_{k=1}^{m} \lambda_k \leq m\lambda_{\max}$. En combinant les deux expressions, on obtient le résultat $\lambda_{\min} \leq 1 \leq \lambda_{\max}$. Ensuite, selon le théorème des disques de Geršgorin [HJ86], les valeurs propres de la matrice $\boldsymbol{K}$ appartiennent à l'union des m disques, centrés sur les éléments diagonaux de $\boldsymbol{K}$ et de rayons donnés par la somme de la valeur absolue des $m-1$ autres éléments de la $k^{\text{ème}}$ colonne, $\sum_{i \neq k}^{m} |\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_k})|$, pour $k = 1, \ldots, m$. En d'autres termes, chacune des m valeurs propres vérifie l'expression $|\lambda_k - \kappa(\boldsymbol{x}_{\omega_k}, \boldsymbol{x}_{\omega_k})| \leq \sum_{i \neq k}^{m} |\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_k})|$. En considérant des fonctions noyau de norme unité, et de fonction de Babel $\mu_1(\cdot)$, on obtient l'expression $|\lambda - 1| \leq \mu_1(m-1)$. Pour conclure, les résultats sont obtenus directement, puisque $\lambda_{\min} \leq 1$ et $\lambda_{\max} \geq 1$, par conséquent $1 - \lambda_{\min} \leq \mu_1(m-1)$ et $\lambda_{\max} - 1 \leq \mu_1(m-1)$. ■

Les bornes proposées par cette proposition permettent une étude approfondie sur la stabilité de la solution en fonction de l'ensemble d'apprentissage, à partir de sa cohérence pour un noyau donné. En particulier, pour la fonction coût quadratique, le problème d'optimisation dual s'exprime sous la forme $\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}} \|\boldsymbol{d} - \boldsymbol{K}\boldsymbol{\alpha}\|^2 + \eta\, \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}$, et la solution est donnée par la résolution du système de n équations à n inconnues $(\boldsymbol{K} + \eta \boldsymbol{I}_n)\, \boldsymbol{\alpha}^* = \boldsymbol{d}$. Bien que l'existence et l'unicité de la solution $\boldsymbol{\alpha}^*$ sont garanties par le terme de régularisation, on ne dispose d'aucune indication quant à sa stabilité par rapport à de faibles perturbations des données. Le conditionnement permet une telle quantification. Dans ce qui suit, on rappelle le principe fondamental de ce concept, avant de présenter le résultat liant conditionnement et paramètre de cohérence.

Le conditionnement d'une matrice carrée $\boldsymbol{K}$ relatif à une norme matricielle $\|\cdot\|$ est défini par la grandeur réelle $\mathrm{cond}(\boldsymbol{K}) = \|\boldsymbol{K}\|\,\|\boldsymbol{K}^{-1}\|$, et par convention $\mathrm{cond}(\cdot) = \infty$ pour les matrices singulières. En considérant la norme 2, on peut montrer que cette grandeur est donnée par

$$\mathrm{cond}(\boldsymbol{K}) = \left| \frac{\lambda_{\max}}{\lambda_{\min}} \right|,$$

où $\lambda_{\min}$ et $\lambda_{\max}$ désignent la plus petite et la plus grande valeur propre de la matrice $\boldsymbol{K}$. Le conditionnement quantifie la sensibilité du problème $\boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{d}$, pour des variations de la matrice $\boldsymbol{K}$ ou du second membre, le vecteur $\boldsymbol{d}$. La solution $\boldsymbol{\alpha}$ est robuste vis-à-vis de perturbations quand le conditionnement de la matrice $\boldsymbol{K}$ est faible, proche de 1. Dans le cas contraire, les matrices à grand conditionnement produisent des problèmes dits mal-conditionnés, voire mal-posés. Le concept de conditionnement d'une matrice est

très puissant et permet l'étude des performances des algorithmes de résolution de problèmes linéaires. A titre indicatif, on considère une procédure de descente de gradient pour résoudre $\boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{d}$. Luenberger a présenté dans [Lue89] une borne supérieure sur la réduction de l'erreur à chaque itération. Cette borne dépend uniquement du conditionnement de la matrice $\boldsymbol{K}$, et lui est proportionnelle. En d'autres termes, plus le conditionnement est élevé, et moins la convergence est rapide. De retour aux méthodes à noyau, le concept de conditionnement du problème est étudié plus en détails dans [KS05]. La proposition suivante présente une borne supérieure sur le conditionnement de la matrice de Gram, en fonction de la cohérence du dictionnaire.

Proposition 7.2. *Le conditionnement d'une matrice de Gram d'un dictionnaire de cohérence μ ne peut être supérieure à $\frac{1+(m-1)\,\mu}{1-(m-1)\,\mu}$.*

Démonstration. Pour démontrer cette proposition, il suffit de reprendre la définition du conditionnement, et d'appliquer les bornes sur les valeurs propres de la matrice de Gram, comme explicitées à la Proposition 7.1. ■

En réécrivant cette borne selon $\text{cond}(\boldsymbol{K}) \leq \frac{2}{1-(m-1)\,\mu} - 1$, on obtient de nouveau $\text{cond}(\boldsymbol{K}) = 1$ pour un dictionnaire orthonormal, soit encore $(m-1)\,\mu = 0$, alors qu'elle augmente avec $(m-1)\,\mu$. De plus, on montre que la borne supérieure explicitée à la Proposition 7.2 est aussi valide pour un problème régularisé selon Tikhonov, avec $(\boldsymbol{K} + \eta \boldsymbol{I}_n)\,\boldsymbol{\alpha} = \boldsymbol{d}$. En désignant par λ une valeur propre de $\boldsymbol{K}$, $\lambda + \eta$ est la valeur propre de la matrice $\boldsymbol{K} + \eta \boldsymbol{I}_n$. La définition du conditionnement permet alors d'écrire

$$\text{cond}(\boldsymbol{K} + \eta \boldsymbol{I}_n) = \frac{\lambda_{\max} + \eta}{\lambda_{\min} + \eta} \leq \frac{1 + (m-1)\,\mu + \eta}{1 - (m-1)\,\mu + \eta} \leq \frac{1 + (m-1)\,\mu}{1 - (m-1)\,\mu}.$$

7.2.2 Dépendance linéaire et cohérence

Il est intéressant d'avoir un dictionnaire d'éléments linéairement indépendants, afin que toute fonction de l'espace engendré par ces éléments puisse être représentée de manière unique par une combinaison linéaire de ceux-ci. Pour avoir un dictionnaire d'éléments linéairement indépendants, on présente une condition suffisante sur sa cohérence. Cette condition de dépendance linéaire est principalement due à Gilbert *at al.* [GMS03], et son extension aux dictionnaires de fonctions noyau est proposée par Richard *et coll.* dans [RBH07]. Sa démonstration repose sur la dualité entre indépendence linéaire et non-singularité de la matrice de Gram.

Proposition 7.3. *Soit $\mathcal{D}_m$ un dictionnaire de m fonctions noyau avec une cohérence μ et une fonction de Babel $\mu_1(\cdot)$. Cet ensemble est linéairement indépendant si on a $\mu_1(m-1) < 1$.*

Démonstration. Les m fonctions noyau sont linéairement indépendantes si toute combinaison linéaire de celles-ci, soit $\sum_{i=1}^{m} \alpha_i\, \kappa(\boldsymbol{x}_{\omega_i}, \cdot)$, est nulle si et seulement si ses coefficients α_i sont nuls. Or on a

$$\| \sum_{i=1}^{m} \alpha_i\, \kappa(\boldsymbol{x}_{\omega_i}, \cdot) \|_{\mathcal{H}}^2 = \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha} \geq \lambda_{\min} \|\boldsymbol{\alpha}\|^2 \geq (1 - \mu_1(m-1)) \sum_{i=1}^{m} \alpha_i^2,$$

où la dernière inégalité est due à la Proposition 7.1. A partir de celle-ci, on obtient une condition suffisante pour avoir un ensemble linéairement indépendant avec $1 - \mu_1(m-1) > 0$, soit encore à partir de la Propriété 1 à la Section 7.2 si on a $(m-1)\,\mu < 1$. ■

Cette proposition montre la puissance du paramètre de cohérence pour l'analyse de dictionnaires arbitraires. De manière plus approfondie, on élabore une borne inférieure sur l'erreur d'approximation d'un élément du dictionnaire par les autres éléments en fonction de ce paramètre. On retrouve la condition suffisante ci-dessus en considérant cette borne inférieure nulle, ce qui correspond à des fonctions noyau linéairement indépendantes.

7.2.3 Relation entre les éléments d'un dictionnaire

Dans cette section, on étudie le problème d'approximation d'un élément du dictionnaire par les autres éléments. En élaborant une borne inférieure pour le résidu d'une telle approximation, on peut conclure que tout élément ne peut pas être représenté avec une erreur arbitrairement faible par les autres éléments du dictionnaire. La proposition suivante fournit une telle borne.

Proposition 7.4. *Soit $\mathcal{D}_m$ un dictionnaire de m fonctions noyau de cohérence μ, avec $(m-1)\,\mu < 1$. L'erreur quadratique d'approximation d'un élément quelconque de ce dictionnaire par les m autres éléments est supérieure à $1 - \sqrt{(m-1)\,\mu^2/(1-(m-2)\,\mu)}$.*

Démonstration. On désigne par $P_{\mathcal{D}_{m-1}}$ l'opérateur de projection sur l'espace engendré par les éléments de $\mathcal{D}_{m-1} = \{\kappa(\boldsymbol{x}_{\omega_1},\cdot), \kappa(\boldsymbol{x}_{\omega_2},\cdot), \ldots, \kappa(\boldsymbol{x}_{\omega_{m-1}},\cdot)\}$. La projection de $\kappa(\boldsymbol{x}_{\omega_m},\cdot)$ sur cet espace est alors donnée par $P_{\mathcal{D}_{m-1}}\kappa(\boldsymbol{x}_{\omega_m},\cdot)$. La norme au carré de cette dernière correspond au maximum du produit scalaire $\langle \kappa(\boldsymbol{x}_{\omega_m},\cdot), \psi(\cdot)\rangle_{\mathcal{H}}$ sur l'ensemble de toutes les fonctions $\psi(\cdot)$ de norme unité de cet espace. En écrivant $\psi(\cdot) = \sum_{i=1}^{m-1} \alpha_i\,\kappa(\boldsymbol{x}_{\omega_i},\cdot)/\|\sum_{i=1}^{m-1}\alpha_i\,\kappa(\boldsymbol{x}_{\omega_i},\cdot)\|_{\mathcal{H}}$, le problème s'exprime formellement par

$$\begin{aligned}\|P_{\mathcal{D}_{m-1}}\kappa(\boldsymbol{x}_{\omega_m},\cdot)\|^2_{\mathcal{H}} &= \max_{\alpha} \frac{\langle\sum_{i=1}^{m-1}\alpha_i\,\kappa(\boldsymbol{x}_{\omega_i},\cdot), \kappa(\boldsymbol{x}_{\omega_m},\cdot)\rangle_{\mathcal{H}}}{\|\sum_{i=1}^{m-1}\alpha_i\,\kappa(\boldsymbol{x}_{\omega_i},\cdot)\|_{\mathcal{H}}} \\ &= \max_{\alpha} \frac{(\sum_{i=1}^{m-1}\alpha_i\,\kappa(\boldsymbol{x}_{\omega_i},\boldsymbol{x}_{\omega_m}))}{\|\sum_{i=1}^{m-1}\alpha_i\,\kappa(\boldsymbol{x}_{\omega_i},\cdot)\|_{\mathcal{H}}}, \end{aligned} \tag{7.8}$$

où la seconde égalité est due à la propriété reproduisante du noyau considéré. Pour expliciter une borne supérieure de (7.8), on procède en deux temps. D'une part, le numérateur est borné supérieurement par

$$\begin{aligned}\Big(\sum_{i=1}^{m-1}\alpha_i\kappa(\boldsymbol{x}_{\omega_i},\boldsymbol{x}_{\omega_m})\Big)^2 &\leq \Big(\sum_{i=1}^{m-1}|\alpha_i\kappa(\boldsymbol{x}_{\omega_i},\boldsymbol{x}_{\omega_m})|\Big)^2 \\ &\leq \sum_{i=1}^{m-1}\alpha_i^2\sum_{i=1}^{m-1}|\kappa(\boldsymbol{x}_{\omega_i},\boldsymbol{x}_{\omega_m})|^2 \\ &\leq \mu_2^2(m-1)\sum_{i=1}^{m-1}\alpha_i^2,\end{aligned}$$

où la seconde inégalité est obtenue à partir de l'inégalité de Cauchy-Schwartz, alors que la dernière découle de la définition de $\mu_2(\cdot)$. Cette borne supérieure peut s'exprimer à partir de la fonction de Babel du dictionnaire, avec $\mu_1^2(m-1)\sum_{i=1}^{m-1}\alpha_i^2$, en ayant recours à la Propriété 4 à la Section 7.2. En l'exprimant en fonction de la cohérence du dictionnaire, on obtient

$$\Big(\sum_{i=1}^{m-1}\alpha_i\kappa(\boldsymbol{x}_{\omega_i},\boldsymbol{x}_{\omega_m})\Big)^2 \leq (m-1)\,\mu^2\sum_{i=1}^{m-1}\alpha_i^2$$

D'autre part, une borne inférieure sur le dénominateur est obtenue à partir de

$$\frac{\|\sum_{i=1}^{m-1}\alpha_i\kappa(\boldsymbol{x}_{\omega_i},\cdot)\|^2_{\mathcal{H}}}{\sum_{i=1}^{m-1}\alpha_i^2} = \frac{\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|^2} \geq \lambda_{\min} \geq 1-\mu_1(m-2),$$

où la dernière inégalité découle de la Proposition 7.1 appliquée à la plus faible valeur propre de $\boldsymbol{K}$, la matrice de Gram des $m-1$ éléments de $\mathcal{D}_{m-1}$. En considérant la cohérence du dictionnaire, on écrit alors $\|\sum_{i=1}^{m-1}\alpha_i\kappa(\boldsymbol{x}_{\omega_i},\cdot)\|^2_{\mathcal{H}}/\sum_{i=1}^{m-1}\alpha_i^2 \geq 1-(m-2)\,\mu$

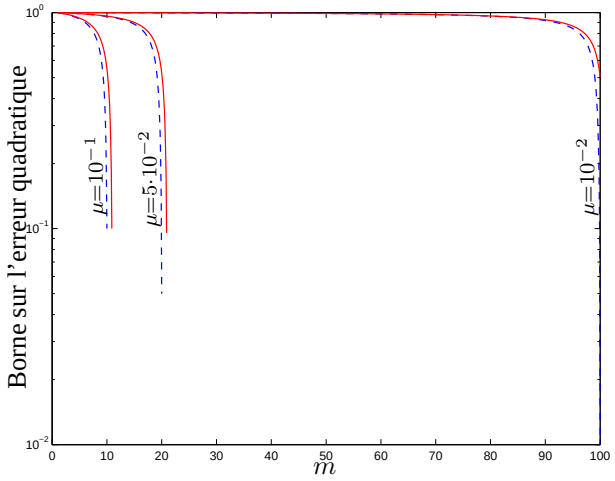


FIG. 7.1 – Borne inférieure sur l'erreur quadratique d'approximation, en pointillés (bleu) pour les précédents travaux [GMS03, RBH07] et en ligne continue (rouge) pour celle dérivée dans ce manuscrit.

Finalement, en combinant les deux bornes présentées ci-dessus, on obtient la borne supérieure suivante sur l'expression (7.8), avec

$$\|P_{\mathcal{D}_{m-1}}\kappa(\boldsymbol{x}_{\omega_m},\cdot)\|^2_{\mathcal{H}} \leq \sqrt{\frac{(m-1)\,\mu^2}{1-(m-2)\,\mu}}. \tag{7.9}$$

On peut expliciter une borne supérieure plus précise en fonction de μ_1 avec $\mu_1^2(m-1)/(1-\mu_1(m-2))$, ou encore en introduisant μ_2 avec $\mu_2^2(m-1)/(1-\mu_1(m-2))$. Toutefois, on se contente dans ce qui suit de considérer la borne (7.9), dépendant de la cohérence du dictionnaire et de sa taille. Pour terminer, la norme au carré du résidu admet une borne inférieure telle que

$$\begin{aligned}\|(I-P_{\mathcal{D}_{m-1}})\,\kappa(\boldsymbol{x}_{\omega_m},\cdot)\|^2_{\mathcal{H}} &= \|\kappa(\boldsymbol{x}_{\omega_m},\cdot)\|^2_{\mathcal{H}} - \|P_{\mathcal{D}_{m-1}}\kappa(\boldsymbol{x}_{\omega_m},\cdot)\|^2_{\mathcal{H}} \\ &\geq 1-\sqrt{\frac{(m-1)\,\mu^2}{1-(m-2)\,\mu}}.\end{aligned} \tag{7.10}$$

Cette borne est valide puisque $1-(m-2)\,\mu > 0$ est vérifiée pour $(m-1)\,\mu < 1$, la condition d'indépendance linéaire. ■

Comme prévu, plus m et μ sont faibles, plus l'erreur d'approximation est élevée. La borne (7.10) est intéressante au sens où elle couvre tout l'intervalle $]0,1]$, la limite supérieure étant atteinte pour $\mu = 0$, et la borne inférieure pour $\mu = 1/(m-1)$. Une fois de plus, on retrouve la condition suffisante $(m-1)\,\mu < 1$ de l'indépendance linéaire. Elle correspond au cas où il n'existe aucun élément qui puisse être représenté, sans erreur d'approximation, par une combinaison linéaire des autres éléments. Cette borne est plus précise que la borne $1-\sqrt{(m-1)\,\mu^2/(1-(m-1)\,\mu)}$, introduite implicitement dans [GMS03] et proposé dans [RBH07] dans le cadre de dictionnaire de fonctions noyau. De plus, la condition de validité de la borne proposée ici, $(m-1)\,\mu < 1$, est moins restrictive que la condition initiale, $(m-1)\,\mu < 1/2$ proposée dans [RBH07, Proposition 2]. On trace à la Figure 7.1 ses deux bornes, en fonction de la taille m du dictionnaire pour plusieurs valeurs du paramètre de cohérence μ.

7.3 Critère de cohérence pour le contrôle de la complexité du modèle

7.3.1 Critère de cohérence

Le résultat précédent stipule qu'un dictionnaire de cohérence μ est forcément formé par des fonctions noyau vérifiant le critère d'approximation linéaire (7.6), avec $\eta_0^2 > 1 - \sqrt{(m-1)\,\mu^2/(1-(m-2)\,\mu)}$. Plutôt que de résoudre le problème quadratique (7.6), on propose un critère de parcimonie en-ligne basé sur la cohérence en construisant un dictionnaire de cohérence inférieure à un seuil donné. Ainsi introduit-on $\kappa(\boldsymbol{x}_n, \cdot)$ dans le dictionnaire si la cohérence de ce dernier après ajout ne dépasse pas un seuil μ_0 donné, à savoir

$$\max_{j=1,\ldots,m} |\kappa(\boldsymbol{x}_n, \boldsymbol{x}_{\omega_j})| \leq \mu_0, \tag{7.11}$$

où $\mu_0 \in [0, 1]$ détermine la cohérence du dictionnaire et le niveau de parcimonie du modèle résultant. Cette règle nécessite à chaque itération $\mathcal{O}(m)$ opérations.

On peut aussi proposer d'autres critères de parcimonie associés aux quantités cousines de la cohérence. Dans le cas particulier de la fonction de Babel, on a la règle d'insertion

$$\max_{J_k} \sum_{j \in J_k} |\kappa(\boldsymbol{x}_n, \boldsymbol{x}_{\omega_j})| \leq \mu_0,$$

où J_k est un ensemble de k indices de $\{1, \ldots, m\}$. On a alors une classe de m critères possibles basés sur la fonction de Babel, le premier correspond au critère de cohérence (7.11). Néanmoins, il faut préciser que la fonction de Babel est monotone, $\mu_1(k) \geq \mu_1(k-1)$ pour tout $k = 2, \ldots, m-1$. Il suffit donc d'imposer une borne supérieure à $\mu_1(m-1)$ pour avoir borné les autres instances de $\mu_1(\cdot)$. Le critère résultant correspond alors à l'insertion de $\kappa(\boldsymbol{x}_n, \cdot)$ dans le dictionnaire si

$$\sum_{j=1}^{m} |\kappa(\boldsymbol{x}_n, \boldsymbol{x}_{\omega_j})| \leq \mu_0. \tag{7.12}$$

On appelle ce critère le critère de Babel. Bien que l'on revienne vers ce critère à la Section 8.4.1, son étude sort du contexte de cette thèse.

On s'intéresse désormais à la taille du dictionnaire résultant du critère de cohérence en montrant que, sous un faible condition sur $\mathcal{X}$, sa taille est finie quand n tend vers l'infini. La proposition suivante est équivalente à un résultat proposé dans [EMM04] pour le critère d'approximation linéaire (7.6).

Proposition 7.5. *Soient $\mathcal{X}$ un sous-espace compact d'un espace de Banach, et $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ un noyau reproduisant. Pour toute séquence $\{\boldsymbol{x}_i\}_{i=1}^{\infty}$ et $0 \leq \mu_0 < 1$, le dictionnaire résultant de la règle de cohérence (7.11) est de taille finie.*

Démonstration. Le sous-espace $\mathcal{X}$ étant compact, la continuité de $\kappa(\boldsymbol{x}, \cdot)$ produit un ensemble $\{\kappa(\boldsymbol{x}, \cdot)\}_{\boldsymbol{x} \in \mathcal{X}}$ compact. Il existe alors un ensemble de boules de rayons non-nuls, définies selon la norme ℓ_2, pouvant couvrir ces fonctions noyau. Or pour toute paire de fonctions noyau du dictionnaire de cohérence inférieure à μ_0, on a $\|\kappa(\boldsymbol{x}_{\omega_i}, \cdot) - \kappa(\boldsymbol{x}_{\omega_j}, \cdot)\|_{\mathcal{H}}^2 = 2 - 2\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j}) \geq 2 - 2\mu_0$. Cette borne impose alors un nombre fini de ces boules. ■

Cette proposition stipule que tout algorithme basé sur la stratégie (7.11) admet une complexité calculatoire fixe à chaque itération, indépendante du temps après une phase transitoire. Cette complexité ne dépend que de la taille m du dictionnaire, celle-ci étant déterminée par le choix du seuil μ_0. A partir de ce résultat, on propose d'exploiter la cohérence afin de construire un dictionnaire de fonctions noyau linéairement indépendantes. Pour cela, il suffit d'imposer une borne sur sa cohérence ne dépassant pas le seuil $\mu_0 < 1/(m-1)$. Dans ce qui suit, on montre que le critère de cohérence admet une relation directe avec d'autres règles de parcimonie, en particulier le critère d'approximation linéaire et l'entropie.

7.3.2 Critère de cohérence comme critère d'approximation linéaire

Dans un premier temps, la Proposition 7.4 montre que tout dictionnaire résultant du critère de cohérence avec un seuil μ_0 vérifie la critère d'approximation linéaire (7.6) avec

$$\eta_0^2 = 1 - \sqrt{\frac{(m-1)\,\mu_0^2}{1-(m-2)\,\mu_0}}.$$

Ce résultat n'offre toutefois aucune information sur le procédé de rejet par la règle de cohérence, et en particulier l'erreur d'approximation des fonctions noyau exclues du dictionnaire. Cette dernière étant bornée supérieurement par η_0 avec la règle d'approximation linéaire, la proposition suivante fournit un résultat similaire pour la règle de cohérence.

Proposition 7.6. *Soient $\mathcal{D}_m$ un dictionnaire obtenu par la règle (7.11), et $\kappa(\boldsymbol{x}_n, \cdot)$ une fonction noyau ne vérifiant pas cette règle. L'erreur quadratique d'approximation de $\kappa(\boldsymbol{x}_n, \cdot)$ par les éléments de $\mathcal{D}_m$ est inférieure à $1 - \mu_0$.*

Démonstration. On considère la projection de $\kappa(\boldsymbol{x}_n, \cdot)$ sur l'espace engendré par les éléments de $\mathcal{D}_m$. La norme au carré du résidu est alors donnée par l'expression (7.8), soit

$$\begin{aligned}
\|(I - P_{\mathcal{D}_m})\kappa(\boldsymbol{x}_n, \cdot)\|_{\mathcal{H}}^2 &= \|\kappa(\boldsymbol{x}_n, \cdot)\|_{\mathcal{H}}^2 - \|P_{\mathcal{D}_m}\kappa(\boldsymbol{x}_n, \cdot)\|_{\mathcal{H}}^2 \\
&= 1 - \max_{\boldsymbol{\alpha}} \frac{\sum_{i=1}^m \alpha_i \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_n)}{\|\sum_{i=1}^m \alpha_i \kappa(\boldsymbol{x}_{\omega_i}, \cdot)\|_{\mathcal{H}}} \\
&\leq 1 - \max_{k \in \{1,\ldots,m\}} \frac{|\kappa(\boldsymbol{x}_{\omega_k}, \boldsymbol{x}_n)|}{\|\kappa(\boldsymbol{x}_{\omega_k}, \cdot)\|_{\mathcal{H}}}.
\end{aligned}$$

L'inégalité dans cette expression correspond à un cas particulier des coefficients α_i, avec $\alpha_1, \ldots, \alpha_m = 0$ sauf pour $\alpha_k = \pm 1$ selon le signe de $\kappa(\boldsymbol{x}_{\omega_k}, \boldsymbol{x}_n)$. De plus, puisque $\kappa(\boldsymbol{x}_n, \cdot)$ ne vérifie pas la condition (7.11), on a alors $\max_{j \in \{1,\ldots,m\}} |\kappa(\boldsymbol{x}_n, \boldsymbol{x}_{\omega_j})| > \mu_0$. En combinant les deux inégalités, on obtient l'expression finale

$$\|(I - P_{\mathcal{D}_m})\kappa(\boldsymbol{x}_n, \cdot)\|_{\mathcal{H}}^2 < 1 - \mu_0,$$

pour les fonctions noyau de norme unité. ■

En combinant cette borne avec celle obtenue en Proposition 7.4, on conclut par la remarque suivante sur l'approximation d'une fonction noyau par un dictionnaire de m éléments de cohérence inférieure à μ_0. Si la règle de cohérence (7.11) est vérifiée, $\kappa(\boldsymbol{x}_n, \cdot)$ est alors ajoutée au dictionnaire. L'erreur quadratique de son approximation dépasse $1 - \sqrt{m\,\mu_0^2/(1-(m-1)\,\mu_0)}$, dans le dictionnaire de $m+1$ éléments. Si la règle n'est pas satisfaite, elle est écartée du dictionnaire. Son erreur quadratique d'approximation est inférieure à $1 - \mu_0$. On précise que la première borne est inférieure à la seconde, pour tout μ_0 et m. Tandis que ces bornes sont confondues en une seule η_0^2 pour le critère d'approximation linéaire, elles sont distinctes pour le critère de cohérence, comme illustré à la Figure 7.2.

7.3.3 Lien avec l'entropie quadratique de Rényi

On s'intéresse au lien entre le critère de cohérence et l'entropie quadratique de Rényi. Cette mesure de la quantité de désordre dans un système est définie par $H_R = -\log \int p(\boldsymbol{x})^2 \, \mathrm{d}\boldsymbol{x}$ pour une densité de probabilité p. N'étant pas connue, on estime souvent cette dernière par l'estimateur de Parzen de la forme

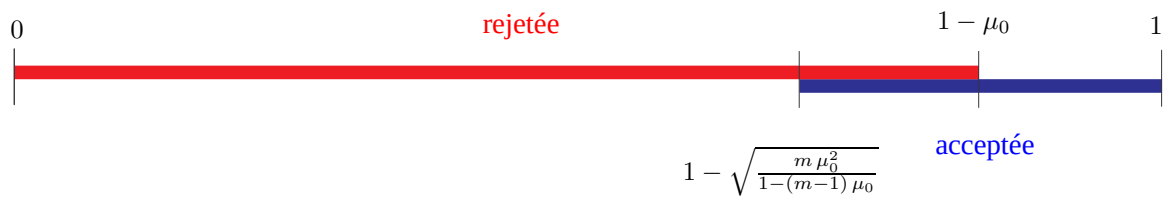


FIG. 7.2 – Bornes sur l'erreur quadratique d'approximation d'une fonction noyau avec un dictionnaire de m fonctions noyau de cohérence inférieure à μ_0. La région bleue (au dessous) correspond à la fonction noyau vérifiant (7.11), alors que celle rouge (au dessus) correspond à son opposé.

$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} w(\boldsymbol{x},\boldsymbol{x}_i)$ pour une fonction fenêtre w donnée que l'on centre sur chaque $\boldsymbol{x}_i$ disponible. En considérant la fenêtre Gaussienne dans un premier temps, l'estimateur de Parzen est défini par

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{(\sqrt{\pi}\beta_0)^l}\ \mathrm{e}^{-\|\boldsymbol{x}-\boldsymbol{x}_i\|^2/\beta_0^2}.$$

En vertu du théorème de convolution appliqué aux distributions Gaussiennes, on obtient

$$H_R \approx -\log\int \big(\hat{p}(\boldsymbol{x})\big)^2\ \mathrm{d}\boldsymbol{x} = -\log\left(\frac{1}{m^2}\sum_{i,j=1}^{m}\frac{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)}{(2\pi\beta_0^2)^{l/2}}\right),$$

où $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp\big(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\beta_0^2\big)$ est le noyau Gaussien. Cette expression montre que la somme des termes de la matrice de Gram caractérise la diversité des éléments du dictionnaire [Gir02]. Dans [SGB+02], les auteurs ont recours à cette propriété afin de proposer une technique d'élagage baptisée *fixed-size least-squares support vector machines*. Pour un dictionnaire construit à partir du critère (7.11), donc ayant une cohérence inférieure à μ_0, on obtient une borne inférieure à l'entropie, selon

$$H_R \geq \log(2\pi\beta_0^2)^{l/2} - \log\left(\frac{1+(m-1)\,\mu_0}{m}\right).$$

Comme prévu, la borne inférieure sur l'entropie augmente quand m augmente (plus d'éléments donc plus de désordre), ou quand μ_0 décroît (moins de cohérence implique plus de désordre). Dans le cas plus général d'un estimateur de Parzen dans un RKHS, l'intégrale $\int \hat{p}(\boldsymbol{x})^2\ \mathrm{d}\boldsymbol{x}$ est aussi donnée par la norme quadratique $\|\hat{p}\|_{\mathcal{H}}^2$ de $\hat{p}(\cdot) = \frac{1}{m}\sum_{i=1}^{m}\kappa(\boldsymbol{x}_i,\cdot)$. On peut alors montrer que

$$H_R \approx -\log\|\hat{p}\|_{\mathcal{H}}^2 = -\log\left(\frac{1}{m^2}\sum_{i,j=1}^{m}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)\right).$$

Dans le cas où le noyau κ n'est pas de norme unité, la règle de cohérence est alors définie à partir des fonctions noyau $\kappa(\boldsymbol{x}_k,\cdot)/\sqrt{\kappa(\boldsymbol{x}_k,\boldsymbol{x}_k)}$. On obtient alors la borne inférieure suivante sur l'entropie quadratique de Rényi

$$H_R \geq -\log\left(\frac{1}{m^2}\sum_{i=1}^{m}\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i) + \frac{\mu_0}{m^2}\sum_{\substack{i,j=1\\ i\neq j}}^{m}\sqrt{\kappa(\boldsymbol{x}_i,\boldsymbol{x}_i)\,\kappa(\boldsymbol{x}_j,\boldsymbol{x}_j)}\right).$$

Comme la borne précédente, celle-ci augmente quand μ_0 décroît et quand m augmente. Ces résultats soulignent l'intérêt de la cohérence pour caractériser la diversité des fonctions noyau dans un dictionnaire.

7.3.4 Connection avec l'ACP-à-noyau

En sélectionnant les m fonctions noyau, déterminant ainsi un sous-espace représentatif des données disponibles, notre approche peut être perçue comme une technique de réduction de dimensionnalité. Il est alors naturel d'étudier son lien avec l'ACP-à-noyau, dont l'algorithme est présenté au Chapitre 4. Par soucis de clarté, on suppose que les données sont centrées dans le RKHS.

On rappelle que l'ACP consiste à déterminer les axes principaux qui absorbent la plus grande variance des données $\boldsymbol{x}_j$, ce qui correspond à l'information utile comparée au bruit. Les axes principaux sont les vecteurs propres Ψ_k associés aux plus grandes valeurs propres λ_k de la matrice de covariance $\boldsymbol{C}$ des données, à savoir $\boldsymbol{C}\,\Psi_k = \lambda_k \Psi_k$. L'ACP-à-noyau est la mise en œuvre de l'ACP dans un espace induit par un noyau reproduisant κ. Les données étudiées sont alors les fonctions noyau $\kappa(\boldsymbol{x}_j, \cdot)$ et le $k^{\text{ème}}$ axe principal est donné par $\Psi_k = \sum_{j=1}^{n} \beta_{j,k} \kappa(\boldsymbol{x}_j, \cdot)$, où les $\beta_{j,k}$ sont les termes du $k^{\text{ème}}$ vecteur propre de la matrice de Gram. Afin d'aboutir à des axes principaux de norme unité, les coefficients $\beta_{j,k}$ sont normalisés de sorte que $\sum_{j=1}^{n} \beta_{j,k}^2 = 1/n\lambda_k$. La proposition suivante met en évidence le lien entre l'ACP-à-noyau et notre approche.

Proposition 7.7. *Soit $\mathcal{D}_m$ un dictionnaire construit par la règle de cohérence (7.11) appliquée à n fonctions noyau. On désigne par Ψ_k le $k^{ème}$ axe principal de ces n fonctions noyau, de valeur propre λ_k. L'erreur quadratique d'approximation de Ψ_k par les m éléments de $\mathcal{D}_m$ est inférieure à $(1-\mu_0)\lambda_k$.*

Démonstration. Pour le montrer, on explicite une borne supérieure du résidu $\|(I - P_{\mathcal{D}_m})\,\Psi_k\|_{\mathcal{H}}$. On désigne par $\kappa(\boldsymbol{x}_{\omega_j}, \cdot)$, pour $j = 1, \ldots, m$, les éléments de $\mathcal{D}_m$, et par $\mathcal{J}_n$ l'ensemble des indices $\{\omega_1, \omega_2, \ldots, \omega_m\}$. En développant Ψ_k selon $\Psi_k = \sum_{j=1}^{n} \beta_{j,k} \kappa(\boldsymbol{x}_j, \cdot)$, on peut écrire

$$\begin{aligned}
\|(I - P_{\mathcal{D}_m})\,\Psi_k\|_{\mathcal{H}} &= \Big\| \sum_{j=1}^{n} \beta_{j,k} (I - P_{\mathcal{D}_m})\,\kappa(\boldsymbol{x}_j, \cdot) \Big\|_{\mathcal{H}} \\
&\leq \sum_{j=1}^{n} |\beta_{j,k}|\, \|(I - P_{\mathcal{D}_m})\,\kappa(\boldsymbol{x}_j, \cdot)\|_{\mathcal{H}} \\
&= \sum_{\substack{j=1 \\ \omega_j \notin \mathcal{J}_n}}^{n} |\beta_{j,k}|\, \|(I - P_{\mathcal{D}_m})\,\kappa(\boldsymbol{x}_j, \cdot)\|_{\mathcal{H}}.
\end{aligned}$$

L'inégalité est obtenue en vertu de l'inégalité triangulaire généralisée, alors que la dernière égalité découle de fait que $\|(I - P_{\mathcal{D}_m})\,\kappa(\boldsymbol{x}_{\omega_j}, \cdot)\| = 0$ pour tout $\kappa(\boldsymbol{x}_{\omega_j}, \cdot)$ du dictionnaire $\mathcal{D}_m$. Les fonctions noyau du dernier membre de l'expression ci-dessus étant toutes exclues du dictionnaire $\mathcal{D}_m$ par la règle de cohérence (7.11), on sait alors en vertu de la Proposition 7.6 que $\|(I - P_{\mathcal{D}_m})\,\kappa(\boldsymbol{x}_j, \cdot)\|_{\mathcal{H}}^2 < 1 - \mu_0$. Par conséquent, on peut écrire

$$\|(I - P_{\mathcal{D}_m})\,\Psi_k\|_{\mathcal{H}}^2 < (1 - \mu_0) \Big(\sum_{\substack{j=1 \\ \omega_j \notin \mathcal{J}_n}}^{n} |\beta_{j,k}| \Big)^2. \tag{7.13}$$

La sommation dans cette expression peut être bornée supérieurement par

$$\Big(\sum_{\substack{j=1 \\ \omega_j \notin \mathcal{J}_n}}^{n} |\beta_{j,k}| \Big)^2 \leq \Big(\sum_{j=1}^{n} |\beta_{j,k}| \Big)^2 \leq n \sum_{j=1}^{n} \beta_{j,k}^2 = \frac{1}{\lambda_k},$$

où la seconde inégalité est due à l'inégalité de Cauchy-Schwartz, et l'égalité à la normalisation des axes principaux de l'ACP-à-noyau. Pour terminer, on injecte ce résultat dans l'expression (7.13), ce qui permet d'écrire

$$\|(I - P_{\mathcal{D}_m})\,\Psi_k\|_{\mathcal{H}}^2 < \frac{1-\mu_0}{\lambda_k}.$$

■

De ce résultat, on conclut que les axes principaux associés aux plus grandes valeurs propres admettent de faibles erreurs d'approximation. On peut alors dire qu'ils appartiennent, à une faible erreur près, à l'espace engendré par les éléments du dictionnaire. On peut voir le critère de cohérence comme un moyen d'approximer des composantes principales sans le surcoût calculatoire occasionné par une inversion matricielle comme pour les algorithmes de l'ACP et de l'ACP-à-noyau. Un résultat semblable est obtenu pour le critère d'approximation linéaire dans [EMM04]. Tandis que ce dernier admet une complexité calculatoire quadratique avec la taille du dictionnaire, le critère de cohérence est à complexité linéaire.

Chapitre 8

Méthodes d'identification adaptatives non-linéaires

Sommaire

Au chapitre précédent, nous avons défini la cohérence d'un dictionnaire de fonctions noyau, et utilisé celle-ci pour le contrôle en-ligne de l'ordre des modèles. Diverses propriétés ont alors été étudiées. En particulier, on a établi le lien entre ce critère et différents critères de parcimonie, dont l'erreur d'approximation linéaire et l'entropie quadratique. On propose dans ce chapitre d'adopter la cohérence pour la résolution en-ligne de problèmes d'estimation fonctionnelle par méthodes à noyau. On profite alors de la vaste littérature sur le filtrage adaptatif linéaire pour proposer de nouveaux algorithmes non-linéaires. On conclut en décrivant des algorithmes séquentiels pour l'ACP-à-noyau et l'AFD-à-noyau, permettant ainsi de résoudre des problèmes d'apprentissage à grand nombre de données.

8.1 Introduction

Une caractéristique fondamentale des méthodes à noyau est le fait que le modèle résultant est une combinaison linéaire de fonctions noyau, dont l'ordre est égal au nombre d'éléments de l'ensemble d'apprentissage. On peut cependant choisir, avec le critère de cohérence exposé au chapitre précédent, de travailler avec des modèles réduits d'ordre plus faible. Il reste à présent à traiter le problème de la mise-à-jour des coefficients.

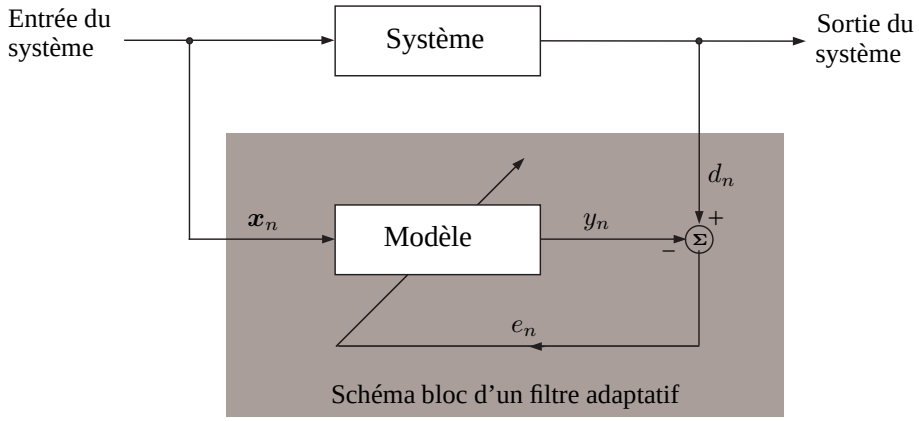


FIG. 8.1 – Identification de système variant dans le temps à l'aide d'une technique de filtrage adaptatif.

Dans le cadre de problèmes d'apprentissage en-ligne, où les données sont présentées de façon séquentielle à l'algorithme, il convient de proposer une méthode d'optimisation apte à mettre à jour les coefficients du modèle et susceptible de s'adapter aux fluctuations du système non-stationnaire considéré. A chaque itération, la méthode d'apprentissage comporte deux étapes élémentaires. Dans un premier temps, le critère de cohérence permet éventuellement d'augmenter l'ordre du modèle en complétant le dictionnaire avec la donnée courante. Dans un second temps, on opère une mise-à-jour des coefficients du modèle de sorte à minimiser l'erreur quadratique de prédiction. Certains algorithmes procèdent par adaptation des coefficients du modèle, sans qu'il soit nécessaire de résoudre le problème d'optimisation initial. Ceux-ci s'inscrivent dans le cadre général des algorithmes de filtrage adaptatif. Ces derniers ont été largement étudiés sous une forme linéaire dans la littérature ayant trait au traitement du signal [Say03, Hay02]. On compte en revanche peu de travaux encore sur l'usage de modèles à noyau pour conférer un caractère non-linéaire à ces techniques. Ce chapitre vise à remédier à ce manque.

Éléments de filtrage adaptatif linéaire

Le concept de filtrage linéaire a conduit au développement d'une armada de techniques pour le traitement des signaux, stationnaires et non-stationnaires. Ces techniques visent essentiellement à extraire des informations pertinentes d'une séquence d'observations noyées dans un bruit. Un critère particulièrement utilisé pour la synthèse de filtres est l'erreur quadratique entre la sortie obtenue et la réponse désirée. Dans le cas de données stationnaires, la solution optimale est donnée par le filtre de Wiener. En environnement non-stationnaire, on a recours aux techniques de filtrage adaptatif car elles permettent de suivre les variations du système étudié pourvu qu'elles soient lentes. La Figure (8.1) illustre ce principe par un système bouclé permettant l'identification du système étudié à partir des entrées et des sorties correspondantes. Une grande variété d'algorithmes de filtrage adaptatif a été proposée dans la littérature, que l'on peut regrouper principalement en deux catégories :

Les algorithmes des moindres carrés récursifs : On cherche à minimiser une somme pondérée des carrés des erreurs, où une pondération permet éventuellement de négliger les erreurs correspondant aux données les plus anciennes. Ces algorithmes ont souvent recours à des lemmes d'inversion matricielle pour la mise à jour des paramètres du filtre, sans la nécessité de résoudre à chaque instant le problème d'optimisation initial. Le plus connu de ces algorithmes est sans doute l'algorithme RLS, pour *recursive-least-squares algorithm*.

Les algorithmes du gradient stochastique : On s'appuie sur une technique de descente du gradient pour minimiser un critère de performance, après avoir estimé son gradient. Parmi ces algorithmes, on recense l'algorithme LMS, pour *least-mean-squares algorithm*, sa version normalisée NLMS, ou encore l'algorithme de projection affine, APA pour *affine projection algorithm*.

Les algorithmes de moindres carrés récursifs convergent plus rapidement que ceux basés sur le gradient stochastique, au prix toutefois d'une complexité calculatoire plus élevée. La linéarité du filtrage adaptatif stipule que la réponse du système est une fonction linéaire des observations appliquées à son entrée. Les algorithmes répondant à une telle propriété profitent d'une simplicité conceptuelle et d'une facilité d'implémentation remarquables. Cependant, un grand nombre d'applications exigent un traitement non-linéaire des données, un fait couvrant plusieurs disciplines, comme décrit dans l'article [GS01] dédié à la théorie des systèmes non-linéaires. Contrairement aux systèmes linéaires qui se définissent uniquement par leur réponse impulsionnelle, il existe une grande variété de représentations décrivant les systèmes non-linéaires, dont les filtres polynômiaux [MS00] et les réseaux de neurones [Hay99]. Les filtres polynômiaux basés sur les séries de Volterra [MS00] permettent de modéliser une grande famille de systèmes non-linéaires. S'ils peuvent bénéficier des classiques algorithmes de moindres carrés récursifs et du gradient stochastique, ils nécessitent en revanche l'estimation d'un nombre important de coefficients. Les réseaux de neurone ont fait quant à eux l'objet de vastes études et sont considérés à juste titre comme des approximateurs universels [Kol57]. Toutefois, les algorithmes d'apprentissage associés souffrent de plusieurs maux, tels que la présence de nombreux minima locaux dans la fonction coût, une complexité calculatoire très élevée ainsi qu'une lente convergence. Dans ce contexte, les méthodes à noyau ouvrent de nouvelles perspectives.

8.2 Algorithme de moindres carrés récursif à noyau (KRLS)

On rappelle que le problème d'optimisation traité est $\min_{\psi\in\mathcal{H}} \sum_{i=1}^{n} |d_i - \psi(\boldsymbol{x}_i)|^2 + \eta\|\psi\|^2_{\mathcal{H}}$. Pour une résolution récursive de ce problème, on réécrit le problème sous la forme

$$\min_{\psi\in\mathcal{H}} \sum_{i=1}^{n} \theta^{n-i}|d_i - \psi(\boldsymbol{x}_i)|^2 + \eta\,\theta^n\|\psi\|^2_{\mathcal{H}}, \tag{8.1}$$

où $\theta \in]0,1]$ est un facteur d'oubli qui permet de négliger les données anciennes au profit des plus récentes, fournissant ainsi un mécanisme de suivi de l'évolution des observations quand le filtre opère en environnement non-stationnaire. Le choix d'une pondération $\eta\,\theta^n$ dans le second terme de l'expression (8.1) permet un amoindrissement des effets de la régularisation avec le temps. En vertu du Théorème de Représentation, la solution optimale s'écrit sous la forme $\psi(\cdot) = \sum_{j=1}^{n} \alpha_j\,\kappa(\boldsymbol{x}_j, \cdot)$. En injectant celle-ci dans (8.1), on obtient le problème dual

$$\min_{\boldsymbol{\alpha}} (\boldsymbol{d} - \boldsymbol{K}\boldsymbol{\alpha})^\top \boldsymbol{\Theta}(\boldsymbol{d} - \boldsymbol{K}\boldsymbol{\alpha}) + \eta\,\theta^n \boldsymbol{\alpha}^\top \boldsymbol{K}\boldsymbol{\alpha}, \tag{8.2}$$

où $\boldsymbol{\Theta}$ est une matrice diagonale dont le $i^{\text{ème}}$ terme est θ^{n-i}.

8.2.1 Critère de cohérence et algorithme KRLS

La résolution du problème d'optimisation (8.2) n'est pas compatible avec un filtrage en-ligne puisqu'elle nécessite l'inversion d'une matrice dont la taille $n \times n$ augmente indéfiniment avec le nombre d'observations. Pour remédier à ce problème, on propose de profiter du paramètre de cohérence qui fournit une information sur l'indépendance linéaire des éléments d'un dictionnaire de fonctions noyau. On

désigne par $\psi_n(\cdot)$ le modèle à l'instant n d'ordre $m \leq n$, avec

$$\psi_n(\cdot) = \sum_{j=1}^{m} \alpha_{n,j}\, \kappa(\boldsymbol{x}_{\omega_j}, \cdot), \tag{8.3}$$

où les fonctions noyau $\kappa(\boldsymbol{x}_{\omega_j}, \cdot)$ constituent un dictionnaire obtenu par la règle (7.11) pour un seuil μ_0 donné. En injectant cette expression dans (8.1), on retrouve conformément à (8.2) le vecteur $\boldsymbol{\alpha}_n$ des coefficients optimaux à l'instant n. Il est donné par la résolution du problème

$$\boldsymbol{\alpha}_n = \arg\min_{\boldsymbol{\alpha}} (\boldsymbol{d}_n - \boldsymbol{H}_n\boldsymbol{\alpha})^\top \boldsymbol{\Theta}_n (\boldsymbol{d}_n - \boldsymbol{H}_n\boldsymbol{\alpha}) + \eta\, \theta^n \boldsymbol{\alpha}^\top \boldsymbol{K}_n \boldsymbol{\alpha}. \tag{8.4}$$

Dans cette expression, $\boldsymbol{K}_n$ désigne la matrice de Gram du dictionnaire de taille $m \times m$ dont le $(i,j)^{\text{ème}}$ terme correspond à $\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$, $\boldsymbol{\Theta}_n$ est la matrice diagonale de taille $n \times n$ avec θ^{n-i} pour le $i^{\text{ème}}$ élément diagonal, et $\boldsymbol{H}_n$ la matrice de taille $n \times m$ dont le $(i,j)^{\text{ème}}$ élément est $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_{\omega_j})$. En supposant que

$$\boldsymbol{P}_n = (\boldsymbol{H}_n^\top \boldsymbol{\Theta}_n \boldsymbol{H}_n + \eta\, \theta^n \boldsymbol{K}_n)^{-1}$$

existe, la solution du problème (8.4) s'exprime par

$$\boldsymbol{\alpha}_n = \boldsymbol{P}_n \boldsymbol{H}_n^\top \boldsymbol{\Theta}_n \boldsymbol{d}_n. \tag{8.5}$$

A l'arrivée d'une nouvelle donnée à l'instant $n+1$, deux cas peuvent se présenter selon la règle de cohérence (7.11), $\max_{j=1\ldots m} |\kappa(\boldsymbol{x}_n, \boldsymbol{x}_{\omega_j})| \leq \mu_0$. Si $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ ne satisfait par cette règle, le dictionnaire reste inchangé, et on met à jour les matrices $\boldsymbol{H}_n$ et $\boldsymbol{\Theta}_n$. En revanche, si la fonction noyau $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ vérifie cette règle, elle est insérée dans le dictionnaire, et $\boldsymbol{K}_n$ doit aussi être mise-à-jour. En considérant cette stratégie, on développe un algorithme récursif qui permet d'estimer à l'instant $n+1$, à partir d'une solution de moindres carrés $\boldsymbol{\alpha}_n$ à l'instant n, la solution $\boldsymbol{\alpha}_{n+1}$. L'algorithme, baptisé KRLS pour *kernel-based recursive least-squares*, est exposé ci-dessous.

Premier cas : $\max_{j=1,\ldots,m} |\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})| > \mu_0$

Ce test indique que la nouvelle fonction noyau $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ peut être représentée avec une faible erreur d'approximation par les fonctions noyau déjà dans le dictionnaire. Il n'est donc pas nécessaire de l'inclure dans le dictionnaire. Une ligne supplémentaire est ajoutée à $\boldsymbol{H}_n$, et un élément supplémentaire est ajouté à $\boldsymbol{d}_n$, menant aux expressions

$$\boldsymbol{H}_{n+1} = \begin{bmatrix} \boldsymbol{H}_n \\ \boldsymbol{h}_{n+1}^\top \end{bmatrix} \qquad \boldsymbol{d}_{n+1} = \begin{bmatrix} \boldsymbol{d}_n \\ d_{n+1} \end{bmatrix},$$

avec $\boldsymbol{h}_{n+1}^\top = [\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_1}) \cdots \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_m})]$. Comme le dictionnaire reste inchangé, on a $\boldsymbol{K}_{n+1} = \boldsymbol{K}_n$. Soit $\boldsymbol{\Theta}_{n+1}$ la matrice diagonale de taille $n+1 \times n+1$ dont le $i^{\text{ème}}$ terme est θ^{n+1-i}. En supposant que la matrice $\boldsymbol{H}_{n+1}^\top \boldsymbol{\Theta}_{n+1} \boldsymbol{H}_{n+1} + \eta\theta^{n+1} \boldsymbol{K}_{n+1}$ n'est pas singulière, la solution à l'instant $n+1$ est donnée par

$$\boldsymbol{\alpha}_{n+1} = \boldsymbol{P}_{n+1} \boldsymbol{H}_{n+1}^\top \boldsymbol{\Theta}_{n+1} \boldsymbol{d}_{n+1}, \tag{8.6}$$

où $\boldsymbol{P}_{n+1} = (\boldsymbol{H}_{n+1}^\top \boldsymbol{\Theta}_{n+1} \boldsymbol{H}_{n+1} + \eta\, \theta^{n+1} \boldsymbol{K}_{n+1})^{-1} = (\theta \boldsymbol{P}_n^{-1} + \boldsymbol{h}_{n+1} \boldsymbol{h}_{n+1}^\top)^{-1}$. Cette expression est similaire à la formulation classique des moindres carrés récursifs, et ne nécessite aucune inversion matricielle pour être évaluée. Pour cela, on propose d'appliquer l'identité de Woodbury connue aussi sous le nom de lemme d'inversion matricielle,

$$(\boldsymbol{A} + \boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{C}^{-1} + \boldsymbol{DA}^{-1}\boldsymbol{B})\boldsymbol{DA}^{-1},$$

à l'expression de $\boldsymbol{P}_{n+1}$. On aboutit à l'expression récursive suivante

$$\boldsymbol{P}_{n+1} = \theta^{-1}\left[\boldsymbol{P}_n - \frac{\theta^{-1}\boldsymbol{P}_n\boldsymbol{h}_{n+1}\boldsymbol{h}_{n+1}^\top\boldsymbol{P}_n}{1+\theta^{-1}\boldsymbol{h}_{n+1}^\top\boldsymbol{P}_n\boldsymbol{h}_{n+1}}\right] \tag{8.7}$$

avec pour condition initiale $\eta\boldsymbol{K}_n$. En remplaçant (8.7) dans (8.6), on obtient la mise-à-jour de $\boldsymbol{\alpha}_{n+1}$ en fonction de $\boldsymbol{\alpha}_n$, selon

$$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n + \boldsymbol{P}_{n+1}\boldsymbol{h}_{n+1}\,(d_{n+1} - \boldsymbol{h}_{n+1}^\top\boldsymbol{\alpha}_n). \tag{8.8}$$

Dans cette expression, on retrouve $\boldsymbol{h}_{n+1}^\top\boldsymbol{\alpha}_n$, ce qui correspond à $\psi_n(\boldsymbol{x}_{n+1})$ selon (8.3) et qui n'est autre que l'estimation de la réponse désirée à l'instant $n+1$. L'expression entre parenthèses dans (8.8) s'écrit alors selon $e_{n+1} = d_{n+1} - \psi_n(\boldsymbol{x}_{n+1})$, ce qui correspond à l'erreur *a priori* à l'instant $n+1$.

Second cas : $\max_{j=1,\ldots,m}|\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})| \leq \mu_0$

Dans ce cas, la fonction noyau candidate $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ à l'instant $n+1$ ne peut être représentée d'une manière satisfaisante par les éléments du dictionnaire. Pour cette raison, il faut l'inclure dans le dictionnaire. On désigne désormais cette fonction noyau par $\kappa(\boldsymbol{x}_{\omega_{m+1}}, \cdot)$. Le nombre de termes dans le modèle (8.3) croît de 1. On note alors

$$\boldsymbol{H}_{n+1} = \begin{bmatrix}\boldsymbol{H}_n & \boldsymbol{0}_n\\ \boldsymbol{h}_{n+1}^\top & h_0\end{bmatrix} \qquad \boldsymbol{K}_{n+1} = \begin{bmatrix}\boldsymbol{K}_n & \boldsymbol{h}_{n+1}\\ \boldsymbol{h}_{n+1}^\top & h_0\end{bmatrix} \qquad \boldsymbol{d}_{n+1} = \begin{bmatrix}\boldsymbol{d}_n\\ d_{n+1}\end{bmatrix},$$

avec $\boldsymbol{h}_{n+1}^\top = [\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_1}) \cdots \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_m})]$, $h_0 = \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{n+1})$, et $\boldsymbol{0}_n$ un vecteur colonne de n zéros. Le vecteur des coefficients optimaux $\boldsymbol{\alpha}_{n+1}$ est toujours donné par l'équation (8.6), avec

$$\boldsymbol{P}_{n+1} = \begin{bmatrix}\theta\boldsymbol{P}_n^{-1} + \boldsymbol{h}_{n+1}\boldsymbol{h}_{n+1}^\top & (h_0 + \eta\,\theta^{n+1})\boldsymbol{h}_{n+1}\\ (h_0 + \eta\,\theta^{n+1})\boldsymbol{h}_{n+1}^\top & (h_0 + \eta\,\theta^{n+1})h_0\end{bmatrix}^{-1}.$$

Afin de calculer $\boldsymbol{P}_{n+1}$, on utilise l'identité suivante pour l'inversion d'une matrice bloc :

$$\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{A}^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{bmatrix} + \begin{bmatrix}-\boldsymbol{A}^{-1}\boldsymbol{B}\\ \boldsymbol{I}\end{bmatrix}(\boldsymbol{D} - \boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\begin{bmatrix}-\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{I}\end{bmatrix}.$$

En appliquant cette équation à $\boldsymbol{P}_{n+1}$, on obtient

$$\boldsymbol{P}_{n+1} = \begin{bmatrix}\tilde{\boldsymbol{P}}_{n+1} & \boldsymbol{0}_n\\ \boldsymbol{0}_n^\top & 0\end{bmatrix} + \frac{1}{\zeta}\begin{bmatrix}-\boldsymbol{q}\\ 1\end{bmatrix}\begin{bmatrix}-\boldsymbol{q}^\top & 1\end{bmatrix} \tag{8.9}$$

où

$$\begin{aligned}\tilde{\boldsymbol{P}}_{n+1} &= (\theta\boldsymbol{P}_n^{-1} + \boldsymbol{h}_{n+1}\boldsymbol{h}_{n+1}^\top)^{-1}\\ \boldsymbol{q} &= (h_0 + \eta\,\theta^{n+1})\tilde{\boldsymbol{P}}_{n+1}\boldsymbol{h}_{n+1}\\ \zeta &= (h_0 + \eta\,\theta^{n+1})(h_0 - \boldsymbol{h}_{n+1}^\top\boldsymbol{q}).\end{aligned}$$

Notre approche repose sur la présence de $\tilde{\boldsymbol{P}}_{n+1}$ dans la solution $\tilde{\boldsymbol{\alpha}}_{n+1} = \tilde{\boldsymbol{P}}_{n+1}\tilde{\boldsymbol{H}}_{n+1}^\top\boldsymbol{\Theta}_{n+1}\boldsymbol{d}_{n+1}$ au problème $\min_{\boldsymbol{\alpha}}\ (\boldsymbol{d}_{n+1} - \tilde{\boldsymbol{H}}_{n+1}\boldsymbol{\alpha})^\top\boldsymbol{\Theta}_{n+1}(\boldsymbol{d}_{n+1} - \tilde{\boldsymbol{H}}_{n+1}\boldsymbol{\alpha}) + \eta\,\theta^{n+1}\boldsymbol{\alpha}^\top\boldsymbol{K}_{n+1}\boldsymbol{\alpha}$, dans lequel on a noté $\tilde{\boldsymbol{H}}_{n+1} = \begin{bmatrix}\boldsymbol{H}_n^\top & \boldsymbol{h}_{n+1}\end{bmatrix}^\top$. On peut alors calculer efficacement $\tilde{\boldsymbol{P}}_{n+1}$ et $\tilde{\boldsymbol{\alpha}}_{n+1}$ avec les equations de mise-à-jour (8.7) et (8.8), soit

$$\tilde{\boldsymbol{P}}_{n+1} = \theta^{-1}\left[\boldsymbol{P}_n - \frac{\theta^{-1}\boldsymbol{P}_n\boldsymbol{h}_{n+1}\boldsymbol{h}_{n+1}^\top\boldsymbol{P}_n}{1+\theta^{-1}\boldsymbol{h}_{n+1}^\top\boldsymbol{P}_n\boldsymbol{h}_{n+1}}\right] \tag{8.10}$$

$$\tilde{\boldsymbol{\alpha}}_{n+1} = \hat{\boldsymbol{\alpha}}_n + \tilde{\boldsymbol{P}}_{n+1}\boldsymbol{h}_{n+1}\,(d_{n+1} - \boldsymbol{h}_{n+1}^\top\hat{\boldsymbol{\alpha}}_n). \tag{8.11}$$

En d'autres termes, la procédure décrite dans la section précédente, dite "Premier cas", peut être appliquée ici à $(\boldsymbol{h}_{n+1}, d_{n+1})$ comme une étape de pré-traitement pour déterminer $\tilde{\boldsymbol{P}}_{n+1}$ et $\tilde{\boldsymbol{\alpha}}_{n+1}$. La relation liant $\tilde{\boldsymbol{P}}_{n+1}$ à $\boldsymbol{P}_{n+1}$ permet alors de déterminer le vecteur de solution $\boldsymbol{\alpha}_{n+1}$ à partir de $\tilde{\boldsymbol{\alpha}}_{n+1}$. En multipliant à droite les deux termes de l'équation (8.9) par $\boldsymbol{H}_{n+1}^\top \boldsymbol{\Theta}_{n+1} \boldsymbol{d}_{n+1}$, on obtient

$$\begin{aligned}\boldsymbol{\alpha}_{n+1} &= \underbrace{\begin{bmatrix}\tilde{\boldsymbol{P}}_{n+1} & \boldsymbol{0}_n \\ \boldsymbol{0}_n^\top & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{H}_n^\top & \boldsymbol{h}_{n+1} \\ \boldsymbol{0}_n^\top & h_0\end{bmatrix}\begin{bmatrix}\theta\boldsymbol{\Theta}_n & \boldsymbol{0}_n \\ \boldsymbol{0}_n^\top & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{d}_n \\ d_{n+1}\end{bmatrix}}_{\boldsymbol{\alpha}_{n+1}^{(1)}} \\ &+ \underbrace{\frac{1}{\zeta}\begin{bmatrix}\boldsymbol{q}\boldsymbol{q}^\top & -\boldsymbol{q} \\ -\boldsymbol{q}^\top & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{H}_n^\top & \boldsymbol{h}_{n+1} \\ \boldsymbol{0}_n^\top & h_0\end{bmatrix}\begin{bmatrix}\theta\boldsymbol{\Theta}_n & \boldsymbol{0}_n \\ \boldsymbol{0}_n^\top & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{d}_n \\ d_{n+1}\end{bmatrix}}_{\boldsymbol{\alpha}_{n+1}^{(2)}}.\end{aligned}$$

On désigne par $\boldsymbol{\alpha}_{n+1}^{(1)}$ et $\boldsymbol{\alpha}_{n+1}^{(2)}$ le premier et le second termes de $\boldsymbol{\alpha}_{n+1}$ dans l'équation ci-dessus. On peut développer ceux-ci selon

$$\boldsymbol{\alpha}_{n+1}^{(1)} = \begin{bmatrix}\tilde{\boldsymbol{P}}_{n+1}(\theta \boldsymbol{H}_n^\top \boldsymbol{\Theta}_n \boldsymbol{d}_n + d_{n+1}\boldsymbol{h}_{n+1}) \\ 0\end{bmatrix} = \begin{bmatrix}\tilde{\boldsymbol{P}}_{n+1}\tilde{\boldsymbol{H}}_{n+1}^\top \boldsymbol{\Theta}_{n+1}\boldsymbol{d}_{n+1} \\ 0\end{bmatrix} = \begin{bmatrix}\tilde{\boldsymbol{\alpha}}_{n+1} \\ 0\end{bmatrix}$$

$$\begin{aligned}\boldsymbol{\alpha}_{n+1}^{(2)} = \frac{1}{\zeta}\begin{bmatrix}\boldsymbol{q}\boldsymbol{q}^\top(\theta \boldsymbol{H}_n^\top \boldsymbol{\Theta}_n \boldsymbol{d}_n + d_{n+1}\boldsymbol{h}_{n+1}) - h_0 d_{n+1}\boldsymbol{q} \\ -\boldsymbol{q}^\top(\theta \boldsymbol{H}_n^\top \boldsymbol{\Theta}_n \boldsymbol{d}_n + d_{n+1}\boldsymbol{h}_{n+1}) + h_0 d_{n+1}\end{bmatrix} &= \frac{1}{\zeta}\begin{bmatrix}-\boldsymbol{q}(h_0 d_{n+1} - (h_0 + \eta\theta^{n+1})\boldsymbol{h}_{n+1}^\top \tilde{\boldsymbol{\alpha}}_{n+1}) \\ h_0 d_{n+1} - (h_0 + \eta\theta^{n+1})\boldsymbol{h}_{n+1}^\top \tilde{\boldsymbol{\alpha}}_{n+1}\end{bmatrix} \\ &= \frac{1}{\zeta}\left(h_0 d_{n+1} - (h_0 + \eta\theta^{n+1})\boldsymbol{h}_{n+1}^\top \tilde{\boldsymbol{\alpha}}_{n+1}\right)\begin{bmatrix}-\boldsymbol{q} \\ 1\end{bmatrix}\end{aligned}$$

Finalement, on obtient l'équation de mise-à-jour liant $\boldsymbol{\alpha}_{n+1}$ à $\boldsymbol{\alpha}_n$, par l'intermédiaire de $\tilde{\boldsymbol{\alpha}}_{n+1}$, les équations (8.10), (8.11) et

$$\boldsymbol{\alpha}_{n+1} = \begin{bmatrix}\tilde{\boldsymbol{\alpha}}_{n+1} \\ 0\end{bmatrix} + \frac{h_0 d_{n+1} - (h_0 + \eta\theta^{n+1})\boldsymbol{h}_{n+1}^\top \tilde{\boldsymbol{\alpha}}_{n+1}}{(h_0 + \eta\theta^{n+1})(h_0 - \boldsymbol{h}_{n+1}^\top \boldsymbol{q})}\begin{bmatrix}-\boldsymbol{q} \\ 1\end{bmatrix}. \tag{8.12}$$

Dans la section suivante, on résume l'algorithme final dans sa version non-régularisée afin de simplifier la présentation, et l'on étudie sa complexité.

8.2.2 Algorithme KRLS et complexité

Dans ce qui suit, on présente l'algorithme d'optimisation en ne prêtant pas attention au terme de régularisation afin de simplifier l'exposé. Ceci revient à considérer $\eta = 0$ dans la fonction coût initiale $(\boldsymbol{d}_n - \boldsymbol{H}_n\boldsymbol{\alpha})^\top \boldsymbol{\Theta}_n (\boldsymbol{d}_n - \boldsymbol{H}_n\boldsymbol{\alpha})$. A chaque instant $n+1$, à l'arrivée de nouvelles données $\boldsymbol{x}_{n+1}$ et d_{n+1}, les expressions (8.7) et (8.8) sont évaluées. On vérifie ensuite si $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ satisfait la règle de cohérence (7.11), afin de l'inclure éventuellement dans le dictionnaire. Si tel est le cas, l'ordre du modèle croît et ses paramètres sont à nouveau rafraîchis. On procède pour cela selon les expressions (8.9) et (8.12), qui se réduisent dans le cas non-régularisé à

$$\boldsymbol{P}_{n+1} = \begin{bmatrix}\tilde{\boldsymbol{P}}_{n+1} & \boldsymbol{0}_n \\ \boldsymbol{0}_n^\top & 0\end{bmatrix} + \frac{1}{1 - \boldsymbol{h}_{n+1}^\top \tilde{\boldsymbol{P}}_{n+1}\boldsymbol{h}_{n+1}}\begin{bmatrix}-\tilde{\boldsymbol{P}}_{n+1}\boldsymbol{h}_{n+1} \\ 1/h_0\end{bmatrix}\begin{bmatrix}-(\tilde{\boldsymbol{P}}_{n+1}\boldsymbol{h}_{n+1})^\top & 1/h_0\end{bmatrix} \tag{8.13}$$

$$\boldsymbol{\alpha}_{n+1} = \begin{bmatrix}\tilde{\boldsymbol{\alpha}}_{n+1} \\ 0\end{bmatrix} + \frac{d_{n+1} - \boldsymbol{h}_{n+1}^\top \tilde{\boldsymbol{\alpha}}_{n+1}}{1 - \boldsymbol{h}_{n+1}^\top \tilde{\boldsymbol{P}}_{n+1}\boldsymbol{h}_{n+1}}\begin{bmatrix}-\tilde{\boldsymbol{P}}_{n+1}\boldsymbol{h}_{n+1} \\ 1/h_0\end{bmatrix}. \tag{8.14}$$

Instructions	Expressions
Paramètres	
0.Seuil de cohérence	μ_0
0.Facteur d'oubli	θ
0.Paramètre de régularisation	$\eta = 0$
Initialisation	
1. Définition du dictionnaire	$m = 1,\ \omega_1 = 1,\ \mathcal{D}_1 = \{\kappa(\boldsymbol{x}_{\omega_1}, \cdot)\}$
2. Initialisation	$\boldsymbol{P}_1 = 1/\kappa(\boldsymbol{x}_1, \boldsymbol{x}_1)$
3. Coefficient initial	$\boldsymbol{\alpha}_1 = \boldsymbol{P}_1 d_1$
A chaque instant $n+1 \geq 2$	
4. Acquisition de nouvelles données	$(\boldsymbol{x}_{n+1}, d_{n+1})$
5. Calcul de $\boldsymbol{h}_n$	$(\boldsymbol{h}_{n+1})_{(i,1)} = \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_i})$
6. Calcul de $\boldsymbol{P}_{n+1}$, selon (8.7)	$\boldsymbol{P}_{n+1} = \theta^{-1}\left[\boldsymbol{P}_n - \frac{\theta^{-1}\boldsymbol{P}_n \boldsymbol{h}_{n+1}\boldsymbol{h}_{n+1}^\top \boldsymbol{P}_n}{1+\theta^{-1}\boldsymbol{h}_{n+1}^\top \boldsymbol{P}_n \boldsymbol{h}_{n+1}}\right]$
8. Calcul de $\boldsymbol{\alpha}_{n+1}$, selon (8.8)	$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n + \boldsymbol{P}_{n+1}\boldsymbol{h}_{n+1}\left(d_{n+1} - \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_n\right)$
9. Si règle de cohérence vérifiée :	$\max_{j=1,\ldots,m} \lvert\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})\rvert \leq \mu_0$
a. Incrémentation de l'ordre	$m = m+1,\ \boldsymbol{\alpha}_{n+1} = [\boldsymbol{\alpha}_{n+1}^\top \quad 0]^\top$
b. Mise-à-jour du dictionnaire	$\omega_m = n+1,\ \mathcal{D}_m = \mathcal{D}_{m-1} \cup \{\kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}$
c. Calcul de $\boldsymbol{P}_{n+1}$, selon (8.13)	$\boldsymbol{p}_{n+1} = [-(\boldsymbol{P}_{n+1}\boldsymbol{h}_{n+1})^\top \quad 1/h_0]^\top$
	$\boldsymbol{P}_{n+1} = \begin{bmatrix} \boldsymbol{P}_{n+1} & \boldsymbol{0}_n \\ \boldsymbol{0}_n^\top & 0 \end{bmatrix} + \frac{1}{1-\boldsymbol{h}_{n+1}^\top \boldsymbol{P}_{n+1}\boldsymbol{h}_{n+1}} \boldsymbol{p}_{n+1}\boldsymbol{p}_{n+1}^\top$
d. Calcul de $\boldsymbol{\alpha}_{n+1}$, selon (8.14)	$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_{n+1} + \frac{d_{n+1} - \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_{n+1}}{1-\boldsymbol{h}_{n+1}^\top \boldsymbol{P}_{n+1}\boldsymbol{h}_{n+1}} \boldsymbol{p}_{n+1}$

TAB. 8.1 – Pseudocode de l'algorithme de moindres carrés récursif à noyau (sans régularisation) avec critère de cohérence.

L'algorithme est illustré en pseudo-code au Tableau 8.1. Le coût calculatoire de $\boldsymbol{h}_n$ dépend du noyau reproduisant choisi, et est proportionnel à la taille m du dictionnaire. Chaque itération de l'algorithme RLS nécessite $\mathcal{O}(m^2)$ opérations. Le critère de cohérence est beaucoup plus simple à mettre en œuvre que le critère d'approximation linéaire puisqu'il consiste à comparer le plus grand élément de $\boldsymbol{h}_n$ en valeur absolue à un seuil μ_0. Cette règle détermine l'incrémentation de l'ordre qui nécessite $\mathcal{O}(m^2)$ multiplications et additions par itération. On rappelle que la taille finale du dictionnaire obtenu par le critère de cohérence est fini, une propriété démontrée à la Proposition 7.5. Ceci implique qu'après une période transitoire d'incrémentation de l'ordre du modèle, l'étape 9. dans le Tableau 8.1 devient obsolète et la complexité calculatoire de l'algorithme se trouve réduite à celle de l'algorithme RLS classique. En pratique, le nombre de fonctions noyau retenues reste très raisonnable, une dizaine, comme illustré par les expérimentations de la section suivante.

L'algorithme exposé ici est original. Il faut toutefois préciser qu'il existe un autre algorithme récursif d'approximation fonctionnelle développé dans [EMM04], en considérant la résolution du problème d'approximation linéaire (7.6). Au delà de la différence des critères de parcimonie adoptés, les deux algorithmes sont distincts en plusieurs points. En particulier, en comparant [EMM04, Eq. (IV.11)] avec (8.5), on remarque que les deux équations ont recours à des matrices différentes. Alors que la première fait intervenir deux inversions matricielles, notre approche n'en nécessite qu'une seule. De plus, le facteur d'oubli, présent dans (8.2) et absent dans [EMM04, Eq. (IV.8)], permet d'opérer en environnement non-stationnaire.

8.3 Méthode du gradient stochastique : l'algorithme KAPA

On reprend le problème de minimisation de l'erreur quadratique (7.1) en abandonnant pour le moment le terme de régularisation, soit

$$\min_{\psi\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(d_i - \psi(\boldsymbol{x}_i))^2. \tag{8.15}$$

Afin de résoudre ce problème d'optimisation en-ligne, on propose à nouveau d'utiliser le critère de cohérence (7.11) comme stratégie de contrôle de l'ordre du modèle. En désignant par $\psi_n(\cdot)$ le modèle à l'instant n, on a

$$\psi_n(\cdot) = \sum_{j=1}^{m} \alpha_{n,j}\,\kappa(\boldsymbol{x}_{\omega_j},\cdot), \tag{8.16}$$

où les $\kappa(\boldsymbol{x}_{\omega_j},\cdot)$ forment un dictionnaire de cohérence μ_0 obtenu par la règle (7.11). En injectant cette forme réduite dans (8.15), le vecteur de coefficients optimaux $\boldsymbol{\alpha}_n = [\alpha_{n,1}\ \ \alpha_{n,2}\ \cdots\ \alpha_{n,m}]^\top$ est alors donné par la résolution du problème dual

$$\boldsymbol{\alpha}_n = \arg\min_{\boldsymbol{\alpha}}\ \|\boldsymbol{d}_n - \boldsymbol{H}_n\boldsymbol{\alpha}\|^2, \tag{8.17}$$

où $\boldsymbol{d}_n = [d_1\ \ d_2\ \cdots\ d_n]^\top$ désigne le vecteur des réponses désirées, et $\boldsymbol{H}_n$ la matrice de taille $n\times m$ dont le $(i,j)^{\text{ème}}$ élément correspond à $\kappa(\boldsymbol{x}_i,\boldsymbol{x}_{\omega_j})$. En supposant que la matrice $\boldsymbol{H}_n^\top\boldsymbol{H}_n$ est non-singulière, la solution optimale est donnée par

$$\boldsymbol{\alpha}_n = (\boldsymbol{H}_n^\top\boldsymbol{H}_n)^{-1}\boldsymbol{H}_n^\top\boldsymbol{d}_n.$$

A la section précédente, on a développé un algorithme récursif pour déterminer le vecteur des coefficients $\boldsymbol{\alpha}_{n+1}$ à partir de $\boldsymbol{\alpha}_n$ dès l'arrivée d'une nouvelle données $\boldsymbol{x}_{n+1}$. On propose à présent une méthode plus simple pour résoudre (8.17) à partir d'une procédure de gradient stochastique.

8.3.1 Algorithme de projection affine à noyau (KAPA)

Pour résoudre le problème d'optimisation (8.17), on considère le principe de fluctuation minimale pour déterminer à l'instant $n+1$ le vecteur des paramètres $\boldsymbol{\alpha}_{n+1}$ à partir de $\boldsymbol{\alpha}_n$, soit en minimisant $\|\boldsymbol{\alpha}_{n+1} - \boldsymbol{\alpha}_n\|^2$. A cette fonction objectif s'ajoute une contrainte sur l'erreur *a posteriori* nulle, des q dernières données. Pour cela, à l'instant $n+1$, seules les q plus récentes entrées $\{\boldsymbol{x}_{n+1},\boldsymbol{x}_n,\ldots,\boldsymbol{x}_{n-q+2}\}$ et observations $\{d_{n+1},d_n,\ldots,d_{n-q+2}\}$ sont considérées. On désigne par $\boldsymbol{d}_{n+1}$ le vecteur colonne des sorties désirées de q éléments dont le $i^{\text{ème}}$ élément est d_{n-i+2}, et par $\boldsymbol{H}_{n+1}$ la matrice de taille $q\times m$ suivante

$$\boldsymbol{H}_{n+1} = \begin{bmatrix} \kappa(\boldsymbol{x}_{n+1},\boldsymbol{x}_{\omega_1}) & \kappa(\boldsymbol{x}_n,\boldsymbol{x}_{\omega_1}) & \cdots & \kappa(\boldsymbol{x}_{n-q+2},\boldsymbol{x}_{\omega_1}) \\ \kappa(\boldsymbol{x}_{n+1},\boldsymbol{x}_{\omega_2}) & \kappa(\boldsymbol{x}_n,\boldsymbol{x}_{\omega_2}) & \cdots & \kappa(\boldsymbol{x}_{n-q+2},\boldsymbol{x}_{\omega_2}) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\boldsymbol{x}_{n+1},\boldsymbol{x}_{\omega_m}) & \kappa(\boldsymbol{x}_n,\boldsymbol{x}_{\omega_m}) & \cdots & \kappa(\boldsymbol{x}_{n-q+2},\boldsymbol{x}_{\omega_m}) \end{bmatrix}. \tag{8.18}$$

On réécrit alors le problème d'optimisation avec contrainte suivant

$$\min_{\boldsymbol{\alpha}_{n+1}} \quad \|\boldsymbol{\alpha}_{n+1} - \boldsymbol{\alpha}_n\|^2 \tag{8.19}$$

$$\text{sous contrainte} \quad \boldsymbol{d}_{n+1} = \boldsymbol{H}_{n+1}\boldsymbol{\alpha}_{n+1} \tag{8.20}$$

Ceci correspond au problème de projection affine [Say03] dans l'espace des paramètres, $\boldsymbol{\alpha}_{n+1}$ étant obtenu par la projection de $\boldsymbol{\alpha}_n$ sur l'intersection des q sous-espaces affines $\mathcal{S}_i$ définis par

$$\mathcal{S}_i = \{\boldsymbol{\alpha} \in \mathbb{R}^m\colon\ \boldsymbol{h}_{n-i+2}^\top\, \boldsymbol{\alpha} - d_{n-i+2} = 0\}, \qquad i = 1, \ldots, q,$$

avec $\boldsymbol{h}_{n-i+2}^\top = [\kappa(\boldsymbol{x}_{n-i+2}, \boldsymbol{x}_{\omega_1}) \cdots \kappa(\boldsymbol{x}_{n-i+2}, \boldsymbol{x}_{\omega_m})]$. On désigne cet algorithme par KAPA, pour *kernel-based affine projection algorithm,* et q par ordre de projection affine[10]. Dans ce qui suit, on développe l'algorithme KAPA. A l'instant $n+1$, à l'arrivée d'une nouvelle donnée, on est confronté à l'une des deux possibilités suivantes, selon que $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ vérifie ou non la règle de cohérence (7.11).

Premier cas : $\max_{j=1,\ldots,m} |\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})| > \mu_0$

La fonction noyau $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ n'est pas ajoutée au dictionnaire puisqu'elle peut être raisonnablement représentée par les éléments de ce dernier. Le dictionnaire reste inchangé. La solution au problème d'optimisation sous contrainte (8.19)–(8.20) est alors donnée par la minimisation du Lagrangien

$$\|\boldsymbol{\alpha}_{n+1} - \boldsymbol{\alpha}_n\|^2 + \boldsymbol{\lambda}^\top(\boldsymbol{d}_{n+1} - \boldsymbol{H}_{n+1}\boldsymbol{\alpha}_{n+1})$$

où $\boldsymbol{\lambda}$ est le vecteur de multiplicateurs de Lagrange. En annulant les dérivées de cette expression par rapport à $\boldsymbol{\alpha}_{n+1}$ et $\boldsymbol{\lambda}$, on obtient

$$\begin{aligned} 2\,(\boldsymbol{\alpha}_{n+1} - \boldsymbol{\alpha}_n) &= \boldsymbol{H}_{n+1}^\top\, \boldsymbol{\lambda} \\ \boldsymbol{H}_{n+1}\boldsymbol{\alpha}_{n+1} &= \boldsymbol{d}_{n+1}. \end{aligned} \tag{8.21}$$

En supposant que $\boldsymbol{H}_{n+1}\boldsymbol{H}_{n+1}^\top$ est non-singulière, ces équations produisent l'expression

$$\boldsymbol{\lambda} = 2\,(\boldsymbol{H}_{n+1}\boldsymbol{H}_{n+1}^\top)^{-1}(\boldsymbol{d}_{n+1} - \boldsymbol{H}_{n+1}\boldsymbol{\alpha}_n).$$

En plaçant ce résultat dans (8.21), on obtient l'équation de mise-à-jour de $\boldsymbol{\alpha}_{n+1}$:

$$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n + \varrho\, \boldsymbol{H}_{n+1}^\top(\eta \boldsymbol{I}_q + \boldsymbol{H}_{n+1}\boldsymbol{H}_{n+1}^\top)^{-1}(\boldsymbol{d}_{n+1} - \boldsymbol{H}_{n+1}\boldsymbol{\alpha}_n), \tag{8.22}$$

Comme préconisé dans [Say03], on a introduit dans cette équation un paramètre de contrôle du pas de convergence ϱ, et le terme $\eta \boldsymbol{I}_q$ relatif à la régularisation. L'évaluation de (8.22) nécessite l'inversion de la matrice $(\eta \boldsymbol{I}_q + \boldsymbol{H}_{n+1}\boldsymbol{H}_{n+1}^\top)$, d'une taille $q \times q$ habituellement faible.

Second cas : $\max_{j=1,\ldots,m} |\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})| \leq \mu_0$

Dans ce cas, la nouvelle fonction noyau $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ ne peut être efficacement représentée par les éléments du dictionnaire. Elle est alors ajouté à ce dernier, et on la désigne par $\kappa(\boldsymbol{x}_{\omega_{m+1}}, \cdot)$. L'ordre m du modèle réduit (8.16) est incrémenté, et $\boldsymbol{H}_{n+1}$ devient une matrice de taille $q \times (m+1)$, en incluant dans l'ancienne matrice une nouvelle colonne $[\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_{m+1}}) \cdots \kappa(\boldsymbol{x}_{n-q+2}, \boldsymbol{x}_{\omega_{m+1}})]^\top$. Pour prendre en considération le nouvel élément de $\boldsymbol{\alpha}_{n+1}$, on transforme le problème (8.19)–(8.20) ainsi

$$\begin{aligned} \min_{\boldsymbol{\alpha}_{n+1}} \quad & \|\boldsymbol{\alpha}_{n+1,1\cdots m} - \boldsymbol{\alpha}_n\|^2 + \alpha_{n+1,m+1}^2 \\ \text{sous contrainte} \quad & \boldsymbol{d}_{n+1} = \boldsymbol{H}_{n+1}\boldsymbol{\alpha}_{n+1}, \end{aligned}$$

où $\boldsymbol{\alpha}_{n+1,1\cdots m}$ désigne les m premiers éléments du vecteur $\boldsymbol{\alpha}_{n+1}$. La $(m+1)^{\text{ème}}$ composante de $\boldsymbol{\alpha}_{n+1}$, notée $\alpha_{n+1,m+1}$, fait fonction de terme de régularisation dans la fonction objectif. En reprenant les développements effectués dans le premier cas, on obtient l'équation de mise-à-jour suivante :

$$\boldsymbol{\alpha}_{n+1} = \begin{bmatrix} \boldsymbol{\alpha}_n \\ 0 \end{bmatrix} + \varrho\, \boldsymbol{H}_{n+1}^\top(\eta \boldsymbol{I}_q + \boldsymbol{H}_{n+1}\boldsymbol{H}_{n+1}^\top)^{-1} \left(\boldsymbol{d}_{n+1} - \boldsymbol{H}_{n+1} \begin{bmatrix} \boldsymbol{\alpha}_n \\ 0 \end{bmatrix} \right). \tag{8.23}$$

On retrouve alors une forme similaire à l'équation (8.22).

[10] Ne pas confondre cet ordre de projection affine q avec l'ordre du modèle qui n'est autre que la taille m du dictionnaire

Instructions	Expressions
Paramètres	
0.Seuil de cohérence	μ_0
0.Paramètre de régularisation	η
0.Ordre de projection affine	q
0.Pas de convergence	ϱ
Initialisation	
1. Définition du dictionnaire	$m = 1,\ \omega_m = 1,\ \mathcal{D}_m = \{\kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}$
2. Coefficient initial	$\boldsymbol{\alpha}_1 = d_1/\kappa(\boldsymbol{x}_1, \boldsymbol{x}_1)$
A chaque instant $n+1 \geq 2$	
3. Acquisition de nouvelles données	$(\boldsymbol{x}_{n+1}, d_{n+1})$
4. Si règle de cohérence vérifiée :	$\max_{j=1,\ldots,m} \lvert\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})\rvert \leq \mu_0$
a. Incrémentation de l'ordre	$m = m+1,\ \boldsymbol{\alpha}_n = [\boldsymbol{\alpha}_n^\top \ 0]^\top$
b. Mise-à-jour du dictionnaire	$\omega_m = n+1,\ \mathcal{D}_m = \mathcal{D}_{m-1} \cup \{\kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}$
5. Matrice $\boldsymbol{H}_{n+1}$, selon (8.18)	$(\boldsymbol{H}_{n+1})_{(i,j)} = \kappa(\boldsymbol{x}_{n-i+2}, \boldsymbol{x}_{\omega_j})$
6. Erreur *a priori*	$\boldsymbol{e}_{a,n+1} = \boldsymbol{d}_{n+1} - \boldsymbol{H}_{n+1}\boldsymbol{\alpha}_n$
7. Mise-à-jour des coefficients	$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n + \varrho\, \boldsymbol{H}_{n+1}^\top(\eta \boldsymbol{I}_q + \boldsymbol{H}_{n+1}\boldsymbol{H}_{n+1}^\top)^{-1}\boldsymbol{e}_{a,n+1}$

TAB. 8.2 – Pseudocode de l'algorithme de projection affine à noyau avec critère de cohérence.

8.3.2 Algorithme KAPA et complexité

L'algorithme KAPA est illustré en pseudo-code au Tableau 8.2. Il est principalement composé de la règle de cohérence (7.11), et des expressions de mise-à-jour (8.22) et (8.23). Le coût calculatoire de la matrice $\boldsymbol{H}_{n+1}$, de taille $q \times m$, dépend du noyau reproduisant κ considéré. Il est proportionnel à la taille du dictionnaire m ainsi qu'à l'ordre de projection q. En pratique, ce dernier est souvent choisi de sorte à être inférieur à 10. Chaque itération de l'algorithme nécessite l'inversion de $(\eta \boldsymbol{I}_q + \boldsymbol{H}_{n+1}\boldsymbol{H}_{n+1}^\top)$, ce qui représente $\mathcal{O}(q^2)$ opérations. Le coût calculatoire du critère de cohérence est similaire à celui étudié à la Section 8.2.2.

L'algorithme KAPA admet donc un coût calculatoire linéaire relativement à la taille du dictionnaire, alors qu'il est quadratique pour l'algorithme KRLS présenté à la Section 8.2. Cependant, le prix à payer est une convergence moins rapide. Une configuration particulière de l'algorithme KAPA est le choix $q = 1$, que l'on présente dans la section suivant.

8.3.3 Approximation instantanée : l'algorithme KNLMS

On considère dans ce qui suit la configuration particulier où $q = 1$. Le problème d'optimisation reste inchangé, avec le principe de fluctuation minimale (8.19). La contrainte (8.20) correspondant à l'annulation de l'erreur *a posteriori* instantanée est de la forme $d_{n+1} = \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_{n+1}$ où $\boldsymbol{h}_{n+1}^\top = [\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_1}) \cdots \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_m})]$. Les relations (8.22) et (8.23) se réduisent à

- Si $\max_{j=1,\ldots,m} \lvert\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})\rvert > \mu_0$,

$$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n + \frac{\varrho}{\eta + \|\boldsymbol{h}_{n+1}\|^2} \left(d_{n+1} - \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_n\right) \boldsymbol{h}_{n+1}. \tag{8.24}$$

- Si $\max_{j=1,\ldots,m} \lvert\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})\rvert \leq \mu_0$,

$$\boldsymbol{\alpha}_{n+1} = \begin{bmatrix} \boldsymbol{\alpha}_n \\ 0 \end{bmatrix} + \frac{\varrho}{\eta + \|\boldsymbol{h}_{n+1}\|^2} \left(d_{n+1} - \boldsymbol{h}_{n+1}^\top \begin{bmatrix} \boldsymbol{\alpha}_n \\ 0 \end{bmatrix} \right) \boldsymbol{h}_{n+1}. \tag{8.25}$$

Ces expressions correspondent à une approximation instantanée du gradient. Les deux équations de mise-à-jour de $\boldsymbol{\alpha}_n$, (8.24) et (8.25), ne sont autre qu'une forme d'algorithme LMS normalisé à noyau. L'erreur *a priori* correspond au terme entre parenthèses, soit $e_{a,n+1} = d_{n+1} - \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_n$. On désigne cet algorithme particulier par KNLMS, pour *kernel-based least-mean-squares*. Dans ce qui suit, on développe des variantes de cet algorithme.

8.4 Variantes

On développe dans cette section d'autres variantes de l'algorithme KAPA. En particulier, on propose de remplacer le critère de cohérence par le critère de Babel, ce dernier étant défini selon (7.12). Puis on étudie l'usage d'un modèle d'ordre fixe, en définissant à l'avance la taille du dictionnaire. On termine par la mise en œuvre d'algorithmes séquentiels d'ACP-à-noyau et d'AFD-à-noyau.

8.4.1 Contrôle de complexité par le critère de Babel

Tout au long de ce chapitre, on a développé des algorithmes de filtrage adaptatifs en contrôlant l'ordre du modèle à l'aide du critère de cohérence. Les fonctions noyau ainsi sélectionnées et formant un dictionnaire de faible cohérence vérifient plusieurs propriétés, exposées au chapitre précédent. Ce critère est facile à évaluer, avec un faible coût calculatoire comparé aux critères classiques. Néanmoins, il ne dépend que de la corrélation entre la fonction noyau candidate et l'élément du dictionnaire qui lui est le plus corrélé. La fonction de Babel permet une description plus fine de la structure du dictionnaire.

Le critère de Babel est défini ainsi. A l'instant $n+1$, on ajoute la nouvelle fonction noyau $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ au dictionnaire $\mathcal{D}_m$ si la somme des corrélations en valeur absolue entre elle et les éléments du dictionnaire ne dépasse pas un seuil donné μ_0, à savoir

$$\sum_{j=1}^{m} |\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})| \leq \mu_0. \tag{8.26}$$

Le niveau de parcimonie du modèle est déterminé par le seuil μ_0. Comme tous les éléments du dictionnaire vérifient cette règle, le seuil μ_0 détermine une borne supérieure de la fonction de Babel du dictionnaire. D'un point de vue algorithmique et calculatoire, la règle de Babel nécessite m additions supplémentaires à chaque itération comparée à la règle de cohérence (7.11). Cependant, il faut préciser que les termes de cette opération d'addition correspondent, à un signe près, aux éléments du vecteur colonne $\boldsymbol{h}_{n+1}$ déjà utilisés par les algorithmes KRLS, KAPA et KNLMS. Pour ce dernier à titre d'exemple, on calcule à chaque itération $\boldsymbol{h}_{n+1}^\top = [\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_1}) \cdots \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_m})]$ et on évalue la somme de ses éléments en valeur absolue. Si cette dernière est supérieure à un seuil μ_0 donné, on ajuste les coefficients du modèle selon $\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n + \frac{\varrho}{\eta + \|\boldsymbol{h}_{n+1}\|^2} (\boldsymbol{d}_{n+1} - \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_n) \boldsymbol{h}_{n+1}$. Dans le cas contraire, on incrémente l'ordre du modèle en augmentant la taille de $\boldsymbol{\alpha}_n$ par l'ajout d'une composante nulle, avant d'appliquer la règle ci-dessus pour calculer $\boldsymbol{\alpha}_{n+1}$.

8.4.2 Modèle d'ordre fixe

Nous avons montré que le critère de cohérence assure le caractère fini du dictionnaire. Cependant, il peut être toutefois intéressant de fixer l'ordre du modèle au préalable. A cet effet, on inclut une étape supplémentaire dans l'algorithme consistant à retirer un élément du dictionnaire à chaque fois que l'on en ajoute un. Il semble naturel d'utiliser le même critère pour ajouter et retirer une fonction noyau, la cohérence en l'occurrence. Dans ce contexte, plusieurs stratégies sont envisageables pour conserver un ordre de modèle fixé au préalable à m_0.

La première est définie ainsi. On commence par appliquer les algorithmes présentés précédemment, jusqu'à ce que le dictionnaire atteigne une taille m. On incorpore alors une étape supplémentaire d'élagage à chaque instant. On propose pour cela de retirer l'élément à-même de faire décroître la cohérence du dictionnaire , soit la fonction noyau $\kappa(\boldsymbol{x}_{\omega_k}, \cdot)$ telle que

$$\omega_k = \arg\max_{i \neq j} |\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})|.$$

On note que cette expression implique simultanément deux éléments du dictionnaire susceptibles d'être supprimés, les entrées $\boldsymbol{x}_{\omega_i}$ et $\boldsymbol{x}_{\omega_j}$) de $\kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$. Il est toutefois judicieux de retirer l'élément le plus ancien afin de prendre en compte la non-stationnarité du système étudié. Cette stratégie ne garantit pas une décroissance monotone de la cohérence du dictionnaire au cours des itérations.

Afin d'aboutir à une décroissance monotone de la cohérence, on propose une stratégie consistant à combiner les deux étapes, l'ajout et la suppression de fonctions noyau, en une seule étape dite de substitution. On commence l'apprentissage avec un dictionnaire de taille m_0 en y incluant les m_0 premières observations. La stratégie consiste alors à diminuer la cohérence du dictionnaire à chaque instant. Le critère de substitution est le suivant. En désignant par $\mu(n)$ la cohérence du dictionnaire à l'instant n (celle-ci étant arbitraire au début de l'apprentissage), on ajoute $\kappa(\boldsymbol{x}_n, \cdot)$ dans le dictionnaire et on en retire un élément $\kappa(\boldsymbol{x}_{\omega_k}, \cdot)$ si

$$\max_{\substack{j=1,\ldots,m_0 \\ j \neq k}} |\kappa(\boldsymbol{x}_n, \boldsymbol{x}_{\omega_j})| < \mu(n).$$

Par ce critère, on obtient une diminution monotone de la cohérence. Toutefois, nous avons noté que cette stratégie ne garantit pas de bonnes performances de l'algorithme. En effet, pour un système variant dans le temps, les éléments du dictionnaire obtenus dans un premier temps peuvent bloquer l'ajout de nouvelles fonctions noyau pertinentes par la suite.

Afin de profiter des avantages des deux stratégies précédentes, on propose une stratégie combinant les deux. L'algorithme nécessite deux paramètres, la taille du dictionnaire m_0 et le seuil de cohérence μ_0. On commence avec un dictionnaire constitué des m_0 premières fonctions noyau, et de cohérence arbitraire. A l'instant $n+1$, on ajoute la nouvelle fonction noyau $\kappa(\boldsymbol{x}_{n+1}, \cdot)$ au dictionnaire si le critère de cohérence (7.11) est vérifié, soit $\max_{j=1,\ldots,m_0} |\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})| \leq \mu_0$. Afin d'avoir un modèle d'ordre fixe, on propose ensuite de retirer la plus ancienne fonction noyau du dictionnaire. L'algorithme KNLMS à ordre de modèle fixe est illustré au Tableau 8.3[11].

8.4.3 Algorithmes séquentiels de méthodes à noyau

Nous avons présenté en Section 7.1 comment la minimisation d'un coût quadratique pouvait être mise au service de diverses méthodes à noyau classiques. On propose maintenant des versions séquentielles de celles-ci, à partir de l'algorithme KNLMS et du critère de cohérence développé dans ce chapitre.

Dans le cadre de la discrimination entre deux classes, on considère la modélisation d'un système dont la sortie correspond à l'étiquette désirée. En particulier, on pose $d_i \in \{+n/n_+, -n/n_-\}$ avec n_+ et n_- désignant le nombre d'observations de chacune des deux classes et n le nombre total. Ceci correspond au schéma (7.2), et la minimisation de l'erreur quadratique (7.1) produit un résultat similaire à celui obtenu par AFD-à-noyau. On peut alors exploiter les algorithmes développés tout au long de ce chapitre pour proposer une version séquentielle de l'algorithme AFD-à-noyau. On considère en particulier la mise en œuvre d'un algorithme de descente de gradient stochastique à cet effet, l'algorithme KNLMS

[11] Une initialisation plus correcte des coefficients du modèle consisterait à remplacer l'étape 2. d'initialisation aléatoire par $\boldsymbol{\alpha}_{m_0} = \boldsymbol{d}/(\boldsymbol{K} + \eta \boldsymbol{I}_{m_0})$, avec $(\boldsymbol{K})_{(i,j)} = \kappa(\boldsymbol{x}_{\omega_i}, \boldsymbol{x}_{\omega_j})$.

Instructions	Expressions
Paramètres	
0.Ordre du modèle	m_0
0.Seuil de cohérence	μ_0
0.Paramètre de régularisation	η
0.Pas de convergence	ϱ
Initialisation	
1. Définition du dictionnaire	$\omega_j = j$ pour $j = 1, \ldots, m_0$ $\mathcal{D} = \{\kappa(\boldsymbol{x}_{\omega_1}, \cdot), \ldots, \kappa(\boldsymbol{x}_{\omega_{m_0}}, \cdot)\}$
2. Coefficients du modèle	$\boldsymbol{\alpha}_j =$ rand pour $j = 1, \ldots, m_0$
A chaque instant $n+1 \geq m_0+1$	
3. Acquisition de nouvelles données	$(\boldsymbol{x}_{n+1}, d_{n+1})$
4. Si règle de cohérence vérifiée :	$\max_{j=1,\ldots,m_0} \lvert\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})\rvert \leq \mu_0$
a. Substitution dans le dictionnaire	$\omega_k = n+1$ avec $\omega_k = \arg\min_{j=1,\ldots,m_0} \omega_j$
b. Modification du $k^{\text{ème}}$ coefficient	$(\boldsymbol{\alpha}_n)_k = 0$
5. Vecteur de noyau	$\boldsymbol{h}_{n+1}^\top = [\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_1}) \cdots \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_m})]$
6. Erreur *a priori*	$e_{a,n+1} = d_{n+1} - \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_n$
7. Mise-à-jour des coefficients	$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n + \frac{\varrho}{\eta + \lVert\boldsymbol{h}_{n+1}\rVert^2} e_{a,n+1} \boldsymbol{h}_{n+1}$

TAB. 8.3 – Pseudocode de l'algorithme KNLMS à ordre fixe avec critère de cohérence.

avec le critère de cohérence. Cet algorithme n'impose pas de contrainte de norme sur les coefficients du modèle, comme le nécessite l'algorithme classique d'AFD-à-noyau dont le vecteur de coefficients est de norme unité. Pour introduire cette contrainte dans l'algorithme KNLMS, on propose une approche d'optimisation alternée. A chaque instant $n+1$, on applique dans un premier temps l'algorithme KNLMS comme exposé à la Section 8.3.3, avec les expressions (8.24) et (8.25), pour déterminer le vecteur $\boldsymbol{\alpha}_{n+1}$. Dans un second temps, on opère une projection de ce dernier sur l'ensemble des vecteurs admissibles, $\|\boldsymbol{\alpha}\|^2 = 1$, avec une normalisation selon $\boldsymbol{\alpha}_{n+1}/\|\boldsymbol{\alpha}_{n+1}\|^2$. L'algorithme final est présenté au Tableau 8.4.

Dans un cadre non-supervisé, on rappelle que le critère de cohérence permet de sélectionner les fonctions noyau les plus pertinentes parmi un ensemble de fonctions noyau disponibles. Cette pertinence admet une signification au sens de l'ACP, comme exposé au Chapitre 7 par la Proposition 7.7. On rappelle que l'on y montre que le sous-espace engendré par les fonctions noyau retenues permet une bonne approximation des composantes principales les plus pertinentes. Dans le cadre de l'ACP-à-noyau, on cherche à déterminer la fonction qui maximise la variance des données projetées, parmi les fonctions de norme unité. On est alors amené à la résolution du problème d'optimisation sous contrainte suivant :

$$\psi^* = \arg\max_{\psi \in \mathcal{H}} \sum_i |\psi(\boldsymbol{x}_i)|^2,$$

sous la contrainte $\|\psi^*\|_{\mathcal{H}} = 1$. Pour résoudre ce problème, on propose d'adapter l'algorithme KNLMS en particulier. Pour cela, on reprend le problème de minimisation du risque quadratique (7.1) où la sortie désirée est fixée à zero pour chaque élément de l'ensemble d'apprentissage. On aboutit à la minimisation de $\frac{1}{n}\sum_{i=1}^n (\psi(\boldsymbol{x}_i))^2$ sans le terme de régularisation. La maximisation de cette fonctionnelle correspond à une ACP. On peut alors proposer à cet effet un algorithme séquentiel par montée de gradient stochastique, en modifiant les équations de mise-à-jour (8.24)-(8.25). Une normalisation de l'axe principal est également nécessaire. Pour cela, on a recours à l'optimisation alternée, comme présentée au paragraphe précédent. On illustre le pseudo-code de cet algorithme au Tableau 8.5. Il faut noter que l'algorithme fournit un axe principal. Pour les autres, il convient de procéder par déflation. Ceci sort du contexte du présent travail.

Instructions	Expressions
Paramètres	
0.Seuil de cohérence	μ_0
0.Paramètre de régularisation	η
0.Pas de convergence	ϱ
Initialisation	
1. Définition du dictionnaire	$m = 1,\ \omega_m = 1,\ \mathcal{D}_m = \{\kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}$
2. Coefficient initial	$\boldsymbol{\alpha}_1 = d_1/\kappa(\boldsymbol{x}_1, \boldsymbol{x}_1)$
A chaque instant $n+1 \geq 2$	
3. Acquisition de nouvelles données	$(\boldsymbol{x}_{n+1}, d_{n+1})$
4. Si règle de cohérence vérifiée :	$\max_{j=1,\ldots,m} \lvert\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})\rvert \leq \mu_0$
a. Incrémentation de l'ordre	$m = m+1,\ \boldsymbol{\alpha}_n = [\boldsymbol{\alpha}_n^\top\ 0]^\top$
b. Mise-à-jour du dictionnaire	$\omega_m = n+1,\ \mathcal{D}_m = \mathcal{D}_{m-1} \cup \{\kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}$
5. Vecteur de noyau	$\boldsymbol{h}_{n+1}^\top = [\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_1}) \cdots \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_m})]$
6. Erreur *a priori*	$e_{a,n+1} = d_{n+1} - \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_n$
7. Mise-à-jour des coefficients	$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n + \frac{\varrho}{\eta + \lVert\boldsymbol{h}_{n+1}\rVert^2}\, e_{a,n+1}\, \boldsymbol{h}_{n+1}$
8. Normalisation des coefficients	$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_{n+1}/\lVert\boldsymbol{\alpha}_{n+1}\rVert^2$

TAB. 8.4 – Pseudocode de l'algorithme séquentiel d'AFD-à-noyau à partir de l'algorithme KNLMS avec critère de cohérence, avec $d_n \in \{+n/n_+, -n/n_-\}$ selon l'appartenance du $n^{\text{ème}}$ élément à l'une des deux classes.

Instructions	Expressions
Paramètres	
0.Seuil de cohérence	μ_0
0.Paramètre de régularisation	η
0.Pas de convergence	ϱ
Initialisation	
1. Définition du dictionnaire	$m = 1,\ \omega_m = 1,\ \mathcal{D}_m = \{\kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}$
2. Coefficient initial	$\boldsymbol{\alpha}_1 = \texttt{rand}/\kappa(\boldsymbol{x}_1, \boldsymbol{x}_1)$
A chaque instant $n+1 \geq 2$	
3. Acquisition d'une nouvelle donnée	$\boldsymbol{x}_{n+1}$
4. Si règle de cohérence vérifiée :	$\max_{j=1,\ldots,m} \lvert\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_j})\rvert \leq \mu_0$
a. Incrémentation de l'ordre	$m = m+1,\ \boldsymbol{\alpha}_n = [\boldsymbol{\alpha}_n^\top\ 0]^\top$
b. Mise-à-jour du dictionnaire	$\omega_m = n+1,\ \mathcal{D}_m = \mathcal{D}_{m-1} \cup \{\kappa(\boldsymbol{x}_{\omega_m}, \cdot)\}$
5. Vecteur de noyau	$\boldsymbol{h}_{n+1}^\top = [\kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_1}) \cdots \kappa(\boldsymbol{x}_{n+1}, \boldsymbol{x}_{\omega_m})]$
6. Mise-à-jour des coefficients	$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_n - \frac{\varrho}{\eta + \lVert\boldsymbol{h}_{n+1}\rVert^2}\, \boldsymbol{h}_{n+1}^\top \boldsymbol{\alpha}_n\, \boldsymbol{h}_{n+1}$
7. Normalisation des coefficients	$\boldsymbol{\alpha}_{n+1} = \boldsymbol{\alpha}_{n+1}/\lVert\boldsymbol{\alpha}_{n+1}\rVert^2$

TAB. 8.5 – Pseudocode de l'algorithme séquentiel d'ACP-à-noyau à partir de l'algorithme KNLMS avec critère de cohérence.

Chapitre 9

Applications

Sommaire

L'objet de ce chapitre est d'illustrer les algorithmes développés tout au long de ce manuscrit par le traitement de signaux synthétiques et réels. On s'intéresse tout d'abord à la question de l'identification de systèmes non-linéaires et non-stationnaires avant de revenir vers des considérations temps-fréquence.

9.1 Filtrage adaptatif non-linéaire

9.1.1 L'algorithme KRLS – Première application

En guise de première application, on considère le système dynamique non-linéaire défini par les entrées x_n et les sorties désirées d_n telles

$$\begin{cases} v_n = 1.1 \, \exp(-|v_{n-1}|) + x_n \\ d_n = v_n^2 \end{cases} \tag{9.1}$$

et la condition initiale $v_0 = 0.5$. Les entrées x_n suivent une distribution Gaussienne de moyenne nulle et d'écart type 0.25. Les sorties du système sont noyées dans un bruit blanc Gaussien de moyenne nulle et de variance 1, correspondant à un rapport signal-sur-bruit de l'ordre de -4.0 dB. On se propose ici d'identifier un modèle non-linéaire de la forme $d_n = \psi(x_n)$ à l'aide de l'algorithme KRLS à mémoire infinie, soit $\theta = 1$. Dans une étape dite hors-ligne, nous avons déterminé la meilleure configuration de l'algorithme à partir de 10 séquences indépendantes de $2\,000$ échantillons. Les performances ont été estimées par l'erreur quadratique moyenne de prédiction sur les 500 derniers échantillons de chaque séquence, et moyennées sur toutes les séquences. Le noyau de Laplace nous est apparu le mieux adapté,

pour une largeur $\sigma_0 = 0.35$ et un seuil de cohérence $\mu_0 = 0.3$. Ces paramètres ont été obtenus par une recherche sur la grille définie par $(0.1 \leq \sigma_0 \leq 1) \times (0.05 \leq \mu_0 \leq 0.5)$, avec les pas $\Delta\sigma_0 = 0.05$ et $\Delta\mu_0 = 0.05$. Nous avons déterminé dans le même temps le paramètre de régularisation $\eta = 10^{-4}$ par une exploration de l'intervalle $[10^{-6}, 10^{-1}]$ avec un pas logarithmique.

Nous avons déterminé les caractéristiques du modèle résultant sur 50 séquences indépendantes de $10\,000$ échantillons. Ceci nous a conduit à un ordre m du modèle à noyau égal en moyenne à 5.4, et à une valeur moyenne de la fonction de Babel de 0.56. En vertu de la Proposition 7.3, il apparaît que les fonctions noyau du dictionnaire étaient très souvent, voire toujours, linéairement indépendantes. Pour étudier les performances en régime statique, nous avons mesuré l'erreur quadratique moyenne normalisée de prédiction sur les derniers $5\,000$ échantillons de chaque séquence, soit

$$\frac{\sum_{n=5001}^{10000}(d_n - \psi_{n-1}(\boldsymbol{x}_n))^2}{\sum_{n=5001}^{10000} d_n^2}.$$

L'erreur moyenne que nous avons obtenue sur les 50 séquences indépendantes est de 0.0711. La Figure 9.1 illustre la convergence de l'algorithme KRLS pour différentes valeurs de largeur σ_0 du noyau de Laplace. Les courbes de l'erreur quadratique ont été lissées en les moyennant sur 20 échantillons consécutifs.

On se propose de comparer notre méthode KRLS à celle de Engel *et coll.* [EMM04] qui repose sur le critère d'approximation linéaire (7.6). Pour déterminer les meilleurs paramètres, nous avons opéré comme précédemment par une recherche sur la grille $(0.1 \leq \sigma_0 \leq 1) \times (0.6 \leq \eta_0^2 \leq 0.95)$, par pas $\Delta\sigma_0 = 0.05$ et $\Delta\eta_0^2 = 0.05$ pour l'ajustement de la largeur σ_0 du noyau et du seuil d'approximation linéaire η_0^2. Nous avons abouti à $\sigma_0 = 0.35$ et $\eta_0^2 = 0.9$. L'erreur quadratique moyenne normalisée de prédiction que nous avons obtenue est de 0.0733, légèrement supérieure à celle obtenue ci-dessus. Comme illustrée à la Figure 9.2, la vitesse de convergence est similaire pour les deux méthodes. Toutefois, le modèle de Engel *et coll.* est légèrement plus parcimonieux, avec un ordre de 4.46 en moyenne. Avec la même largeur de noyau dans les deux cas, on précise que 62% des fonctions noyau ont été communes en moyenne aux deux approches pour chaque séquence de données. Ce résultat est à rapprocher du lien établi entre les deux critères de parcimonie, précisée par la Proposition 7.4 et étudiée à la Section 7.3.2, en remarquant que les seuils utilisés ont été déterminés séparément. On rappelle finalement que la règle d'approximation linéaire (7.6) nécessite un calcul plus élaboré pour un coût calculatoire plus important par rapport à la règle de cohérence (7.11).

Dans une seconde série d'expérimentations, on se propose de comparer les performances des deux approches dans le cas où les entrées x_n sont échantillonnées selon un processus Gaussien coloré. Pour cela, nous avons repris les mêmes configurations d'expérimentation que ci-dessus, à l'exception de x_n que nous avons filtré à l'aide d'un système linéaire de réponse impulsionnelle $[0.9045\ 1\ 0.9045\ 0]$. Ce filtrage conduit à un rapport signal-sur-bruit de -3.78 dB. Avec la méthode KRLS, l'ordre des modèles auxquels on a abouti est de 8.64 en moyenne, et l'erreur quadratique moyenne normalisée de prédiction de 0.044. Avec l'approche de Engel *et coll.*, l'ordre du modèle et l'erreur quadratique moyenne normalisée de prédiction ont été respectivement égaux à 7.74 et 0.048. La Figure 9.2 montre que les deux méthodes admettent une vitesse de convergence très similaire.

9.1.2 L'algorithme KRLS – Seconde application

Dans une second application, on considère le système non-linéaire décrit par l'équation

$$d_n = \left(0.8 - 0.5\,\exp(-d_{n-1}^2)\right) d_{n-1} - \left(0.3 + 0.9\,\exp(-d_{n-1}^2)\right) d_{n-2} + 0.1\,\sin(d_{n-1}\pi) \qquad (9.2)$$

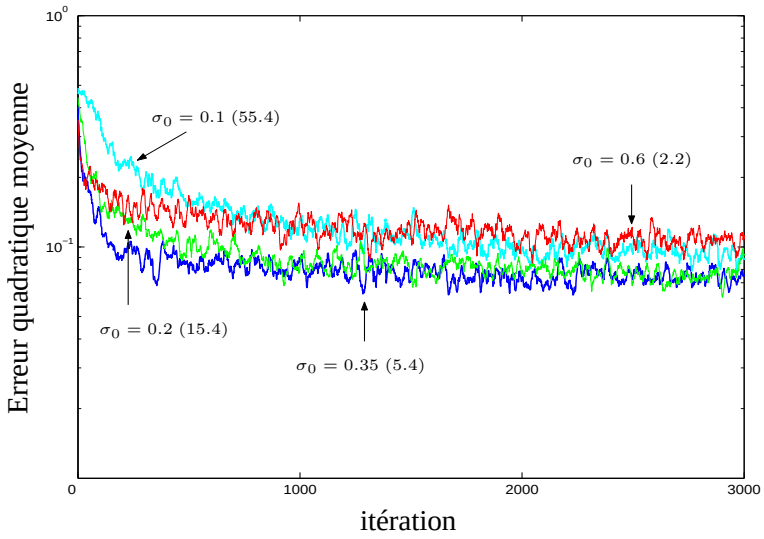


FIG. 9.1 – Convergence de l'algorithme KRLS pour différentes largeurs du noyau de Laplace. L'ordre de chaque modèle est donné par sa valeur moyenne entre parenthèses.

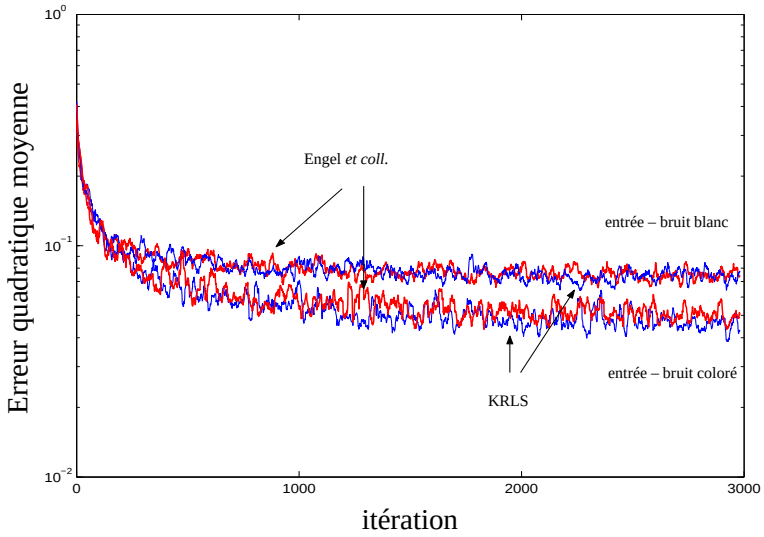


FIG. 9.2 – Comparaison des vitesses de convergence de l'algorithme KRLS et de celui proposé par Engel *et coll.* pour des entrées Gaussiennes blanches ou colorées.

où d_n désigne la sortie désirée. Cette série temporelle a fait l'objet de plusieurs travaux dont [CB86, DH02, DKH03]. Afin d'identifier le système, nous avons utilisé la même configuration que celle proposée dans [DKH03]. Les données ont été générées par l'équation (9.2) à partir de l'état initial $(0.1, 0.1)$. Les sorties désirées ont été noyées dans un bruit de mesure, échantillonné selon une distribution Gaussienne de moyenne nulle et de variance égale à 0.01. Il en a résulté un rapport signal-bruit de 17.2 dB. A partir de ces données, nous avons estimé un modèle non-linéaire de la forme $d_n = \psi(d_{n-1}, d_{n-2})$. Nous avons choisi le noyau Gaussien comme dans [DKH03]. Des expérimentations hors-ligne ont été conduites sur des séquences de $1\,500$ échantillons afin de déterminer les meilleurs seuils μ_0 et η_0 pour chaque algorithme. Nous avons observé qu'un bon compromis entre l'ordre du modèle et son erreur de prédiction était obtenu pour un seuil de cohérence $\mu_0 = 0.25$. Le paramètre de régularisation $\eta = 10^{-4}$

a été conjointement déterminé par une recherche sur l'intervalle $[10^{-6}, 10^{-1}]$ avec un pas logarithmique. Dans un premier temps, nous avons considéré le cas $\theta = 1$, correspondant à une mémoire infinie. Pour étudier le modèle résultant, nous avons généré un ensemble de 50 séquences de $10\,000$ échantillons chacune. L'erreur quadratique moyenne normalisée de prédiction que l'on a obtenue sur les $5\,000$ derniers échantillons de chaque séquence est égale à 0.0193. L'ordre des modèles correspondant était de 11.68 en moyenne, et la fonction de Babel inférieure à 1 ce qui constitue une condition suffisante pour que les fonctions noyau du dictionnaire soient linéairement indépendantes. On illustre à la Figure 9.3 le comportement de l'algorithme en terme de convergence pour différentes valeurs de seuil μ_0. Ces courbes ont été lissées en les moyennant sur 20 observations consécutives. L'algorithme de Engel *et coll.* [EMM04] a été configuré et testé dans les même conditions, le meilleur seuil ϵ_0 étant égal à 0.9. Nous avons obtenu une erreur quadratique moyenne normalisée de prédiction de 0.0186, et un ordre de modèle de 12.32. La valeur moyenne de la fonction de Babel à laquelle on a abouti était égale à 0.98 et, en moyenne, 65% des fonctions noyau retenues étaient communes aux deux approches à la fin de chaque séquence d'apprentissage. En comparant le comportement des deux méthodes à la Figure 9.4, on constate que leurs vitesses de convergence sont très proches.

On se propose d'étudier le comportement de l'algorithme KRLS pour un suivi de système variant dans le temps, passant d'un modèle à un autre. Pour cela, nous avons considéré un signal de $15\,000$ échantillons d_n répartis en trois séquences, la première de $5\,000$ échantillons donnés par l'équation (9.2), suivie d'une séquence de $5\,000$ observations générées selon l'équation

$$d_n = 1.5\left(0.8 - \exp(-d_{n-1}^2)\right) d_{n-1} - 1.2\left(0.3 + 0.9\exp(-2\,d_{n-1}^2)\right) d_{n-2} + 0.2\,\sin(d_{n-1}\pi + \pi). \quad (9.3)$$

La dernière séquence était formée de $5\,000$ échantillons donnés par le premier modèle. Chacune de ces trois séquences a été générée à partir de l'état initial $(0.1, 0.1)$, causant ainsi un changement brusque du signal d_n à chaque modification d'état. Les sorties d_n ont été noyées dans un bruit de mesure échantillonné selon une distribution Gaussienne de moyenne nulle et de variance égale à 0.01. Nous avons utilisé les mêmes noyaux Gaussiens, seuils de cohérence μ_0 et paramètres de régularisation η que ci-dessus pour estimer un modèle non-linéaire de la forme $d_n = \psi(d_{n-1}, d_{n-2})$. Les résultats obtenus ont été moyennés sur 50 évaluations. La Figure 9.5 illustre la vitesse de convergence de l'algorithme KRLS lorsque $\theta = 0.99$. On représente sur la même figure le résultat obtenu pour les systèmes (9.2) et (9.3) séparément. L'algorithme proposé offre manifestement des propriétés de poursuite, et par conséquent est adapté aux environnements non-stationnaires. A la fin de la première séquence de la série d_n, l'ordre du modèle était de 12.13, en moyenne. Seules 3.70 fonctions noyau ont été ajoutées au dictionnaire durant la seconde séquence, produisant ainsi un modèle d'ordre égal à 15.83 en moyenne. La troisième séquence n'a pas contribué à une augmentation significative de l'ordre du modèle, celui-ci valant 15.95 à la fin de l'apprentissage. A titre comparatif, nous avons repris les mêmes configurations que ci-dessus, avec 50 séquences de $15\,000$ observations bruitées obtenues par le second système (9.3). L'ordre du modèle que nous avons obtenu est égal en moyenne à 14.40. Ce résultat montre que la technique proposée pour la sélection de fonctions noyau est capable de détecter et réutiliser l'information pertinente de l'état courant et des états précédents. Le prix à payer est un accroissement de l'ordre du modèle, et par suite une augmentation de la complexité calculatoire de l'algorithme. Afin de combler ce défaut, on peut recourir à des techniques à ordre fixe, comme celles développées dans la Section 8.4.2, avec l'inconvénient du choix préalable de l'ordre du modèle.

9.1.3 L'algorithme KAPA

Pour étudier l'algorithme KAPA, nous avons repris le système défini par l'équation (9.1) afin de le comparer à l'algorithme KRLS. Nous avons utilisé la même configuration du système, le noyau de

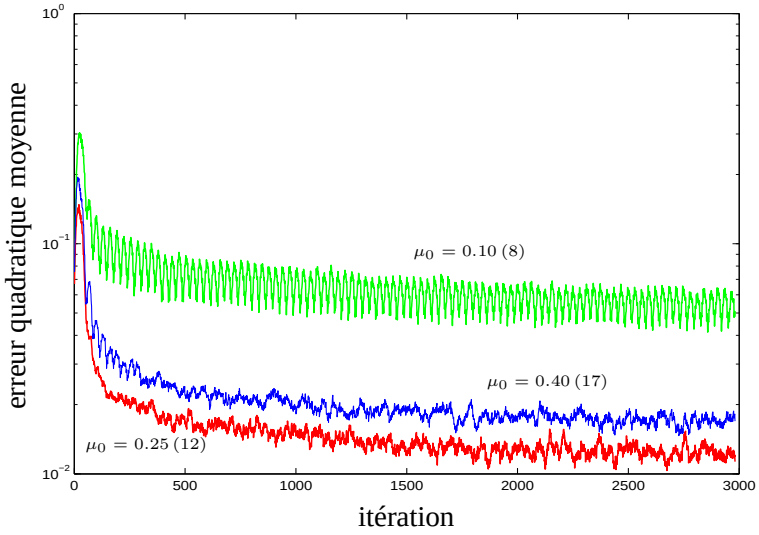


FIG. 9.3 – Convergence de l'algorithme KRLS pour différentes valeurs du seuil μ_0. L'ordre de chaque modèle est donné par sa valeur moyenne entre parenthèses.

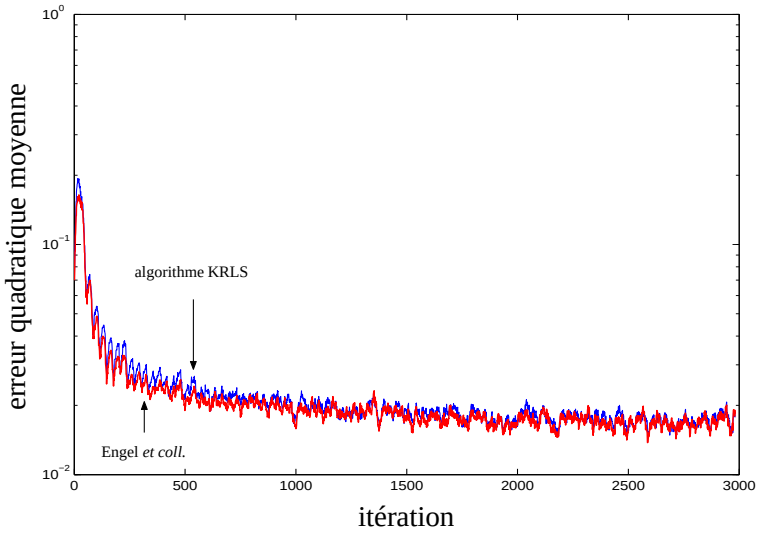


FIG. 9.4 – Comparaison des vitesses de convergence entre l'algorithme KRLS et celui de Engel *et coll.*.

Laplace avec $\sigma_0 = 0.35$ et le seuil de cohérence $\mu_0 = 0.3$. Le modèle à estimer d_n est de la forme $d_n = \psi(\boldsymbol{x}_n)$. Une expérimentation hors-ligne sur des séquences de $5\,000$ échantillons a été conduite afin de déterminer les paramètres de l'algorithme KAPA, en particulier l'ordre de projection affine p, le pas de convergence ϱ et le paramètre de régularisation η. La mesure de performance retenue est l'erreur quadratique moyenne sur les 500 derniers échantillons de chaque séquence, moyennée sur les 50 séquences. L'ordre de projection a été fixé à $p = 5$ car les valeurs inférieures produisaient des modèles moins performants. Les meilleurs paramètres ϱ et η correspondent aux valeurs 9×10^{-4} et 7×10^{-3}. Ces valeurs ont été obtenues par une recherche dans $(10^{-4} \leq \varrho \leq 10^{-1}) \times (10^{-3} \leq \eta \leq 10^{-2})$ avec un pas logarithmique 2×10^{-k} dans chaque intervalle $[10^{-k},\, 10^{-k+1}]$.

Nous avons étudié l'algorithme KAPA sur une série de 50 séquences de $5\,000$ observations. L'ordre des modèles que nous avons obtenus est égal à 5.4, en moyenne, comme dans le cas de l'algorithme

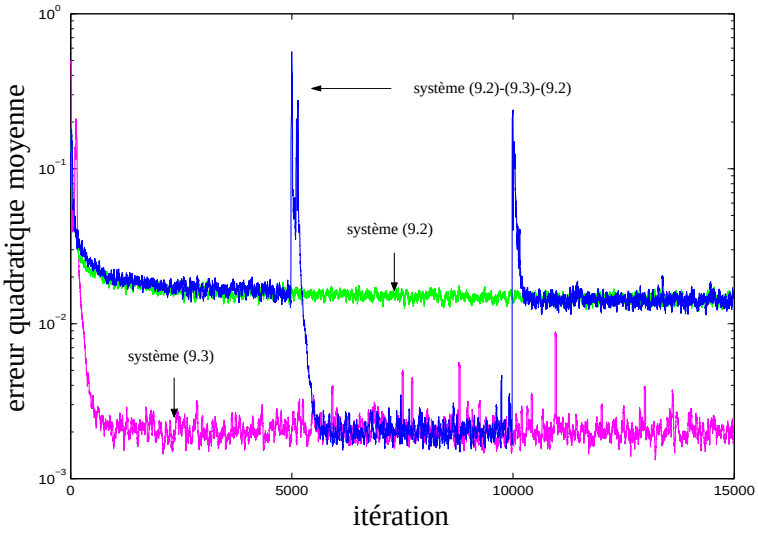


FIG. 9.5 – Convergence de l'algorithme KRLS pour un suivi de système.

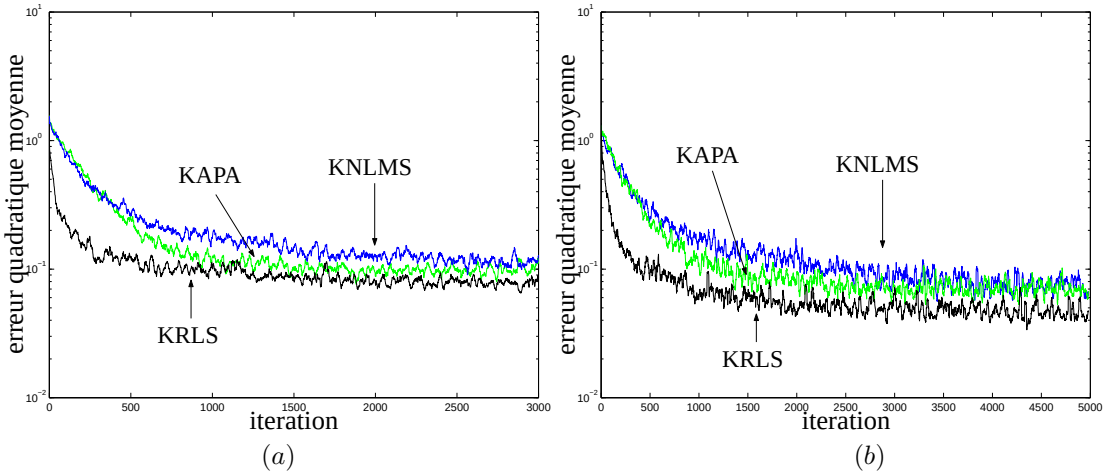


FIG. 9.6 – Convergence des algorithmes KAPA, KNLMS et KRLS pour des entrées (a) blanches et (b) colorées.

KRLS puisque nous avons utilisé le même critère de parcimonie. Les performances en régime statique ont été évaluées grâce à l'erreur quadratique moyenne de prédiction, en considérant les 500 dernières observations, soit $\frac{1}{500}\sum_{n=4501}^{5000}(d_n - \psi_{n-1}(\boldsymbol{x}_n))^2$. Avec l'algorithme KAPA, nous avons obtenu une valeur moyenne de 0.0839 sur les 50 séquences. Cette erreur est supérieure à celle de l'algorithme KRLS, estimée à 0.0711. La complexité algorithmique de l'algorithme KAPA est en revanche moindre. Pour l'algorithme KNLMS, correspondant à $p = 1$, nous avons abouti aux meilleures performances avec $\varrho = 5 \times 10^{-3}$ et $\eta = 9 \times 10^{-3}$. L'ordre du modèle n'a pas varié, 5.4 en moyenne. Avec la même mesure de performance, nous avons obtenu 0.0896 pour l'algorithme KNLMS. Comme prévu, la complexité calculatoire réduite de ce dernier pèse sur ses propriétés de convergence par rapport aux algorithmes KAPA et KRLS. Pour comparer le comportement transitoire de KAPA et KNLMS, nous avons reconfiguré les paramètres du premier afin qu'il ait les mêmes performances en état statique que le second. Nous avons abouti alors à $p = 3$, $\varrho = 3 \times 10^{-3}$ et $\eta = 5 \times 10^{-2}$. La Figure 9.6 (a) illustre l'évolution de l'erreur quadratique moyenne pour les trois algorithmes, après les avoir lissées sur 20 échantillons consécutifs. La vitesse de convergence de KAPA est intermédiaire entre celles de KRLS et de KNLMS. Nous avons

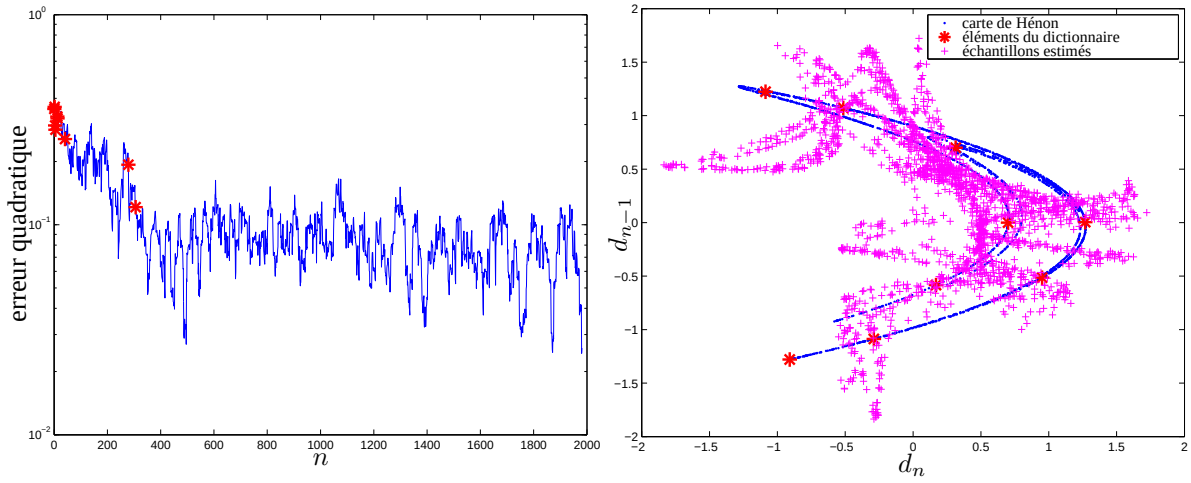


FIG. 9.7 – Evolution de l'erreur quadratique (a) entre la sortie désirée d_n et celle estimée $\psi(d_{n-1}, d_{n-2})$, pour $\mu_0 = 0.1$. Les deux séries sont comparées en (b) dans le plan de phase. Les éléments du dictionnaire sont indiqués par des étoiles.

repris l'expérience où x_n est donnée par un processus Gaussien coloré, obtenue en filtrant un bruit blanc Gaussien à l'aide d'un filtre de réponse impulsionnelle $[0.9045\ 1\ 0.9045\ 0]$. L'ordre des modèles auxquels on a abouti était en moyenne de 8.64 pour KAPA et KNLMS, comme c'est le cas avec KRLS. La Figure 9.6 (b) illustre l'évolution de l'erreur quadratique moyenne pour les trois algorithmes.

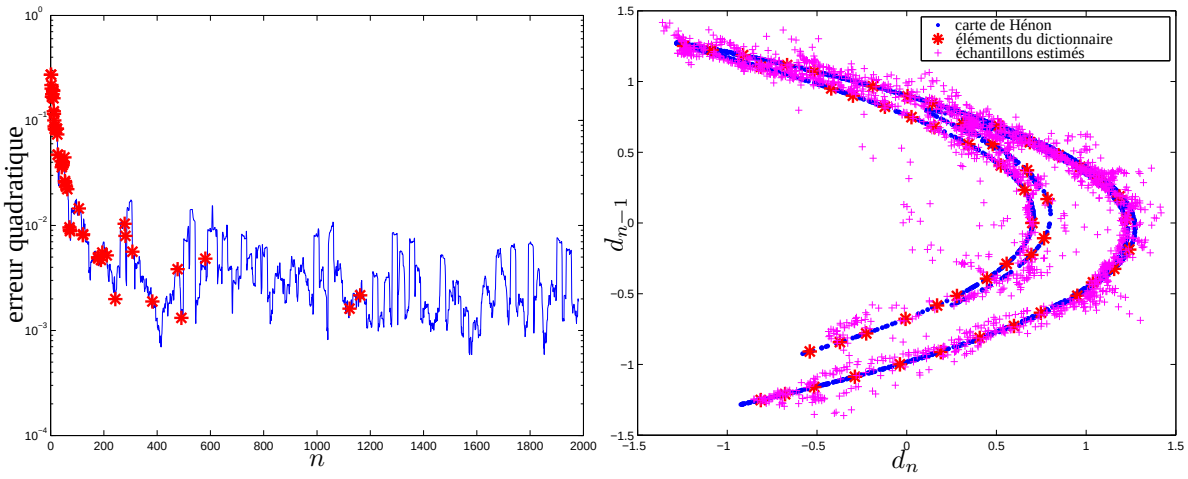
9.2 Applications diverses

9.2.1 Modélisation de systèmes chaotiques – la carte de Hénon

La modélisation de systèmes non-linéaires a fait l'objet de plusieurs travaux, et en particulier les systèmes dits à comportement chaotique. L'un des plus étudiés est le système non-linéaire de Hénon, décrit par l'équation

$$d_n = 1 - a_1\, d_{n-1}^2 + a_2\, d_{n-2}.$$

On s'intéresse en particulier à la configuration $a_1 = 1.4$ et $a_2 = 0.3$, et à l'état initial donné par $d_0 = -0.3$ et $d_1 = 0$, pour laquelle la série temporelle présente un comportement chaotique. On souhaite modéliser cette série par une expression non-linéaire de la forme $d_n = \psi(d_{n-1}, d_{n-2})$ avec l'algorithme KRLS, à partir de $2\,000$ échantillons. Après avoir testé plusieurs noyaux, dont on a ajusté les paramètres par validation croisée, on a observé que le noyau Gaussien avec $\sigma_0 = 0.35$ menait à des performances satisfaisantes en terme d'erreur de prédiction. Dans un premier temps, on a fixé le seuil de cohérence à $\mu_0 = 0.1$, ce qui a conduit à un dictionnaire de $m = 9$ éléments vérifiant (7.11). On note que le dictionnaire obtenu satisfait à la condition suffisante d'indépendance linéaire, soit $\mu_0 < 1/(m-1)$. L'évolution de l'erreur quadratique de prédiction est représentée à la Figure 9.7 (a), où l'on indique par des étoiles les éléments retenus pour le dictionnaire. On présente à la Figure 9.7 (b) la carte de Hénon, c'est-à-dire les données dans le plan (d_n, d_{n-1}), les prédictions ainsi que les éléments du dictionnaire. Si ces derniers semblent choisis de manière pertinente sur les paraboles marquant les lieux de Hénon, on remarque que les prédictions s'écartent sensiblement de ce qui est attendu. Afin d'améliorer les résultats, nous avons choisi d'augmenter la taille du dictionnaire en rehaussant le seuil de cohérence μ_0 à 0.6. Bien que la condition suffisante d'indépendance linéaire ne soit plus vérifiée, on note une amélioration



(a) (b)

FIG. 9.8 – Résultats obtenus pour un seuil de cohérence $\mu_0 = 0.6$

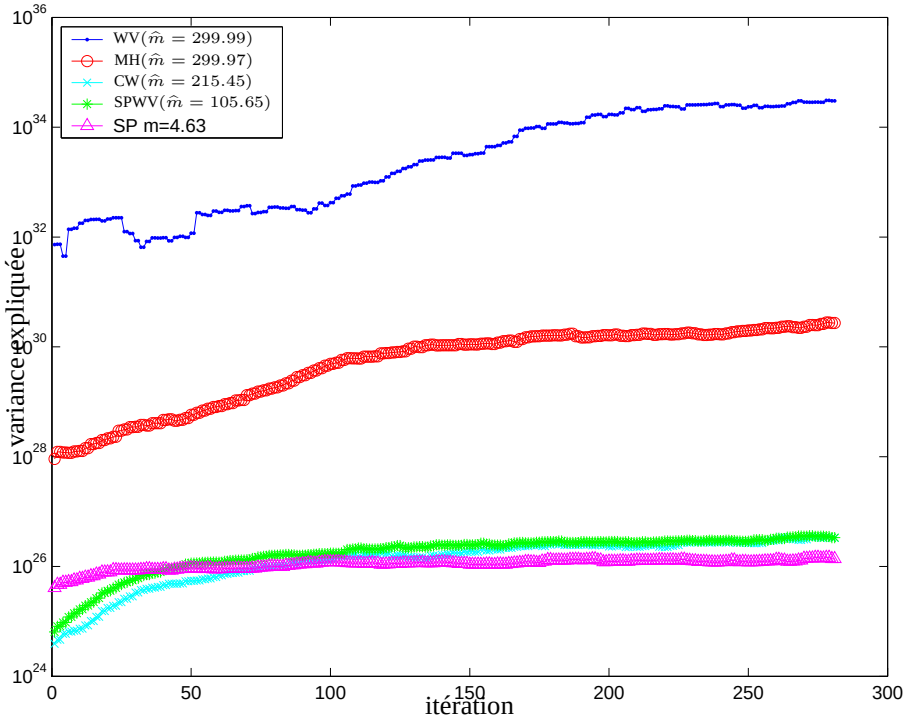


FIG. 9.9 – Evolution de la variance expliquée par la signature principale obtenue pour différentes distributions temps-fréquence de la classe de Cohen. Le seuil de Cohérence étant fixé à $\mu_0 = 0.46$, le nombre d'éléments dans le dictionnaire est donné dans la légende, moyenné sur 100 réalisation. L'algorithme utilisé est le KNLMS

des performances grâce aux 52 éléments retenus pour le dictionnaire. En Figure 9.8, on représente (a) l'évolution de l'erreur quadratique résultante, et (b) les cartes de Hénon des observations et de leur prédiction. On aboutit à une erreur de prédiction plus faible, au prix d'un ordre de modèle plus élevé, donc d'une complexité plus importante.

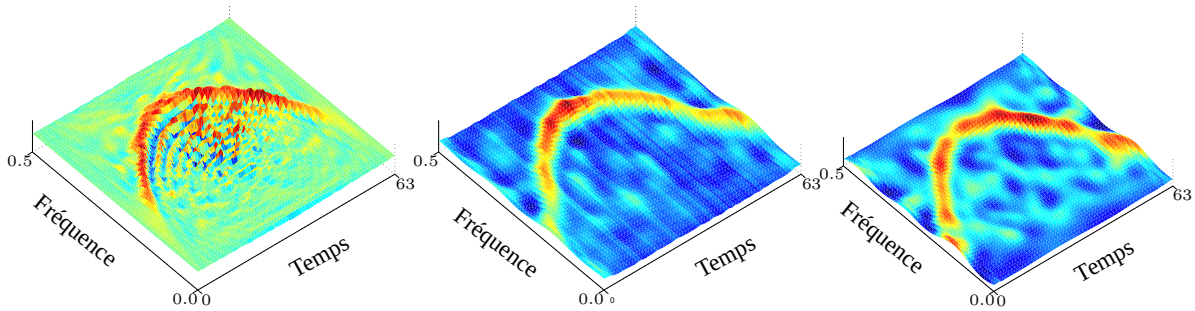


FIG. 9.10 – Signatures temps-fréquences obtenues par l'usage d'un algorithme d'ACP en-ligne avec critère de cohérence dans la même configuration que celle de la Figure 9.9 avec (a) la distribution de Wigner, (b) la distribution de Choï-Williams, et (c) la pseudo-Wigner lissée.

9.2.2 ACP en-ligne pour l'analyse de systèmes non-stationnaires

Tout au long de ces 3 derniers chapitres, nous avons présenté le problème d'estimation fonctionnelle dans le cadre général de la régression, où l'on cherche une fonction qui produit la réponse désirée associée à une entrée donnée. Nous nous sommes également intéressés aux liens de cette problématiques avec celles d'autres méthodes de reconnaissance des formes, en particulier l'AFD et l'ACP à noyau. Dans cette section, on se contente d'étudier la seconde pour l'analyse de signaux non-stationnaires dans le domaine temps-fréquence.

On considère un système produisant des signaux non-stationnaires, à modulation fréquentielle parabolique noyée dans un bruit blanc Gaussien avec un rapport signal-sur-bruit de l'ordre de -9 dB. On cherche alors à déterminer cette signature principale à partir de 300 signaux de taille 64 échantillons, par la mise en œuvre de l'algorithme séquentiel d'ACP-à-noyau pour différentes distributions temps-fréquence[12]. Dans un premier temps, nous avons fixé le seuil de cohérence à $\mu_0 = 0.46$, produisant un dictionnaire d'environ 100 éléments dans le cas de la pseudo-Wigner-lissée. On illustre à la Figure 9.9 l'évolution de la variance expliquée, moyennée sur 100 réalisations, dans le cas des distributions de Wigner (WV), de Margeneau-Hill (MH), de Choï-Williams (CW), de pseudo-Wigner-lissée (SPWV), et du spectrogramme (S). Le caractère non-monotone observé sur certaines de ces tracés est principalement dû à la technique d'optimisation alternée considérée dans l'algorithme d'ACP en-ligne. En effet, on opère à chaque itération une maximisation de la variance puis une normalisation par projection dans l'ensemble des fonctions admissibles. On représente à la Figure 9.10 certaines des signatures temps-fréquence résultantes, en précisant la valeur moyenne de l'ordre du modèle pour chacune des distributions temps-fréquence. Ces résultats concordent avec ceux obtenus tout au long de ce manuscrit sur la comparaison entre différentes distributions temps-fréquence, voir par exemple la Figure 6.3. Il faut toutefois préciser que le seuil de cohérence choisi n'a pas conduit à une solution parcimonieuse dans le cas de la distribution de Wigner, avec $\widehat{m} = 299.99$, alors que le spectrogramme a permis une sélection assez stricte des données d'apprentissage, avec seulement $\widehat{m} = 4.63$ signaux sélectionnés parmi les 300 disponibles. Pour remédier à cela, il convient de diminuer le seuil de cohérence dans le premier cas, et de le relever dans le second cas. Ceci est illustré à la Figure 9.11, où l'on a fixé le seuil μ_0 à 0.15 pour le premier, et à 0.65 pour le second.

[12] Les paramètres de lissage sont considérés par défaut dans la Toolbox Temps-Fréquence [AFGL05]

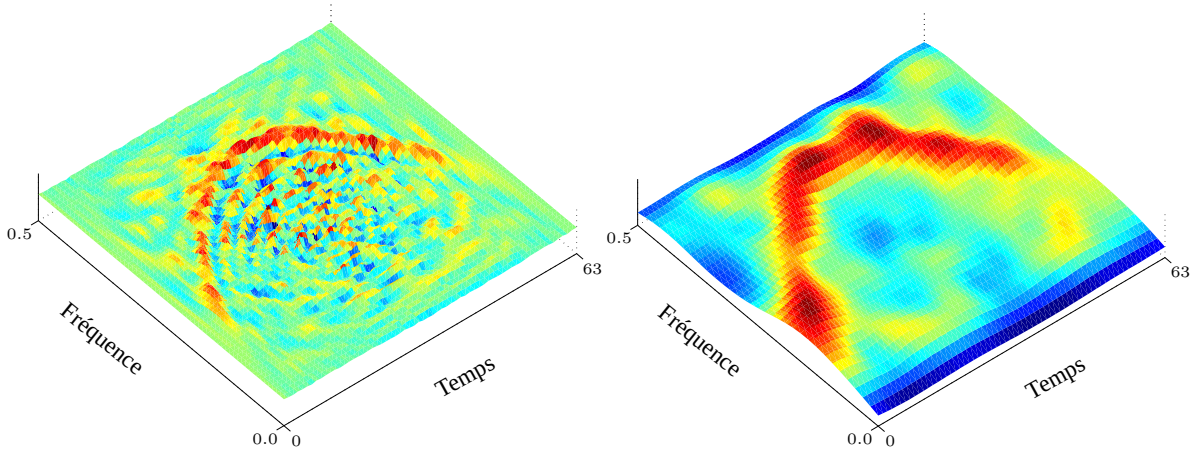


FIG. 9.11 – Signatures temps-fréquence obtenues dans le cas (a) de la distribution de Wigner et (b) du spectrogramme, pour un seuil de cohérence de 0.15 et 0.65 respectivement.

9.3 Applications à des signaux réels

Cette section est destinée à mettre l'approche proposée en œuvre au travers d'applications biomédicales[13]. On s'intéresse d'une part au débruitage d'un signal magnétoencéphalographique (MEG) corrompu par une activité cardiaque importante, ainsi qu'à la modélisation de cette dernière à partir d'un enregistrement électrocardiographique (ECG). On analyse d'autre part les signaux *complexe K* de l'électroencéphalogramme (EEG) de sommeil, dans le domaine temps-fréquence.

9.3.1 Débruitage d'un signal MEG

L'électrencéphalographie (EEG) et la magnétoencéphalographie (MEG) sont des techniques non-invasives permettant d'enregistrer, avec une grande résolution temporelle, les champs électriques et les champs magnétiques résultant d'une activité neuronale dans le cerveau. Puisque les champs magnétiques ne sont pas déformés à travers le crâne et le scalp (cuir chevelu), une résolution spatiale élevée est souvent obtenue par le MEG. Ceci rend cette technique particulièrement attractive pour la localisation de source. Toutefois, les mesures EEG et MEG sont souvent corrompues par des artéfacts, comme le mouvement des yeux, les activités musculaires et le rythme cardiaque. Car son amplitude est assez élevée pour pouvoir masquer les signaux porteur d'information, on souhaite extraire l'activité cardiaque par des techniques dites de filtrage de signal MEG, voir [CRLS06] pour une revue des stratégies existantes. Dans cette dernière communication, plusieurs algorithmes de filtrage à noyau reproduisant ont été proposés afin de résoudre le problème d'optimisation (8.1) sans le terme d'oubli, $\theta = 1$, et sans terme de régularisation, $\eta = 0$. La solution est donnée par $\boldsymbol{\alpha}^* = (\boldsymbol{K}^\top \boldsymbol{K})^{-1} \boldsymbol{K}^\top \boldsymbol{d} = \boldsymbol{K}^{-1} \boldsymbol{d}$, où $\boldsymbol{K}$ est la matrice de Gram des données d'apprentissage. Afin d'éviter des problèmes numériques liés à l'inversion de cette matrice, souvent mal-conditionnée, les auteurs ont recours à des techniques de régularisation à noyau, dont l'ACP-à-noyau (KPCA), l'algorithme *kernel partial least squares* (KPLS) [RT02], et l'algorithme *kernel ridge regression* (KRR) [SGV98a]. Les résultats montrent que les filtres obtenus permettent de réduire efficacement les artéfacts cardiaques du signal MEG. Cette approche n'étant toutefois pas adaptative par essence, elle ne permet pas un suivi dans un environnement non-stationnaire contrairement aux

[13] Les signaux abordés dans cette section ont été enregistrés dans les laboratoires de FORENAP (FOndation pour la REcherche en Neuro-sciences Appliquées à la Psychiatrie)

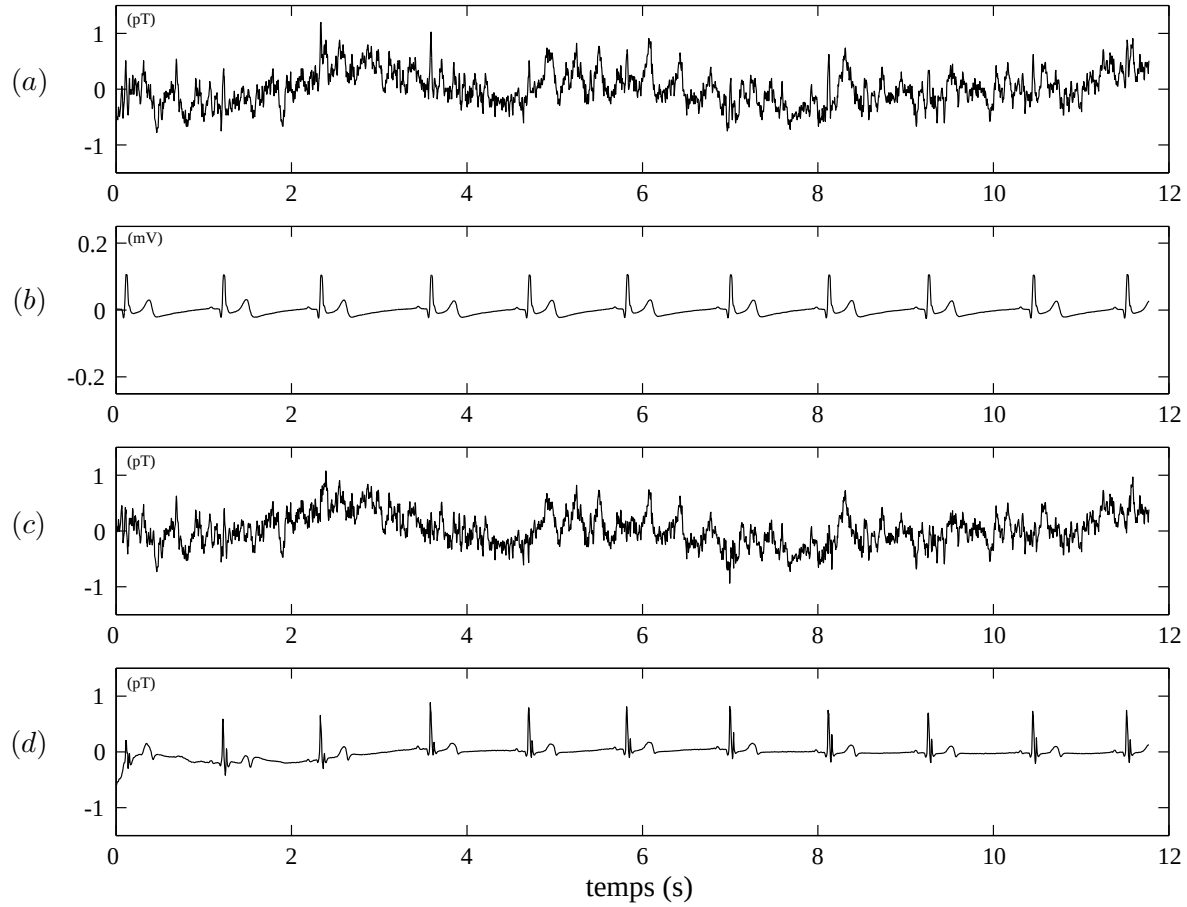


FIG. 9.12 – Résultats avec le noyau Gaussien : (a) signal MEG corrompu, (b) signal ECG de référence, (c) signal MEG débruité, et (d) contribution estimée de l'ECG dans le signal MEG.

algorithmes adaptatifs proposés dans ce manuscrit.

Des données MEG fortement corrompues par l'activité cardiaque ont été enregistrées simultanément avec le signal électrocardiographique (ECG). Ces données ont été échantillonnées à la fréquence de 254.31 Hz, avant un passage par un filtre passe bande $[0.1\ 50]$ Hz. On considère la même base de données que celle étudiée dans [CRLS06], soit deux signaux de $3\,000$ échantillons, l'un pour l'apprentissage et l'autre pour le test. Voir l'article en question pour plus de détails. Les signaux ECG et MEG ont été utilisés, respectivement, comme entrée et sortie désirée du filtre afin que l'erreur résiduelle corresponde au signal MEG débruité. On désigne dans la suite par $\boldsymbol{x}_i$ et $\boldsymbol{x}_j$ les séquences ECG de longueur p. On se contente d'utiliser les configurations optimales des noyaux précisées dans [CRLS06], soit le noyau Gaussien $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\bigl(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2\sigma_0^2\bigr)$ avec $\sigma_0 = 0.05$ et $p = 12$, et le noyau polynômial $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = (1 + \boldsymbol{x}_i{}^\top \boldsymbol{x}_j)^q$ de degré $q = 9$ et de longueur d'entrée $p = 6$. On a eu recours à l'erreur quadratique moyenne normalisée de prédiction sur les données d'apprentissage afin de déterminer le seuil optimal de cohérence, soit $\mu_0 = 0.26$ pour le noyau Gaussien, et $\mu_0 = 0.75$ pour le noyau polynômial. Dans chaque cas, le paramètre de régularisation η a été obtenu par une recherche sur l'intervalle $[10^{-6}, 1]$, avec un pas logarithmique. On a obtenu 10^{-1} et 10^{-3} pour les noyaux Gaussien et polynômial, respectivement. Le choix $\theta = 1$ correspondant à une mémoire de taille infinie a été considéré. L'erreur quadratique moyenne normalisée de prédiction sur les données test a été de 0.884 pour le noyau Gaussien, et de 0.879 pour le noyau polynômial. L'ordre du modèle était de $m = 12$ pour le premier, avec

	Noyau Gaussien				Noyau polynômial			
	KPCA	KPLS	KRR	**KRLS**	KPCA	KPLS	KRR	**KRLS**
Erreur quadratique moyenne	0.866	0.867	0.884	0.884	0.887	0.885	0.890	0.879

TAB. 9.1 – Comparaison des résultats obtenus sur les données MEG, pour ACP-à-noyau, KPLS, KRR et l'algorithme adaptatif KRLS proposé dans la présente étude.

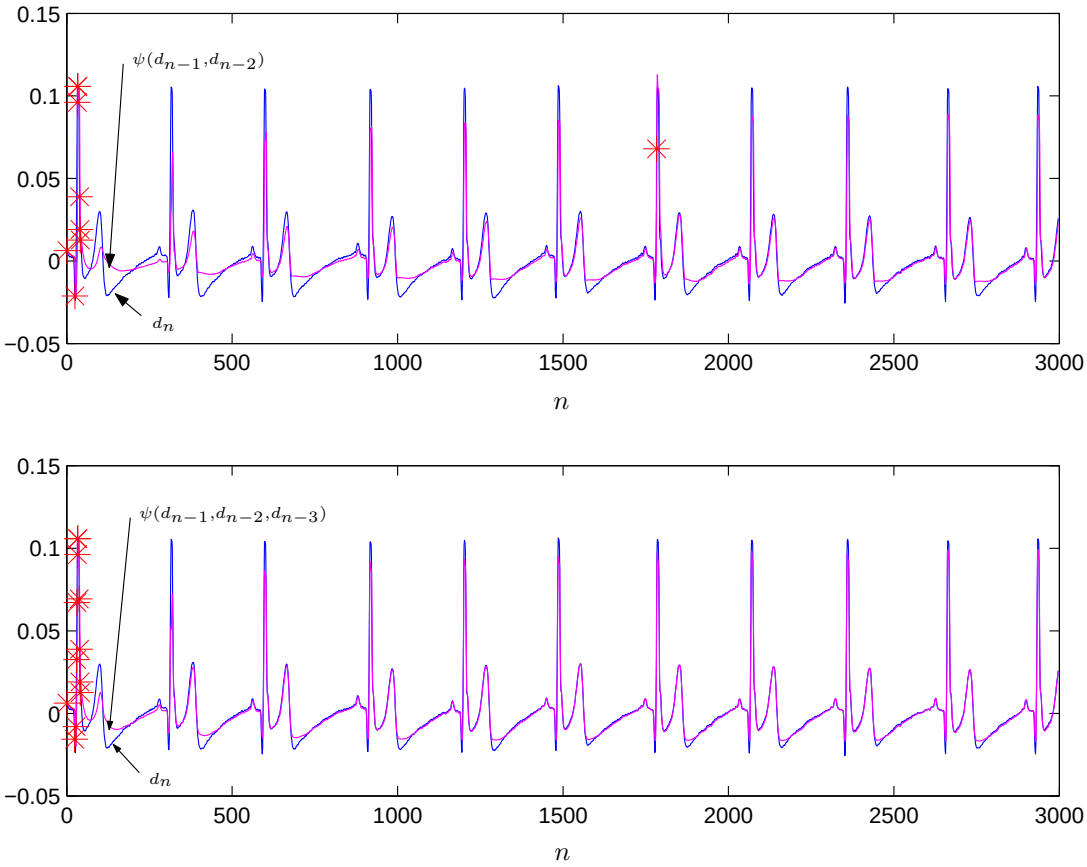


FIG. 9.13 – On représente le signal ECG à identifier, et le signal résultant du modèle retenu, de la forme $\psi(d_{n-1}, d_{n-2})$ en haut et $\psi(d_{n-1}, d_{n-2}, d_{n-3})$ en bas. Les éléments du dictionnaire sont représentés par les étoiles.

une fonction de Babel de $\mu_1(m) = 0.73$. Pour le second noyau, nous avons obtenu un ordre de $m = 6$ et une fonction de Babel de $\mu_1(m) = 2.73$. On illustre à la Figure 9.12 les résultats obtenus pour le noyau Gaussien. Les deux premières courbes, (a) et (b), montrent le signal MEG contaminé et le signal ECG de référence. Les courbes (c) et (d) montrent le signal MEG débruité et le signal correspondant aux artéfacts. Ces résultats confirment ceux obtenus dans [CRLS06] dans un cadre d'apprentissage hors-ligne. Comme illustré au Tableau 9.1, notre approche offre de meilleures performances que la stratégie basée sur KPLS pour le noyau polynômial, et est légèrement inférieure à celle basée sur l'ACP-à-noyau pour le noyau Gaussien. Cependant, une caractéristique commune des méthodes décrites dans [CRLS06] est que l'ordre du modèle résultant correspond à la taille de l'ensemble d'apprentissage, soit $3\,000$ termes, les rendant inappropriées pour une application en-ligne. Avec le critère de cohérence pour mode de contrôle de l'ordre du modèle, ce dernier ne dépasse pas 12.

Modélisation de l'ECG

Dans le cadre d'une autre application, on reprend le signal électrocardiographique (ECG) afin de le modéliser à partir des $3\,000$ échantillons disponibles correspondant à environ 10 cycles cardiaques. Dans cette expérimentation, on propose d'utiliser le noyau de Laplace avec $\sigma_0 = 0.35$, en fixant le seuil de cohérence à $\mu_0 = 0.9$. On considère l'algorithme KRLS, les différents paramètres étant obtenus par validation croisée.

Dans un premier temps, notre approche a consisté à identifier un modèle autorégressif de ce signal de la forme $d_n = \psi(d_{n-1}, d_{n-2})$. Ceci a produit un dictionnaire de $m = 9$ éléments. On représente à la Figure 9.13 en haut, le signal ECG et le signal obtenu selon le modèle. Au-delà d'une erreur de convergence élevée pour les premiers 500 échantillons, un décrochage pour les ondes négatives de l'ECG est visible. Dans un second temps, on s'est intéressé à un modèle de la forme $d_n = \psi(d_{n-1}, d_{n-2}, d_{n-3})$. Ceci a conduit à une augmentation de la taille du dictionnaire, portée à $m = 12$. En étudiant les résultats obtenus dans la Figure 9.13 en bas, on remarque que l'on modélise mieux le signal ECG que dans le premier cas. On illustre à la Figure 9.14 le signal ECG, les observations obtenues à partir du modèle $\psi(d_{n-1}, d_{n-2}, d_{n-3})$ et les éléments du dictionnaire dans l'espace de phase, à savoir (d_n, d_{n-1}, d_{n-2}). Dans les deux cas, il faut noter que la condition suffisante d'indépendance linéaire n'est pas satisfaite.

9.3.2 Analyse des complexes K dans l'EEG de sommeil

Dans une deuxième application, on étudie un événement transitoire particulier observé dans l'électroencéphalogramme (EEG) de sommeil, le *complexe K*. Le complexe K est souvent caractérisé par un aspect biphasique, avec une onde positive rapide suivie d'une onde négative de grande amplitude. Il dure environ 1 seconde, avec un spectre limité au support fréquentiel compris entre 1 et 4 Hz. Il se distingue naturellement de l'électroencéphalogramme de fond en stade 2 de sommeil par son amplitude, tandis

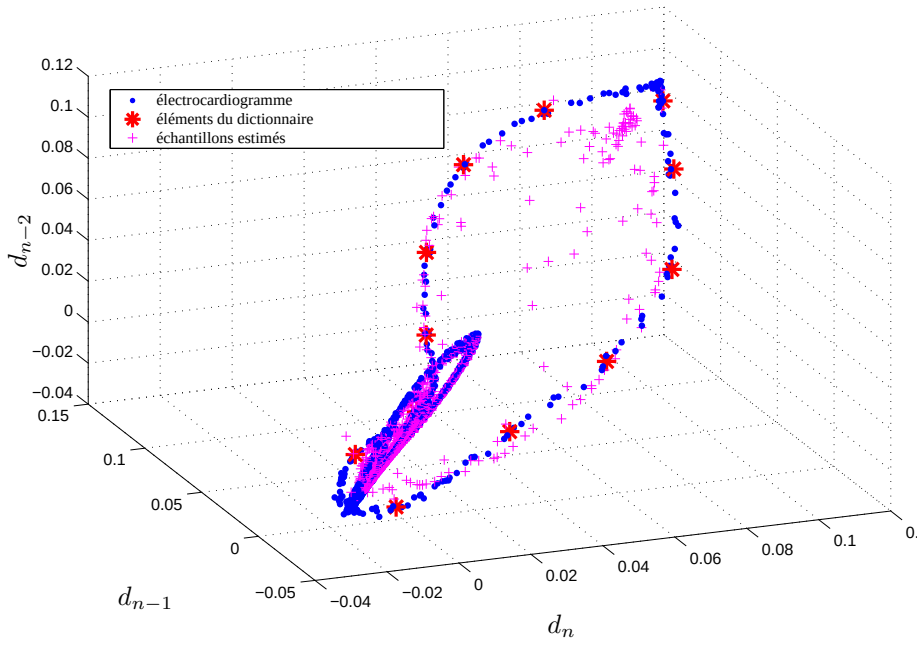


FIG. 9.14 – On représente dans l'espace de phase les données ECG à modéliser, et les observations obtenues à partir du modèle de la forme $\psi(d_{n-1}, d_{n-2}, d_{n-3})$. Les éléments du dictionnaire sont représentés par les étoiles.

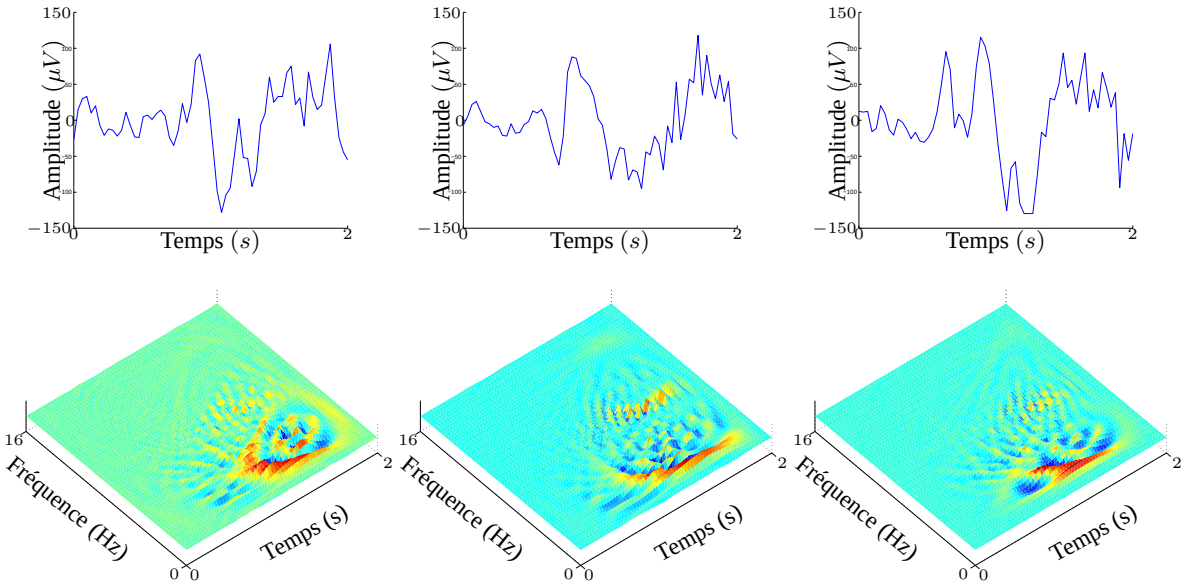


FIG. 9.15 – Représentations de 3 signaux de complexes K dans les domaines temporel et temps-fréquence (distribution de Wigner).

qu'il demeure difficile à isoler en stades 3 et 4, principalement en raison de la présence d'autres phénomènes non-stationnaires présents dans l'EEG. On propose d'étudier la structure temps-fréquence de ces signaux transitoires. Pour cela, on dispose d'un ensemble de 292 signaux de complexe K, de durée 2 secondes avec une fréquence d'échantillonnage de 32 Hz, donc compatible avec le support fréquentiel du complexe K. Ces signaux sont synchronisés au passage par zéro de l'onde transitoire. On représente à la Figure 9.16 trois signaux de complexe K choisis (aléatoirement) parmi l'ensemble de données disponibles. Les représentations dans le domaine temporel et temps-fréquence montrent la diversité de sa signature.

On propose dans un premier temps de déterminer la signature temps-fréquence principale associée au complexe K, au sens de l'analyse en composantes principales. Pour cela, on considère la mise en œuvre de l'ACP-à-noyau dans le domaine de la distribution de Wigner, avec le noyau reproduisant quadratique donné par l'expression (3.4), soit $\kappa_W(\boldsymbol{x}_i, \boldsymbol{x}_j) = |\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle|^2$. La (première) signature temps-fréquence principale est donnée par l'expression (4.11), soit $\Psi = \sum_{i=1}^{n} \alpha_i \, W_{\boldsymbol{x}_i}$, où les n coefficients α_i sont donnés par le vecteur propre associé à la plus grande valeur propre de la matrice Gram (centrée) de l'en-

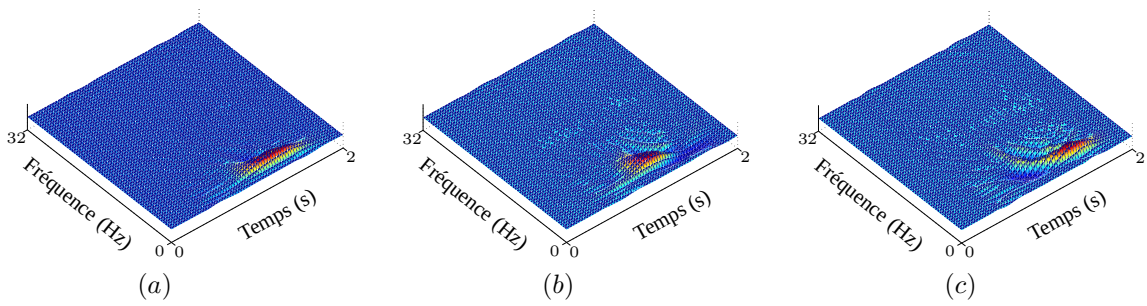
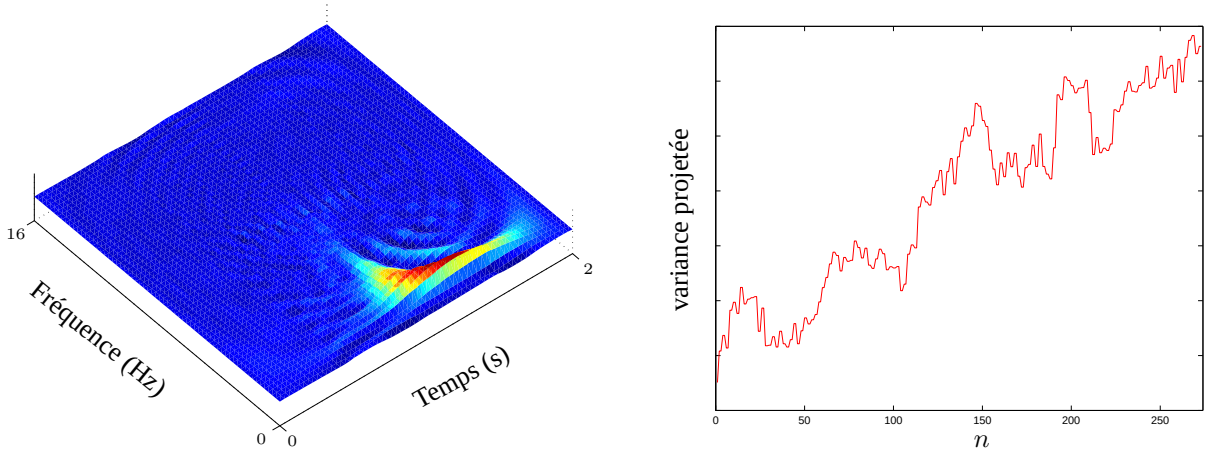


FIG. 9.16 – Les 3 premières signatures temps-fréquence principales obtenues par l'ACP-à-noyau et la distribution de Wigner, pour l'ensemble des 292 signaux de complexe K.



(a) (b)

FIG. 9.17 – Résultats obtenus par l'algorithme séquentiel d'ACP-à-noyau avec le seuil de cohérence fixé à 0.2. Dans (a) on représente la signature temps-fréquence principale associée au complexe K, et en (b) on illustre l'évolution de la variance expliquée, moyennée sur 100 évaluations, et lissée selon l'axe des abscisses par une fenêtre de taille 19.

semble d'apprentissage. La Figure 9.16 illustre les 3 premières signatures principales, où l'on aperçoit les signatures temps-fréquence les plus pertinentes. On peut y distinguer une composante tonale (à fréquence constante) dans (a), et dans (b) et (c) deux composantes contribuant à une forme hyperbolique dans le plan temps-fréquence. Ces signatures ne sont pas parcimonieuses, au sens où elles nécessitent l'évaluation de la distribution de Wigner pour les 292 signaux de l'ensemble d'apprentissage. Toutefois, il faut préciser que certains signaux ne contribuent pas essentiellement à la signature principale, alors qu'il est toujours nécessaire de les considérer dans l'ACP-à-noyau classique.

Dans un second temps, on propose d'aborder le problème de la parcimonie de la signature principale. L'usage du critère de cohérence permet de sélectionner un sous-ensemble pertinent des distributions de Wigner. Le modèle résultant est de la forme $\Psi = \sum_{j=1}^{m} \alpha_j \, W_{\boldsymbol{x}_{\omega_j}}$, où l'ordre m du modèle dépend du seuil de cohérence μ_0 fixé et les ω_j forment un sous-ensemble de $\{1, \cdots, n\}$. Les m distributions de Wigner ainsi retenues sont optimales au sens de l'ACP-à-noyau, comme préconisé par la Proposition 7.7. Les coefficients α_j du modèle sont alors donnés par l'algorithme séquentiel présenté au Tableau 8.5, permettant ainsi de prendre en compte tous les éléments de l'ensemble d'apprentissage, retenus et écartés du modèle réduit. Les résultats obtenus sont moyennés sur une base de 100 évaluations à partir de séquences aléatoires de l'ensemble des signaux disponibles. On fixe le seuil de cohérence à $\mu_0 = 0.2$, sans terme de régularisation $\eta = 0$. Le pas de convergence est obtenu par recherche sur une grille de valeurs, et est fixé à 0.0075. La Figure 9.17 illustre en (a) la signature temps-fréquence principale, où l'on retrouve la signature hyperbolique dans le plan temps-fréquence avec une composante à fréquence constante plus prononcée, ainsi que des termes hyperboliques que l'on peut qualifier d'interférences. On trace en (b) l'évolution, en fonction du nombre de données disponibles, de la variance des distributions projetées sur celle-ci, lissée sur 19 échantillons selon l'axe des abscisses. L'ordre du modèle résultant est en moyen de 15.92. Au-delà de la distribution classique de Wigner, on peut aussi considérer d'autres distributions temps-fréquence de la classe de Cohen, en particulier la pseudo-Wigner-lissée et le spectrogramme, sous contrainte d'augmenter le seuil de cohérence afin d'obtenir des modèles d'ordre similaire. On fixe alors le seuil de cohérence à 0.55 pour la première et 0.7 pour la seconde, conduisant à des modèles d'ordres respectifs 16.54 et 13.25. Les signatures principales résultantes et l'évolution de la variance expliquée

sont illustrées à la Figure 9.18. La signature hyperbolique est apparente dans le cas de la pseudo-Wigner-lissée, avec une composante tonale plus prononcée et une suppression des termes interférentiels. Dans le cas du spectrogramme, le lissage élevé cache cette structure au profit de la composante tonale, similaire à celle obtenue avec la première signature principale dans le cas de l'ACP-à-noyau, voir Figure 9.16 (a).

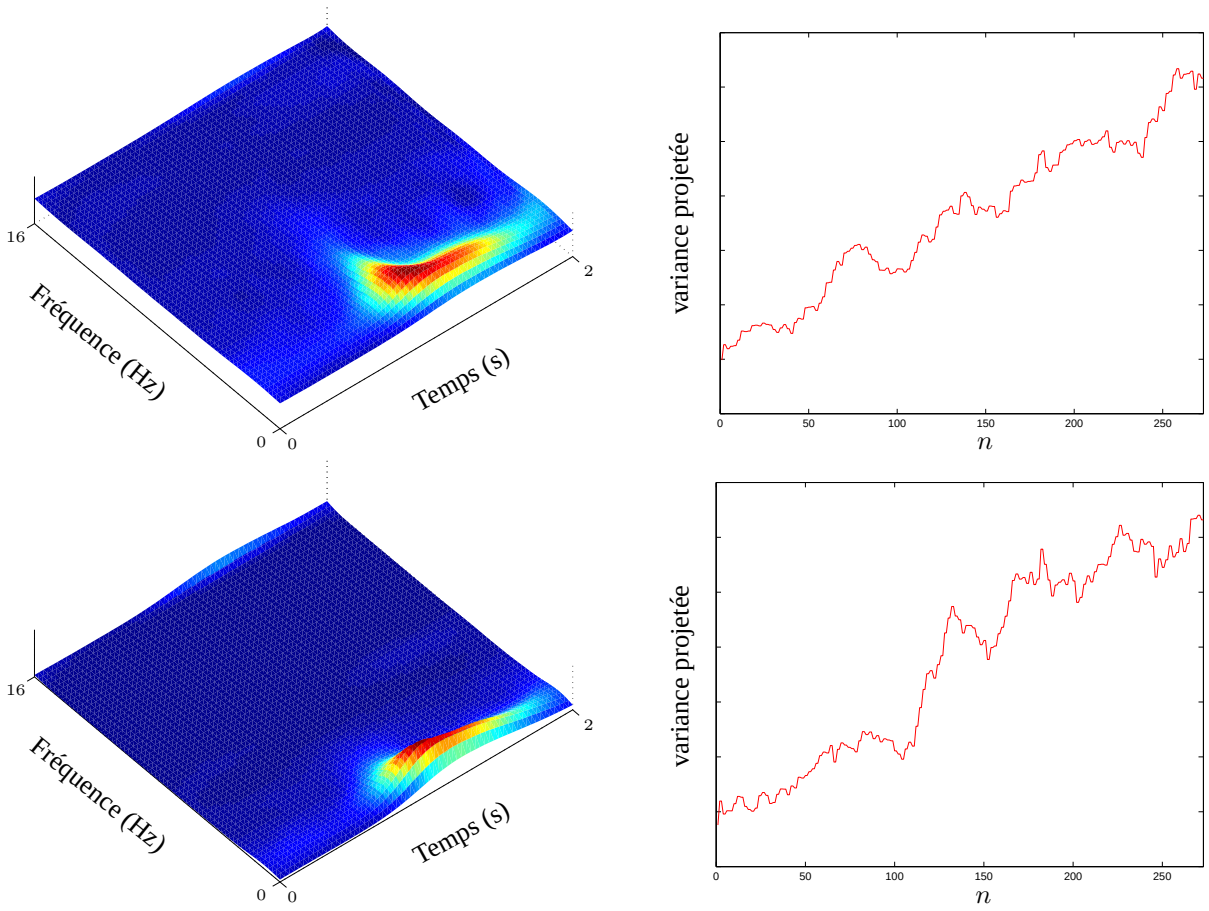


FIG. 9.18 – Résultats obtenus comme pour la Figure 9.17, pour la distribution pseudo-Wigner-lissée (en haut) et pour le spectrogramme (en bas), en fixant les seuils de cohérence à 0.55 et 0.7 respectivement.

Conclusion générale et perspectives

Dans ce mémoire de thèse, nous avons proposé des outils flexibles et puissants pour l'analyse de signaux non-stationnaires. Deux approches complémentaires ont été présentées : d'une part la mise en œuvre des méthodes à noyau dans le domaine temps-fréquence, et d'autre part le développement d'algorithmes d'apprentissage en-ligne reposant sur des techniques classiques de filtrage adaptatif.

Nous avons introduit dans un premier temps un cadre général pour bénéficier des plus récents développements des méthodes à noyau dans le domaine de l'analyse temps-fréquence grâce à un choix approprié de noyau reproduisant. Nous avons commencé par la mise en œuvre de l'analyse en composantes principales à noyau dans cet espace. L'efficacité calculatoire d'une telle approche face à une technique d'ACP conventionnelle a été étudiée, et validée par des expérimentations. Sa pertinence a été démontrée au travers d'applications sur des signaux synthétiques et réels issus d'un environnement biomédical. Puis nous avons étendu ce spectre de méthodes aux techniques d'apprentissage supervisé, pour la discrimination et la classification de signaux non-stationnaires. L'analyse factorielle discriminante et les Support Vector Machines ont été successivement considérées. Nous avons montré comment ces différentes techniques peuvent profiter de la diversité des espaces de représentation de signaux que procurent les distributions temps-fréquence de la classe de Cohen. La sélection d'une représentation adaptée à la résolution d'un problème de classification demeurait toutefois une question récurrente. Nous avons donc proposé une solution reposant sur un critère généralement utilisé pour la sélection de noyau reproduisant : l'alignement noyau-cible.

Si la complexité des méthodes à noyau ne dépend pas de l'espace de représentation adopté, l'ordre des modèles résultants correspond néanmoins au nombre de données d'apprentissage disponibles. Ceci réduit à néant toute perspective de traitement en-ligne des données sans dispositif complémentaire. Nous avons donc proposé dans un second temps de travailler sur la parcimonie des solutions à l'aide d'un critère directement inspiré de la littérature relative à l'approximation parcimonieuse de fonctions : la cohérence d'un dictionnaire. Nous en avons étudié les nombreuses propriétés et établi un lien avec d'autres critères tels que l'erreur d'approximation linéaire et l'entropie de Rényi. Nous avons enfin fait un rapprochement avec l'ACP. Dans un dernier temps, nous avons mis en œuvre ces principes dans le cadre de problèmes d'identification de systèmes non-linéaires et non-stationnaires. Des algorithmes originaux de filtrage adaptatif non-linéaire ont ainsi été développés, puis testés sur des signaux synthétiques et réels.

Discussion et perspectives

Nous avons essentiellement proposé de nouveaux outils d'analyse pour les signaux non-stationnaires dans le domaine temps-fréquence. Nous nous sommes contentés d'étudier les distributions temps-fréquence de la classe de Cohen, en évoquant simplement de possibles extensions vers d'autres classes de représentations. Les méthodes temps-échelle se prêteraient également à ce type de développements, en particulier l'analyse en ondelettes. On pourrait alors envisager celle-ci et ses liens avec les méthodes à noyau dans le cadre d'autres applications telles que le traitement d'images. Des critères tels que l'ali-

gnement noyau-cible pourraient également être utilisés pour le choix de l'ondelette optimale pour un problème donné.

Tout au long de ce manuscrit, nous avons cherché à diversifier les méthodes à noyau présentées afin de couvrir une vaste famille d'applications potentielles, sans prétendre à l'exhaustivité. Toutefois, certaines techniques n'ont pas été abordées. En particulier, la séparation de sources a fait l'objet de travaux dans le domaine temps-fréquence dans [BA98, AD05]. Des techniques d'analyse en composantes indépendantes et séparation aveugle de sources ont également été proposées dans le cadre général des méthodes à noyau [BJ03]. On pourrait à présent envisager de croiser ces travaux à la lumière des développements présentés dans ce manuscrit. D'autres applications sont également envisageables, telles que le détection de changement et le mise en œuvre de techniques de régression dans le domaine temps-fréquence. Enfin, les méthodes proposent différentes techniques dans le cadre générique de la *reconstruction de pré-images*. Elles visent à déterminer un échantillon x de $\mathcal{X}$ qui produit à travers l'application ϕ associée au noyau reproduisant considéré, une image $\phi(x)$ proche d'un élément de $\mathcal{H}$ donné [MSS$^+$99, KT03]. 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 conviendrait alors de comparer ces stratégies avec celles développées dans un contexte purement temps-fréquence [HK92, Hla98].

Autour du thème de l'optimalité d'un espace de représentation, il serait intéressant d'étudier celle d'une distribution temps-fréquence dans le cadre de problèmes d'analyse non-supervisée tels que l'ACP. Le critère adopté dans ce manuscrit, l'alignement noyau-cible, ne répond en effet pas à cette question et la littérature ayant trait aux méthodes à noyau offre peu de solutions. Une piste possible pourrait être celle suivie dans de récents travaux relatifs à la consistence de la matrice de Gram pour un tel problème [ZBB04, STWCK05].

Nos perspectives dans le cadre de l'apprentissage en-ligne sont aujourd'hui de plusieurs ordres. La première concerne une étude théorique et pratique plus approfondie des différentes variantes d'algorithmes proposées et de leurs performances. Le domaine du traitement adaptatif offre déjà de nombreux outils dont il faudrait s'inspirer. Enfin, il serait également souhaitable de lancer des investigations plus poussées vers d'autres critères de parcimonie, en s'appuyant sur l'abondante littérature ayant trait aux méthodes de représentation creuse.

Annexes

Annexe A

Classification dans le domaine temps-fréquence par l'algorithme des *support vector machines*

Sommaire

Le domaine de la reconnaissance des formes connaît une révolution depuis le milieu des années 90 avec la théorie de l'apprentissage statistique et l'avènement des *Support Vector Machines* (SVM) pour la résolution de problèmes de détection, de classification et de régression. Clairement, les techniques conventionnelles de décision en environnement non-stationnaire n'ont pas encore profité de ces avancées récentes, ni de la théorie qui les accompagne. On propose dans cette annexe la mise en œuvre des SVM dans le domaine temps-fréquence.

Cette annexe est organisée ainsi. On commence par un rappel sur les fondements de la théorie de l'apprentissage statistique, plus particulièrement la VC-dimension et les principes de minimisation des risques empirique et structurel. Puis on présente l'algorithme classique des SVM dans les cas de données linéairement et non-linéairement séparables. On adapte cette approche au domaine temps-fréquence pour la classification de signaux non-stationnaires, avant de conclure par des expérimentations.

A.1 Introduction

Pour un problème d'apprentissage donné, les meilleures performances en généralisation peuvent être atteintes lorsqu'on trouve un compromis satisfaisant entre les performances atteintes sur l'ensemble d'apprentissage et la capacité d'apprentissage de la famille de statistiques considérée. Ce concept s'est concrétisé par la théorie de l'apprentissage statistique, élaborée au milieu des années 90 par Vapnik [Vap95] mais dont les premiers éléments remontent aux années 70 [VC71]. Les deux piliers de cette théorie sont les principes inductifs de la minimisation du risque empirique et du risque structurel. Cette théorie a mis en évidence une nouvelle classe de méthodes d'apprentissage pour la reconnaissance des formes, les *Support Vector Machines* [BGV92]. Ces structures, couramment appelées SVM, constituent des solutions à marge maximale entre l'hyperplan séparateur qu'elles définissent et les échantillons d'un ensemble d'apprentissage. Des considérations fondamentales sur les espaces de Hilbert à noyau reproduisant, regroupées sous le principe du coup du noyau, leur assure une extension non-linéaire.

Différentes approches ont été récemment proposées dans la littérature, combinant le domaine temps-fréquence pour le traitement non-linéaire du signal et les SVM pour la synthèse de classifieurs. On peut les regrouper principalement en deux catégories. D'une part, on trouve les techniques classiques de construction d'un vecteur de descripteurs (ou de caractéristiques), ceux-ci étant extraits du plan temps-fréquence, comme par exemple la variance de la fréquence instantanée. Dans [GGR04], les auteurs recourent aux SVM classiques associés au noyau Gaussien qu'ils appliquent à des vecteurs de descripteurs globaux, tandis que l'évolution de vecteurs de descripteurs locaux est prise en compte [RDR$^+$07] et un algorithme de SVM à une classe [Tax01] est considéré. Ces techniques nécessitent toutefois une sélection appropriée de l'ensemble de descripteurs à utiliser. D'autre part, des techniques plus globales consistent à considérer le plan temps-fréquence tout entier. Celles-ci ont été introduites par les travaux de Davy *et coll.* [GDDR01, DGDR02], amenant à combiner distributions temps-fréquence et SVM. L'approche proposée consiste à utiliser comme espace d'entrée, non pas les signaux mais leurs distributions temps-fréquence, et appliquer les SVM associées au noyau Gaussien. Cette approche exige cependant le calcul explicite des différentes distributions temps-fréquence de l'ensemble des signaux d'apprentissage afin de déterminer la matrice de Gram correspondante.

On présente dans cette annexe l'algorithme classique de SVM, comme initialement proposé dans [BGV92] pour un problème de classification à deux classes. Toutefois, il existe une grande variété d'algorithmes selon les applications envisagées. Parmi ceux-ci, on trouve l'apprentissage à une classe, avec *one class SVM*, et les ν-SVM. Différents algorithmes de SVM pour les problèmes multi-classes ont été proposés tout au long de la dernière décennie, dont des techniques de types un-contre-tous [Vap95] (et amélioré dans [Vap98]) et un-contre-un [Kre99] ; Voir [Abe03] pour une comparaison des récentes techniques proposées. On peut aussi citer les travaux de Crammer *et coll.* qui proposent un cadre général de codage permettant d'adapter les algorithmes classiques de SVM au cas multi-classes [CS00, CS02a]. A l'exception de [EPM00], peu de travaux ont considéré les principes inductifs de minimisation des risque empirique et structurel afin d'élaborer un algorithme de SVM multi-classes approprié.

A.2 Éléments de théorie de l'apprentissage statistique

On reprend la théorie de l'apprentissage statistique, comme proposée au Chapitre 1, en introduisant des résultats clés sur les capacités et les performances en généralisation des méthodes d'apprentissage. Les éléments de cette section sont principalement extraits de [Vap95, Vap98].

A.2.1 Position du problème

On rappelle le schéma de l'apprentissage statistique, comme introduit au Chapitre 1, dans le cadre d'un problème supervisé. On cherche à déterminer la relation qui lie un espace d'entrée $\mathcal{X}$ à un espace de sortie $\mathcal{Y}$. Cette relation, définie par un modèle ψ, minimise le risque réel

$$R(\psi) = \int_{\mathcal{X}\times\mathcal{Y}} V(\psi(x), y)\, \mathrm{d}P(x, y), \tag{A.1}$$

avec $\mathrm{d}P(x, y) = P(x, y)\, \mathrm{d}x\, \mathrm{d}y$, et V la fonctionnelle coût qui vise à pénaliser l'erreur entre la sortie estimée $\psi(x)$ et celle désirée y. On désigne par ψ^* la fonction minimisant le risque réel (A.1). La résolution de ce problème nécessite la connaissance de la distribution de probabilité conjointe $P(x, y)$, supposée inaccessible. On dispose au lieu de cela d'un ensemble de n couples échantillonnés aléatoirement selon cette distribution, que l'on désigne par $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, où $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. A partir de ces données, on cherche à minimiser le risque empirique

$$R_n(\psi) = \frac{1}{n} \sum_{i=1}^{n} V(\psi(x_i), y_i). \tag{A.2}$$

La minimisation du risque empirique est un problème mal-posé, puisqu'il existe une infinité de fonctions faisant état d'un apprentissage exacte. Toutefois, la plupart de ces fonctions ne déterminent pas les bonnes étiquettes sur de nouvelles données, pourtant obtenues à partir de la même distribution de probabilité P. Pour remédier à cet inconvénient, on a recours à une restriction du domaine des fonctions candidates. On considère par exemple des fonctions régulières reflétant la régularité du problème traité. Parmi un ensemble $\mathcal{F}$ de fonctions, on détermine alors la fonction ψ_n^* minimisant le risque empirique (A.2), selon

$$\psi_n^* = \arg\min_{\psi\in\mathcal{F}} R_n(\psi).$$

Rien ne garantit que la solution ainsi obtenue, au travers du processus d'apprentissage, ne soit optimale au sens de la minimisation du risque réel (A.1) dans la famille $\mathcal{F}$ des fonctions candidates. L'optimalité dans ce dernier cas peut s'exprimer selon l'expression

$$\psi_{\mathcal{F}}^* = \arg\min_{\psi\in\mathcal{F}} R(\psi).$$

Les fonctions ψ_n^*, $\psi_{\mathcal{F}}^*$ et ψ^* correspondent respectivement à la fonction obtenue par un processus d'apprentissage au sein de la famille $\mathcal{F}$, à celle minimisant le risque réel dans $\mathcal{F}$, et à la fonction optimale au sens de Bayes. Ces fonctions sont comparées entre elles à partir des risques en question. En considérant les deux premières, on désigne par erreur d'estimation, notée E_{estim}, la différence des risques réel et empirique au sein de $\mathcal{F}$, selon

$$E_{\text{estim.}} = R_n(\psi_n^*) - \inf_{\psi\in\mathcal{F}} R(\psi).$$

Cette erreur dépend de l'ensemble d'apprentissage et du processus d'apprentissage, ainsi que de $\mathcal{F}$. D'autre part, en considérant les deux dernières, on obtient l'erreur d'approximation, $E_{\text{approx.}}$, définie par

$$E_{\text{approx.}} = \inf_{\psi\in\mathcal{F}} R(\psi) - R(\psi^*).$$

Ne dépendant que du choix de $\mathcal{F}$, elle détermine la difficulté d'approximer le modèle optimal ψ^* à partir de cet ensemble de fonctions. En combinant les deux erreurs, l'erreur d'estimation et l'erreur d'approximation, on obtient l'erreur dite de modélisation définie selon l'expression

$$E_{\text{modél.}} = E_{\text{estim.}} + E_{\text{approx.}} = R_n(\psi_n^*) - R(\psi^*).$$

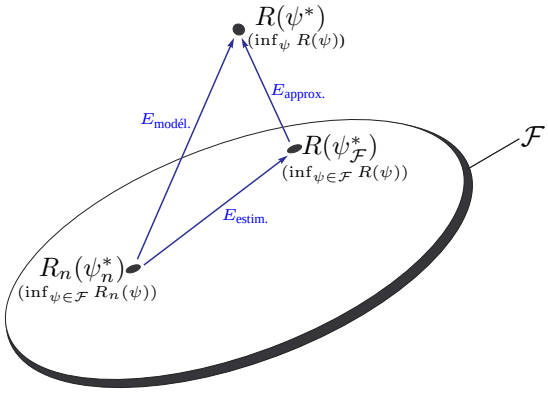


FIG. A.1 – Représentation schématique des différentes erreurs dues à la minimisation du risque empirique sur un ensemble $\mathcal{F}$ de fonctions.

On illustre à la Figure A.1 les différentes sources d'erreur induites par une procédure d'apprentissage visant à minimiser le risque empirique sur un ensemble d'apprentissage, en considérant une famille $\mathcal{F}$ de fonctions candidates. La minimisation de l'erreur de modélisation fait intervenir deux termes antagonistes, $E_{\text{estim.}}$ et $E_{\text{approx.}}$. Afin de diminuer la première, on a recours à une famille $\mathcal{F}$ plus riche en augmentant le nombre de fonctions candidates, au détriment de la seconde. Réciproquement, l'usage d'une famille $\mathcal{F}$ plus réduite entraîne une diminution de $E_{\text{estim.}}$ tandis que $E_{\text{approx.}}$ croît. Ce phénomène n'est autre que le fameux compromis biais-variance.

Établie par les travaux précurseurs de Vapnik et Chervonenkis [VC71], la théorie de l'apprentissage statistique visent à traiter ce compromis en considérant deux problèmes. Dans un premier temps, la famille de modèles étant fixée, l'étude concerne la convergence uniforme du risque empirique vers le risque le plus proche du risque réel pour ces fonctions candidates. Il s'agit du principe inductif de minimisation du risque empirique. Dans un second temps, la question du choix de la famille de modèles est alors posée. La sélection de la famille optimale est alors considérée en recherchant le meilleur compromis possible. Il s'agit du principe inductif de minimisation du risque structurel.

A.2.2 Dimension de Vapnik-Chervonenkis

La dimension de Vapnik-Chervonenkis, ou encore VC-dimension, d'une famille de fonctions est une mesure de sa capacité d'apprentissage. Soit $\mathcal{F}$ une famille de fonctions définies sur $\mathcal{X}$.

Définition 7 (VC-dimension)**.** *La dimension h de Vapnik-Chervonenkis d'une famille $\mathcal{F}$ donnée est le plus grand nombre d'éléments de l'ensemble des réalisations $\mathcal{X}$ dont les fonctions de $\mathcal{F}$ peuvent réaliser toutes les 2^h dichotomies.*

Le sens d'une dichotomie est naturel pour une famille de fonctions à valeur binaire, soit $\{0, 1\}$ dans le cas d'une famille de détecteurs, ou encore $\{-1, +1\}$ pour une classification. Cette dichotomie s'étend aisément aux familles de fonctions à valeurs réelles, au sens suivant : pour toute partition de l'espace des réalisations, $\mathcal{X}_+ \cup \mathcal{X}_- = \mathcal{X}$, il existe au moins une fonction ψ de $\mathcal{F}$ telle que l'hypothèse $\psi(x) \gtrless_{\mathcal{X}_-}^{\mathcal{X}_+} 0$ est vérifiée pour tout $x \in \mathcal{X}$. Il faut toutefois noter que cette expression est souvent considérée en remplaçant le seuil 0 par un seuil arbitraire.

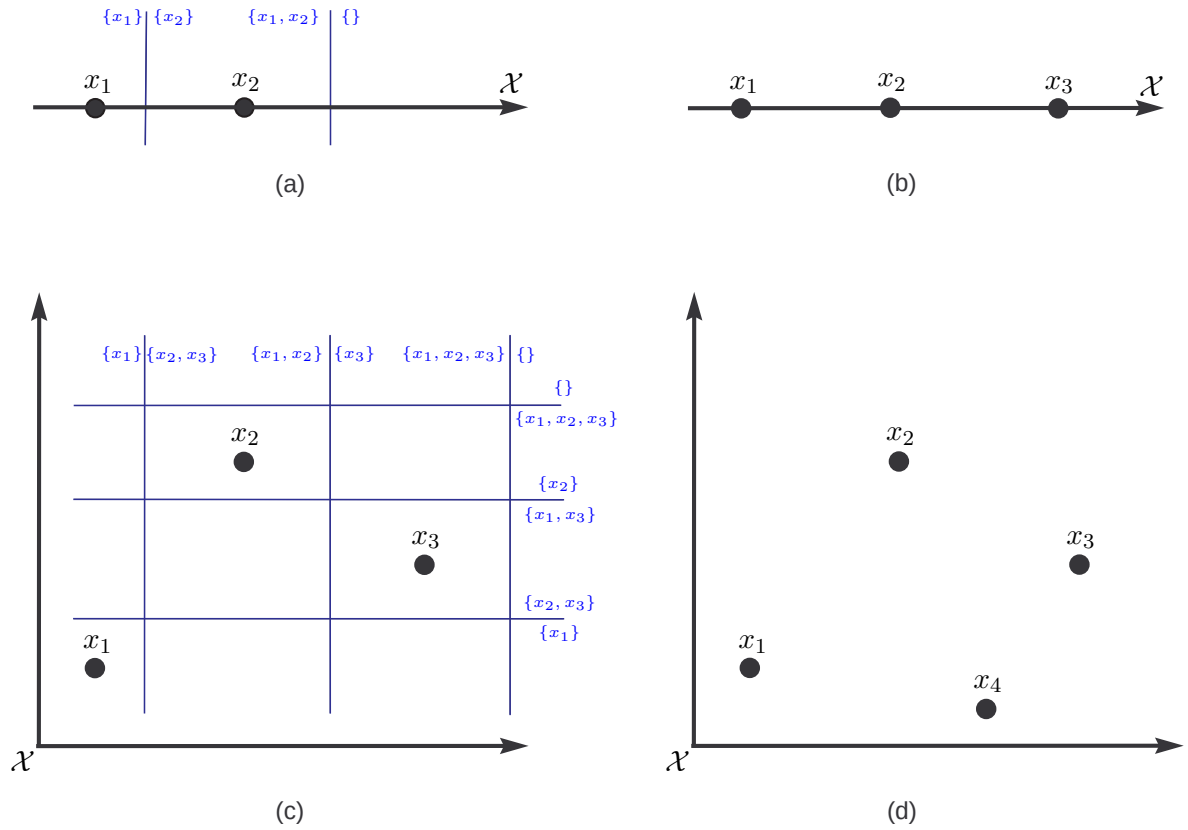


FIG. A.2 – Détermination de la VC-dimension des discriminants linéaires, de $\mathbb{R}$ (première ligne) et $\mathbb{R}^2$ (deuxième ligne).

Bien avant les développements proposés par Vapnik et Chervonenkis [VC71], l'intérêt d'une caractérisation de la capacité d'apprentissage d'une famille de classifieurs a été souligné par Cover dans le cadre des discriminants linéaires [Cov65]. Afin d'illustrer la notion de VC-dimension, on s'intéresse précisément à présent à celle des discriminants linéaires dans l'espace $\mathbb{R}^l$, en commençant par $l = 1$ et $l = 2$. Dans $\mathbb{R}$, toute fonction appartenant à $\mathcal{F}$ est de la forme $\psi(x) = wx + w_0$. Comme illustré à la Figure A.2 (a), deux points $\{x_1, x_2\}$ peuvent être discriminés par des fonctions de $\mathcal{F}$, en réalisant toutes les dichotomies que l'on représente sous forme de couples $\{\}/\{x_1, x_2\}, \{x_1\}/\{x_2\}, \{x_2\}/\{x_1\}, \{x_1, x_2\}/\{\}$. Cependant, pour trois points $\{x_1, x_2, x_3\}$ on ne peut pas obtenir les 2^3 dichotomies comme le montre la Figure A.2 (b) avec le couple $\{x_1, x_3\}/\{x_2\}$. On dit alors que les discriminants linéaires de $\mathbb{R}$ admettent une VC-dimension égale à 2. En considérant $\mathbb{R}^2$, on peut avoir que toutes les dichotomies de 3 points peuvent être réalisées comme cela est présenté en Figure A.2 (c). Il n'en est en revanche pas de même pour 4 points, la dichotomie $\{x_1, x_4\}/\{x_2, x_3\}$ étant impossible à obtenir comme le montre la Figure A.2 (d). La VC-dimension des discriminants linéaires de $\mathbb{R}^2$ est donc égale à 3. On peut étendre aisément ce résultat à $\mathbb{R}^l$, où l'ensemble des discriminants linéaires $\psi(x) = \langle \boldsymbol{w}, x \rangle_{\mathcal{X}} + w_0$ admet une VC-dimension égale à $l + 1$. Bien que la VC-dimension des discriminants linéaires corresponde au nombre de paramètres libres, cette analogie n'est pas vraie dans le cas général comme illustré par plusieurs contre-exemples dans [Vap95].

A.2.3 Principe de minimisation du risque empirique

La pertinence de la substitution du risque réel par le risque empirique est donné par le principe de minimisation du risque empirique. Le théorème suivant est un résultat clé de la théorie de l'apprentissage.

Théorème 8. *Soit $\mathcal{F}$ une famille de fonctions ψ. Le principe de minimisation du risque empirique est consistant si et seulement si le risque empirique converge uniformément vers le risque réel au sens probabiliste suivant :*

$$P(\sup_{\psi\in\mathcal{F}}\{R(\psi) - R_n(\psi)\} > \epsilon) \underset{n\to\infty}{\longrightarrow} 0$$

pour tout $\epsilon > 0$.

En pratique, on ne peut pas invoquer ce théorème puisqu'il fait appel à la distribution de probabilité inconnue $P(x, y)$ dans le calcul du risque réel R. La théorie de l'apprentissage statistique offre non seulement un résultat qualitatif sur la convergence uniforme du risque empirique vers le risque réel pour tout un ensemble des fonctions candidates, mais aussi des bornes explicites sur la vitesse de convergence de l'expression du Théorème 8. De plus, on montre qu'il existe une borne dite universelle ne dépendant pas de la distribution de probabilité des données. Plus encore, il est possible de démontrer que la borne suivante est satisfaite pour toute famille $\mathcal{F}$ de fonctions à VC-dimension h :

$$P(\sup_{\psi\in\mathcal{F}}\{R(\psi) - R_n(\psi)\} > \epsilon) \leq 4\,(2\,\mathrm{e}n/h)^h\ \mathrm{e}^{-n\epsilon^2/8}, \tag{A.3}$$

pour tout $\epsilon > 0$. Ce résultat permet de préciser le principe de minimisation du risque empirique présenté au Théorème 8. Avec une VC-dimension finie, la convergence uniforme est alors satisfaite, et le principe de minimisation du risque empirique est consistant. Il s'avère que cette condition est également nécessaire. Ce résultat, valable pour toute famille de fonctions, constitue une généralisation des bornes proposées séparément par Kolmogorov et Smirnov, largement répandues en statistique classique. Ces bornes classiques sont de plus valables asymptotiquement, comparées à l'expression (A.3) qui est satisfaite pour un nombre fini d'observations. Vapnik énonce dans [Vap95] d'autres inégalités, comme par exemple une borne supérieure sur l'écart entre les risques empirique et réel. Depuis sa formulation originelle, plusieurs travaux ont revisité l'expression (A.3), en proposant des bornes de plus en plus fines. On pourra se référer à [Vay00] pour un aperçu sur ces différentes améliorations, dont

$$P(\sup_{\psi\in\mathcal{F}}\{R(\psi) - R_n(\psi)\} > \epsilon) \leq 4\,(2\,\mathrm{e}n/h)^h\ \mathrm{e}^{-n\epsilon^2}. \tag{A.4}$$

A partir de l'expression (A.4), il est possible de déduire un intervalle de confiance liant les risques empirique et réel pour un fonction appartenant à famille $\mathcal{F}$ donnée. Le résultat est donné dans le théorème suivant.

Théorème 9. *Pour toute fonction ψ appartenant à une famille $\mathcal{F}$, l'inégalité suivante est satisfaite avec une probabilité au moins égale à $1 - \epsilon$*

$$R(\psi) \leq R_n(\psi) + C(h, n, \epsilon), \tag{A.5}$$

où C désigne la largeur de l'intervalle de confiance qui dépend de la taille de l'ensemble d'apprentissage n, de la VC-dimension h de $\mathcal{F}$, et du niveau de confiance accordé ϵ, vérifiant l'expression

$$C(h, n, \epsilon) = \sqrt{\frac{h}{n}\left(1 - \log 2\frac{h}{n}\right) - \frac{1}{n}\log\frac{\epsilon}{4}}. \tag{A.6}$$

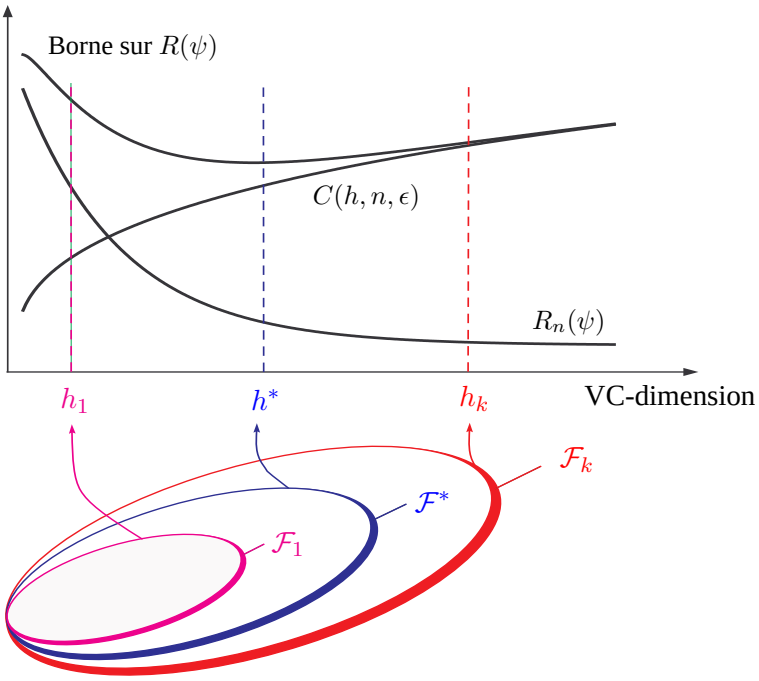


FIG. A.3 – Représentation schématique du principe de minimisation du risque structurel.

Pour les faibles valeurs du quotient h/n, la largeur C de l'intervalle de confiance est proche de zero, signifiant que le risque empirique $R_n(\psi)$ tend vers le risque réel $R(\psi)$. Il s'agit là de l'essence même de ce principe d'induction. En revanche, pour les valeurs de h/n élevées (proche de 1), la largeur de l'intervalle de confiance est considérable, et $R_n(\psi)$ ne permet plus une bonne estimation de $R(\psi)$. Le compromis entre ces deux quantités est étudié par le principe de minimisation du risque structurel.

A.2.4 Principe de minimisation du risque structurel

Le principe de minimisation du risque empirique incite à la minimisation de celui-ci à tout prix. Toutefois, rien ne garantit que les performances atteintes soient proches du risque optimal, au sens de Bayes. Un contrôle de la capacité en généralisation est souvent nécessaire. Ceci est en particulier indispensable pour les ensembles d'apprentissage de faible taille ayant un quotient h/n supérieur à 5%, où ce quotient correspond à la VC-dimension rapportée par donnée d'apprentissage. Dans ce paragraphe, on présente le principe inductif pour contrôler la capacité en généralisation, en contrôlant la VC-dimension du modèle. C'est le principe de minimisation du risque structurel.

Le Théorème 9 détermine une borne supérieure sur le risque réel pour toute fonction d'une famille donnée $\mathcal{F}$, avec une probabilité au moins égale à $1 - \epsilon$. La borne supérieure fait intervenir deux termes antagonistes, le risque empirique R_n et la largeur de l'intervalle de confiance C. Étant donné un ensemble d'apprentissage, la VC-dimension du modèle permet de contrôler ce compromis comme préconisé par Vapnik avec le principe de minimisation du risque structurel. Pour cela, au lieu d'envisager une seule

famille de fonctions, on considère une séquence de plusieurs familles $\mathcal{F}_k$ imbriquées, selon

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_k \subset \ldots,$$

en tenant compte de certaines contraintes techniques énoncées dans [Vap95]. Un exemple particulier d'une telle structure correspond aux familles $\mathcal{F}_k$ de fonctions polynômiales de degré k. En notant h_k la VC-dimension supposée finie associée à $\mathcal{F}_k$, on peut alors déduire que

$$h_1 \leq h_2 \leq \ldots \leq h_k \leq \ldots.$$

Cette séquence croissante des VC-dimensions a deux conséquences indissociables. On a d'une part une séquence décroissante de risques empiriques correspondant à la fonction optimale de chaque famille, soit

$$\inf_{\psi \in \mathcal{F}_1} R_n(\psi) \geq \inf_{\psi \in \mathcal{F}_2} R_n(\psi) \geq \ldots \geq \inf_{\psi \in \mathcal{F}_k} R_n(\psi) \geq \ldots.$$

D'autre part, on a une séquence croissante du terme relatif à l'intervalle de confiance dans (A.5) tel que

$$C(h_1, n, \epsilon) \leq C(h_2, n, \epsilon) \leq \ldots C(h_k, n, \epsilon) \leq \ldots.$$

En représentant ces résultats sur la Figure A.3, on résume le principe de minimisation du risque structurel par les deux étapes suivantes :

1. Pour chaque famille $\mathcal{F}_k$, déterminer la fonction optimale minimisant l'erreur empirique

$$\psi^*_{n,\mathcal{F}_k} = \arg \min_{\psi \in \mathcal{F}_k} R_n(\psi).$$

2. Parmi toutes les fonctions optimales obtenues, opter pour celle garantissant la borne supérieure $R_n(\psi^*_{n,\mathcal{F}_k}) + C(h_k, n, \epsilon)$ la plus favorable,

$$\psi^*_n = \arg \min_{k \geq 1} \{R_n(\psi^*_{n,\mathcal{F}_k}) + C(h_k, n, \epsilon)\}.$$

La viabilité théorique de ce principe se heurte à quelques difficultés lors de sa mise en œuvre. Parmi celles-ci, notons la nécessité de connaître la VC-dimension des familles $\mathcal{F}$. De plus, en élaborant un résultat indépendamment de la distribution de probabilité conjointe $P(x, y)$, le Théorème 9 propose une borne supérieure souvent surestimée. Afin de pallier ces inconvénients, des techniques de validation croisée ou de ré-échantillonnage sont souvent considérées, comme étudié par exemple dans [SBSS99] pour les SVM.

A.3 Support vector machines

Au paragraphe précédent, on a posé les fondations de la théorie de l'apprentissage statistique à partir du concept de VC-dimension. Ce paragraphe est dédié à la mise en œuvre de ce cadre théorique à l'aide des SVM. Le critère considéré correspond à la maximisation de la marge entre l'hyperplan séparateur recherché et les éléments de chaque classe de l'ensemble d'apprentissage, comme illustré à la Figure A.4.

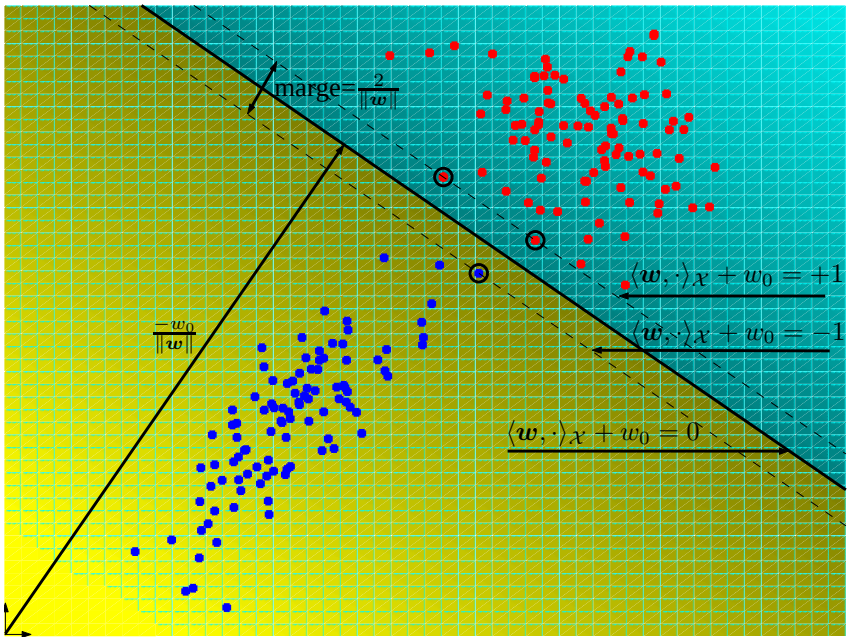


FIG. A.4 – Schéma d'un séparateur avec les différentes quantités associées aux SVM. Les support vectors sont identifiés par des cercles

A.3.1 VC-dimension et discrimination linéaire

Depuis leur introduction au début des années 90 dans [BGV92], les SVM ont permis des avancées considérables en reconnaissance des formes, d'une part avec des propriétés de généralisation soutenues par la théorie de l'apprentissage statistique, et d'autre part avec des algorithmes sans cesse plus performants grâce à l'usage du coup du noyau et l'évolution des techniques d'optimisation.

Soit $\mathcal{A}_n$ un ensemble d'apprentissage de n données $x_i \in \mathcal{X}$, avec leurs étiquettes $y_i = \pm 1$. En supposant que ces données sont linéairement séparables, il existe alors une infinité d'hyperplans séparateurs définis par des expressions de la forme $\psi(x) = \langle \boldsymbol{w}, x \rangle_{\mathcal{X}} + w_0$ telles que $y_i \psi(x_i) \geq 1$. On désigne par $\mathcal{F}_\delta$ l'ensemble des hyperplans séparateurs distants d'au moins δ des éléments des deux classes, soit

$$\mathcal{F}_\delta = \{\psi(\cdot) = \langle \boldsymbol{w}, \cdot \rangle_{\mathcal{X}} + w_0 \ : \ |\psi(x)| \geq 1, \|\boldsymbol{w}\| \leq 1/\delta\}.$$

Le théorème suivant est dû à Vapnik [Vap82].

Théorème 10. *La VC-dimension $h_{\mathcal{F}_\delta}$ de la famille $\mathcal{F}_\delta$ est bornée selon la relation*

$$h_{\mathcal{F}_\delta} \leq \min \left\{ \frac{R^2}{\delta^2}, l \right\} + 1,$$

où les données d'apprentissage sont contenues dans une boule de rayon R centrée sur l'origine.

Ce théorème permet la construction d'une séquence de familles imbriquées selon $\mathcal{F}_{\delta_1} \subset \ldots \subset \mathcal{F}_{\delta_k}$, obtenue en considérant différentes valeurs de largeurs de marge regroupées par ordre décroissant. On peut alors obtenir de bonnes propriétés de généralisation en considérant la marge optimale, à partir du principe de la minimisation du risque structurel évoqué à la section précédente.

A.3.2 Cas de données linéairement séparables

Soit $\mathcal{A}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un ensemble d'apprentissage de données linéairement séparables, représentées par $x_i \in \mathcal{X}$, avec leurs étiquettes $y_i = \pm 1$. Un hyperplan est défini dans cet espace par $\boldsymbol{w}$ et w_0 selon l'expression $\langle \boldsymbol{w}, x \rangle_{\mathcal{X}} + w_0 = 0$. La distance d'un élément x_i à cet hyperplan est alors donnée par

$$\delta_i = \frac{|\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0|}{\|\boldsymbol{w}\|} \tag{A.7}$$

Afin de construire un hyperplan séparateur, on contraint la solution recherchée en imposant

$$\min_{i=1,\ldots,n} |\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0| = 1.$$

Celle-ci se traduit par les inégalités suivantes

$$\begin{aligned} \langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0 \geq +1, \quad \text{si} \quad y_i = +1 \\ \langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0 \leq -1, \quad \text{si} \quad y_i = -1. \end{aligned}$$

En arrangeant ces relations, on obtient pour tout $i = 1, \ldots, n$ la contrainte

$$y_i(\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0) \geq 1. \tag{A.8}$$

En combinant cette dernière avec l'expression (A.9), on obtient une contrainte sur la distance des éléments de l'ensemble d'apprentissage de l'hyperplan séparateur, selon

$$\delta_i \geq \frac{1}{\|\boldsymbol{w}\|}, \tag{A.9}$$

pour tout $i = 1, \ldots, n$. Par conséquent, l'hyperplan optimal au sens de la marge maximale est donné par la minimisation de $\frac{1}{2}\|\boldsymbol{w}\|^2$ sous la contrainte (A.8). Par l'usage de ce critère convexe, soumis à des contraintes linéaires, la solution optimale est obtenue par le point selle du Lagrangien donné par

$$\frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \big(y_i(\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0) - 1\big), \tag{A.10}$$

où $\alpha_1, \ldots, \alpha_n \geq 0$ désignent les multiplicateurs de Lagrange. Ce point-selle correspond à l'annulation des dérivées partielles par rapport aux différentes variables recherchées et aux multiplicateurs de Lagrange. L'optimalité de ceux-ci étant désignée par $\boldsymbol{w}^*$, w_0^* et α_i^*, on aboutit alors aux conditions suivantes :

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i \tag{A.11}$$

$$\sum_{i=1}^{n} \alpha_i^* y_i = 0 \tag{A.12}$$

En injectant ces deux résultats dans l'expression du Lagrangien (A.10), le problème d'optimisation se traduit alors par la maximisation de la forme duale du Lagrangien

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, \langle x_i, x_j \rangle_{\mathcal{X}}, \tag{A.13}$$

sous la contrainte $\sum_{i=1}^{n} \alpha_i y_i = 0$, ainsi que la positivité des multiplicateurs de Lagrange α_i. On sait que les coefficients ainsi obtenus vérifient la condition de Kühn-Tucker

$$\alpha_i \big(y_i(\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0) - 1 \big) = 0, \tag{A.14}$$

pour $i = 1, \ldots, n$. En conséquence, seules les données x_i satisfaisant l'égalité dans la contrainte (A.8), $y_i(\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0) = 1$, admettent des multiplicateurs de Lagrange α_i non-nuls. Ces données vérifient $\delta_i = 1$, ce qui signifie qu'il n'existe pas de donnée plus proche de l'hyperplan optimal que x_i. De tels échantillons sont appelés *support vectors* puisqu'ils définissent à eux seuls l'hyperplan optimal, caractérisé par $\boldsymbol{w}^*$ et w_0^*. Le premier est donné par l'expression (A.11), les échantillons correspondant à α_i non-nul n'y contribuant pas. Le second paramètre w_0^* permet de situer l'hyperplan à mi-distance des deux classes. On exprime alors le seuil selon

$$w_0^* = \frac{1}{2} \big(\langle \boldsymbol{w}^*, x_j \rangle_{\mathcal{X}} + \langle \boldsymbol{w}^*, x_k \rangle_{\mathcal{X}} \big),$$

où x_j et x_k désignent deux support vectors appartenant à deux classes différentes. Plus généralement, les support vectors regroupent l'information discriminante de l'ensemble d'apprentissage. Les autres échantillons pourraient être retirés de ce dernier et la répétition de l'algorithme à partir des support vectors uniquement produirait le même hyperplan optimal.

Quelques avantages de l'approche SVM

Les support vectors contribuent à l'attrait pour les SVM en raison du caractère parcimonieux des solutions obtenues. Ceci est dû au critère de maximisation de la marge, qui a par ailleurs d'autres effets positifs pour l'approche SVM. On a d'une part un problème d'optimisation quadratique, n'admettant donc pas de minima local mais une solution unique si la matrice formée par les quantités $y_i y_j \langle x_i, x_j \rangle_{\mathcal{X}}$ est de rang plein. D'autre part, le résultat obtenu est relativement robuste aux faibles variations des paramètres en question. Ces avantages constituent autant d'avancées considérables pour les méthodes de reconnaissance des formes classiques. L'approche des réseaux de neurones par exemple nécessite la résolution d'un problème d'optimisation admettant des optimum locaux. De plus, la parcimonie de la solution nécessite une étape supplémentaire d'élagage. Au-delà de ces propriétés intéressantes, on rappelle que la maximisation de la marge est motivée par le principe de minimisation du risque structurel. Celui-ci correspond à la recherche d'un compromis entre l'erreur empirique et la richesse de l'ensemble des fonctions candidates, comme illustré à la Figure A.3. Les deux classes étant linéairement séparables, on se restreint à l'ensemble des hyperplans séparateurs, produisant donc une erreur empirique nulle. Parmi ceux-ci, on cherche celui avec les meilleurs propriétés de généralisation, mesurée par une faible VC-dimension conformément au Théorème 10.

Plus généralement, le principe de minimisation du risque structurel s'applique dans le cas de classes non-linéairement séparables. La mise en œuvre des SVM pour ce type de problèmes est traité dans la section suivante.

A.3.3 Cas de classes non-linéairement séparables

Précédemment, on s'est restreint aux problèmes où les données d'apprentissage sont linéairement séparables. Ceci n'est pas le cas en général. On peut toutefois considérer un hyperplan séparant les deux classes en question, en tolérant des erreurs de classification sur certains échantillons. L'erreur de classification de l'échantillon x_i est quantifiée par les variables positives de pénalisation ξ_i, dites *slack variables* ou variables de relaxation, ce qui permet de généraliser les contraintes dans (A.8) ainsi

$$y_i(\langle \boldsymbol{w}, x_i \rangle_{\mathcal{X}} + w_0) \geq 1 - \xi_i, \tag{A.15}$$

pour tout $i = 1, \ldots, n$, avec $\xi_1, \ldots, \xi_n \geq 0$. La minimisation de l'erreur totale $\sum_{i=1}^{n} \xi_i$, combinée avec le critère de la la marge, correspond à la minimisation du critère

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i,$$

sous les contraintes (A.15) avec C un paramètre permettant une pondération relative des deux termes antagonistes : une solution très régulière est obtenue pour $C = 0$ tandis qu'on aboutit à une classification parfaite pour $C = \infty$. En d'autres termes, il s'agit d'un paramètre de régularisation permettant un contrôle supplémentaire de la capacité du classifieur résultant. On retrouve la même formulation que celle de la théorie de la régularisation selon Tikhonov présentée dans la Section 1.1.2, avec $\eta = 1/C$.

La solution de ce problème d'optimisation avec contraintes est donnée par le point selle du Lagrangien

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i - \sum_{i=1}^{n} \alpha_i \big(y_i(\langle \boldsymbol{w}, x_i\rangle_{\mathcal{X}} + w_0) - 1 + \xi_i\big) - \sum_{i=1}^{n} \beta_i \xi_i,$$

α_i et β_i désignant les multiplicateurs de Lagrange. La minimisation du Lagrangien par rapport à $\boldsymbol{w}$, w_0 et les ξ_i produit les conditions suivantes :

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

$$\sum_{i=1}^{n} \alpha_i^* y_i = 0$$

$$\alpha_i^* + \beta_i^* = C$$

Le problème dual est alors obtenu en injectant ces expressions dans la forme primaire, et la solution est donnée par

$$\boldsymbol{\alpha} = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, \langle x_i, x_j\rangle_{\mathcal{X}} - \sum_{i=1}^{n} \alpha_i, \tag{A.16}$$

sous les contraintes

$$0 \geq \alpha_i \geq C \qquad i = 1, \ldots, n$$

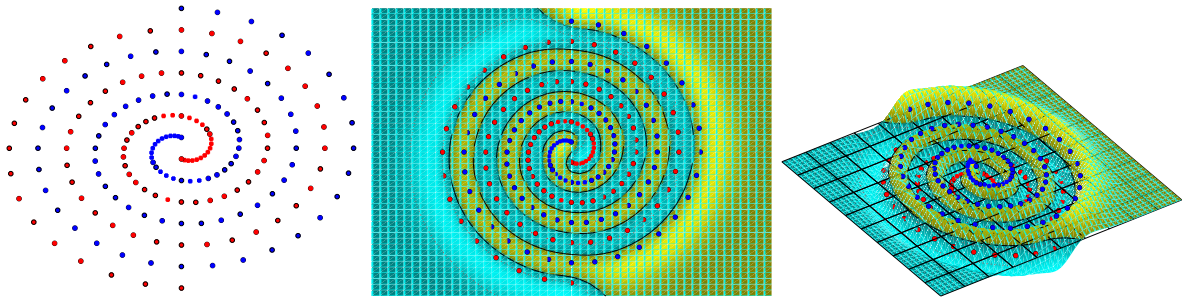
$$\sum_{i=1}^{n} \alpha_i y_i = 0.$$

La résolution de ce problème est similaire à celle obtenue dans le cas de classes linéairement séparables, avec une contrainte supplémentaire sur les α_i.

A.4 Mise en œuvre des SVM dans le domaine temps-fréquence

A.4.1 SVM dans un RKHS

Les différentes expressions montrent que la méthode présentée dans les Sections A.3.2 et A.3.3 se prête aisément à une généralisation dans un espace transformé à noyau reproduisant. On considère le noyau reproduisant κ et l'espace de Hilbert associé $\mathcal{H}$, l'échantillon x_i est alors représenté dans cet



(a) (b) (c)

FIG. A.5 – Résultat obtenu par les SVM pour la résolution du problème de classification de deux classes sous forme de spirales enroulées l'une autour de l'autre.

espace par κ_{x_i}, et le produit scalaire est désigné par $\kappa(x_i, x_j)$, pour tout $x_i, x_j \in \mathcal{X}$. La forme duale du Lagrangien dans (A.13) ou (A.16), s'écrit alors

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \, \kappa(x_i, x_j), \tag{A.17}$$

sous des contraintes inchangées. La résolution de ce problème d'optimisation avec contrainte détermine les multiplicateurs de Lagrange, et par conséquent l'hyperplan optimal dans l'espace fonctionnel $\mathcal{H}$, défini par les paramètres

$$\boldsymbol{w}^* = \sum_{i=1}^{n} \alpha_i^* y_i \kappa_{x_i} \tag{A.18}$$

et le seuil par

$$w_0^* = \frac{1}{2} \sum_{i=1}^{n} \alpha_i^* y_i \big(\kappa(x_i, x_j) + \kappa(x_i, x_k) \big), \tag{A.19}$$

où x_j et x_k désignent deux support vectors appartenant à deux classes différentes. On retrouve les effets du Théorème de Représentation dans l'expression de $\boldsymbol{w}^*$, et ceux de la propriété reproduisante dans l'expression du seuil. La règle de décision consiste à comparer la statistique $\langle \boldsymbol{w}^*, \kappa_x \rangle_{\mathcal{H}}$ au seuil w_0^*, soit

$$\Lambda(x) = \sum_{i=1}^{n} \alpha_i^* y_i \kappa(x, x_i) \underset{H_0}{\overset{H_1}{\gtrless}} \frac{1}{2} \sum_{i=1}^{n} \alpha_i^* y_i \big(\kappa(x_i, x_j) + \kappa(x_i, x_k) \big).$$

L'évaluation de la règle de décision ne nécessite à aucun moment d'exhiber l'espace de représentation $\mathcal{H}$, ni les fonctions noyau de l'ensemble de données disponibles, une fois que leur produit scalaire est connu. Afin d'illustrer l'efficacité des SVM avec un noyau reproduisant non-linéaire, on considère le problème classique de classification de deux classes prenant la forme de spirales enroulées l'une autour de l'autre, voir Figure A.5 (a). Un classifieur linéaire ne peut évidemment pas mener à une discrimination satisfaisante. On considère alors le noyau Gaussien, défini par $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / 2\sigma_0^2)$ où σ_0 est le paramètre de largeur de bande. Comme illustré à la Figure A.5 $(b) - (c)$, ce noyau permet une discrimination entre les deux classes.

A.4.2 SVM dans le domaine temps-fréquence

Dans le cadre d'un problème de classification de signaux non-stationnaires, on propose de profiter des avancées théoriques et des résultats algorithmiques présentés dans cette annexe. Pour cela, on considère

la mise en œuvre de l'approche des SVM dans le domaine temps-fréquence, par un choix approprié du noyau reproduisant κ. Dans ce cas, les données représentées à la Figure A.4 représentent les distributions temps-fréquence des signaux d'apprentissage, soit κ_{x_i} représentant x_i pour $i = 1, \ldots, n$. L'hyperplan recherché est défini par $\boldsymbol{w}^*$ selon l'expression (A.18), ce qui correspond à une combinaison linéaire des distributions temps-fréquence disponibles. Au lieu de parler d'hyperplan, on désigne cette combinaison par la signature séparatrice, $\Psi(t, f)$, tout en rappelant qu'il ne s'agisse pas forcément d'une distribution temps-fréquence valide. La maximisation de la marge admet une interprétation temps-fréquence. En effet, maximiser la distance de cette signature aux représentations des signaux d'apprentissage, selon l'expression (A.9), correspond à minimiser la norme de $\boldsymbol{w}^*$, ce qui revient dans le domaine temps-fréquence à minimiser la norme définie dans le RKHS de $\Psi(t, f)$. La minimisation de ce terme impose alors une certaine régularité à la signature résultante, donc une certaine robustesse. Le produit scalaire de cette signature avec la distribution temps-fréquence d'un signal arbitraire permet, en le comparant au seuil donné par l'expression (A.19), de déterminer son appartenance à l'une des deux classes. On présente au paragraphe 5.3.3 des applications sur la mise en œuvre des SVM dans le domaine temps-fréquence pour plusieurs distributions de la classe de Cohen, ainsi que l'approche hybride. On compare les performances résultantes avec celles obtenues par l'AFD-à-noyau.

Annexe B

Noyaux (reproduisants) classiques

	Nom du noyau[14]	Expression mathématique
Noyau projectif	monomial	$(\langle x_i, x_j\rangle)^p$
	polynomial	$(c + \langle x_i, x_j\rangle)^p$
	polynomial de Vovk	$\frac{1-\langle x_i,x_j\rangle^p}{1-\langle x_i,x_j\rangle}$
	polynomial inifini	$(1 - \langle x_i, x_j\rangle)^{-p}$
	exponentiel	$\exp(\langle x_i, x_j\rangle/2\sigma^2)$
	sigmoïde (perceptron)	$\tanh(\langle x_i, x_j\rangle/\sigma + c)$
Noyau radial	Gaussien	$\exp(-\|x_i - x_j\|^2/2\sigma^2)$
	Laplacien	$\exp(-\|x_i - x_j\|/2\sigma^2)$
	multiquadrique	$(\|x_i - x_j\|^2 + c^2)^{1/2}$
	multiquadrique inverse	$(\|x_i - x_j\|^2 + c^2)^{-1/2}$
	splines (*thin plate*)	$\|x_i - x_j\|^{2n+1}$
	splines log (*thin plate*)	$\|x_i - x_j\|^{2n} \log(\|x_i - x_j\|)$
	B-splines sur $\mathbb{R}$	$B_{2n+1}(x_i - x_j)$
	trigonométrique polynomial sur $\mathbb{R}$	$\frac{\sin(d+1/2)(x_i-x_j)}{\sin(x_i-x_j)/2}$
	de Fourier sur $\mathbb{R}$ (régularisation faible)	$\frac{\pi}{2\gamma}\frac{\cosh\frac{\pi-\|x_i-x_j\|}{\gamma}}{\sinh\pi/\gamma}$
	de Fourier sur $\mathbb{R}$ (régularisation forte)	$\frac{1-\gamma^2}{2(1-2\gamma\cos(x_i-x_j)+\gamma^2)}$
	Anova 1	$\left(\sum_{k=1}^n \exp((x_i(k)-x_j(k))^2\theta)\right)^2$
	quadratique rationnel	$1 - \frac{\|x_i-x_j\|^2}{c+\|x_i-x_j\|^2}$
	circulaire sur $\mathbb{R}^2$	$\frac{2}{\pi}\arccos\frac{\|x_i-x_j\|}{\theta} - \frac{2}{\pi}\frac{\|x_i-x_j\|}{\theta}\sqrt{1-\frac{\|x_i-x_j\|^2}{\theta^2}}$
	sphérique sur $\mathbb{R}^3$	$1-\frac{3}{2}\frac{\|x_i-x_j\|}{\theta}+\frac{1}{2}\left(\frac{\|x_i-x_j\|}{\theta}\right)^3$
	onde $\mathbb{R}^3$	$\frac{\theta}{\|x_i-x_j\|}\sin\|x_i - x_j\|/\theta$

[14] Certains noyaux ne sont pas définis positifs que pour des valeurs particulières de leurs paramètres, comme c'est le cas de la fonction sigmoïde.

Annexe C

Méthodes à noyau les plus connues

	Algorithme	Fonction coût $V(\psi(x_i), y_i)$
Non-supervisés	Estimation de densité [Vap95]	$-\log(\psi(x_i))$
	Analyse en composantes principales [VTS04]	$-\lvert\psi(x_i)\rvert^2/\lVert\psi\rVert^2_{\mathcal{H}}$
	Projection pursuit [Sun98]	skewness$\{\psi(x_i)\}$, kurtosis$\{\psi(x_i)\}$, entropie$\{\psi(x_i)\}$, ...
	Variétés principales régularisées [SWMS99]	
	Detection de nouveauté (ν-SVM) [KSW04, SWS$^+$00]	$\max\{\rho - \psi(x_i), 0\} - \nu\rho$
Moindres carrés	Régression ridge [SGV98b, HTF01]	$(y_i - \psi(x_i))^2$
	Least square support vector machine [SV99, SGB$^+$02]	$(y_i - \psi(x_i))^2$
	Regularized least square classification [CB04, Rif02]	$(y_i - \psi(x_i))^2$
	Réseaux de régularisation [EPP99]	$(y_i - \psi(x_i))^2$
	Proximal support vector machine [FM01]	$(y_i - \psi(x_i))^2$
	Moindres carrés modifiés [Zha04]	$\max\{y_i - \psi(x_i), 0\}^2$
Régression	Support Vector Regression [Vap95]	
	ϵ-insensitive	$\max\{\lvert y_i - \psi(x_i)\rvert - \epsilon, 0\}$
	ϵ-insensitive quadratique	$\max\{\lvert y_i - \psi(x_i)\rvert - \epsilon, 0\}^2$
	Régression en quantile [TLSS06]	$\tau(y_i - \psi(x_i))$ si $y_i \geq \psi(x_i)$ $(\tau - 1)(y_i - \psi(x_i))$ sinon
	AdaBoost [Vap95, HTF01]	$\exp(-y_i\psi(x_i))$
	Régression robuste de Huber [Hub81, Vap95]	$\frac{1}{4\epsilon}\lvert y_i - \psi(x_i)\rvert^2$ si $\lvert y_i - \psi(x_i)\rvert \leq 2\epsilon$ $\lvert y_i - \psi(x_i)\rvert - \epsilon$ sinon
Classification	Support Vector Classification : [Vap95]	
	Marge dure (indicateur)	$\mathbb{1}_{1-y_i\psi(x_i)}$
	Mal-classification (indicateur)	$\mathbb{1}_{-y_i\psi(x_i)}$
	Marge diffuse (hinge)	$\max\{1 - y_i\psi(x_i), 0\}$
	Marge diffuse (mal-classification)	$\max\{-y_i\psi(x_i), 0\}$
	Marge diffuse (hinge) quadratique	$\max\{1 - y_i\psi(x_i), 0\}^2$
	Régression logistique (à noyau)	$\log(1 + \exp(-y_i\psi(x_i)))$
	Import Vector Machine [ZH02]	$y_i\psi(x_i) - \log(1 + \exp(\psi(x_i)))$
	ν-SVM [KSW04]	$\max\{\rho - y_i\psi(x_i), 0\} - \nu\rho$

Bibliographie

[Abe03] S. Abe. Analysis of multiclass support vector machines. In *Proc. International Conference on Computational Intelligence for Modelling Control and Automation (CIMCA'2003)*, pages 385–396, Vienna, Austria, 2003.

[ABR64] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25 :821–837, 1964.

[AD05] F. Abrard and Y. Deville. A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources. *Signal Processing*, 85(7) :1389–1403, 2005.

[AF95] F. Auger and P. Flandrin. Improving the readability of time-frequency and time-scale representations by the reassignment method. *IEEE Trans. Signal Processing*, 43(5) :1068–1089, 1995.

[AFGL05] F. Auger, P. Flandrin, P. Gonçalves, and O. Lemoine. *Time-Frequency Toolbox Reference Guide (for use with MATLAB)*. CNRS, 2005. Publié sous les auspices du Centre National de la Recherche Scientifique (CNRS), France et de la Rice University, USA.

[Ami94] M.G. Amin. Spectral decomposition of time-frequency distribution kernels. *IEEE Trans. Signal Processing*, 42 :1156–1165, May 1994.

[Aro50] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68 :337–404, 1950.

[AW99] S. Amari and S. Wu. Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, 12(6) :783–789, July 1999.

[BA98] A. Belouchrani and M.G. Amin. Blind source separation based on time-frequency signal representations. *IEEE Trans. Signal Processing*, 46 :2888–2897, November 1998.

[BA00] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural Computation*, 12(10) :2385–2404, 2000.

[BA01] G. Baudat and F. Anouar. Kernel-based methods and function approximation. In *In International Joint Conference on Neural Networks (IJCNN)*, volume 5, pages 1244–1249, Washington, DC, USA, July 2001.

[Bal81] R. Balian. Un principe d'incertitude fort en théorie du signal ou en mécanique quantique. *C. R. Acad. Sci. Paris*, 292 :1357–1362, 1981.

[Bar92] R.G. Baraniuk. Shear madness : Signal-dependent and metaleptic time-frequency representation. Technical Report UILU-ENG-92-2226, Coordinated Science Laboratory, University of Illinois, Urbana, 1992.

[BD95] J.B. Buckheit and D.L. Donoho. Improved linear discrimination using time-frequency dictionaries. In *Proc. SPIE-Wavelet Applications in Signal and Image Processing III*, volume 2569 (2), pages 540–551, sept 1995.

[BGV92] B. Boser, I. Guyon, and V.N. Vapnik. An training algorithm for optimal margin classifiers. In *Proc. 5th Annual Workshop on Computational Learning Theory*, pages 144–152, 1992.

[BJ93] R.G. Baraniuk and D.L. Jones. Signal-dependent time-frequency analysis using a radially gaussian kernel. *Signal Processing*, 32(3) :263–284, 1993.

[BJ03] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3 :1–48, 2003.

[BN91] D. Beyerbach and H. Nawab. Principal components analysis of the short-time fourier transform. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 0 :1725–1728, 1991.

[BO90] B. Boashash and P. O'Shea. A methodology for detection and classification of some underwater acoustic signals using time-frequency analysis techniques. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11) :1829–1841, 1990.

[Bur96] C.J.C. Burges. Simplified support vector decision rules. In *International Conference on Machine Learning*, pages 71–77, 1996.

[Bur98] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2) :121–167, 1998.

[Cal64] A.P. Calderón. Intermediate spaces and interpolation, the complex method. *Studia Math.*, 24 :113–190, 1964.

[Cap06] A. Caponnetto. Optimal rates for regularization operators in learning theory. Technical Report Technical Report MIT-CSAIL-TR-2006-062, Computer Science and Artificial Intelligence Laboratory, massachusetts institute of technology, Cambridge, MA, USA, September 2006.

[CB86] S. Chen and S. A. Billings. Neural networks for nonlinear dynamic system modeling and identification. *International Journal of Control*, 56(2) :319–346, 1986.

[CB04] N. Cesa-Bianchi. Applications of regularized least squares to classification problems. In S. Ben-David, J. Case, and A. Maruoka, editors, *Proc. 15th International Conference ALT 2004 : Algorithmic Learning Theory, October 2-5, 2004*, volume 3244, pages 14–18, Padova, Italy, 2004. Springer.

[CDS98] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1) :33–61, 1998.

[CF99] M.J. Coates and W.J. Fitzgerald. Regionally optimised time-frequency distributions using finite mixture models. *Signal Processing*, 77(3) :247–260(14), September 1999.

[CH02] Z. Chen and S. Haykin. On different facets of regularization theory. *Neural Comput.*, 14(12) :2791–2846, 2002.

[CKEST06] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor. On kernel target alignment. In D. Holmes and L. Jain, editors, *Innovations in Machine Learning : Theory and Application*, pages 205–255. Springer Verlag, 2006.

[CKS+03] K-M. Chung, W-C. Kao, C-L. Sun, L-L. Wang, and C-J. Lin. Radius margin bounds for support vector machines with the rbf kernel. *Neural Comput.*, 15(11) :2643–2681, 2003.

[CM80] T.A.C.M. Claasen and W.F.G. Mecklenbrauker. The wigner distribution — a tool for time-frequency signal analysis. part i : Continuous-time signals. *Philips Journal of Research*, 35 :217–250, 1980.

[CM05] E. Chassande-Mottin. Géométrie des ensembles de chirps et détection des ondes gravitationnelles. In *Actes du XX^{ème} Colloque GRETSI sur le Traitement du Signal et des Images*, pages 261–264, Louvain-la-Neuve (Belgique), 2005.

[CMAF05] E. Chassande-Mottin, F. Auger, and P. Flandrin. La réallocation. In F. Hlawatsch and F. Auger, editors, *Temps-Fréquence - Concepts et Outils*, Traité IC2 "Information, Communications, Contrôle", pages 259–288. Hermes, 2005.

[CMF99] E. Chassande-Mottin and P. Flandrin. On the time-frequency detection of chirps. *Appl. Comp. Harm. Anal.*, 6(9) :252–281, 1999.

[CMP05] E. Chassande-Mottin and A. Pai. Discrete time and frequency wigner-ville distribution : Moyal's formula and aliasing. *IEEE Signal Processing Letters*, 12(7) :508–511, 2005.

[CO01] Lehel Csató and Manfred Opper. Sparse representation for gaussian process models. In Todd K. Leen, Thomas G. Dietterich, and Volker Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 444–450. MIT Press, 2001.

[Cov65] T.M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14 :326–334, 1965.

[CRLS06] I. Constantin, C. Richard, R. Lengellé, and L. Soufflet. Nonlinear regularized Wiener filtering with kernels. Application in denoising MEG data corrupted by ECG. *IEEE Trans. Signal Processing*, 2006.

[CS00] K. Crammer and Y. Singer. On the learnability and design of output codes for multiclass problems. In *Proc. the Thirteenth Annual Conference on Computational Learning Theory (COLT '00)*, pages 35–46, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[CS02a] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2 :265–292, 2002.

[CS02b] F. Cucker and S. Smale. Best choices for regularization parameters in learning theory : on the bias-variance problem. *Found. Comput. Math.*, 2(4) :413–428, 2002.

[CS02c] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39(1) :1–49, 2002.

[CSS02] M. Collins, R.E. Schapire, and Y. Singer. Logistic regression, adaboost and Bregman distances. *Mach. Learn.*, 48(1-3) :253–285, 2002.

[CSTEK01] N. Cristianini, J. Shawe-Taylor, A. Elisseeff, and J. Kandola. On kernel-target alignment. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Proc. Neural Information Processing Systems (NIPS) 14*, pages 367–373. MIT Press, December 2001.

[CT03] G.C. Cawley and N.L.C. Talbot. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. *Pattern Recognition*, 36(11) :2585–2592, November 2003.

[CVBM02] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1-3) :131–159, 2002.

[Dav00] M. Davy. *Noyaux optimisés pour la classification dans le plan temps-fréquence – Proposition d'un algorithme constructif et d'une référence bayésienne basés sur les méthodes MCMC – Application au diagnostic d'enceintes acoustiques*. Thèse de doctorat, Université de Nantes, September 2000.

[Dav04] M. Davy. Classification. In C. Doncarli and N. Martin, editors, *Décision dans le plan temps-fréquence*, pages 147–175, Paris, 2004. Hermès Sciences, Traité IC2.

[DD98] M. Davy and C. Doncarli. Optimal kernels of time-frequency representations for signal classification. In *Proc. of the IEEE International Symposium on Time-Frequency and Time-Scale analysis*, pages 581–584, Pittsburgh, USA, October 1998. IEEE Signal Processing Society.

[DD99] M. Davy and C. Doncarli. Distances et critères de contraste dans le plan temps-fréquence. In *Actes du XVII^{ème} Colloque GRETSI sur le Traitement du Signal et des Images*, volume 2, pages 287–290, Vannes, France, 13-17 septembre 1999.

[DDBB01] M. Davy, C. Doncarli, and G.F. Boudreaux-Bartels. Improved optimization of time-frequency based signal classifiers. *IEEE Signal Processing Letters*, 8(2) :52–57, 2001.

[DE03] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proceedings - National Academy of Sciences (PNAS)*, 100(5) :2197–2202, March 2003.

[DGDR02] M. Davy, A. Gretton, A. Doucet, and P.W.J. Rayner. Optimised support vector machines for nonstationary signal classification. *IEEE Signal Processing Letters*, 9(12) :442–445, 2002.

[DH01] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Information Theory*, 47(7) :2845–2862, March 2001.

[DH02] T. J. Dodd and C. J. Harris. Identification of nonlinear time series via kernels. *International Journal of System Science*, 33(9) :737–750, 2002.

[DHS00] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[DK96] K.I. Diamantaras and S.Y. Kung. *Principal component neural networks : theory and applications*. John Wiley & Sons, Inc., New York, NY, USA, 1996.

[DKH03] T. J. Dodd, V. Kadirkamanathan, and R. F. Harrison. Function estimation in Hilbert space using sequential projections. In *Proc. IFAC Conference on Intelligent Control Systems and Signal Processing*, pages 113–118, 2003.

[DKP03] K. Duan, S.S. Keerthi, and A.N. Poo. Evaluation of simple performance measures for tuning svm hyperparameters. *Neurocomputing*, 51 :41–59, 2003.

[DMCB00] M. Dutat, I. Magrin-Chagnolleau, and F. Bimbot. Analyse en composantes principales temps-fréquence : application à la reconnaissance de la langue. In *Proceedings of JEP 2000*, June 2000. Aussois, France.

[EBG05] W.J. Williams E.M. Bernat and W.J. Gehring. Decomposing erp time-frequency energy using pca. *Clinical Neurophysiology*, 116(6) :1314–1334, June 2005.

[EMM04] Y. Engel, S. Mannor, and R. Meir. The kernel recursive least squares algorithm. *IEEE Trans. Signal Processing*, 52(8) :2275–2285, 2004.

[EPM00] A. Elisseeff and H. Paugam-Moisy. A new multi-class svm based on a uniform convergence result. In *Proc. the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)*, volume 4, page 4183, Washington, DC, USA, 2000. IEEE Computer Society.

[EPP99] T. Evgeniou, M. Pontil, and T. Poggio. A unified framework for regularization networks and support vector machines. Technical report, Massachusetts Institute of Technology, Cambridge, MA, USA, 1999.

[FCC98] T.-T. Frieß, N. Cristianini, and C. Campbell. The kernel-adatron algorithm : A fast and simple learning procedure for support vector machines. In *Proc. Fifteenth International Conference on Machine Learning (ICML'98)*, pages 188–196, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[FDBR04] G. Fung, M. Dundar, J. Bi, and B. Rao. A fast iterative algorithm for fisher discriminant using heterogeneous kernels. In *Proc. twenty-first international conference on Machine learning (ICML'04)*, page 40, New York, NY, USA, 2004. ACM Press.

[Fis36] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 :179–188, 1936.

[Fla86] P. Flandrin. On detection-estimation procedures in the time-frequency plane. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2331–2334, Tokyo, Japan, 1986.

[Fla88] P. Flandrin. A time-frequency formulation of optimum detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(9) :1377–1384, 1988.

[Fla98] P. Flandrin. *Time-Frequency/Time-Scale Analysis*. Academic Press, San Diego, CA, USA, 1998.

[FM01] G. Fung and O.L. Mangasarian. Proximal support vector machine classifiers. In *Proc. seventh ACM SIGKDD International Conference on Knowledge Discovery and data mining (KDD '01)*, pages 77–86, New York, NY, USA, 2001. Asscociation for Computing Machinery Press.

[FS91] M.L. Fowler and L.H. Sibul. A unified formulation for detection using time-frequency and time-scale methods. In *Proc. 25th Asilomar Conference on Signals, Systems and Computers*, pages 637–642, Pacific Grove, CA, USA, 1991.

[FS97] G.B. Folland and A. Sitaram. The uncertainty principle : A mathematical survey. *Journal of Fourier Analysis and Applications*, 3(3) :207–238, 1997.

[FS02] S. Fine and K. Scheinberg. Efficient svm training using low-rank kernel representations. *Journal of Machine Learning Research*, 2 :243–264, 2002.

[GA01] B. Gillespie and L. Atlas. Optimizing time-frequency kernels for classification. *IEEE Trans. Signal Processing*, 49(3) :1341–1344, 2001.

[GDDR01] A. Gretton, M. Davy, A. Doucet, and P.W.J. Rayner. Nonstationary signal classification using support vector machine. In *Proc. IEEE Workshop on Statistical Signal Processing*, Singapore, August 2001.

[GGR04] M. Gandetto, M. Guainazzo, and C.S. Regazzoni. Use of time-frequency analysis and neural networks for mode identification in a wireless software-defined radio approach. *EURASIP Journal on Applied Signal Processing*, 2004(12) :1778–1790, 2004.

[Gir98] F. Girosi. An equivalence between sparse approximation and support vector machines. *Neural Computation*, 10(6) :1455–1480, 1998.

[Gir02] M. Girolami. Orthogonal series density estimation and the kernel eigenvalue problem. *Neural Computation*, 14 :669–688, 2002.

[GM84] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal of Mathematical Analysis*, 15 :723–736, 1984.

[GMS03] A. C. Gilbert, S. Muthukrishnan, and M. J. Strauss. Approximation of functions over redundant dictionaries using coherence. In *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms (SODA'03)*, pages 243–252, Philadelphia, PA, USA, 2003. Society for Industrial and Applied Mathematics.

[GMST03] A. C. Gilbert, S. Muthukrishnan, M. J. Strauss, and J. Tropp. Improved sparse approximation over quasi-incoherent dictionaries. In *International Conference on Image Processing (ICIP'03)*, volume 1, pages 37–40, Barcelona, Spain, Sept. 2003.

[GRG05] J. Gosme, C. Richard, and P. Gonçalvès. Adaptive diffusion as a versatile tool for time-frequency and time-scale representations processing : a review. *IEEE Trans. Signal Processing*, 53(11) :4136–4146, November 2005.

[GRSV07] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite ! average case analysis of multi-channel sparse recovery using greedy algorithms. Publication interne 1848, IRISA, May 2007.

[GS01] G. B. Giannakis and E. Serpedin. A bibliography on nonlinear system identification. *Signal Processing*, 81 :553–580, 2001.

[GSV07] S. Günter, N.N. Schraudolph, and S.V.N. Vishwanathan. Fast iterative kernel principal component analysis. *Journal of Machine Learning Research*, 2007. submitted.

[GV06] R. Gribonval and P. Vandergheynst. On the exponential convergence of Matching Pursuits in quasi-incoherent dictionaries. *IEEE Transactions on Information Theory*, 52(1) :255–261, 2006.

[Hay99] S. Haykin. *Neural networks : a comprehensive foundation*. Prentice Hall, Englewood Cliffs, NJ, 1999.

[Hay02] S. Haykin. *Adaptive filtering theory*. Prentice Hall, NJ, USA, fourth edition edition, 2002.

[HBB92] F. Hlawatsch and G.F. Boudreaux-Bartels. Linear and quadratic time-frequency signal representations. *IEEE Signal Processing Magazine*, 9(2) :21–67, April 1992.

[Hei95] C. Heitz. Optimum time-frequency representations for the classification and detection of signals. *Applied Signal Proceedings*, 3 :124–143, 1995.

[Her02] R. Herbrich. *Learning kernel classifiers. Theory and algorithms*. The MIT Press, Cambridge, MA, USA, 2002.

[HJ86] R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, New York, NY, USA, 1986.

[HK92] F. Hlawatsch and W. Krattenthaler. Bilinear signal synthesis. *IEEE Trans. Signal Processing*, 40(2) :352–363, February 1992.

[Hla98] F. Hlawatsch. *Time-Frequency Analysis and Synthesis of Linear Signal Spaces : Time-Frequency Filters, Signal Detection and Estimation, and Range-Doppler Estimation*. Kluwer Academic Publishers, Boston, MA, USA, 1998.

[HM01] F. Hlawatsch and G. Matz. Quadratic time-frequency analysis of linear time-varying systems. In L. Debnath, editor, *Wavelet transforms and time-frequency signal analysis*, pages 235–287. Birkhäuser, 2001.

[Hoe05] L. Hoegaerts. *Eigenspace Methods and Subset Selection in Kernel based Learning*. Phd thesis, Faculty of Engineering, K.U.Leuven, Leuven, Belgium, Jun. 2005.

[Hof06] H. Hoffmann. Kernel pca for novelty detection. *Pattern Recognition*, 2006. to appear.

[HR07a] P. Honeine and C. Richard. Distribution temps-fréequence 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. in press.

[HR07b] P. Honeine and C. Richard. Signal-dependent time-frequency representations for classification using a radially gaussian kernel and the alignment criterion. In *Proc. IEEE Statistical Signal Processing (SSP)*, Madison, WI, USA, August 2007. in press.

[HRB07a] 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. in press.

[HRB07b] P. Honeine, C. Richard, and J. C. M. Bermudez. On-line nonlinear sparse approximation of functions. In *Proc. IEEE International Symposium on Information Theory (ISIT)*, Nice, France, June 2007.

[HRF05] 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*, Louvain-la-Neuve, Belgium, 2005.

[HRF07] P. Honeine, C. Richard, and P. Flandrin. Time-frequency learning machines. *IEEE Trans. Signal Processing*, 55 :3930–3936, July 2007.

[HRFP06] 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. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[HT97] C. Heitz and J. Timmer. Using optimized time-frequency representations for acoustic quality control of motors. *Acustica*, 83 :1053–1064, 1997.

[HTF01] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer, New York, August 2001.

[Hub81] P.J. Huber. *Robust Statistics*. Wiley, New York, NY, USA, 1981.

[Iva76] V.V. Ivanov. The theory of approximate methods and their application to the numerical solution of singular integral equations (translated from the Russian by A. Ideh). In R.S. Anderssen and D. Elliott, editors, *Monographs and Textbooks on Mechanics of Solids and Fluids, Mechanics : Analysis*, 2, Leyden, Netherlands, 1976. Noordhoff International Publishing.

[JB95] D.L. Jones and R.G. Baraniuk. An adaptive optimal-kernel time-frequency representation. *IEEE Trans. Signal Processing*, 43(10) :2361–2371, 1995.

[Joa99] T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, chapter 11, pages 169–184. MIT Press, Cambridge, MA, 1999.

[Jol86] I.T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, NY, USA, 1986.

[Kai71] T. Kailath. RKHS approach to detection and estimation problems–i : Deterministic signals in gaussian noise. *IEEE Transactions on Information Theory*, 17(5) :530–549, September 1971.

[KFS03] K.I. Kim, M.O. Franz, and B. Schölkopf. Kernel Hebbian algorithm for iterative kernel principal component analysis. Technical Report 109, Max-Planck-Institut für biologische Kybernetik, Tübingen, Germany, 2003.

[KFS05] K.I. Kim, M.O. Franz, and B. Schölkopf. Iterative kernel principal component analysis for image modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9) :1351–1366, 2005.

[KMB06] S.-J. Kim, A. Magnani, and S. Boyd. Optimal kernel selection in kernel Fisher disciminant analysis. In *Proc. of the 23rd International Conference on Machine Learning (ICML'06)*, pages 465–472, New York, NY, USA, 2006. ACM Press.

[Kol57] A. N. Kolmogorov. On the representation of continuous functions of many variables by superpositions of continuous functions of one variable and addition. *Doklady Akademii Nauk USSR*, 114 :953–956, 1957.

[Kre99] U.H.-G. Kreßel. Pairwise classification and support vector machines. *Advances in kernel methods : support vector learning*, pages 255–268, 1999.

[KS05] V. Kurková and M. Sanguineti. Learning with generalization capability by kernal methods of bounded complexity. *J. Complex.*, 21(3) :350–367, 2005.

[KSTC02a] J. Kandola, J. Shawe-Taylor, and N. Cristianini. On the extensions of kernel alignment. Technical Report 120, Department of Computer Science, University of London, 2002.

[KSTC02b] J. Kandola, J. Shawe-Taylor, and N. Cristianini. Optimizing kernel alignment over combinations of kernels. Technical Report 121, Department of Computer Science, University of London, 2002.

[KSW04] J. Kivinen, A. J. Smola, and R. C. Williamson. Online learning with kernels. *IEEE Trans. Signal Processing*, 52(8), Aug 2004.

[KT03] James T. Kwok and Ivor W. Tsang. The pre-image problem in kernel methods. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003)*, pages 408–415, Washington, DC, USA, August 2003. AAAI Press.

[Kur04] V. Kurková. Learning from data as an inverse problem. In J. Antoch, editor, *Proc. Computational Statistics (CompStat'04)*, pages 1377–1384, Heidelberg,Germany, 2004. Physica-Verlag/SpringerAcademic Press, Inc.

[KW71] G. Kimeldorf and G. Wahba. Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33 :82–95, 1971.

[LCB+02] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semi-definite programming. In *Proc. 19th International Conference on Machine Learning*, pages 323–330, 2002.

[LCB+04] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M.I. Jordan. Learning the kernel matrix with semi-definite programming. *Journal of Machine Learning Research*, 5 :27–72, 2004.

[Lem95] O. Lemoine. *Détection de Signaux Non Stationnaires par Représentation Temps-Fréquence*. Ph.d. thesis, Université de Nice – Sophia Antipolis, 1995.

[Lue89] D.G. Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, second edition, 1989.

[Mac03] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. Available from `http ://www.inference.phy.cam.ac.uk/mackay/itila/`.

[Mal98] S. Mallat. *A wavelet tour of signal processing*. Academic Press, 1998.

[Mar97] N.M. Marinovich. The singular value decomposition of the wigner distribution and its applications. In W. Mecklenbräuker and F. Hlawatsch, editors, *The Wigner distribution - theory and applications in signal processing*, Amsterdam, The Netherlands, 1997. Elsevier.

[MBF94] O. Michel, R.G. Baraniuk, and P. Flandrin. Time-frequency based distance and divergence measures. In *Proc. of the IEEE International Symposium on Time-Frequency and Time-Scale analysis*, pages 64–67, Philadelphia, PA, USA, October 1994. IEEE Signal Processing Society.

[MCA06] Z. Hamou Mamar, P. Chainais, and A. Aussem. Probabilistic classifiers and time-scale representations : application to the monitoring of a tramway guiding system. In *Proc. of the 14th European Symposium on Artificial Neural Networks (ESANN 2006)*, pages 659–664, Bruges, Belgium, April 2006.

[MCD00] I. Magrin-Chagnolleau and G. Durou. Application of time-frequency principal component analysis to speaker verification. *Digital Signal Processing*, 10(1–3) :226–236, January/April/July 2000.

[McL97] J. McLaughlin. *Applications of the operator theory to time-frequency analysis and classification*. Ph.d. thesis, University of Washington, Seattle, USA, 1997.

[MDA97] J. McLaughlin, J. Droppo, and L. Atlas. Class-dependent time-frequency distributions via operator theory. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 3, pages 2045–2048, 1997.

[ME85] N. Marinovic and G. Eichmann. An expansion of wigner distribution and its applications. *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 10 :1021–1024, March 1985.

[Mik98] S. Mika. Kernalgorithmen zur nichtlinearen signalverarbeitung in merkmalsräumen (kernel algorithms for nonlinear signal processing in feature spaces). Master's thesis, Technische Universität Berlin, Berlin, Germany, November 1998.

[Mik02] S. Mika. *Kernel Fisher Discriminants*. Phd thesis, University of Technology, Berlin, October 2002.

[MLH03] D. Meyer, F. Leisch, and K. Hornik. The support vector machine under test. *Neurocomputing*, 55 :169–186, September 2003.

[MM01] O.L. Mangasarian and D.R. Musicant. Lagrangian support vector machines. *Journal of Machine Learning Research*, 1 :161–177, 2001.

[MNY06] H.Q. Minh, P. Niyogi, and Y. Yao. Mercer's theorem, feature maps, and smoothing. In *Proc. 19th Annual Conference on Learning Theory (COLT 2006)*. Springer, 2006.

[Mor02] M. Morvidone. *Etude et comparaison d'algorithmes de détection optimale pour les signaux modulés en amplitude et en fréquence ; application aux ondes gravitationnelles*. Ph.d. thesis, Ecole Doctorale de Mathématiques et Informatique de Marseille, Univertsité de Provence, Marseille, France, Décembre 2002.

[Moy49] J.E. Moyal. Quantum mechanics as a statistical theory. In *Proc. Cambridge Philosophical Society*, volume 45, pages 99–124, 1949.

[MR97] P. Maaß and A. Rieder. Wavelet-accelerated tikhonov–phillips regularization with applications. In A.K. Louis and W. Rundell, editors, *Inverse Problems in Medical Imaging and Nondestructive Testing*, pages 134–158, Springer, Vienna, 1997.

[MRM00] S. Mika, G. Rätsch, and K.-R. Müller. A mathematical programming approach to the kernel fisher algorithm. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS-2000)*, pages 591–597. MIT Press, 2000.

[MRW+99] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, and K.R. Müller. Fisher discriminant analysis with kernels. In Y. H. Hu, J. Larsen, E. Wilson, and S. Douglas, editors, *Advances in neural networks for signal processing*, pages 41–48, San Mateo, CA, USA, 1999. Morgan Kaufmann.

[MS00] V. J. Mathews and G. L. Sicuranze. *Polynomial signal processing*. John Wiley & Sons, New York, NY, 2000.

[MSS+99] S. Mika, B. Schölkopf, A. Smola, K.R. Müller, M. Scholz, and G. Rätsch. Kernel pca and de-noising in feature spaces. In *Proceedings of the 1998 conference on advances in neural information processing systems II*, pages 536–542, Cambridge, MA, USA, 1999. MIT Press.

[MSS01] S. Mika, A. Smola, and B. Schölkopf. An improved training algorithm for kernel Fisher discriminants. In T. Jaakkola and T. Richardson, editors, *Artificial Intelligence and Statistics*, pages 98–104, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers. Also : Microsoft Research TR-2000-77.

[Muk04] S. Mukherjee. Statistical learning : Algorithms and theory. Course notes for STA270 : statistical methods for computational biology, Institute of Statistics and Decision Sciences (ISDS), Duke University, Durham, NC, USA, November 2004.

[MZ93] S. Mallat and Z. Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Trans. Signal Processing*, 41(12) :3397–3415, December 1993.

[NBD93] H. Nawab, D. Beyerbach, and E. Dorken. Principal decomposition of time-frequency distributions. *IEEE Trans. Signal Processing*, 41 :3182–3186, November 1993.

[OFG97] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, *Neural Networks for Signal Processing VII — Proceedings of the 1997 IEEE Workshop*, pages 276–285, New York, NY, USA, 1997. IEEE.

[OG99] E.E. Osuna and F. Girosi. Reducing the run-time complexity in support vector machines. *Advances in kernel methods : support vector learning*, pages 271–283, 1999.

[Ong05] C.S. Ong. *Kernels : Regularization and Optimization*. Ph.d. thesis, The Australian National University, 2005.

[OW99] J. O'Neill and W. Williams. Shift covariant time-frequency distributions of discrete signals. *IEEE Trans. Signal Processing*, 47(1) :133–146, 1999.

[Par70] E. Parzen. Statistical inference on time series by RKHS methods. In R. Pyke, editor, *Proc. 12th Biennial Seminar*, pages 1–37, Montreal, Canada, 1970. Canadian Mathematical Congress.

[PF86] B. Porat and B. Friedlander. Adaptive detection of transient signals. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(6) :1410–1418, 1986.

[PG90] T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. IEEE*, 78 :1481–1497, 1990.

[Phi62] D.L. Phillips. A technique for the numerical solution of certain integral equations of the first kind. *J. ACM*, 9(1) :84–97, 1962.

[Pla99] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 185–208, Cambridge, MA, USA, 1999. MIT Press.

[PP86] F. Peyrin and R. Prost. A unified definition for the discrete-time, discrete-frequency and discrete-time/frequency wigner distributions. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(4) :858–867, 1986.

[PR05] J.-B. Pothin and C. Richard. Kernel machines : une nouvelle méthode pour l'optimisation de l'alignement des noyaux et l'amélioration des performances. In *Actes du XX^{ème} Colloque GRETSI sur le Traitement du Signal et des Images*, Louvain-la-Neuve, Belgium, 2005.

[PR06] J.-B. Pothin and C. Richard. A greedy algorithm for optimizing the kernel alignment and the performance of kernel machines. In *Proc. EUSIPCO*, Florence, Italy, 2006.

[RA03] V.C. Raykar and A. Ankur. Fast kernel principal component analysis for the polynomial and gaussian kernels. Technical Report Project report CMSC878R, Perceptual Interfaces and Reality Laboratory, Institute for Advanced Computer Studies, University of Maryland, MD, USA, 2003.

[Rak04] A. Rakotomamonjy. Optimizing area under roc curve with svms. In *Proc. of European Conference on Artificial Intelligence Workshop on ROC Curve and AI*, Valencia, Spain, 2004.

[RB93] B. Ristic and B. Boashash. Kernel design for time-frequency signal analysis using the radon transform. *IEEE Trans. Signal Processing*, 41 :1996–2008, May 1993.

[RBH07] C. Richard, J. C. M. Bermudez, and P. Honeine. Nonlinear kernel-based adaptive filtering with order controlled by a coherence criterion. *submitted to IEEE Transactions on Signal Processing*, 2007.

[RC05] A. Rakotomamonjy and S. Canu. Frames, reproducing kernels, regularization and learning. *J. Mach. Learn. Res.*, 6 :1485–1515, 2005.

[RDR$^+$07] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze. Using one-class svms and wavelets for audio surveillance systems. *IEEE Trans. on Information Forensic and Security*, 2007. submitted.

[Ric01] C. Richard. Linear redundancy of information carried by the discrete wigner distribution. *IEEE Trans. Signal Processing*, 49(11) :2536–2544, November 2001.

[Ric04] C. Richard. Détection par représentations temps-fréquence discrètes. In C. Doncarli and N. Martin, editors, *Décision dans le plan temps-fréquence*, pages 127–146, Paris, 2004. Hermès Sciences, Traité IC2.

[Rif02] R.M. Rifkin. *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, September 2002.

[RMC05] A. Rakotomamonjy, X. Mary, and S. Canu. Non-parametric regression with wavelet kernels. *Applied Stochastic Models in Business and Industry*, 21(2) :153–163, 2005.

[RPS98] M.S. Richman, T.W. Parks, and R.G. Shenoy. Discrete-time, discrete-frequency, time-frequency analysis. *IEEE Trans. Signal Processing*, 46(6) :1517–1527, 1998.

[RT94] R.G.Shenoy and T.W.Parks. The Weyl correspondence and time-frequency analysis. *IEEE Trans. Signal Processing*, 42(2) :318–331, February 1994.

[RT02] R. Rosipal and L.J. Trejo. Kernel partial least squares regression in reproducing kernel hilbert space. *Journal of Machine Learning Research*, 2 :97–123, 2002.

[RYP03] R. Rifkin, G. Yeo, and T. Poggio. Regularized least squares classification. In J.A.K. Suykens, G. Horvath, S. Basu, C. Micchelli, and J. Vandewalle, 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, May 2003. VIOS Press.

[SABO99] S. Saitoh, D. Alpay, J.A. Ball, and T. Ohsawa, editors. *Reproducing Kernels and Their Applications*, volume 3, Dordrecht, The Netherlands, 1999. International Society for Analysis, Applications and Computation (ISAAC'97), Kluwer Academic Publishers.

[Say03] A.H. Sayed. *Fundamentals of adaptive filtering*. Wiley-IEEE Press, NY, USA, June 2003.

[SB88] J. Stapleton and S. Bass. Synthesis of musical tones based on the karhunen-loeve transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36 :305–319, March 1988.

[SB00] A.J. Smola and P.L. Bartlett. Sparse greedy gaussian process regression. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *NIPS*, pages 619–625. MIT Press, 2000.

[SBSS99] A.J. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans. *Advances in Large Margin Classifiers*. MIT Press, Cambridge, MA, USA, 1999.

[SGB+02] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific Pub. Co., Singapore, 2002.

[SGV98a] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proc. of the Fifteenth International Conference on Machine Learning (ICML'98)*, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[SGV98b] C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proc. of the Fifteenth International Conference on Machine LearningICML '98*, pages 515–521, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

[SHW00] B. Schölkopf, R. Herbrich, and R. Williamson. A generalized representer theorem. Technical Report NC2-TR-2000-81, NeuroCOLT, Royal Holloway College, University of London, UK, 2000.

[SJ95] A.M. Sayeed and D.L. Jones. Optimal detection using bilinear time-frequancy and time-scale representations. *IEEE Trans. Signal Processing*, 43(12) :2872–2883, December 1995.

[SK87] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, 4(3) :519–524, March 1987.

[SN96] G. Strang and T.Q. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Cambridge, MA, USA, 1996.

[SS85] B.E.A. Saleh and N.S. Subotic. Time-variant filtering of signals in the mixed time-frequency domain. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(6) :1479–1485, December 1985.

[SS00] A.J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. In P. Langley, editor, *17th International Conference on Machine Learning, Stanford, 2000*, pages 911–918, San Francisco, CA, USA, 2000. Morgan Kaufman.

[SS01] B. Schölkopf and A.J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[SSM98] B. Schölkopf, A.J. Smola, and K.R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5) :1299–1319, 1998.

[SSM99] B. Schölkopf, A.J. Smola, and K.-R. Müller. Kernel principal component analysis. *Advances in kernel methods : support vector learning*, pages 327–352, 1999.

[STC04a] K. Saadi, N.L.C. Talbot, and G.C. Cawley. Optimally regularised kernel fisher discriminant analysis. In *Proc. of the 17th International Conference on Pattern Recognition,(ICPR'04)*, volume 2, pages 427–430, Washington, DC, USA, 2004. IEEE Computer Society.

[STC04b] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[Ste02] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2 :67–93, 2002.

[STWCK05] J. Shawe-Taylor, C.K.I. Williams, N. Cristianini, and J. Kandola. On the eigenspectrum of the gram matrix and the generalization error of kernel-pca. *IEEE Transactions on Information Theory*, 51(7) :2510–2522, 2005.

[Suc04] V. Sucic. *Parameters Selection for Optimising Time-Frequency Distributions and Measurements of Time-Frequency Characteristics of Nonstationary Signals*. Ph.d. thesis, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, Australia, March 2004.

[Sun98] J. Sun. Projection pursuit. In S. Kotz, C. Read, D. Banks, and N. Johnson, editors, *Encyclopedia of Statistical Sciences*, volume 2, pages 554–560. Wiley, 1998.

[SV99] J.A.K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9(3) :293–300, 1999.

[SWMS99] A. Smola, R.C. Williamson, S. Mika, and B. Schölkopf. Regularized principal manifolds. In *Computational Learning Theory : 4th European Conference*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 214–229. Springer, 1999.

[SWS+00] B. Schölkopf, R.C. Williamson, A.J. Smola, J. Shawe-Taylor, and J.C. Platt. Support vector method for novelty detection. In S.A. Solla, T.K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 582–588, Cambridge, MA, USA, 2000. MIT Press.

[SY06] S.Smale and Y. Yao. Online learning algorithms. *Found. Comput. Math.*, 6(2) :145–170, 2006.

[TA77] A.N. Tikhonov and V.Y. Arsenin. *Solutions of ill-posed problems*. John Wiley, New York, 1977.

[Tax01] D.M.J. Tax. *One-class classification ; Concept-learning in the absence of counter-examples*. Phd thesis, Advanced School for Computing and Imaging – Delft University of Technology, June 2001.

[TGMS03] J. A. Tropp, A. C. Gilbert, S. Muthukrishnan, and M. Strauss. Improved sparse approximation over quasi-incoherent dictionaries. In *ICIP (1)*, pages 37–40, 2003.

[Tik63] A.N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Sov. Math. Dokl.*, 4 :1035–1038, 1963.

[TK02] T. Takahashi and T. Kurita. Robust de-noising by kernel pca. In *International Conference on Artificial Neural Networks(ICANN2002) In Artificial Neural Networks*, pages 739–744. Springer, 2002.

[TLSS06] I. Takeuchi, Q.V. Le, T.D. Sears, and A.J. Smola. Nonparametric quantile regression. *Journal of Machine Learning Research*, 7 :1231–1264, 2006.

[TP91] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1) :71–86, 1991.

[Tro04] J. A. Tropp. Greed is good : algorithmic results for sparse approximation. *IEEE Trans. Information Theory*, 50(10) :2231–2242, 2004.

[Tut89] F.B. Tuteur. Wavelet transformations in signal detection. In J.M. Combes, A. Grossmann, and Ph. Tchamitchian, editors, *Wavelets : Time-Frequency Methods and Phase Space*, pages 132–138, Berlin, Germany, 1989. Springer-Verlag.

[Vap82] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer-Verlag New York, Inc., Secaucus, NJ, USA, springer edition, 1982.

[Vap95] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[Vap98] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, NY, USA, September 1998.

[Vay00] N. Vayatis. *Inégalités de Vapnik-Chervonenkis et mesures de complexité*. Ph.d. thesis, Ecole Polytechnique, Palaiseau, France, jan 2000. in english.

[VC71] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2) :264–280, 1971.

[Ver93] D. Verkindt. *Etude d'algorithmes rapides de recherche d'un signal d'onde gravitationnelle provenant de coalescences d'étoiles binaires*. Phd thesis, Université de Savoie, 1993.

[Vil48] J. Ville. Théorie et applications de la notion de signal analytique. *Câbles et Transmission*, 2éme. A.(1) :61–74, 1948.

[VSS06] S. V.N. Vishwanathan, Nicol N. Schraudolph, and Alex J. Smola. Step size adaptation in reproducing kernel hilbert space. *Journal of Machine Learning Research*, 7 :1107–1133, 2006.

[VTS04] J.P. Vert, K. Tsuda, and B. Schölkopf. A primer on kernel methods. In B. Schölkopf, K. Tsuda, and J.P. Vert, editors, *Kernel Methods in Computational Biology*, pages 35–70, Cambridge, MA, USA, 2004. MIT Press.

[Wah90] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Math (SIAM), Philadelphia, 1990.

[Wat85] S. Watanabe. *Pattern recognition : human and mechanical*. John Wiley & Sons, Inc., New York, NY, USA, 1985.

[WB05] L. Wolf and S. Bileschi. Combining variable selection with dimensionality reduction. In *CVPR '05 : Proc. of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 801–806, Washington, DC, USA, 2005. IEEE Computer Society.

[WC02] L. Wang and K.L. Chan. Learning kernel parameters by using class separability measure. In *Sixth kernel machines workshop, In conjunction with Neural Information Processing Systems (NIPS)*, Whistler, Canada, 2002.

[WCP05] G. Wu, E. Y. Chang, and N. Panda. Formulating distance functions via the kernel trick. In *Proc. 11th ACM International conference on knowledge discovery in Data mining*, pages 703–709, 2005.

[Wig32] E.P. Wigner. On the quantum correction for thermodynamic equilibrium. *Physics Review*, 40 :749–759, 1932.

[WLZ00] G. Wahba, Y. Lin, and H. Zhang. Generalized approximate cross validation for support vector machines. In A. Smola, P. Bartlett, B. Schölkopf, and C. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 297–309, Cambridge, MA, USA, 2000. MIT Press.

[WM97] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1) :67–82, April 1997.

[Wol01] D.H. Wolpert. The supervised learning no-free-lunch theorems. In *Proc. 6th Online World Conf. on Soft Computing in Industrial Applications*, 2001.

[WS01] C.K.I. Williams and M. Seeger. Using the Nyström method to speed up kernel machines. In T. Leen, T. Dietterich, and V. Tresp, editors, *Neural Information Processing Systems 13*, pages 682–688. MIT Press, 2001.

[XZL01] J. Xu, X. Zhang, and Y. Li. Kernel MSE algorithm : A unified framework for KFD, LS-SVM and KRR. In *Proc. International Joint Conference on Neural Networks (IJCNN'01)*, pages 1486–1491, Washington, DC, USA, 2001.

[ZBB04] L. Zwald, O. Bousquet, and G. Blanchard. Statistical properties of kernel principal component analysis. In *Proc. 17th. Conference on Learning Theory (COLT)*, pages 594–608, 2004.

[ZH02] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS)*, Cambridge, MA, USA, 2002. MIT Press.

[Zha04] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32 :56–134, March 2004.

[ZR05] P. Zhang and N. Riedel. Discriminant analysis : A unified approach. In *Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM '05)*, pages 514–521, Washington, DC, USA, 2005. IEEE Computer Society.