



Méthodes à noyaux en reconnaissance de formes, prédition et classification. Applications aux biosignaux

Maya Kallas

► To cite this version:

Maya Kallas. Méthodes à noyaux en reconnaissance de formes, prédition et classification. Applications aux biosignaux. Sciences de l'ingénieur [physics]. Université de Technologie de Troyes, 2012. Français. tel-01088936

HAL Id: tel-01088936

<https://hal.archives-ouvertes.fr/tel-01088936>

Submitted on 29 Nov 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

Présentée par

Mlle Maya KALLAS

soutenue le

23 novembre 2012

en vue de l'obtention du

DOCTORAT de

I'UNIVERSITÉ DE TECHNOLOGIE DE TROYES

Spécialité : OPTIMISATION ET SÛRETÉ DES SYSTÈMES

Titre :

**Méthodes à noyaux en reconnaissance de formes, prédition
et classification. Applications aux biosignaux**

Membres du Jury

Mme Florence D'ALCHE-BUC	Professeur des Universités	Université d'Evry-Val d'Essone	Rapporteur
M. Stéphane CANU	Professeur des Universités	INSA Rouen	Rapporteur
M. Clovis FRANCIS	Professeur	Université Libanaise	Directeur de thèse
M. Paul HONEINE	Maitre de conférences	Université de Technologie de Troyes	Directeur de thèse
M. Régis LENGELLÉ	Professeur des Universités	Université de Technologie de Troyes	Examinateur
M. Nabil NASSIF	Professeur	American University of Beirut	Examinateur
M. Cédric RICHARD	Professeur des Universités	Université de Nice-Sophia Antipolis	Examinateur

Université de Technologie de Troyes

T H È S E

Présentée par

Mlle Maya KALLAS

soutenue le

23 novembre 2012

en vue de l'obtention du

DOCTORAT de

I'UNIVERSITÉ DE TECHNOLOGIE DE TROYES

Spécialité : OPTIMISATION ET SÛRETÉ DES SYSTÈMES

Titre :

Méthodes à noyaux en reconnaissance de formes, prédition et classification. Applications aux biosignaux

Membres du Jury

Mme Florence D'ALCHE-BUC	Professeur des Universités	Université d'Evry-Val d'Essone	Rapporteur
M. Stéphane CANU	Professeur des Universités	INSA Rouen	Rapporteur
M. Clovis FRANCIS	Professeur	Université Libanaise	Directeur de thèse
M. Paul HONEINE	Maitre de conférences	Université de Technologie de Troyes	Directeur de thèse
M. Régis LENGELLÉ	Professeur des Universités	Université de Technologie de Troyes	Examinateur
M. Nabil NASSIF	Professeur	American University of Beirut	Examinateur
M. Cédric RICHARD	Professeur des Universités	Université de Nice-Sophia Antipolis	Examinateur

Remerciements

“La recherche, c'est l'acte par lequel une société avancée exprime sa foi en un avenir ouvert.”

Claude Détraz

L'écriture d'une telle page n'est pas une tâche aisée. Déjà trois ans depuis le tout début. Pendant ces trois ans, j'ai rencontré des personnes qui ont contribué à ce projet et à qui j'adresse mes remerciements les plus sincères. Certes des noms viennent immédiatement à l'esprit, mais combien d'anonymes ou d'oubliés y ont, indirectement, contribué. Que ceux et celles que je n'ai pas nommés reçoivent aussi l'expression de toutes mes pensées.

Tout d'abord, je tiens à exprimer mes plus vifs remerciements et ma gratitude à mes directeurs de thèse, Monsieur Clovis FRANCIS, Professeur à l'Université Libanaise, et Monsieur Paul HONEINE, Maître de conférences à l'Université de Technologie de Troyes, pour leurs encadrements continus, pour les remarques constructives qu'ils m'ont fournies ainsi que pour leurs précieux conseils durant toute la période de mon travail. Je les remercie également pour la confiance qu'ils m'ont accordée. En dehors de leurs apports scientifiques, je n'oublierai pas aussi de les remercier pour leurs qualités humaines, leur hospitalité et leur soutien qui m'ont permis de mener à bien cet ouvrage. Leur confiance, leur disponibilité, et leur sens critique ont été pour moi toujours très précieux.

Je remercie Madame Florence D'ALCHE-BUC, Professeur à l'Université d'Evry-Val d'Essone et Monsieur Stéphane CANU, Professeur à l'INSA de Rouen, qui ont accepté la responsabilité de juger ce travail en qualité de rapporteurs. Leurs critiques et leurs suggestions m'ont permis d'améliorer mon travail. Ma gratitude, mon profond respect et mes remerciements vont à Monsieur Régis LENGELLÉ, Professeur à l'UTT, et à Monsieur Nabil NASSIF, Professeur à American University of Beirut, qui ont accepté d'être membres de ce jury de thèse. Je remercie également Monsieur Cédric RICHARD, Professeur à l'Université de Nice-Sophia Antipolis pour sa présence et son encouragement durant mes travaux de thèse.

J'adresse aussi mes remerciements à Madame Zeinab SAAD, Professeur à l'Université Libanaise et Doyenne de l'école doctorale de l'UL, pour avoir mis en place la convention de thèse en cotutelle du réseau UT-INSA avec l'UL, permettant ainsi des séjours entre Le Liban et La France et une collaboration entre les deux pays et précisément les deux universités, et pour son équipe talentueuse qui a su gérer nos demandes. Je remercie aussi Monsieur Hassan AMOUD, responsable du Campus Numérique Francophone du Liban Nord, pour sa présence encourageante.

Mes remerciements vont également à Pascale DENIS, gestionnaire à l'école doctorale de l'UTT, pour sa disponibilité, sa compétence, et pour toute aide qu'elle a apportée. J'associe à mes remerciements Marie-José ROUSSELET et Véronique BANSE, secrétaires du Pôle ROSAS à l'UTT, pour leur gentillesse, leur efficacité, leur disponibilité, et le soutien qu'elles m'ont apporté.

À ces remerciements, j'associe tous mes collègues de travail Elias KHOURY, Nathalie MATTA, Roy AAD, Imane MAATOUK, Chafic SAIDE, Zineb NOUMIR. Je ne peux également manquer l'occasion d'adresser une pensée particulière à mes amis : Widad SLEIMAN et Anthony SAWAYA. Je remercie également mes parents Christiane et Youssef, pour leur soutien et leurs encouragements.

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
mon père et ma mère*

“La recherche procède par des moments distincts et durables, intuition, aveuglement, exaltation et fièvre. Elle aboutit un jour à cette joie, et connaît cette joie celui qui a vécu des moments singuliers.”

Albert Einstein, “*Comment je vois le monde*”

Table des matières

Résumé	1
Abstract	3
Introduction	5
1 Les méthodes à noyaux et le problème de la pré-image	11
1.1 Introduction	12
1.2 Noyaux et espace de Hilbert à noyau reproduisant	13
1.2.1 Noyau défini positif et RKHS	13
1.2.2 Théorème de Moore-Aronszajn	14
1.3 Du modèle linéaire au modèle à noyaux	17
1.3.1 Astuce du noyau	17
1.3.2 Théorème de Représentation	18
1.4 Exemple du linéaire au non-linéaire	19
1.4.1 Analyse en composantes principales	19
1.4.2 Analyse en composantes principales à noyaux	20
1.5 L'ACP-à-noyaux pour la reconnaissance des formes	22
1.5.1 Extraction des caractéristiques	22
1.5.2 Débruitage	22
1.5.3 Une vue unifiée	23
1.6 Définition du problème de pré-image	23
1.7 Méthodes de résolution	25
1.7.1 Méthode de la descente du gradient	25
1.7.2 Méthode itérative du point fixe	26
1.7.3 Apprentissage de la carte de pré-image	28
1.7.4 Méthode de l'échelle multidimensionnelle	28
1.7.5 Approche conforme	30
1.7.6 Pré-image régularisée ou pénalisée	31
1.8 Formulation de la pré-image	31
1.9 Conclusion	33
2 Le problème de pré-image avec contraintes de non-négativité	35
2.1 Introduction	35
2.2 Méthodes à noyaux, pré-image et non-négativité	36
2.3 Pré-image avec contraintes de non-négativité	37

2.3.1	Contraintes de non-négativité sur la pré-image	38
2.3.2	Contraintes de non-négativité sur les coefficients du modèle	41
2.4	Expérimentations	43
2.4.1	Extraction des caractéristiques de signaux ERP	43
2.4.2	Débruitage de données et des images	46
2.5	Conclusion	52
3	Modèles autorégressifs-à-noyaux : technique de prédiction	55
3.1	Introduction	55
3.2	Séries temporelles	57
3.2.1	Processus stochastique	58
3.2.2	Propriétés des séries temporelles	58
3.3	Prédiction des séries temporelles avec un modèle autorégressif	59
3.3.1	Méthode des moindres carrés	60
3.3.2	Équations de Yule-Walker	60
3.4	Modèle autorégressif à noyaux pour la prédiction des séries temporelles	61
3.4.1	Modèle autorégressif dans l'espace fonctionnel	62
3.4.2	Le problème de la pré-image comme technique de prédiction	65
3.5	Modèles autorégressifs-à-noyaux sans pré-image	67
3.5.1	Modèle autorégressif sur les valeurs du noyau	68
3.5.2	Un modèle autorégressif hybride	69
3.5.3	Lien entre le modèle AR hybride et les modèles proposés auparavant	70
3.6	Expérimentations	70
3.6.1	Comparaison les techniques prédictives non-linéaires	72
3.6.2	Comparaison entre les différentes techniques de pré-image	74
3.6.3	Comparaison entre les différentes techniques proposées	74
3.7	Conclusion	77
4	Étude de cas : les classifications binaire et multi-classes avec les méthodes à noyaux	79
4.1	Introduction	79
4.2	Classification binaire par SVM	80
4.3	Classification multi-classes	82
4.3.1	Un contre-tous	82
4.3.2	Un contre-un	83
4.4	Carte d'auto-organisation	84
4.5	Expérimentations	87
4.5.1	Critères d'évaluation de la classification	87
4.5.2	Classification binaire	88
4.5.3	Classification multi-classes	91

4.5.4 Carte d'auto-organisation	93
4.6 Conclusion	96
Conclusion générale et perspectives	99
A Annexe	103
A.1 Noyaux projectifs	103
A.2 Noyaux radiaux	104
Bibliographie	106

Glossaire des notations et abréviations

Notation	Signification
\mathbb{N}	ensemble des nombres entiers
\mathbb{R}	ensemble des nombres réels
\mathbb{E}	espérance mathématique d'une série statistique
Cov	covariance
dim	dimension d'un espace
$[\cdot]_j$	$j^{\text{ième}}$ composante de $[\cdot]$
diag $[\cdot]$	opérateur indiquant la matrice diagonale
κ	noyau reproduisant
\mathcal{X}	espace des observations
\mathcal{H}	espace de Hilbert à noyau reproduisant, associé à la fonction Φ : $x \in \mathcal{X} \mapsto \Phi(x) \in \mathcal{H}$
Φ	fonction de mapping associée à l'espace de représentation \mathcal{H}
$\langle \cdot, \cdot \rangle_{\mathcal{H}}$	produit scalaire dans l'espace \mathcal{H}
$\ \cdot\ _{\mathcal{H}}$	norme associée au produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$
$J(\cdot)$	fonction coût
ACP	Analyse en Composantes Principales (<i>Principal Component Analysis</i>)
RKHS	espace de Hilbert à noyau reproduisant (<i>Reproducing Kernel Hilbert Space</i>)
MDS	Échelle multidimensionnelle (<i>MultiDimensional Scaling</i>)
ERP	Potentiels évoqués (<i>Event Related Potentials</i>)
MAE	Erreur absolue moyenne (<i>Mean Absolute Error</i>)
PSNR	Rapport signal sur bruit (<i>Peak Signal-to-Noise Ratio</i>)
AR	AutoRégessif
ECG	ElectroCardiogGramme
MIT-BIH	Massachusetts Institute of Technology-Beth Israel Hospital
MSE	Erreur quadratique moyenne (<i>Mean Square Error</i>)
SVM	machine à vecteurs supports (<i>Support Vector Machine</i>)
SOM	carte d'auto-organisation (<i>Self-Organizing Map</i>)
BMU	unité correspondant le mieux (<i>Best Matching Unit BMU</i>)
PTB	Physikalisch-Technische Bundesanstalt

Résumé

Durant les deux dernières décennies, les méthodes dites à noyaux ont favorisé l'essor de l'apprentissage statistique pour le traitement de systèmes non-linéaires, souvent en pénurie d'information *a priori*. L'idée principale de ces méthodes non-linéaires réside dans *l'astuce du noyau*, qui permet de transformer l'espace des observations en un espace plus pertinent, souvent de plus grande dimension. Ainsi, en appliquant dans ce dernier les techniques classiques de traitement linéaire, ces méthodes sont-elles aménagées pour une vaste classe de systèmes non-linéaires. Dans le foisonnement de ces méthodes, dont les Machines à Vecteurs Supports sont le fer de lance, peu de travaux ont été menés sur le problème inverse, c'est-à-dire le retour à l'espace des observations. Paradoxalement, bien que la transformation non-linéaire induite par le noyau est fondamentale, le retour inverse à l'espace des observations est souvent crucial. La résolution de ce problème, dit de la pré-image, permet de nouveaux domaines d'application pour les méthodes à noyaux, dont la reconnaissance des formes, l'extraction de caractéristiques, le débruitage de signaux, ou encore l'analyse de séries temporelles.

L'objectif de cette thèse est de montrer que les récentes avancées en théorie de l'apprentissage statistique apportent des solutions pertinentes à plusieurs problèmes soulevés en traitement du signal et des images, et plus précisément dans le cas de signaux issus de capteurs physiologiques où la nonlinéarité est extrêmement courante. Diverses méthodes à noyaux sont élaborées pour proposer des solutions en reconnaissance des formes, analyse de séries temporelles, et classification d'anomalies.

La première partie de cette thèse porte sur la résolution du problème de la *pré-image* avec contraintes. Nous étudions des contraintes imposées par la physiologie, et en particulier la non-négativité. Nous nous intéressons dans un premier temps à la non-négativité du résultat. Dans un second temps, nous tenons compte de l'additivité des contributions, induisant une certaine parcimonie dans le résultat. En couplant ces développements avec l'analyse en composantes principales à noyaux, nous proposons une extraction de caractéristiques avec contraintes pour des signaux électroencéphalographiques et le débruitage des images.

La deuxième partie porte sur l'analyse de séries temporelles, selon une approche prédictive. Nous élaborons alors des *modèles autorégressifs* dans l'espace transformé, la prédiction nécessitant la résolution du problème de pré-image. Deux modèles prédictifs à noyaux sont étudiés en détail : une méthode basée sur le problème de moindres carrés, et une sur la résolution des équations de Yule-Walker. Une investigation de l'impact de la résolution du problème de pré-image est effectuée. Une étude empirique montre la pertinence de ces modèles pour l'analyse d'électrocardiogrammes.

La dernière partie de cette thèse traite le problème de *classification* de signaux électrocardiogrammes, afin de détecter des anomalies présentes dans les enregistrements. Nous étudions les performances des machines à vecteurs supports, avec et sans extraction de caractéristiques, pour réaliser un détecteur d'anomalies. Une étude multi-classes est menée pour l'analyse de différents types d'anomalies, en utilisant d'une part les cartes auto-organisatrices et d'autre part les

machines à vecteurs supports.

Abstract

Over the past two decades, kernel-based methods have favored the development of the statistical learning for the analysis of nonlinear systems, often with a lack of *prior* information. The main idea behind these nonlinear methods is the *kernel trick*, which allows the transformations from the input space into a high-dimensional feature space. Thus, by applying in this space standard linear techniques, these methods provide nonlinear processing in the input space. During the proliferation of these methods, where Support Vector Machines are the spearhead, little work has been done on the inverse problem of the kernel trick, that is the return to the input space. Paradoxically, although the nonlinear transformation induced by the kernel is fundamental, the return to the input space is often critical. This problem is called the pre-image problem. Its resolution allows a new class of kernel-based machines.

The purpose of this thesis is to show that recent advances in statistical learning theory provide relevant solutions to several issues in signal and image processing, and more specifically the case of signals from physiological sensors where nonlinearities are extremely common. Several kernel-based methods are elaborated to provide solutions in pattern recognition, time series analysis, and classification of anomalies.

The first part of this thesis covers the resolution of the *pre-image* problem with constraints. We study the constraints imposed by physiology, and in particular the non-negativity. We are interested in the first place with the non-negativity of the result. In a second step, we take into account the additivity of the contributions, inducing some sparsity in the result. By combining these developments with the Kernel Principal Component Analysis, we propose a constrained feature extraction for event related potential signals and for denoising images.

The second part focuses on the analysis of time series, according to a predictive approach. Thus, we develop *autoregressive models* in the feature space, where the prediction requires solving the pre-image problem. Two kernel-based models for prediction are studied in detail : the first one is based on the least-squares problem, and the second one is based on solving the Yule-Walker equations. An investigation of the impact of the resolution of the pre-image problem is made. An empirical study demonstrates the relevance of these models for the analysis of electrocardiograms.

The last part of this thesis deals with the problem of *classification* of electrocardiogram signals to detect anomalies. We study the performance of support vector machines, with and without feature extraction, to provide an anomaly detector. A study for multi-class is conducted to analyze various types of anomalies using, on the one hand, self-organizing maps, and on the other hand the support vector machines.

Introduction

Présentation générale

Durant ces deux dernières décennies, le domaine du traitement du signal a connu des progrès notables grâce à l'émergence d'outils nouveaux pour le traitement des problèmes non-linéaires. En apprentissage automatique, nous pouvons citer la reconnaissance des formes, l'extraction de caractéristiques, le débruitage de signaux ou des images, l'analyse de séries temporelles, sans oublier les problèmes décisionnels tels que la détection et la discrimination à deux ou plusieurs classes. Les données disponibles dans ces domaines d'applications, provenant de systèmes naturels, sont très complexes et ne peuvent hélas être expliquées par des modèles linéaires traditionnels. Par suite, les chercheurs ont éprouvé la nécessité de proposer des algorithmes non-linéaires permettant d'appréhender une classe étendue de problèmes. Les nouvelles approches visent à utiliser les méthodes à noyaux. Ces techniques exploitent la théorie des *noyaux reproduisants*. L'idée principale est l'*astuce du noyau*, permettant de transformer les données par le biais d'une application non-linéaire, dans un espace de dimension élevée, où des méthodes linéaires peuvent être appliquées. Ces méthodes à noyaux ont été appliquées avec succès pour de nombreuses applications et ont montré des performances remarquables. L'objectif de cette thèse est de montrer que les méthodes à noyaux apportent des solutions pertinentes à plusieurs problèmes soulevés en traitement du signal et des images, plus précisément dans le cas de signaux biomédicaux. Différentes méthodes non-linéaires sont élaborées : la résolution du problème de la pré-image sous contraintes de non-négativité pour la reconnaissance des formes, l'extraction des caractéristiques et le débruitage, le modèle autorégressif à noyaux pour la prédiction des séries temporelles, et finalement les machines à vecteurs supports pour la discrimination afin d'améliorer les performances en classification.

En reconnaissance des formes, dont l'extraction des caractéristiques et le débruitage des données, parfois nous cherchons à identifier les formes ou les données afin de les interpréter ou de les représenter dans l'espace des données où elles sont décrites. Ainsi, devons-nous faire le retour inverse de l'espace de projection, désigné par espace fonctionnel, à l'espace des observations. Or, ce retour n'est pas toujours aussi évident, puisque nous avons recours à utiliser des noyaux afin de faire la transformation vers cet espace. La fonction inverse permettant le retour à l'espace d'entrée n'existe pas, ce problème mal posé est le *problème de la pré-image*. Pour ce faire, nous cherchons à trouver une solution dans l'espace des observations ayant comme image la fonction calculée dans l'espace fonctionnel. Une nouvelle écriture de la pré-image est détaillée dans la thèse afin de présenter la solution en question. Tenant compte de l'aspect physique des observations disponibles, le problème de la pré-image est sujet aux contraintes imposées. Nous étudions en particulier les contraintes de non-négativité. Nous distinguons entre les contraintes sur les données elles-mêmes des contraintes sur des coefficients du modèle définissant la pré-image. Une propriété importante, obtenue pour le second cas, est la parcimonie des

résultats.

Confronté à des problèmes de vie, l'humanité a toujours été intéressée par l'avenir. Comme la civilisation avance, avec une sophistication croissante dans toutes les phases de la vie, la nécessité de regarder vers l'avenir a grandi avec elle. De nos jours, chaque agence gouvernementale, entreprise, ou industrie ainsi que le citoyen veulent et doivent être en mesure de prévoir et planifier des événements futurs. En effet, pour prendre une décision dans le présent, nous devons avoir des plans pour l'avenir. Aujourd'hui, les techniques liées à la prédiction des séries temporelles constituent un outil d'aide à la décision. La recherche en matière de méthodes de prédiction est très intense. Le développement de ces techniques a été attribué aux progrès réalisés en statistique et probabilités. Parmi les méthodes les plus connues, nous citons le modèle autorégressif. Il prédit un échantillon futur à partir d'un certain nombre d'échantillons de son passé. En utilisant les méthodes à noyaux, nous proposons alors d'étendre l'usage de ce modèle pour les signaux pris sur des systèmes non-linéaires.

Autre que la prédiction, l'humanité a intérêt de classer les données. Il est indispensable de classer et regrouper les données présentant les mêmes caractéristiques. Les méthodes de classification ont pour objectif de classer un objet, dans une catégorie donnée suivant certains traits descriptifs. Elles peuvent être appliquées pour différents problèmes, en particulier, le problème de la santé qui est un secteur concernant tout le monde, comme le cas des problèmes cardiaques qui sont la première cause de mort de nos jours. Dans ce manuscrit, nous proposons d'appliquer les méthodes à noyaux pour discriminer les personnes présentant des anomalies cardiaques des personnes saines. Pour ce faire, les machines à vecteurs supports, (SVM pour *Support Vector Machines*) sont utilisées pour la classification binaire qui se compose soit d'une personne saine soit d'une personne malade. L'étude comporte alors une décision entre deux hypothèses possibles. Or, les anomalies cardiaques sont présentes sous différentes formes. Il est nécessaire de les classer en plusieurs catégories. Même si les SVM sont conçues principalement pour une discrimination binaire, cependant elles peuvent être appliquées pour distinguer entre différentes classes. À cette fin, deux stratégies sont détaillées dans ce mémoire pour une discrimination multi-classes, en décomposant le problème multi-classes en un ensemble de problèmes de discrimination binaire. Une autre technique pour une telle classification est la carte d'auto-organisation (SOM pour *Self-Organizing Map*). Sa fonction principale est de faire correspondre les éléments de l'espace d'entrée avec des unités ordonnées sur une carte qui est une représentation graphique où chaque unité est entourée de ses voisins, les voisinages ayant été définis a priori. Une SOM est réalisée pour discriminer différentes anomalies cardiaques.

Dans ce mémoire, nous étudions différentes techniques pour la reconnaissance des formes et le traitement du signal en les combinant aux noyaux reproduisants. Les travaux sont divisés en quatre axes :

- Résolution du problème de pré-image
- Contraintes de non-négativité de la pré-image
- Analyse et prédiction de séries temporelles
- Discrimination binaire et multi-classes

Plan du document

Le manuscrit se compose de quatre chapitres. La progression suivante est adoptée pour exposer les travaux réalisés.

Le premier chapitre a pour objectif de définir le cadre du travail. Dans un premier temps, il introduit la théorie des noyaux reproduisants, tout en décrivant leur caractérisation. Il présente alors les principes fondamentaux de cette théorie, qui sont le théorème de représentation et l'astuce du noyau. Ces différents concepts sont illustrés pour l'analyse en composantes principales (ACP), tout en l'étendant pour les systèmes non-linéaires par le biais des méthodes à noyaux. Finalement, le problème de la pré-image confronté lors de l'usage des noyaux reproduisants est détaillé tout en décrivant les différentes techniques présentes pour sa résolution. Une formulation de la pré-image est élaborée dans ce chapitre.

Le second chapitre de cette thèse tient compte de la résolution du problème de la *pré-image* sous contrainte. Motivés par des contraintes physiologiques, nous étudions alors en particulier la non-négativité. Dans un premier temps, la contrainte étudiée est celle du résultat. En d'autres termes, la *pr-image* devra être non-négative. Ensuite, nous proposons la résolution du problème de la *pré-image* sous contraintes de non-négativité des coefficients du modèle. Nous parlons alors de l'additivité des contributions, induisant une certaine parcimonie au niveau du résultat. Finalement, ces deux approches sont utilisées avec l'analyse en composantes principales à noyaux en extraction de caractéristiques et en débruitage des images et de chiffres manuscrits.

Le troisième chapitre porte sur l'analyse de séries temporelles par un *modèle autorégressif (AR)*. Ainsi une présentation de ce modèle est-elle faite et une extension du cas linéaire à celui non-linéaire est réalisée au moyen de noyaux reproduisants. Deux techniques pour la définition d'un modèle AR-à-noyaux sont détaillées pour l'analyse de séries temporelles selon un modèle prédictif : la méthode des moindres carrés et les équations de Yule-Walker. Trois approches différentes du modèle AR-à-noyaux sont proposées. La première est basée sur l'idée principale des méthodes à noyaux pour lesquelles une transformation de l'espace des observations à l'espace fonctionnel est réalisée. Cette approche nécessite la résolution du problème de *pré-image* pour la prédiction d'un échantillon. Afin de surmonter le problème de la *pré-image*, deux autres techniques à noyaux sont alors élaborées : la première est basée sur les valeurs du noyau et la seconde représente un modèle hybride. Finalement, ces méthodes proposées sont appliquées sur des séries temporelles unidimensionnelles et chaotiques multidimensionnelles. Ces applications comportent de même une étude comparative entre les différentes techniques de *pré-image*.

Le quatrième chapitre de cette thèse tient compte de la *classification* de signaux électrocardiogrammes basée sur ces méthodes à noyaux, plus précisément, nous détectons des anomalies présentes dans ces signaux. Dans un premier temps, une comparaison entre les performances des machines à vecteurs supports appliquées sur les caractéristiques extraites à l'aide, d'une part de l'ACP classique, et de l'autre part de l'ACP-à-noyaux, est réalisée pour une discrimination binaire détectant les signaux présentant des anomalies des signaux de sujets sains. Ensuite, une classification multi-classes est faite pour détecter les différentes anomalies présentes en utilisant les

machines à vecteurs supports. Cette discrimination est basée sur deux stratégies : “un-contre-un” et “un-contre-tous”. Nous décrivons aussi la *carte d’auto-organisation* permettant la détection de différents types d’anomalies. Enfin, les résultats d’application de ces techniques pour la discrimination des signaux électrocardiogrammes sont présentés.

Ce document s’achève par une conclusion présentant des nouvelles perspectives de travail.

Produits de la recherche

Cette thèse a fait l'objet de plusieurs publications. Le Tableau suivant résume les références par ordre de leur apparition, ainsi que les chapitres auxquels ils ont trait.

Citation	Référence	Chapitres
[KHR ⁺ 10]	M. Kallas, P. Honeine, C. Richard, H. Amoud, and C. Francis. Nonlinear feature extraction using kernel principal component analysis with non-negative pre-image. <i>In Proc. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)</i> , Buenos Aires, Argentina, 31 Aug. - 4 Sept. 2010.	2
[KHR ⁺ 11a]	M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Kernel-based autoregressive modeling with a pre-image technique. <i>In 16th IEEE Workshop on Statistical Signal Processing</i> , Nice, France, 28-30 June 2011.	3
[KHR ⁺ 11c]	M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Non-negative pre-image in machine learning for pattern recognition. <i>In 19th European Signal Processing Conference</i> , Barcelona, Spain, 29 August - 2 September 2011.	2
[KHR ⁺ 11b]	M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Modèle autorégressif non-linéaire à noyau : une première approche. <i>In 23eme édition du Colloque GRETSI</i> , Bordeaux, France, 5-8 Septembre 2011.	3
[KHAF11]	M. Kallas, P. Honeine, H. Amoud, and C. Francis. Sur le problème de la pré-image en reconnaissance des formes avec contraintes de non-négativité. <i>In 23eme édition du Colloque GRETSI</i> , Bordeaux, France, 5-8 Septembre 2011.	2
[KMK ⁺ 11]	L. Kanaan, D. Merheb, M. Kallas, C. Francis, H. Amoud, and P. Honeine. PCA and KPCA of ECG signals with binary SVM classification. <i>In 25th IEEE Workshop on Signal Processing Systems SiPS</i> , Beirut, Lebanon, 4-7 Octobre 2011.	4
[KHFA11]	M. Kallas, P. Honeine, C. Francis, and H. Amoud. Kernel autoregressive models. a comparative study of pre-image techniques. <i>In 25th IEEE Workshop on Signal Processing Systems SiPS</i> , Beirut, Lebanon, 4-7 Octobre 2011.	3
[KHR ⁺ 12b]	M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Prediction of time series using Yule-Walker equations with kernels. <i>In 37th IEEE International Conference on Acoustics, Speech and Signal Processing</i> , Kyoto, Japan, 25-30 March 2012.	3
[KFK ⁺ 12]	M. Kallas, C. Francis, L. Kanaan, D. Merheb, P. Honeine, and H. Amoud. Multi-class SVM classification combined with kernel PCA feature extraction of ECG signals. <i>In 19th International Conference on Telecommunications</i> , Jounieh, Lebanon, 23-25 April 2012.	4
[KFH ⁺ 12]	M. Kallas, C. Francis, P. Honeine, H. Amoud, and C. Richard. Modeling electrocardiogram using Yule-Walker equations and kernel machines. <i>In 19th International Conference on Telecommunications</i> , Jounieh, Lebanon, 23-25 April 2012.	3
[KHR ⁺ 12a]	M. Kallas, P. Honeine, C. Richard, H. Amoud, and C. Francis. Non-negativity constraints on the pre-image for pattern recognition with kernel machines. <i>Elsevier, Pattern Recognition</i> , Soumis en Avril 2012.	1,2
[KHFA12]	M. Kallas, P. Honeine, C. Francis, and H. Amoud. Kernel autoregressive models using Yule-Walker equations. <i>Elsevier, Signal Processing</i> , Soumis en Juillet 2012	3

Les méthodes à noyaux et le problème de la pré-image

Sommaire

1.1	Introduction	12
1.2	Noyaux et espace de Hilbert à noyau reproduisant	13
1.2.1	Noyau défini positif et RKHS	13
1.2.2	Théorème de Moore-Aronszajn	14
1.3	Du modèle linéaire au modèle à noyaux	17
1.3.1	Astuce du noyau	17
1.3.2	Théorème de Représentation	18
1.4	Exemple du linéaire au non-linéaire	19
1.4.1	Analyse en composantes principales	19
1.4.2	Analyse en composantes principales à noyaux	20
1.5	L'ACP-à-noyaux pour la reconnaissance des formes	22
1.5.1	Extraction des caractéristiques	22
1.5.2	Débruitage	22
1.5.3	Une vue unifiée	23
1.6	Définition du problème de pré-image	23
1.7	Méthodes de résolution	25
1.7.1	Méthode de la descente du gradient	25
1.7.2	Méthode itérative du point fixe	26
1.7.3	Apprentissage de la carte de pré-image	28
1.7.4	Méthode de l'échelle multidimensionnelle	28
1.7.5	Approche conforme	30
1.7.6	Pré-image régularisée ou pénalisée	31
1.8	Formulation de la pré-image	31
1.9	Conclusion	33

1.1 Introduction

Au cours des deux dernières décennies, nous avons assisté à une prolifération des méthodes à noyaux grâce à la diversité des traitements non-linéaires qu'elles autorisent avec un faible coût calculatoire [STC04]. Depuis les Machines à Vecteurs Supports de Vapnik [Vap98, BGV92b, SBS98], elles ont montré leurs performances dans plusieurs domaines aux finalités variées. Bien qu'elles soient appliquées avec succès pour résoudre des problèmes décisionnels non-linéaires, comme la classification, la régression et la détection, souvent elles sont moins adaptées en ce qui concerne la reconnaissance des formes. Cette condition est due essentiellement à la notion de l'astuce du noyau, ou *kernel trick* en anglais, une “arme à double tranchant”. En fait, l'astuce du noyau fournit un moyen de transformer implicitement les données dans un espace de caractéristique non-linéaire de grande dimension, ce qui permet de construire des règles de décision non-linéaires, avec essentiellement le même coût de calcul que celles des cas linéaires. En d'autres termes, l'idée principale réside dans l'interprétation d'un noyau défini positif comme un produit scalaire dans un espace fonctionnel. Ainsi un tel noyau assure-t-il le passage des données de l'espace des observations à l'espace dit de Hilbert à noyau reproduisant, sans la nécessité d'exhiber la fonction de transformation non-linéaire associée. Cet espace de Hilbert est initialement proposé pour les problèmes de régression par Kimeldorf et Wahba dans [KW71, Wah90]. Cette notion de non-linéarité par l'usage de noyau a été proposée par Aizerman *et al.* dans [ABR64] dans le cadre d'un problème de classification, et renforcé par Vapnik dans [Vap98] avec la théorie de l'apprentissage statistique dans un contexte plus général de classification et régression. C'est le cas de l'Analyse en Composantes Principales à noyaux (ACP-à-noyaux). A l'instar de l'ACP classique, cette extension non-linéaire vise à identifier un sous-espace pertinent pour les données en maximisant leur variance projetée. Une telle projection se faisant implicitement dans le RKHS, néanmoins, nous n'avons pas accès à la plupart des éléments de l'espace de Hilbert à noyau reproduisant, comme des éléments ou caractéristiques estimés par l'ACP-à-noyaux [SSM98a].

L'importance du passage de l'espace des observations à l'espace de Hilbert à noyau reproduisant est claire en classification et régression. Cependant, la fonction réciproque, de l'espace transformé à l'espace des observations, est souvent indispensable, surtout pour retrouver le résultat dans l'espace des observations, e.g., l'espace des signaux en traitement du signal. Or, les deux espaces ne sont pas en bijection, et très peu d'éléments de l'espace transformé ont un antécédent dans l'espace des observations. Le problème de la recherche de cet antécédent est connu sous le nom du problème de la *pré-image*. Il consiste à trouver une observation dont l'image, par la fonction noyau considérée, soit la plus proche possible de l'élément en question dans l'espace transformé. Plusieurs méthodes ont été proposées dans la littérature afin de résoudre ce problème mal-posé.

Ce chapitre couvre tout d'abord la notion des noyaux reproduisants. Ensuite, les caractéristiques de tels noyaux sont données. Après, nous introduisons les deux éléments fondamentaux des méthodes à noyaux qui sont l'astuce du noyau et le théorème de Représentation. Dans la section 1.4, un exemple d'algorithme linéaire précisément l'analyse en composantes principales est détaillé, tout en présentant

son extension à l'aide des méthodes à noyaux pour le cas non-linéaire. La section 1.5 montre l'usage d'un tel algorithme pour la reconnaissance des formes. Le problème de la pré-image confronté lors de l'utilisation d'un noyau est décrit dans la section 1.6, tout en présentant les méthodes pour sa résolution dans la section 1.7. Finalement, une nouvelle écriture de la pré-image en fonction des données disponibles est proposée dans la section 1.8.

1.2 Noyaux et espace de Hilbert à noyau reproduisant

Nous considérons l'espace des observations \mathcal{X} , auquel est associé le produit scalaire $\langle \cdot, \cdot \rangle$ et sa norme correspondante $\|\cdot\|^2$. Avant d'étudier des propriétés liées à la notion du noyau, il est nécessaire de le définir.

Définition 1.1. (*Noyau*). *Un noyau désigne une fonction de $\mathcal{X} \times \mathcal{X}$ dans \mathbb{R} à symétrie Hermitienne, c'est-à-dire telle que $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \kappa(\mathbf{x}_j, \mathbf{x}_i)$.*

À partir d'un noyau, nous construisons sa matrice de Gram comme donnée par la Définition 1.2.

Définition 1.2. (*Matrice de Gram*). *Étant donné un noyau $\kappa(\cdot, \cdot)$ et n observations $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$, la matrice définie par*

$$(\mathbf{K})_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j),$$

pour tout $i, j = 1, 2, \dots, n$ est appelée la matrice de Gram de κ associée à l'ensemble $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$. C'est une matrice de dimension $n \times n$.

L'idée principale des méthodes à noyaux réside en l'utilisation de techniques linéaires classiques sur des données transformées. Soit $\Phi(\cdot)$ la transformation des données de l'espace des observations \mathcal{X} , à un espace fonctionnel \mathcal{H} .

Cependant, dans certains cas, la fonction $\Phi(\cdot)$ peut être opérée implicitement en utilisant un noyau afin d'évaluer des produits scalaires dans \mathcal{H} . Afin qu'une fonction $\kappa(\cdot, \cdot)$ représente un produit scalaire dans l'espace fonctionnel \mathcal{H} , elle doit satisfaire des conditions explorées dans la section suivante

1.2.1 Noyau défini positif et RKHS

Commençons tout d'abord par quelques définitions afin de déterminer la condition d'existence d'un espace fonctionnel \mathcal{H} .

Définition 1.3. (*Noyau défini positif*). *Un noyau κ est dit défini positif sur \mathcal{X} si et seulement si, il vérifie*

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \geq 0, \quad (1.1)$$

pour tout $n \in \mathbb{N}$, $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$ et $\alpha_1, \dots, \alpha_n \in \mathbb{R}$.

Définition 1.4. (*Espace de Hilbert*). Un espace vectoriel \mathcal{H} muni d'un produit scalaire $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ est un espace de Hilbert s'il est complet pour la norme associée au produit scalaire $\|\iota\|_{\mathcal{H}}^2 = \langle \iota, \iota \rangle_{\mathcal{H}}$ (en d'autres termes, toutes les suites de Cauchy convergent dans \mathcal{H}).

Définition 1.5. (*Espace de Hilbert à noyau reproduisant - RKHS*). Soit $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ un espace de Hilbert constitué par des fonctions de \mathcal{X} dans \mathbb{R} . La fonction $\kappa(x_i, x_j)$ de $\mathcal{X} \times \mathcal{X}$ dans \mathbb{R} est le noyau reproduisant de \mathcal{H} , sous réserve que celui-ci en admette un, si et seulement si

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

Nous disons que \mathcal{H} est un espace de Hilbert à noyau reproduisant, ou encore RKHS, acronyme de *Reproducing Kernel Hilbert Space*.

Une propriété importante est tirée de cette définition.

Propriété 1.1. (*Reproduction*). La propriété reproduisante, définie par Aronszajn dans [Aro50], du noyau κ induisant un espace de Hilbert \mathcal{H} , est donnée par

$$\langle \kappa(x, \cdot), f(\cdot) \rangle_{\mathcal{H}} = f(x) \quad \text{pour tout } f(\cdot) \in \mathcal{H}.$$

À partir de la Propriété 1.1, nous pouvons déduire facilement le corollaire suivant :

Corollaire 1.1. (*Astuce du noyau*). Tout noyau défini positif κ induisant un espace de Hilbert \mathcal{H} définit le produit scalaire dans cet espace, comme suit :

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

pour chaque x_i, x_j dans \mathcal{X} .

La Figure 1.1 représente l'espace de Hilbert à noyau reproduisant \mathcal{H} associé au noyau κ appliqué sur l'espace des observations \mathcal{X} . Nous définissons une transformation de \mathcal{X} vers l'espace des fonctions de \mathcal{X} , noté \mathcal{H} , ainsi

$$\begin{aligned} \Phi &: \mathcal{X} \rightarrow \mathcal{H} \\ x &\mapsto \kappa(x, \cdot). \end{aligned}$$

Dans cette expression, $\Phi(x) = \kappa(x, \cdot)$ désigne une fonction définie sur \mathcal{X} , obtenue en fixant le premier argument de κ à x .

1.2.2 Théorème de Moore-Aronszajn

Le théorème suivant [Aro50], combiné avec les définitions précédentes, permet de faire le lien entre un noyau défini positif et l'espace de Hilbert à noyau reproduisant.

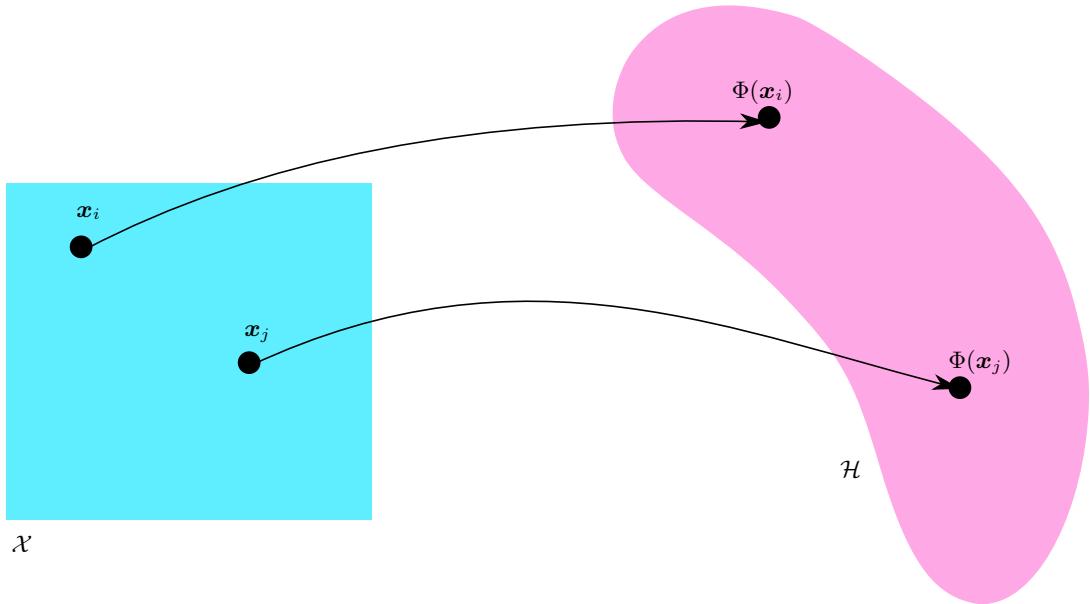


FIGURE 1.1: Espace de Hilbert à noyau reproduisant \mathcal{H} associé au noyau κ appliqué sur l'espace des observations \mathcal{X} .

Théorème 1.1. (Moore-Aronszajn [Aro50]). À tout noyau défini positif κ , il lui correspond un espace de Hilbert à noyau reproduisant \mathcal{H} unique, et réciproquement.

Démonstration. Nous montrons tout d'abord que tout noyau reproduisant est défini positif. À cette fin, il suffit de constater que $\sum_i \sum_j \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) = \|\sum_i \alpha_i \Phi(\mathbf{x}_i)\|^2$ ne peut être négatif. Réciproquement, nous démontrons que tout noyau défini positif κ est le noyau reproduisant d'un espace de Hilbert de fonctions de \mathcal{X} dans \mathbb{R} . Pour ce faire, un espace de Hilbert à noyau reproduisant \mathcal{H} associé à un noyau κ , est construit en considérant une fonction $\Phi(\cdot)$ de \mathcal{X} dans \mathcal{H} , selon $\Phi(\mathbf{x}) = \kappa(\mathbf{x}, \cdot)$. L'espace \mathcal{H} est engendré par les fonctions $\Phi(\mathbf{x})$. Soient deux fonctions dans \mathcal{H}

$$\iota_1 = \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i), \quad \iota_2 = \sum_{j=1}^n \beta_j \Phi(\mathbf{x}_j),$$

où n est un entier naturel, α_i et β_j sont des réels et \mathbf{x}_i et \mathbf{x}_j appartiennent à \mathcal{X} . Le produit scalaire entre ces deux fonctions est donné par :

$$\langle \iota_1, \iota_2 \rangle_{\mathcal{H}} = \left\langle \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i), \sum_{j=1}^n \beta_j \Phi(\mathbf{x}_j) \right\rangle_{\mathcal{H}}.$$

TABLE 1.1: Les noyaux reproduisants couramment utilisés en apprentissage, avec les paramètres $c, \sigma > 0$, et $q \in \mathbb{N}_+$.

	Type	Forme générale
Projectif	Polynomial	$\kappa_q(\mathbf{x}_i, \mathbf{x}_j) = (c + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^q$
	Polynomial de Vovk	$\kappa_{PV}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1 - \langle \mathbf{x}_i, \mathbf{x}_j \rangle^q}{1 - \langle \mathbf{x}_i, \mathbf{x}_j \rangle}$
	Exponentiel	$\kappa_E(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{1}{\sigma} \langle \mathbf{x}_i, \mathbf{x}_j \rangle\right)$
Radial	Laplacien	$\kappa_L(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-1}{\sigma} \ \mathbf{x}_i - \mathbf{x}_j\ \right)$
	Gaussien	$\kappa_G(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-1}{2\sigma^2} \ \mathbf{x}_i - \mathbf{x}_j\ ^2\right)$
	Quadratique rationnel	$\kappa_R(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\ \mathbf{x}_i - \mathbf{x}_j\ ^2}{\ \mathbf{x}_i - \mathbf{x}_j\ ^2 + \sigma}$

En utilisant l'astuce du noyau, l'expression du produit scalaire devient

$$\langle \boldsymbol{\iota}_1, \boldsymbol{\iota}_2 \rangle_{\mathcal{H}} = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \beta_j \kappa(\mathbf{x}_i, \mathbf{x}_j).$$

Nous obtenons alors un espace pré-Hilbertien. Pour aboutir à un espace Hilbertien, il suffit de le compléter conformément à [Aro50] afin que toute suite de Cauchy y converge. \square

La relation associant l'espace de Hilbert à noyau reproduisant à un noyau donné est décrite par le biais du théorème de Moore-Aronszajn. Dans la suite, un noyau défini positif est désigné par un noyau reproduisant. Le Tableau 1.1 résume les noyaux reproduisants les plus utilisés. Ils sont groupés sous deux classes : les noyaux projectifs, de la forme

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = f(\langle \mathbf{x}_i, \mathbf{x}_j \rangle), \quad (1.2)$$

et les noyaux radiaux, de la forme

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = g(\|\mathbf{x}_i - \mathbf{x}_j\|^2). \quad (1.3)$$

Les deux propositions suivantes sont considérées dans cette thèse afin de démontrer d'autres résultats. Soit $f^{(k)}(\zeta)$ la $k^{\text{ème}}$ dérivée de la fonction f par rapport à ζ . Commençons par les noyaux radiaux. Le résultat qui suit est dû à [CS02] et [Bur99, Proposition 7.2].

Proposition 1.1 (Noyaux radiaux). *Une condition suffisante pour une fonction de la forme $\kappa(\mathbf{x}_i, \mathbf{x}_j) =$*

$g(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$ soit un noyau défini positif est sa monotonie complète, c'est-à-dire, ses dérivées satisfont

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

pour tout $\zeta > 0$ et $k \geq 0$.

C'est le cas du noyau Gaussien $\kappa_G(\boldsymbol{x}_i, \boldsymbol{x}_j) = g(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2)$ avec

$$g^{(k)}(\zeta) = \left(-\frac{1}{2\sigma^2}\right)^k g(\zeta),$$

ou encore le noyau quadratique rationnel avec

$$g^{(k)}(\zeta) = (-1)^k k! \frac{\sigma}{(\zeta + \sigma)^{k+1}}.$$

Passons maintenant aux noyaux projectifs, le résultat suivant est donné dans [Bur99, Proposition 7.1].

Proposition 1.2 (Noyaux projectifs). *Trois conditions nécessaires pour qu'une fonction $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = f(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$ soit un noyau défini positif sont, pour tout ζ non-négatif :*

$$\begin{aligned} f(\zeta) &\geq 0 \\ f^{(1)}(\zeta) &\geq 0 \\ f^{(1)}(\zeta) + \zeta f^{(2)}(\zeta) &\geq 0 \end{aligned}$$

Il est facile de montrer ces conditions pour les noyaux projectifs donnés dans le Tableau 1.1.

1.3 Du modèle linéaire au modèle à noyaux

Après avoir introduit les noyaux ainsi que leur caractérisation, nous passons maintenant à leur usage. L'idée principale de l'usage des noyaux reproduisants est le passage de la linéarité à la non-linéarité. Pour ce faire, les algorithmes linéaires sont modifiés à l'aide des deux éléments fondamentaux qui sont : l'astuce du noyau [ABR64] et le théorème de Représentation [Wah90, SHS01].

1.3.1 Astuce du noyau

En utilisant le Corollaire 1.1, nous pouvons écrire le noyau reproduisant avec

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \Phi(\boldsymbol{x}_i), \Phi(\boldsymbol{x}_j) \rangle_{\mathcal{H}},$$

quels que soient \boldsymbol{x}_i et \boldsymbol{x}_j dans \mathcal{X} , où \mathcal{H} est l'espace de Hilbert associé à ce noyau. Cette propriété, dite *astuce du noyau*, permet de transformer les méthodes de traitement linéaire de données en des

méthodes non-linéaires, sous réserve qu'elles puissent s'exprimer en fonction de produits scalaires des observations $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$, qui n'est autre le noyau linéaire. Ce produit scalaire est alors remplacé par un noyau non-linéaire $\kappa(\mathbf{x}_i, \mathbf{x}_j)$. Ainsi la structure des algorithmes demeure-t-elle inchangée, et le surcoût calculatoire dû à l'évaluation des noyaux négligeable.

En fait, l'astuce du noyau fournit un moyen de représenter les observations implicitement dans un espace fonctionnel. Par conséquent, le noyau reproduisant correspond à une généralisation du produit scalaire canonique, et est donc une mesure non-linéaire de la similarité entre les observations. Il s'avère que la plupart des algorithmes linéaires utilisés pour le traitement des données peuvent être facilement reformulés en termes de produit scalaire dans l'espace des observations. Sa substitution par un noyau offre des extensions non-linéaires des algorithmes classiques. Le concept de l'astuce du noyau est illustré dans [SSM98b, RGT00] pour l'ACP-à-noyaux décrite dans la section 1.4.2.

1.3.2 Théorème de Représentation

Nous pouvons constater que sous certaines conditions, la solution optimale d'un problème d'optimisation dans \mathcal{H} peut s'écrire sous la forme d'une combinaison de noyaux, indépendamment de la dimension de \mathcal{H} . Cette constatation est formulée par le théorème suivant :

Théorème 1.2. (*Théorème de Représentation [KW71, SHS01]*). Soient un espace non vide \mathcal{X} , un noyau défini positif κ sur $\mathcal{X} \times \mathcal{X}$, un ensemble d'échantillons d'apprentissage $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n) \in \mathcal{X} \times \mathbb{R}$, une fonction réelle g strictement monotone croissante sur $[0, \infty[$, et une fonction coût arbitraire c . Soit \mathcal{H} l'espace de Hilbert associé au noyau κ , avec $\Phi(\mathbf{x}_i) = \kappa(\mathbf{x}_i, \cdot)$. Toute fonction $f_{\mathcal{H}} \in \mathcal{H}$ minimisant la fonctionnelle de risque régularisée

$$c((\mathbf{x}_1, y_1, f_{\mathcal{H}}(\mathbf{x}_1)), \dots, (\mathbf{x}_n, y_n, f_{\mathcal{H}}(\mathbf{x}_n))) + g(\|f\|), \quad (1.4)$$

admet une représentation de la forme

$$f_{\mathcal{H}} = \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i).$$

Démonstration. Étant donné un ensemble $\mathbf{x}_1, \dots, \mathbf{x}_n$, toute fonction $f_{\mathcal{H}} \in \mathcal{H}$ peut être décomposée en une partie appartenant à l'espace engendré par les $\Phi(\mathbf{x}_i)$, et une partie orthogonale à celui-ci, selon

$$f_{\mathcal{H}} = \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) + v,$$

où les $\alpha_1, \dots, \alpha_n$ sont des réels et $v \in \mathcal{H}$ vérifiant pour tout i

$$\langle v, \Phi(\mathbf{x}_i) \rangle = 0.$$

En utilisant cette équation et la propriété reproduisante, l'évaluation de $f_{\mathcal{H}}$ en tout échantillon arbitraire \mathbf{x}_j donne

$$f_{\mathcal{H}}(\mathbf{x}_j) = \left\langle \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) + v, \Phi(\mathbf{x}_j) \right\rangle = \sum_{i=1}^n \alpha_i \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle.$$

Par conséquence, le premier terme de l'expression (1.4) est indépendant de v . Cependant pour le second terme, puisque v est orthogonal à $\sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i)$, et g est strictement monotone, nous obtenons

$$\begin{aligned} g(\|f_{\mathcal{H}}\|) &= g\left(\left\| \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) + v \right\|\right) = g\left(\sqrt{\left(\left\| \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) \right\|^2 + \|v\|^2\right)}\right) \\ &\geq g\left(\left\| \sum_i \alpha_i \Phi(\mathbf{x}_i) \right\|\right), \end{aligned}$$

où l'égalité se produit si et seulement si $v = 0$. L'égalité entre $\left\| \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) + v \right\| = \sqrt{\left(\left\| \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i) \right\|^2 + \|v\|^2\right)}$ est donnée par le théorème de Pythagore. Mettant v à 0 n'influe pas sur le premier terme de (1.4), tout en réduisant son second terme. Par conséquence, toute solution s'écrit sous la forme

$$f_{\mathcal{H}} = \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i).$$

□

1.4 Exemple du linéaire au non-linéaire

Dans la section précédente, les éléments fondamentaux des méthodes à noyaux sont présentés, en mentionnant le passage de la linéarité à la non-linéarité. Nous présentons alors un exemple de ce passage. Pour ce faire, nous détaillons la méthode de l'analyse en composantes principales en détaillant son extension au cas non-linéaire.

1.4.1 Analyse en composantes principales

L'analyse en composantes principales (ACP) est un outil mathématique puissant pour révéler des formes au sein d'un ensemble de données. Il s'agit d'une approche non-paramétrique, qui ne tient compte d'aucune connaissance préalable du système, à l'exception de sa linéarité. Cette approche est considérée comme une approche globale, par opposition à des méthodes telles que les modèles paramétriques ou décomposition en ondelettes, où les caractéristiques extraites dépendent fortement du type de modèle ou du type d'ondelettes utilisé pour l'analyse.

En ACP, les caractéristiques choisies sont celles qui présentent le maximum de variance des données. Nous pouvons montrer que les vecteurs propres donnent le maximum de variance des données. Pour ce faire, ces caractéristiques sont obtenues par diagonalisation de la matrice de

corrélation des données, tout en conservant seulement les vecteurs propres les plus pertinents, c'est-à-dire, vecteurs propres associés aux plus grandes valeurs propres. Ces vecteurs propres constituent alors un ensemble d'axes orthonormaux présentant la plus grande variance dans les données. Considérons un ensemble de n données $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ dans un espace donné \mathcal{X} . Sans perte de généralité, nous supposons que ces données sont centrées dans cet espace \mathcal{X} . L'ACP cherche les m caractéristiques $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_m$, comme les vecteurs propres du problème aux valeurs propres suivant :

$$\lambda_k \mathbf{v}_k = \mathbf{C} \mathbf{v}_k,$$

où $\mathbf{C} = \frac{1}{n} \sum_{j=1}^n \mathbf{x}_j \mathbf{x}_j^\top$ est la matrice de covariance, avec \mathbf{x}_j représentant un vecteur colonne et \mathbf{x}_j^\top est sa transposée. La pertinence de chaque vecteur propre \mathbf{v}_k est donnée par sa valeur propre correspondante λ_k , qui mesure la proportion de la variance des données capturées. Puisque $\lambda_k \mathbf{v}_k = \mathbf{C} \mathbf{v}_k = \frac{1}{n} \sum_{j=1}^n \langle \mathbf{x}_j, \mathbf{v}_k \rangle \mathbf{x}_j$, les vecteurs propres appartiennent à l'espace engendré par les n données. Il est important à noter que les données sont supposées centrées et que les vecteurs propres résultants sont de norme unité.

1.4.2 Analyse en composantes principales à noyaux

L'un des inconvénients de l'ACP classique est sa linéarité. Elle n'identifie que les structures linéaires dans un ensemble de données. Une technique plus généralisée a été mise en place pour apprendre les non-linéarités en utilisant les noyaux, ladite ACP-à-noyaux. Cette dernière peut révéler les composantes principales non-linéaires par le biais du noyau qui sont plus appropriées aux données complexes tels que les images de visage, les chiffres manuscrits et les signaux naturels. A cet effet, les données sont (implicitement) transformées dans un espace fonctionnel, où l'ACP classique est appliquée. Bien que les vecteurs propres résultant soient obtenus par une technique linéaire dans l'espace fonctionnel, ils décrivent des relations non-linéaires dans l'espace des observations. Afin de résoudre ce problème non-linéaire, il est plus souhaitable d'appliquer l'astuce du noyau, et de ne pas calculer explicitement la fonction non-linéaire de transformation.

Pour ce faire, l'algorithme de l'ACP est reformulé en termes de produit scalaire des données dans l'espace fonctionnel. Soit Φ une transformation non-linéaire de l'espace des observations \mathcal{X} à l'espace fonctionnel \mathcal{H} qui, à chaque \mathbf{x}_i lui fait correspondre son image $\Phi(\mathbf{x}_i)$. Ainsi l'ensemble des observations transformées est $\{\Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2), \dots, \Phi(\mathbf{x}_n)\}$. Nous souhaitons résoudre l'ACP (-à-noyaux), en terme de produit scalaire dans l'espace fonctionnel, $\langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$, pour tout $i, j = 1, 2, \dots, n$. La matrice de covariance dans \mathcal{H} est donnée par :

$$\mathbf{C}^\Phi = \frac{1}{n} \sum_{j=1}^n \langle \Phi(\mathbf{x}_j), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}}.$$

Les axes principaux, $\varphi_k \in \mathcal{H}$, pour $k = 1, 2, \dots, m$, correspondent aux vecteurs propres ayant des

valeurs propres λ_k vérifiant l'expression suivante

$$\lambda_k \psi_k = \mathbf{C}^\Phi \psi_k. \quad (1.5)$$

Par analogie avec l'ACP classique, chaque solution φ_k se situe dans l'espace engendré par les images des données par la fonction $\Phi(\cdot)$. Cette écriture implique qu'il existe des coefficients $\alpha_1, \alpha_2, \dots, \alpha_n$ de sorte que

$$\psi_k = \sum_{i=1}^n \alpha_{k,i} \Phi(\mathbf{x}_i), \quad (1.6)$$

puisque

$$\lambda_k \psi_k = \mathbf{C}^\Phi \psi_k = \frac{1}{n} \sum_{j=1}^n \langle \Phi(\mathbf{x}_j), \psi_k \rangle_{\mathcal{H}} \Phi(\mathbf{x}_j).$$

En remplaçant l'expression de \mathbf{C}^Φ et la représentation des axes principaux ψ_k de (1.6) dans l'équation du problème aux valeurs propres (1.5), nous obtenons une nouvelle expression du problème aux valeurs propres écrite en termes de produit scalaire, avec

$$n \lambda_k \boldsymbol{\alpha}_k = \mathbf{K} \boldsymbol{\alpha}_k, \quad (1.7)$$

où \mathbf{K} est une matrice de taille $n \times n$ d'éléments $\kappa(\mathbf{x}_i, \mathbf{x}_j)$, et $\boldsymbol{\alpha}_k$ est un vecteur regroupant les n coefficients, à savoir $\boldsymbol{\alpha}_k = [\alpha_{k,1} \ \alpha_{k,2} \ \dots \ \alpha_{k,n}]^\top$. Par ailleurs, deux mises au point sont à considérer dans l'algorithme de l'ACP-à-noyaux final. Tout d'abord, les données doivent être centrées dans l'espace fonctionnel. Cette condition est valide en remplaçant la matrice \mathbf{K} par

$$(\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top) \mathbf{K} (\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top),$$

où \mathbf{I} est la matrice identité et $\mathbf{1}_n$ est la matrice unité de taille $n \times n$ telle que $(\mathbf{1}_n)_{i,j} = 1$. Cette condition est démontrée ici¹. Ensuite, et par analogie avec l'ACP classique les vecteurs correspondant dans \mathcal{H} doivent être de norme unitaire, c'est-à-dire $\langle \varphi_k, \varphi_k \rangle_{\mathcal{H}} = 1$. Nous pouvons facilement montrer que cette condition

1. Centrage des données dans l'espace fonctionnel

Bien que le centrage dans l'espace des observations \mathcal{X} est aisément vérifié, il n'est pas le cas dans l'espace fonctionnel \mathcal{H} . Soit $\Phi^c(\mathbf{x}_i)$ la fonction centrée dans \mathcal{H} . Elle est définie par $\Phi^c(\mathbf{x}_i) = \Phi(\mathbf{x}_i) - \frac{1}{n} \sum_{k=1}^n \Phi(\mathbf{x}_k)$. Chaque élément de la matrice de Gram de celles-ci peut s'écrire

$$\begin{aligned} (\mathbf{K}^c)_{i,j} &= \langle \Phi^c(\mathbf{x}_i), \Phi^c(\mathbf{x}_j) \rangle_{\mathcal{H}} \\ &= \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle_{\mathcal{H}} - \frac{1}{n} \sum_{k=1}^n \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_k) \rangle_{\mathcal{H}} - \frac{1}{n} \sum_{k=1}^n \langle \Phi(\mathbf{x}_j), \Phi(\mathbf{x}_k) \rangle_{\mathcal{H}} + \frac{1}{n^2} \sum_{k,k'=1}^n \langle \Phi(\mathbf{x}_k), \Phi(\mathbf{x}_{k'}) \rangle_{\mathcal{H}}. \end{aligned}$$

Une écriture matricielle est alors :

$$\mathbf{K}^c = \mathbf{K} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top - \frac{1}{n} \mathbf{K} \mathbf{1}_n + \frac{1}{n^2} \mathbf{1}_n \mathbf{K} \mathbf{1}_n,$$

ce qui correspond à

$$\mathbf{K}^c = (\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top) \mathbf{K} (\mathbf{I} - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top).$$

se fait par mise à l'échelle des vecteurs de poids α_k suivant l'équation donnée par $\lambda_k(\alpha_k \cdot \alpha_k) = 1$, pour toutes les m caractéristiques $k = 1, 2, \dots, m$.

1.5 L'ACP-à-noyaux pour la reconnaissance des formes

Deux domaines d'application principaux peuvent être considérés avec l'ACP classique : d'une part, considérer les axes principaux pertinents comme des caractéristiques extraites, et d'autre part, projeter les observations bruitées sur ces axes formant ainsi un sous-espace assurant le débruitage. Ces deux domaines d'application sont étudiés dans l'espace fonctionnel, en utilisant l'ACP-à-noyaux avant de proposer une vue unifiée.

1.5.1 Extraction des caractéristiques

Nous étudions dans cette partie l'extraction des caractéristiques dans le RKHS, avec l'ACP-à-noyaux. Ce dernier définit un ensemble d'axes les plus pertinents dans l'espace fonctionnel. Soient $\{\psi_1, \psi_2, \dots, \psi_m \in \mathcal{H}\}$ l'ensemble de ces axes. Alors chaque ψ_k est de la forme (1.6), comme suit

$$\psi_k = \sum_{j=1}^n \alpha_{k,j} \Phi(\mathbf{x}_j),$$

où $\alpha_{k,1}, \alpha_{k,2}, \dots, \alpha_{k,n}$ sont obtenus du vecteur propre associé à la $k^{\text{ème}}$ valeur propre dans (1.7). Par analogie à l'ACP classique, ces axes absorbent la plus grande variance des données. Ainsi, les axes sont-ils considérés comme étant une extraction des caractéristiques de ces données-ci, captant les variations les plus grandes et sont orthonormaux les uns par rapport aux autres. De même, nous pouvons définir le sous-espace pertinent de \mathcal{H} , celui qui est engendré par ces axes principaux, ce qui nous permet de débruiter les données, comme décrit dans la section suivante.

1.5.2 Débruitage

Nous étudions maintenant l'idée de débruitage des données. Soit $x_0 \in \mathcal{X}$ un des échantillons d'apprentissage (en particulier bruité). Alors, son image par la fonction non-linéaire $\Phi(\cdot)$, désignée par $\Phi(x_0)$, est projetée sur le sous-espace pertinent décrit auparavant, donnant ainsi la forme débruite en question. Cette forme est donnée par le produit scalaire entre les transformées des échantillons par la fonction non-linéaire qui sont projetées sur le sous-espace pertinent et les m axes principaux du sous-espace, comme suit

$$\psi = \sum_{k=1}^m \langle \Phi(x_0), \psi_k \rangle_{\mathcal{H}} \psi_k.$$

En substituant cette expression dans (1.6) et en appliquant l'équivalent entre l'opérateur du produit scalaire et la fonction noyau κ , nous obtenons

$$\begin{aligned}\psi &= \sum_{k=1}^m \langle \Phi(\mathbf{x}_0), \sum_{i=1}^n \alpha_{k,i} \Phi(\mathbf{x}_i) \rangle_{\mathcal{H}} \sum_{j=1}^n \alpha_{k,j} \Phi(\mathbf{x}_j) \\ &= \sum_{k=1}^m \sum_{i=1}^n \alpha_{k,i} \kappa(\mathbf{x}_0, \mathbf{x}_i) \sum_{j=1}^n \alpha_{k,j} \Phi(\mathbf{x}_j).\end{aligned}$$

1.5.3 Une vue unifiée

Nous proposons maintenant une vue unifiée pour confronter les deux problèmes de reconnaissance des formes décrits auparavant. Pour ce faire, nous écrivons la caractéristique extraite et la forme débruitée comme une combinaison linéaire des données d'apprentissage projetées, avec $\psi_k = \sum_{j=1}^n \alpha_{k,j} \Phi(\mathbf{x}_j)$ et $\psi = \sum_{j=1}^n [\sum_{k=1}^m \sum_{i=1}^n \alpha_{k,i} \alpha_{k,j} \kappa(\mathbf{x}_0, \mathbf{x}_i)] \Phi(\mathbf{x}_j)$. En regroupant tous ces termes, nous aboutissons à une vue unifiée de ces deux cas, avec

$$\psi = \sum_{j=1}^n \gamma_j \Phi(\mathbf{x}_j). \quad (1.8)$$

D'une part, la caractéristique extraite est donnée par $\psi = \psi_k$ où

$$\gamma_j = \alpha_{k,j},$$

et d'autre part, la forme débruitée est donnée par

$$\gamma_j = \sum_{k=1}^m \sum_{i=1}^n \alpha_{k,i} \alpha_{k,j} \kappa(\mathbf{x}_0, \mathbf{x}_i).$$

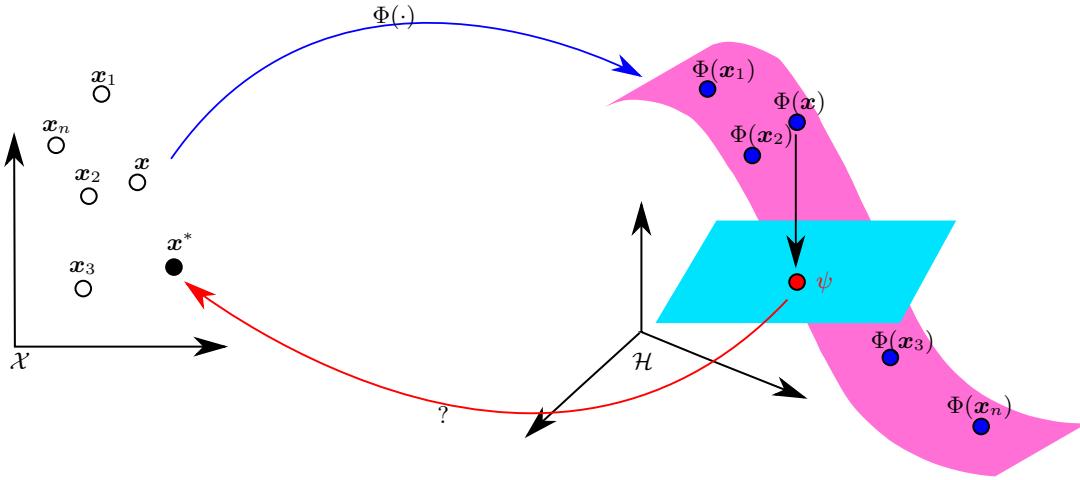
Dans ce dernier cas, les coefficients γ_j dépendent de l'échantillon bruité \mathbf{x}_0 , qui peut être soit une nouvelle observation, soit une des données d'apprentissage. Le Tableau 1.2 résume ces deux domaines d'application, et l'expression unifiée donnée par (1.8) nous permet de définir une forme générale pour le problème d'optimisation pour les deux applications, celle de l'extraction des caractéristiques et celle du débruitage.

1.6 Définition du problème de pré-image

Les méthodes à noyaux permettent la transformation d'un espace des observations à un espace des caractéristiques. Toutefois, souvent en reconnaissance des formes, nous sommes intéressés par la forme elle-même. Nous cherchons plutôt l'équivalent dans l'espace des observations de la caractéristique obtenue dans le RKHS. Cependant le retour inverse de l'espace caractéristiques à l'espace des observations

TABLE 1.2: Vue unifiée pour la définition de γ_j dans $\psi = \sum_{j=1}^n \gamma_j \Phi(\mathbf{x}_j)$

Application	γ_j
Extraction des caractéristiques	$\alpha_{k,j}$
Débruitage de \mathbf{x}_0	$\sum_{k=1}^m \sum_{i=1}^n \alpha_{k,i} \alpha_{k,j} \kappa(\mathbf{x}_0, \mathbf{x}_i)$

FIGURE 1.2: Schéma illustrant le problème de la pré-image qui consiste à trouver un retour inverse de l'espace \mathcal{H} à l'espace des observations \mathcal{X} afin de déterminer les éléments dont leurs images constituent ce RKHS, surtout que la plupart des éléments du RKHS ne sont pas des images d'un élément de l'espace des observations.

n'est pas évident.

Définition 1.6. (*Problème mal-posé au sens de Hadamard*). Un problème est dit mal-posé si l'une des trois conditions caractérisant les problèmes bien posés au sens de Hadamard n'est pas satisfaite, à savoir

- La solution existe
- La solution est unique
- La solution dépend de façon continue des données (condition dite de stabilité)

En fait, retrouver la pré-image est un problème mal-posé. Il suffit de se rappeler que l'espace fonctionnel soit souvent de dimension plus grande que celle de l'espace des observations. Par suite, la solution exacte de ce problème peut ne pas exister, ou si elle existe, elle peut ne pas être unique. Pour résoudre ce problème, il peut s'avérer nécessaire de déterminer un élément x^* de l'espace des observations tel que son image, par la fonction non-linéaire $\Phi(\cdot)$, soit la plus proche possible de ψ . Le retour inverse de l'espace fonctionnel à l'espace des observations est le *problème de la pré-image*, illustrée par la Figure

1.2. Il s'agit donc de résoudre le problème d'optimisation, en cherchant un \mathbf{x}^* vérifiant

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{X}} \frac{1}{2} \|\psi - \Phi(\mathbf{x})\|_{\mathcal{H}}^2, \quad (1.9)$$

où $\|\cdot\|_{\mathcal{H}}$ représente la norme dans le RKHS, donc nous fournit une mesure de distance entre les éléments de l'espace fonctionnel, avec la norme de leur résidu. En particulier, nous considérons une fonction coût $J(\mathbf{x})$ qui mesure l'écart entre l'image $\Phi(\mathbf{x}^*)$ et la fonction ψ

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{X}} J(\mathbf{x}).$$

En développant l'expression de la fonction coût dans 1.9, nous obtenons trois termes dont l'un ne dépend pas de \mathbf{x} , par suite, nous pouvons l'éliminer de l'expression, qui sera donnée par

$$J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i \kappa(\mathbf{x}_i, \mathbf{x}) + \frac{1}{2} \kappa(\mathbf{x}, \mathbf{x}). \quad (1.10)$$

C'est un problème d'optimisation non-linéaire et non-convexe, à cause de la nature du noyau. Récemment, un grand nombre de chercheurs se sont intéressés à ce problème et ont proposé des éléments de solution. Nous présentons les solutions les plus étudiées dans la section suivante.

1.7 Méthodes de résolution

Dans cette section, nous présentons les différentes techniques existantes pour la résolution du problème de la pré-image. Une étude est récemment réalisée dans [HR11] sur le problème de pré-image et son lien avec le problème de réduction de dimension qu'il convient de rappeler ici.

1.7.1 Méthode de la descente du gradient

La descente du gradient est une technique d'optimisation du premier ordre se basant sur le gradient de la fonction coût $J(\mathbf{x})$ par rapport à \mathbf{x} , noté $\nabla_{\mathbf{x}} J(\mathbf{x})$. Son concept est alors de partir d'un point initial et de se déplacer dans la direction opposée au gradient, en remplaçant à chaque itération \mathbf{x}^* par

$$\mathbf{x}^* - \eta_t \nabla_{\mathbf{x}} J(\mathbf{x}^*),$$

où η_t est le pas. Le gradient de (1.10) par rapport à \mathbf{x} est donné par

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i \nabla_{\mathbf{x}} \kappa(\mathbf{x}_i, \mathbf{x}) + \frac{1}{2} \nabla_{\mathbf{x}} \kappa(\mathbf{x}, \mathbf{x}). \quad (1.11)$$

C'est la forme générale pour tous les noyaux, par exemple les noyaux projectifs de la forme (1.2) comme le noyau polynomial, ou encore les noyaux radiaux de la forme (1.3) comme le noyau Gaussien. Voir le

TABLE 1.3: Gradient par rapport à \mathbf{x} de la fonction coût (1.10), pour les noyaux les plus utilisés

Type	$\nabla_{\mathbf{x}} J(\mathbf{x})$
Polynomial	$-\sum_{i=1}^n \gamma_i q \kappa_{q-1}(\mathbf{x}_i, \mathbf{x}) \mathbf{x}_i + q \kappa_{q-1}(\mathbf{x}, \mathbf{x}) \mathbf{x}$
Laplacien	$-\frac{1}{\sigma} \sum_{i=1}^n \gamma_i \kappa_L(\mathbf{x}_i, \mathbf{x})$
Exponentiel	$-\frac{1}{\sigma} \sum_{i=1}^n \gamma_i \kappa_E(\mathbf{x}_i, \mathbf{x}) \mathbf{x}_i + \frac{1}{\sigma} \kappa_E(\mathbf{x}, \mathbf{x}) \mathbf{x}$
Gaussien	$-\frac{1}{\sigma^2} \sum_{i=1}^n \gamma_i \kappa_G(\mathbf{x}_i, \mathbf{x}) (\mathbf{x}_i - \mathbf{x})$

Tableau 1.1 pour les expressions des gradients des noyaux couramment utilisés, et l'Annexe A pour la dérivation des expressions.

Puisque notre fonction coût est non-convexe et non-linéaire, des minima locaux sont présents. Pour remédier à ce problème, cette technique devrait être initialisée aléatoirement à plusieurs reprises pour essayer d'aboutir au résultat optimal. La descente du gradient n'est pas adaptée cependant, nous pouvons utiliser comme alternative à la descente du gradient, la méthode de Newton, mais elle représente une complexité calculatoire plus grande que celle de la descente du gradient.

Le pas devra être bien choisi, sinon le calcul peut diverger ou converger à un minimum local. Donc, si la valeur de η_t est trop grande, l'algorithme n'est pas stable et oscille autour d'une solution, et si la valeur de η_t est trop petite, un très grand nombre d'itérations est nécessaire pour converger vers la solution, et le risque de converger vers une solution locale est plus grand. Plusieurs critères d'arrêt peuvent être définis : le nombre d'itérations maximal, ou l'erreur résiduelle minimale entre deux itérations successives.

1.7.2 Méthode itérative du point fixe

Les fonctions de noyaux reproduisants ont une structure fournissant des indications utiles pour dériver des techniques d'optimisation plus appropriées, au-delà de la descente du gradient classique. A l'optimum, le gradient par rapport à \mathbf{x} s'annule, à savoir $\nabla_{\mathbf{x}} J(\mathbf{x}) = 0$. Le problème d'optimisation est alors simplifié, donnant lieu à une méthode itérative du point-fixe, avec un problème d'optimisation de la forme $\mathbf{x}^* = h(\mathbf{x}^*)$. Il suffit alors de substituer, à chaque itération, la solution candidate \mathbf{x}^* par $h(\mathbf{x}^*)$. Selon le type du noyau reproduisant utilisé, le gradient admet des expressions différentes.

Un calcul détaillé est présenté dans l'Annexe A. Prenons tout d'abord, les noyaux radiaux de la forme

(1.3), le gradient de la fonction coût est donné par

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i g^{(1)}(\|\mathbf{x}_i - \mathbf{x}\|^2) (\mathbf{x}_i - \mathbf{x}).$$

À l'optimum \mathbf{x}^* , l'expression du point fixe est définie par

$$\mathbf{x}^* = \frac{\sum_{i=1}^n \gamma_i g^{(1)}(\|\mathbf{x}_i - \mathbf{x}\|^2) \mathbf{x}_i}{\sum_{i=1}^n \gamma_i g^{(1)}(\|\mathbf{x}_i - \mathbf{x}\|^2)}.$$

Un exemple de cette classe est le noyau Gaussien [MSS⁺99], avec la fonction coût $-\sum_{i=1}^n \gamma_i \exp(-\|\mathbf{x} - \mathbf{x}_i\|^2/2\sigma^2)$, et son gradient donné dans le Tableau 1.3. En annulant ce dernier, nous aboutissons à l'expression itérative du point fixe pour le noyau Gaussien

$$\mathbf{x}^* = \frac{\sum_{i=1}^n \gamma_i \kappa_G(\mathbf{x}_i, \mathbf{x}^*) \mathbf{x}_i}{\sum_{i=1}^n \gamma_i \kappa_G(\mathbf{x}_i, \mathbf{x}^*)}. \quad (1.12)$$

Pour les noyaux projectifs de la forme (1.2), le gradient de la fonction coût est donné par

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i f^{(1)}(\langle \mathbf{x}_i, \mathbf{x} \rangle) \mathbf{x}_i + f^{(1)}(\langle \mathbf{x}, \mathbf{x} \rangle) \mathbf{x}.$$

À l'optimum \mathbf{x}^* , l'expression du point fixe est définie par :

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

Nous évaluons, pour un exemple de cette classe qui est le noyau polynomial de degré q [KT03], la fonction coût et son gradient afin de calculer l'expression itérative pour ce noyau, avec

$$\mathbf{x}^* = \frac{\sum_{i=1}^n \gamma_i \kappa_{q-1}(\mathbf{x}_i, \mathbf{x}^*) \mathbf{x}_i}{\kappa_{q-1}(\mathbf{x}^*, \mathbf{x}^*)}. \quad (1.13)$$

Un calcul détaillé est présenté dans l'annexe A, pour dériver ces expressions du gradient des fonctions radiales et projectives.

L'implémentation de cette technique nécessite le choix du point initial pour le calcul, et un critère d'arrêt. Les résultats varient largement tenant compte des différents points de départ. De même, la technique du point-fixe peut être instable et aboutir à des minima locaux, et parfois peut ne pas converger. On parle d'instabilité numérique lorsque le dénominateur est proche de zéro. Une technique est proposée pour entraver les problèmes d'instabilité en utilisant une solution régularisée, à l'instar de [AH09] donnée dans la section 1.7.6. Les problèmes liés à la méthode itérative du point-fixe sont probablement dus à l'absence d'un paramètre de pas d'adaptation pour permettre un contrôle de la convergence de l'algorithme.

La méthode itérative du point-fixe présente un point positif par rapport à la méthode du gradient. La pré-image obtenue réside dans l'espace des solutions admissibles écrite sous la forme de $\mathbf{x}^* = \sum_i \beta_i \mathbf{x}_i$ pour des coefficients $\beta_1, \beta_2, \dots, \beta_n$ à déterminer. Cette propriété est démontrée par le théorème 1.3. Alors, la plage de recherche des pré-images est contrôlée, à l'opposition de la descente du gradient qui décrit tout l'espace. Passons maintenant aux méthodes basées sur des données d'apprentissage et leurs images transformées.

1.7.3 Apprentissage de la carte de pré-image

Pour évaluer la carte de pré-image, une machine d'apprentissage est construite, avec des couples d'éléments d'entraînement de l'espace fonctionnel et des couples d'éléments dans l'espace des observations, comme suit : nous cherchons à estimer une fonction $\Gamma: \mathcal{H} \mapsto \mathcal{X}$ ayant la propriété, $\Gamma(\Phi(\mathbf{x}_i)) = \mathbf{x}_i$ pour tout $i = 1, 2, \dots, n$. Donc, idéalement, $\Gamma(\psi)$ donne \mathbf{x}^* , la pré-image de ψ . Afin de réaliser cette proposition, deux astuces sont considérés dans [BWS04, Bak05] et récemment dans [BSW07].

Lorsque nous devions utiliser la régression en utilisant les noyaux correspondant à \mathcal{H} , nous recherchons simplement des vecteurs de coefficients dans \mathcal{H} afin d'évaluer Γ . À cette fin, nous utilisons l'astuce du noyau. Un moyen de faire ce calcul est de choisir un repère orthogonal souvent obtenu par la méthode de l'ACP-à-noyaux défini par les axes $\psi_1 \psi_2 \dots \psi_k$. Une fois ce repère défini, chaque $\psi \in \mathcal{H}$ est écrit dans ce repère, selon $[\langle \psi, \psi_1 \rangle \langle \psi, \psi_2 \rangle \dots \langle \psi, \psi_k \rangle]^T$. Ensuite, la carte de pré-image Γ est décomposée suivant $\dim\{\mathcal{X}\}$ afin d'estimer les composantes de \mathbf{x}^* . A partir de ces considérations, nous pouvons écrire Γ sous forme de $\Gamma_1, \Gamma_2, \dots, \Gamma_{\dim\{\mathcal{X}\}}$. Ces fonctions peuvent être obtenues par la résolution du problème d'optimisation

$$\Gamma_m = \arg \min_{\Gamma} \frac{1}{n} \sum_{i=1}^n |[\mathbf{x}_i]_m - \Gamma(\psi)|^2 + \eta \|\Gamma\|^2.$$

La solution à ce problème d'optimisation est donnée par une technique d'inversion de matrice. Cette technique d'apprentissage est de plus en plus étudiée dans la littérature en tenant compte de l'information sur le voisinage [ZL06], et une régularisation avec un apprentissage pénalisé [ZLY10] donnés dans la section 1.7.6.

Ces techniques sont basées sur un ensemble de données dans l'espace des observations et leurs images dans le RKHS. D'autres techniques se basant sur une étude des distances dans chaque espace sont présentées.

1.7.4 Méthode de l'échelle multidimensionnelle

Comme décrit avant, le problème de pré-image cherche à trouver des observations dans l'espace des observations basées sur leurs images dans le RKHS. On parle alors de problème de réduction de dimension des objets appartenant à un espace de haute dimension. Ce problème est souvent traité sous le nom de l'échelle multidimensionnelle ou en anglais *Multi-Dimensional Scaling* (MDS) [CCC00]. Le

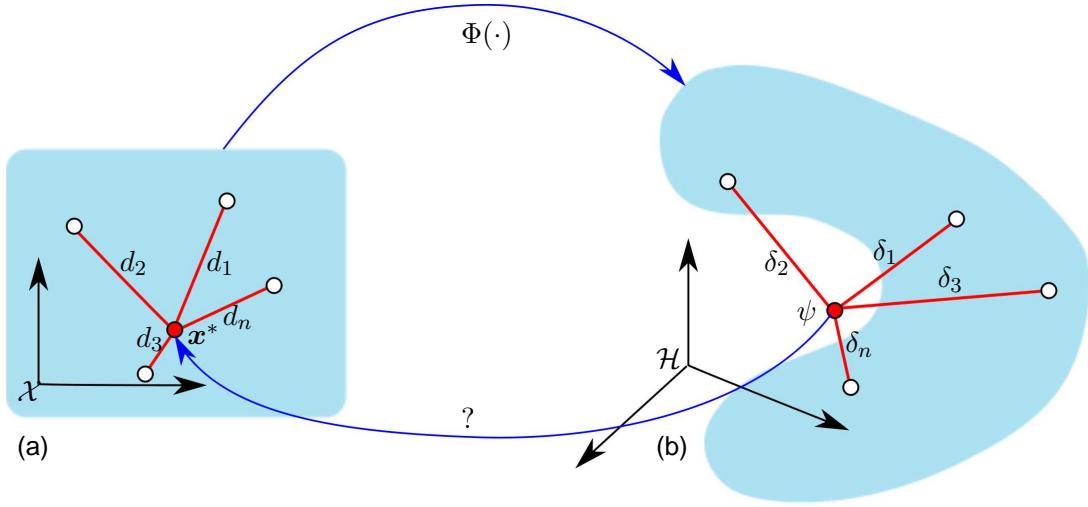


FIGURE 1.3: Schéma illustrant la technique de l'échelle multidimensionnelle, pour laquelle chaque pré-image est identifiée par les paires de distances dans l'espace des observations (a) et l'espace fonctionnel (b).

lien entre la technique MDS et l'ACP-à-noyaux est étudié [Wil01]. Cette technique intègre des données dans un espace de dimension réduite en conservant les distances de chaque couple. Cette approche est utilisée pour résoudre le problème de la pré-image dans [KT03, KT04]. Considérons chacune des distances dans le RKHS $\delta_i = \|\psi - \Phi(\mathbf{x}_i)\|_{\mathcal{H}}$ et son équivalent dans l'espace des observations $\|\mathbf{x}^* - \mathbf{x}_i\|$. Idéalement, ces distances sont conservées, à savoir

$$\|\mathbf{x}^* - \mathbf{x}_i\|^2 = \|\psi - \Phi(\mathbf{x}_i)\|_{\mathcal{H}}^2,$$

pour tout $i = 1, 2, \dots, n$. La Figure 1.3 illustre la méthode de l'échelle multidimensionnelle.

Pour résoudre ce problème, on minimise l'erreur quadratique moyenne entre ces distances, avec

$$\mathbf{x}^* = \arg \min_{\mathbf{x} \in \mathcal{X}} \sum_{i=1}^n \left| \|\mathbf{x} - \mathbf{x}_i\|^2 - \|\psi - \Phi(\mathbf{x}_i)\|_{\mathcal{H}}^2 \right|^2.$$

Une technique itérative du point-fixe est proposée en mettant à zéro le gradient de l'expression au-dessus, aboutissant alors à l'expression

$$\mathbf{x}^* = \frac{\sum_{i=1}^n (\|\mathbf{x}^* - \mathbf{x}_i\|^2 - \delta_i^2) \mathbf{x}_i}{\sum_{i=1}^n (\|\mathbf{x}^* - \mathbf{x}_i\|^2 - \delta_i^2)}.$$

Pour réduire le calcul, un faible nombre de voisins est pris en compte comme dans le cas de la technique de l'intégration localement linéaire (en anglais *locally linear embedding*) en réduction de dimension [RS00]. Cette technique a ouvert la porte à d'autres méthodes en réduction de dimension [ESK07].

1.7.5 Approche conforme

En partant de la technique de conservation de distance (MDS), une autre méthode est proposée se basant sur la préservation de mesure du produit scalaire [HR09, HR10]. Dans ce cas, la mesure angulaire est préservée, puisque $\mathbf{x}_i^\top \mathbf{x}_j / \|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|$ définit le cosinus de l'angle entre \mathbf{x}_i et \mathbf{x}_j dans l'espace Euclidien. De ce fait, cette technique est dite l'approche conforme. Pour ce faire, un système de coordonnées est construit dans le RKHS en isométrie avec celui de l'espace des observations.

Soient Ψ_1, \dots, Ψ_n les n fonctions définissant le système de coordonnées dans le RKHS. Selon le théorème de Représentation 1.3.2, chacune des n fonctions de coordonnées s'écrit sous la forme d'une combinaison linéaire des images, à savoir, $\Psi_l = \sum_{i=1}^n \theta_{l,i} \Phi(\mathbf{x}_i)$, pour $l = 1, 2, \dots, n$, où les paramètres $\theta_{l,i}$ sont à déterminer. Alors, les coordonnées de chaque élément de RKHS est obtenu en les projetant sur ces fonctions, c'est-à-dire, chaque $\Phi(\mathbf{x}_i)$ peut être représenté suivant les n fonctions de coordonnées comme suit : $\Psi_{\mathbf{x}_i} = [\langle \Psi_1, \Phi(\mathbf{x}_i) \rangle \langle \Psi_2, \Phi(\mathbf{x}_i) \rangle \dots \langle \Psi_n, \Phi(\mathbf{x}_i) \rangle]^\top$. Idéalement, le produit scalaire entre les observations dans l'espace Euclidien et entre leur image dans RKHS est conservé, plus précisément

$$\Psi_{\mathbf{x}_i}^\top \Psi_{\mathbf{x}_j} = \mathbf{x}_i^\top \mathbf{x}_j,$$

pour tout $i, j = 1, 2, \dots, n$. Une méthode pour résoudre ce problème est de minimiser l'erreur quadratique entre toutes les paires possibles, selon

$$\min_{\Psi_1, \dots, \Psi_n} \sum_{i,j=1}^n |\mathbf{x}_i^\top \mathbf{x}_j - \Psi_{\mathbf{x}_i}^\top \Psi_{\mathbf{x}_j}|^2 + \eta \sum_{l=1}^n \|\Psi_l\|_{\mathcal{H}}^2,$$

où le second membre de l'expression est un terme de régularisation. En écrivant l'expression ci-dessus pour tous les i, j allant de 1 à n sous forme matricielle et en tenant compte de la conservation du produit scalaire dans les deux espaces, nous obtenons une expression simplifiée $\mathbf{X}^\top \mathbf{x}^* = (\mathbf{X}^\top \mathbf{X} - \eta \mathbf{K}^{-1}) \boldsymbol{\alpha}$, ayant comme solution

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

avec $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_n]$.

De plus, nous pouvons identifier les pré-images d'un ensemble d'élément de RKHS en utilisant cette technique, puisque les termes entre parenthèses sont calculés une seule fois. Cette approche est prise comme un achèvement des matrices (en anglais *matrix-completion*) décrit dans [YV07]. Cet achèvement porte sur un produit matriciel autre que la matrice de Gram, c'est la matrice des valeurs de noyau. Nous passons maintenant à définir un théorème proposé pour résoudre le problème de la pré-image, en l'écrivant sous forme d'une combinaison linéaire des données présentes.

1.7.6 Pré-image régularisée ou pénalisée

Afin de fournir une estimation plus stable de la pré-image [AH09], la fonction coût est régularisée en ajoutant un terme définissant la distance dans l'espace des observations, à savoir $\|\mathbf{x} - \mathbf{x}_0\|^2$. Par suite, la fonction coût est donnée par

$$J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i \kappa(\mathbf{x}_i, \mathbf{x}) + \frac{1}{2} \kappa(\mathbf{x}, \mathbf{x}) + \lambda \|\mathbf{x} - \mathbf{x}_0\|^2,$$

où \mathbf{x}_0 est l'échantillon bruité et λ est un paramètre non-négatif de régularisation non-négative. L'idée principale de trouver la pré-image \mathbf{x}^* est qu'elle soit la plus proche possible de \mathbf{x} .

Une méthodologie est proposée dans [ZLY10] en ajoutant deux types de pénalisations. Tout d'abord, pour assurer l'apprentissage d'une pré-image bien définie, pour laquelle chaque donnée existe dans l'espace des solutions admissibles, une contrainte de convexité est imposée pour l'apprentissage des coefficients de la combinaison. Ensuite, une fonction pénalisée est intégrée comme étant une partie de la fonction d'optimisation lors du processus d'apprentissage de la pré-image. La fonction coût à minimiser est essentiellement de la forme

$$J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i \kappa(\mathbf{x}_i, \mathbf{x}) + \frac{1}{2} \kappa(\mathbf{x}, \mathbf{x}) + \lambda_F F(\mathbf{x}),$$

où $F(\mathbf{x})$ est la fonction de pénalisation. La fonction coût est sujette aux conditions que les coefficients définissant la combinaison linéaire sont tous positifs et de somme unité. Donc $\mathbf{x}^* = \sum_i \beta_i \mathbf{x}_i$ avec $\beta_i > 0$ et $\sum_{i=1}^n \beta_i = 1$.

1.8 Formulation de la pré-image

En prenant toutes les méthodes de résolution du problème de la pré-image, nous trouvons que chacune des solutions proposées correspond à une combinaison linéaire des observations d'apprentissage. Le théorème suivant résume cette propriété pour les noyaux projectifs et radiaux.

Théorème 1.3. (*Modélisation linéaire d'une pré-image*) *La pré-image \mathbf{x}^* est donnée par une combinaison linéaire des données disponibles, sous la forme*

$$\mathbf{x}^* = \sum_{i=1}^n \beta_i^* \mathbf{x}_i. \tag{1.14}$$

Démonstration. Afin de prouver ce théorème, nous considérons les deux classes de noyaux : projectifs et radiaux (voir Tableau 1.1). Utilisant l'expression du gradient (3.7) à l'optimum, nous avons $\nabla_{\mathbf{x}} J(\mathbf{x}^*) = 0$,

à savoir

$$\sum_{i=1}^n \gamma_i \nabla_{\mathbf{x}^*} \kappa(\mathbf{x}_i, \mathbf{x}^*) = \frac{1}{2} \nabla_{\mathbf{x}^*} \kappa(\mathbf{x}^*, \mathbf{x}^*). \quad (1.15)$$

Commençons par les noyaux projectifs, de la forme (1.2). Alors, le premier membre de l'équation s'écrit comme

$$\sum_{i=1}^n \gamma_i \nabla_{\mathbf{x}^*} \kappa(\mathbf{x}_i, \mathbf{x}^*) = \sum_{i=1}^n \gamma_i \nabla_{\mathbf{x}^*} f(\langle \mathbf{x}_i, \mathbf{x}^* \rangle) \mathbf{x}_i,$$

et le second membre est exprimé par

$$\frac{1}{2} \nabla_{\mathbf{x}^*} \kappa(\mathbf{x}^*, \mathbf{x}^*) = \frac{1}{2} \nabla_{\mathbf{x}^*} f(\langle \mathbf{x}^*, \mathbf{x}^* \rangle) 2\mathbf{x}^*.$$

En combinant ces deux expressions (voir l'annexe A pour plus de détail), l'équation (1.15) devient

$$\mathbf{x}^* = \sum_{i=1}^n \gamma_i \frac{f^{(1)}(\langle \mathbf{x}_i, \mathbf{x}^* \rangle)}{f^{(1)}(\langle \mathbf{x}^*, \mathbf{x}^* \rangle)} \mathbf{x}_i, \quad (1.16)$$

de la forme $\mathbf{x}^* = \sum_{i=1}^n \beta_i^* \mathbf{x}_i$.

Nous passons maintenant aux noyaux radiaux, de la forme (1.3). Dans ce cas, le terme $\nabla_{\mathbf{x}} \kappa(\mathbf{x}, \mathbf{x})$ s'annule. Le gradient à l'optimum s'écrit sous la forme

$$\sum_{i=1}^n \gamma_i \nabla_{\mathbf{x}^*} \kappa(\mathbf{x}_i, \mathbf{x}^*) = 0,$$

avec le premier terme donné par

$$\sum_{i=1}^n \gamma_i \nabla_{\mathbf{x}^*} \kappa(\mathbf{x}_i, \mathbf{x}^*) = \sum_{i=1}^n \gamma_i \nabla_{\mathbf{x}^*} g(\|\mathbf{x}_i - \mathbf{x}^*\|^2) 2(\mathbf{x}^* - \mathbf{x}_i).$$

Le résultat final de (1.15) est alors écrit (voir l'annexe A pour plus de détail)

$$\mathbf{x}^* = \sum_{i=1}^n \gamma_i \frac{g^{(1)}(\|\mathbf{x}_i - \mathbf{x}^*\|^2)}{\sum_{j=1}^n \gamma_j g^{(1)}(\|\mathbf{x}_j - \mathbf{x}^*\|^2)} \mathbf{x}_i, \quad (1.17)$$

de nouveau de la forme $\mathbf{x}^* = \sum_{i=1}^n \beta_i^* \mathbf{x}_i$. □

Nous pouvons estimer les coefficients β_i^* au lieu de la pré-image. Nous reprenons alors les différentes méthodes de résolution du problème de la pré-image dans ce chapitre, et les réécrivons comme des problèmes d'estimation des β_i^* . Dans le chapitre suivant, nous utilisons cette formulation afin d'imposer des contraintes sur la pré-image.

1.9 Conclusion

Dans ce chapitre, nous avons introduit les méthodes à noyaux. Tout d'abord, nous avons décrit la caractérisation des noyaux. Dans un second temps, nous avons détaillé l'astuce du noyau et le théorème du représentant qui sont les deux éléments fondamentaux pour le passage du cas linéaire au non-linéaire. Ensuite, l'analyse en composantes principales (ACP) classique est introduite, puis elle est étendue à l'ACP-à-noyaux par le biais des méthodes à noyaux. Ces méthodes à noyaux assurent un passage de l'espace des observations à un espace RKHS de haute dimension permettant ainsi diverses applications comme par exemple la classification. Cependant, pour différents domaines d'applications tels que l'extraction des caractéristiques et la reconnaissance des formes, l'utilisateur est plutôt à la recherche de la forme ou caractéristique en question. Il s'avère alors nécessaire de faire le retour inverse de l'espace RKHS à l'espace des observations, où la forme ou caractéristique est définie, ce problème est dit de pré-image. Nous avons alors décrit ce problème tout en présentant les différentes techniques présentes dans la littérature pour sa résolution, dont certaines sont itératives et d'autres ne le sont pas. Finalement, nous avons proposé une nouvelle formulation de la pré-image.

Le chapitre suivant tient en compte le problème de la pré-image sous contraintes de non-négativité, souvent imposées par la physiologie des données. D'une part, des contraintes sur les données elles-mêmes sont considérées. De l'autre part, des contraintes sur l'additivité des contributions sont étudiées, provoquant une certaine parcimonie dans les résultats.

Le problème de pré-image avec contraintes de non-négativité

Sommaire

2.1 Introduction	35
2.2 Méthodes à noyaux, pré-image et non-négativité	36
2.3 Pré-image avec contraintes de non-négativité	37
2.3.1 Contraintes de non-négativité sur la pré-image	38
2.3.2 Contraintes de non-négativité sur les coefficients du modèle	41
2.4 Expérimentations	43
2.4.1 Extraction des caractéristiques de signaux ERP	43
2.4.2 Débruitage de données et des images	46
2.5 Conclusion	52

2.1 Introduction

Il s'avère que la contrainte de non-négativité est très essentielle dans de nombreux problèmes d'optimisation. Cette propriété incorpore l'équivalence mathématique entre la non-négative et la non-positive. Seules les méthodes itératives peuvent être utilisées pour résoudre les problèmes généraux d'optimisation sous de telles contraintes. En outre, un schéma itératif pour la non-négativité peut servir de base pour des problèmes d'optimisation plus complexes sous contrainte, tels que l'optimisation de contraintes bornées. Depuis les années quatre-vingt, des contraintes de non-négativité ont été imposées pour la déconvolution d'un signal par Thomas dans [Tho83] et Prost *et al.* dans [PG84], et étendues à la déconvolution et le débruitage des images ont été étudiés respectivement par Thomas *et al.* dans [TS91] et Snyder *et al.* dans [SSO92]. Durant la dernière décennie, une méthode plus générale pour l'optimisation itérative sous contraintes de non-négativité a été étudiée, initiée par Lantéri *et al.* [LRCA01], et plus récemment pour l'apprentissage en ligne dans [CRH⁺10b], l'identification des systèmes [CRH⁺10a] et la régression distribuée [CRHB10].

La propriété de la non-négativité est nécessaire dans plusieurs domaines. L'analyse en composantes indépendantes impose une factorisation non-négative des données dans [HO00], comme pour la séparation aveugle des sources avec des sources “positives”. Dans [OP03], une analyse non-négative en composantes principales (ACP) est proposée. Différentes études pour la reconnaissance des formes sous contraintes ont été basées sur des algorithmes linéaires, tels que l'ACP pour le diagnostic du cancer dans [Han10], la parcimonie non-négative dans [ZS07], la recherche de solutions optimales en appliquant la procédure par séparation-évaluation dans [MWA06], et la recherche du vecteur propre dominant en utilisant l'algorithme de l'espérance-maximisation dans [SB08].

Lors de la reconnaissance des formes ou le débruitage, nous sommes souvent à la recherche de formes, ou de représentations dans l'espace des observations où elles sont décrites. De plus, inspirés par la physiologie qui nécessite souvent une contrainte de non-négativité, nous élaborons alors la résolution du problème de la pré-image sous contrainte de non-négativité. Pour ce faire, l'étude est décomposée en deux parties. Dans un premier temps, elle porte sur les conditions de non-négativité sur la pré-image estimée. Et dans un second temps, nous tenons compte de l'additivité des contributions ce qui mène à la non-négativité des coefficients qui les définissent. En couplant ces développements avec l'analyse en composantes principales à noyaux, nous étudions l'extraction de caractéristiques pour des signaux potentiels évoqués, et le débruitage des images.

2.2 Méthodes à noyaux, pré-image et non-négativité

Dans cette section, nous combinons le problème de la pré-image avec la condition de non-négativité en étudiant les conditions pour lesquelles nous aboutissons à des résultats non-négatifs. En utilisant le Théorème 1.3, nous pouvons établir un lien entre les coefficients dans les deux espaces des observations et de caractéristiques.

Lemme 2.1. *Pour des données d'apprentissages non-négatives, si les coefficients dans l'espace fonctionnel sont non-négatifs, $\gamma_1, \gamma_2, \dots, \gamma_n \geq 0$, alors les coefficients de la pré-image correspondante sont aussi non-négatifs, i.e., $\beta_1^*, \beta_2^*, \dots, \beta_n^* \geq 0$. Par ailleurs, la non-négativité des données n'est pas nécessaire pour les noyaux radiaux.*

Démonstration. Pour les noyaux projectifs, les coefficients de la pré-image ont la forme (1.16)

$$\beta_i^* = \gamma_i \frac{f^{(1)}(\langle \mathbf{x}_i, \mathbf{x}^* \rangle)}{f^{(1)}(\langle \mathbf{x}^*, \mathbf{x}^* \rangle)}.$$

Lorsque toutes les données d'apprentissages sont non-négatives, les dérivées ci-dessus sont non-négatives suite à la Proposition 1.2. La même preuve peut être appliquée à l'équation (1.17) pour les noyaux radiaux vérifiant la Proposition 1.1, avec

$$\beta_i^* = \gamma_i \frac{g^{(1)}(\|\mathbf{x}_i - \mathbf{x}^*\|^2)}{\sum_{j=1}^n \gamma_j g^{(1)}(\|\mathbf{x}_j - \mathbf{x}^*\|^2)}.$$

□

La non-négativité des coefficients γ_i est une condition imposée par les machines à vecteurs supports (SVM) pour la classification et la régression, ainsi que d'autres méthodes d'apprentissage. Toutefois, ce n'est pas le cas en général, avec l'ACP-à-noyaux par exemple. Nous ne nous limiterons pas au problème convexe, mais considérons le problème non convexe plus général.

Pour le noyau Gaussien, nous avons

$$\kappa_G(\mathbf{x}_i, \mathbf{x}_j) = g(\|\mathbf{x}_i - \mathbf{x}_j\|^2) = \exp\left(\frac{-1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$$

alors la dérivée première du noyau Gaussien s'écrit comme suit

$$g^{(1)}(\|\mathbf{x}_i - \mathbf{x}_j\|^2) = -\frac{1}{2\sigma^2} \kappa_G(\mathbf{x}_i, \mathbf{x}_j),$$

et la dérivée seconde

$$g^{(2)}(\|\mathbf{x}_i - \mathbf{x}_j\|^2) = \frac{1}{4\sigma^4} \kappa_G(\mathbf{x}_i, \mathbf{x}_j).$$

Lorsque le noyau polynomial est appliqué, avec

$$\kappa_q(\mathbf{x}_i, \mathbf{x}_j) = f(\mathbf{x}_i \cdot \mathbf{x}_j) = (c + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^q,$$

alors la dérivée première du noyau polynomial s'écrit comme suit

$$f^{(1)}(\langle \mathbf{x}_i, \mathbf{x}_j \rangle) = q \kappa_{q-1}(\mathbf{x}_i, \mathbf{x}_j),$$

où $\kappa_{q-1}(\mathbf{x}_i, \mathbf{x}_j) = f(\langle \mathbf{x}_i, \mathbf{x}_j \rangle) = (c + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^{q-1}$.

2.3 Pré-image avec contraintes de non-négativité

En reconnaissance des formes, nous cherchons parfois des solutions avec contraintes. Il s'agit souvent des contraintes de non-négativité. Par exemple, en traitement d'images, les données d'apprentissage sont des images ou des pixels dans une image, comme les données qui sont non-négatives si les images sont codées en niveau de gris. Afin d'extraire une caractéristique ou aboutir à une version débruitée du même type (même espace des observations avec une condition de non-négativité de chaque pixel), nous imposons une contrainte de non-négativité sur la pré-image. Cependant, les contraintes peuvent être appliquées soit sur les données elles-mêmes, soit sur les coefficients du modèle en utilisant la combinaison linéaire définie dans (1.14).

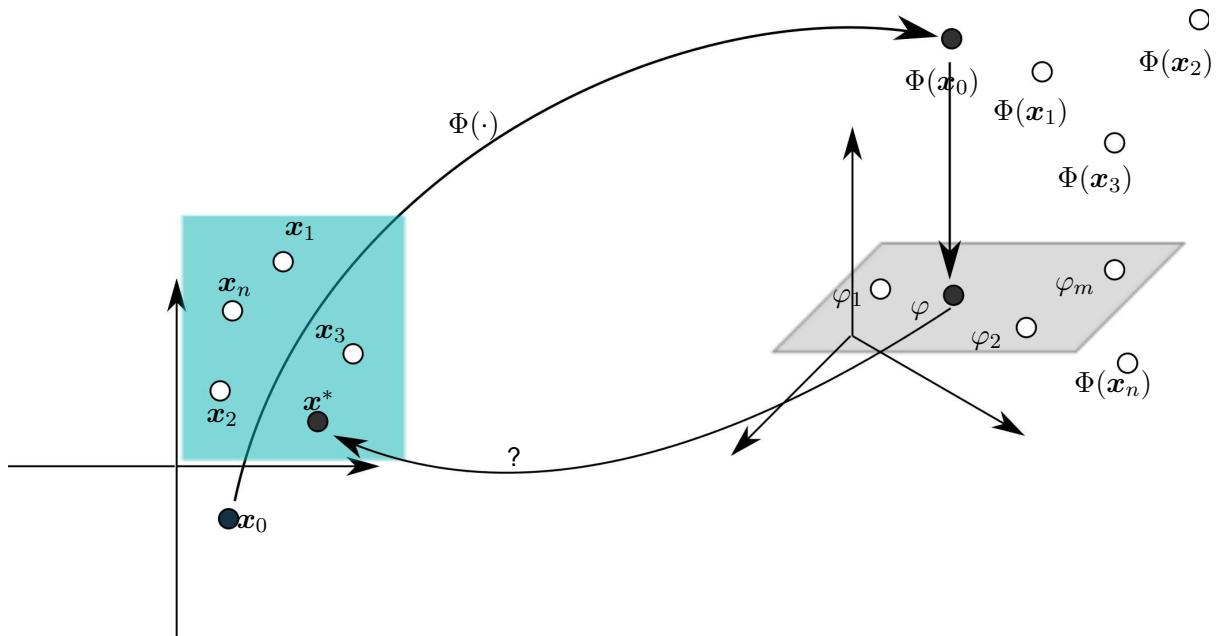


FIGURE 2.1: Schéma illustrant le problème de la pré-image sous contrainte de non-négativité. Étant donné une observation bruitée x_0 , elle est transformée en $\Phi(x_0)$ par la fonction non-linéaire $\Phi(\cdot)$, ensuite projetée sur le sous-espace engendré par les axes principaux les plus pertinents $\varphi_1, \varphi_2, \dots, \varphi_m$. Une fois que la forme débruitée φ est estimée, il est nécessaire de faire le retour inverse vers l'espace des observations, afin de retrouver la pré-image de φ à savoir x^* , où le domaine admissible des résultats est donné par la non-négativité dans l'espace des observations.

2.3.1 Contraintes de non-négativité sur la pré-image

Dans cette section, nous considérons le problème général de résolution du problème de la pré-image avec contrainte de non-négativité. Nous étudions le problème d'optimisation sous contrainte, en utilisant la fonction coût $J(\cdot)$ définie par (1.10), alors notre problème est décrit par

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} J(\mathbf{x}) \quad \text{sous contrainte } \mathbf{x} \geq 0, \quad (2.1)$$

où l'expression $\mathbf{x} \geq 0$ désigne la non-négativité de toutes les composantes du vecteur \mathbf{x} . Le gradient de la fonction coût est donné dans le Tableau 1.3 pour différents types de noyaux. Ensuite, nous dérivons une règle de mise-à-jour itérative qui mène à la non-négativité de la pré-image. L'idée de cette pré-image avec contrainte de non-négativité est illustrée par la Figure 2.1.

Nous considérons le Lagrangien associé à ce problème d'optimisation avec contraintes donné par (2.1). Le Lagrangien de ce problème n'est autre que

$$J(\mathbf{x}) - \boldsymbol{\mu}^\top \mathbf{x},$$

où $\boldsymbol{\mu}$ représente le vecteur des multiplicateurs de Lagrange, tous non-négatifs. Pour la solution optimale

\boldsymbol{x}^* , il lui correspond un vecteur optimal des multiplicateurs de Lagrange soit $\boldsymbol{\mu}^*$. Les conditions du premier ordre de (Karush-)Kuhn-Tucker à l'optimum se traduisent par

$$\begin{aligned}\nabla_{\boldsymbol{x}}[J(\boldsymbol{x}^*) - \boldsymbol{\mu}^{*T} \boldsymbol{x}^*] &= 0 \\ [\boldsymbol{\mu}^*]_i [\boldsymbol{x}^*]_i &= 0 \quad \text{pour tout } i = 1, 2, \dots, \dim\{\mathcal{X}\}\end{aligned}$$

où $[\cdot]_i$ représente la $i^{\text{ème}}$ composante. Nous pouvons facilement voir que la première condition s'écrit sous la forme $\nabla_{\boldsymbol{x}}[J(\boldsymbol{x}^*)]_i - [\boldsymbol{\mu}^*]_i = 0$ pour tout i . La combinaison de toutes ces conditions d'égalité nous donne, pour chaque composante $i = 1, 2, \dots$, une contrainte active pour $[\boldsymbol{x}^*]_i = 0$ ou une contrainte inactive pour $[\nabla_{\boldsymbol{x}} J(\boldsymbol{x}^*)]_i = 0$ avec $[\boldsymbol{x}^*]_i > 0$. Nous proposons de résoudre ce problème avec une méthode itérative en s'inspirant de [LRCA01]. Ainsi, l'expression de mise-à-jour à l'itération $t + 1$ est-elle donnée par

$$[\boldsymbol{x}(t+1)]_i = [\boldsymbol{x}(t)]_i + \eta_i(t) p([\boldsymbol{x}(t)]_i) [-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i,$$

où le signe moins montre une technique de descente du gradient, $p(\boldsymbol{x}(t))$ est une fonction positive sur le domaine admissible du vecteur \boldsymbol{x} , et $\eta_i(t)$ représente un facteur de relaxation. Nous proposons dans la suite deux techniques en variant l'expression de la fonction positive $p([\boldsymbol{x}(t)]_i)$, d'une part pour donner un pas fixe et d'autre part pour donner un pas variable dépendant du vecteur \boldsymbol{x} lui-même.

2.3.1.1 Pas fixe

Dans ce paragraphe, nous fixons la valeur de la fonction positive $p([\boldsymbol{x}(t)]_i)$ à l'unité. L'expression ci-dessus est donnée par

$$[\boldsymbol{x}(t+1)]_i = [\boldsymbol{x}(t)]_i + \eta_i(t) [-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i.$$

Le pas d'adaptation $\eta_i(t)$ est utilisé afin de contrôler la convergence. Ce pas d'adaptation $\eta_i(t)$ doit satisfaire une condition pour assurer cette non-négativité de toutes les composantes $[\boldsymbol{x}(t+1)]_i$ du vecteur $\boldsymbol{x}(t+1)$. Deux cas sont alors distingués : Si le gradient ci-dessus est inférieur ou égal à zéro, c'est-à-dire $[\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i \leq 0$, aucune restriction n'est appliquée sur la valeur que peut prendre le pas ; cependant, si ce gradient est supérieur à zéro, i.e. $[\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i > 0$, alors le pas d'adaptation doit être borné, selon

$$\eta_i(t) \leq \frac{[\boldsymbol{x}(t)]_i}{[\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i}.$$

Même en utilisant une valeur du pas pour chacune des directions de descente du gradient, il est souvent intéressant d'avoir une valeur unique pour ce pas pour chaque itération t . Nous définissons alors le pas d'adaptation $\eta(t)$ par

$$\eta(t) \leq \min_i \frac{[\boldsymbol{x}(t)]_i}{[\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i}. \quad (2.2)$$

À partir de ces expressions, la règle de mise-à-jour sous forme matricielle est alors écrite

$$\boldsymbol{x}(t+1) = \boldsymbol{x}(t) - \eta(t) \nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t)), \quad (2.3)$$

où le pas est adapté comme présenté ci-dessus. Passons maintenant à une autre forme de la fonction positive $p([\boldsymbol{x}(t)]_i)$.

2.3.1.2 Pas variable

Inspirée par les travaux de Lantéri *et al.* [LTBM09], nous proposons une approche qui permet une convergence plus rapide vers les valeurs nulles. À cette fin, nous remplaçons la fonction positive $p([\boldsymbol{x}(t)]_i)$ par le vecteur \boldsymbol{x} en question en tenant compte de chacune des composantes $[\boldsymbol{x}(t)]_i$. Alors, le pas d'adaptation est multiplié par la valeur de $[\boldsymbol{x}(t)]_i$ correspondante, résultant en une rapidité de convergence pour aboutir au zéro en question. Nous aboutissons alors à l'expression

$$[\boldsymbol{x}(t+1)]_i = [\boldsymbol{x}(t)]_i + \eta_i(t) [\boldsymbol{x}(t)]_i [-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i.$$

Notons que la non-négativité des composantes $[\boldsymbol{x}(t+1)]_i$ impose une condition sur le pas $\eta_i(t)$. En ré-écrivant l'expression de mise à jour de $[\boldsymbol{x}(t+1)]_i$ nous obtenons la factorisation suivante

$$[\boldsymbol{x}(t+1)]_i = [\boldsymbol{x}(t)]_i (1 + \eta_i(t) [-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i).$$

Afin de conserver la non-négativité à chaque itération, une condition apparaît sur le terme entre parenthèses, *i.e.*, $1 + \eta_i(t) [-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i$. Lorsque le gradient de la fonction coût est négatif, aucune restriction n'est nécessaire sur le pas. Cependant, lorsque ce gradient est positif, la valeur du pas est maximisée par

$$\eta_i(t) \leq \frac{1}{[-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i}.$$

En pratique, nous pouvons utiliser un pas indépendant de la composante i , à condition qu'il soit maximisé par le minimum sur tous les i de l'inverse du gradient

$$\eta(t) \leq \min_i \frac{1}{[-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))]_i}.$$

Écrite sous une forme matricielle, la règle de mise-à-jour est définie par

$$\boldsymbol{x}(t+1) = \boldsymbol{x}(t) - \eta(t) \operatorname{diag}[\boldsymbol{x}(t)] \nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t)), \quad (2.4)$$

où $\operatorname{diag}[\cdot]$ est l'opérateur pour définir une matrice diagonale, précisément $\operatorname{diag}[\boldsymbol{x}(t)]$ est la matrice diagonale ayant comme éléments les composantes $[\boldsymbol{x}(t)]_i$ du vecteur $\boldsymbol{x}(t)$. Dans cette expression, le terme $-\operatorname{diag}[\boldsymbol{x}(t)] \nabla_{\boldsymbol{x}} J(\boldsymbol{x}(t))$ correspond à la direction de la descente du gradient. Il est clair alors

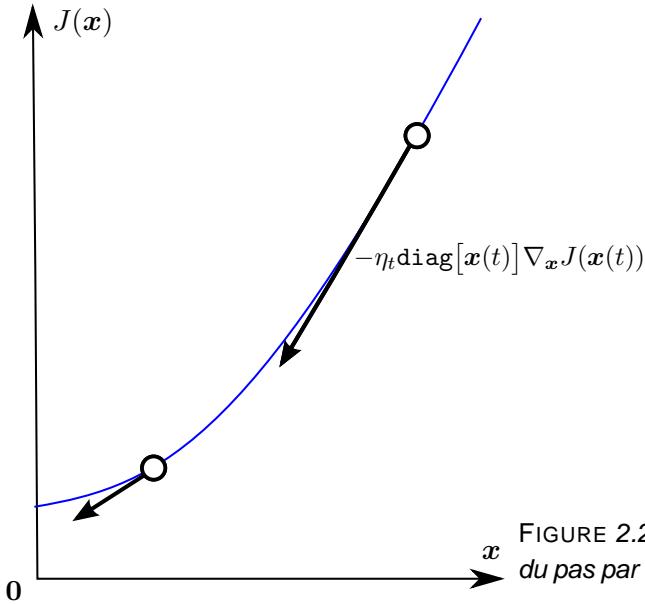


FIGURE 2.2: Illustration de la pondération du pas par la valeur de x correspondante.

que le pas pour les grandes valeurs de $[x(t)]_i$ est multiplié par une valeur plus importante que pour les faibles valeurs. Cette propriété permet alors une convergence plus rapide vers les valeurs nulles, comme illustrée dans la Figure 2.2.

2.3.2 Contraintes de non-négativité sur les coefficients du modèle

En vertu du Théorème 1.3, la pré-image s'écrit selon une combinaison linéaire des observations disponibles, utilisant la forme $\mathbf{x}^* = \sum_{i=1}^n \beta_i^* \mathbf{x}_i$, avec des paramètres β_i^* à déterminer [KHR⁺10]. Donc, nous cherchons la pré-image optimale de la forme matricielle

$$\mathbf{x}^* = \mathbf{X}^\top \boldsymbol{\beta}^*,$$

où les vecteurs des données \mathbf{x}_i sont regroupés dans la matrice $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \cdots \ \mathbf{x}_n]^\top$ et les paramètres β_i^* à déterminer dans $\boldsymbol{\beta}^* = [\beta_1^* \ \beta_2^* \ \cdots \ \beta_n^*]^\top$. Cette écriture nous permet de présenter une autre stratégie pour résoudre le problème de la pré-image, cette fois en imposant une contrainte sur les coefficients dans l'expression ci-dessus. Le problème de la pré-image sous contrainte de non-négativité des coefficients est

$$\mathbf{x}^* = \arg \min_{\mathbf{x}} J(\mathbf{x}) \quad \text{sous contrainte } \boldsymbol{\beta} \geq 0,$$

où $\mathbf{x} = \mathbf{X}^\top \boldsymbol{\beta}$.

Dans cette expression, la fonction coût (1.10) est donnée en fonction des coefficients $\boldsymbol{\beta}$. Elle est définie par

$$J(\mathbf{X}^\top \boldsymbol{\beta}) = - \sum_{i=1}^n \gamma_i \kappa(\mathbf{x}_i, \mathbf{X}^\top \boldsymbol{\beta}) + \frac{1}{2} \kappa(\mathbf{X}^\top \boldsymbol{\beta}, \mathbf{X}^\top \boldsymbol{\beta}). \quad (2.5)$$

TABLE 2.1: Gradient de la fonction coût (1.10) par rapport à β défini par $x = \mathbf{X}^\top \beta$, pour les noyaux les plus utilisés.

Type	$\nabla_\beta J(\mathbf{X}^\top \beta)$
Polynomial	$-\sum_{i=1}^n \gamma_i q \kappa_{q-1}(\mathbf{x}_i, \mathbf{X}^\top \beta) \mathbf{X} \mathbf{x}_i + q \kappa_{q-1}(\mathbf{X}^\top \beta, \mathbf{X}^\top \beta) \mathbf{X} \mathbf{X}^\top \beta$
Laplacien	$-\frac{1}{\sigma} \sum_{i=1}^n \gamma_i \kappa_L(\mathbf{x}_i, \mathbf{X}^\top \beta) \mathbf{X} \mathbf{x}_i$
Exponentiel	$-\frac{1}{\sigma} \sum_{i=1}^n \gamma_i \kappa_E(\mathbf{x}_i, \mathbf{X}^\top \beta) \mathbf{X} \mathbf{x}_i + \frac{1}{\sigma} \kappa_E(\mathbf{X}^\top \beta, \mathbf{X}^\top \beta) \mathbf{X} \mathbf{X}^\top \beta$
Gaussien	$-\frac{1}{\sigma^2} \sum_{i=1}^n \gamma_i \kappa_G(\mathbf{x}_i, \mathbf{X}^\top \beta) \mathbf{X} (\mathbf{x}_i - \mathbf{X}^\top \beta)$

Le gradient de cette expression par rapport β vaut

$$\nabla_\beta J(\mathbf{X}^\top \beta) = \mathbf{X} \nabla_x J(x). \quad (2.6)$$

Le Tableau 2.1 regroupe les expressions des gradients par rapport aux coefficients β des noyaux les plus couramment utilisés. Il est important de noter, qu'il existe une relation entre les expressions du gradient par rapport aux données x (Tableau 1.3) et le gradient par rapport aux coefficients β , et elle est définie dans l'expression (2.6).

En tenant compte de l'expression qui lie les coefficients aux données à savoir $\mathbf{x}^* = \mathbf{X}^\top \beta^*$, et en prenant l'expression de la fonction coût (2.5), nous pouvons établir une nouvelle formulation du problème d'optimisation sous contrainte de non-négativité appliquée sur les coefficients β^* . Pour ce faire, et par analogie avec le problème d'optimisation sous contrainte de non-négativité sur la pré-image (2.1), nous aboutissons alors à ce nouveau problème

$$\beta^* = \arg \min_{\beta} J(\mathbf{X}^\top \beta) \quad \text{sous contrainte } \beta \geq 0.$$

Dans cette expression la fonction coût est définie par (2.5), avec un gradient par rapport aux coefficients β donné par (2.6). Une règle de mise-à-jour est alors établie, par analogie avec (2.4), comme suit

$$\beta(t+1) = \beta(t) - \eta(t) \operatorname{diag}[\beta(t)] \mathbf{X} \nabla_x J(x). \quad (2.7)$$

Une fois les coefficients sont déterminés $\beta_1^*, \beta_2^*, \dots, \beta_n^*$, nous pouvons ainsi déterminer la pré-image avec $\mathbf{x}^* = \mathbf{X}^\top \beta^*$. De cette expression, nous pouvons voir que dans le cas de la non-négativité des données d'apprentissage, c'est-à-dire $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \geq 0$, la pré-image résultante \mathbf{x}^* le sera alors aussi.

2.3.2.1 La parcimonie

Une des propriétés utiles de forcer une contrainte sur les coefficients d'un modèle de combinaison linéaire est la propriété de la parcimonie. En fait, la solution sans contrainte peut unir des contributions additives et soustractives dans la combinaison linéaire ; une grande partie de ces contributions se neutralisent alors entre elles. En fixant des contraintes de non-négativité sur les coefficients, il s'avère que cet équilibre aboutira à un grand nombre de composants inactifs, c'est-à-dire, des coefficients proches de zéro. C'est la propriété de la parcimonie, contribuant à la généralisation des algorithmes des Supports Vecteurs [Vap98] et la littérature du *compressive sampling* [CW08]. Nous insistons sur le fait qu'il s'agit d'un effet secondaire fortuit des contraintes de non-négativité, par opposition à une fonction principale ayant pour objectif la parcimonie, où l'on contrôle le degré de parcimonie de la solution. La parcimonie signifie qu'un grand nombre des coefficients est proche de zéro, ou en d'autres termes, seulement un petit nombre de données d'apprentissage contribue à la solution finale. Cette propriété est probablement due à la redondance dans les observations. En effet, dans la solution sans contrainte, cette redondance entraîne des coefficients additifs et soustractifs permettant de neutraliser leurs contributions.

La parcimonie est une propriété très souhaitable dans la reconnaissance des formes et dans l'apprentissage, ce qui contribue à une meilleure compréhension des résultats, en bioinformatique par exemple. Il est à noter que la parcimonie n'est pas de la pré-image. De plus, l'inclusion explicite de la contrainte de parcimonie, comme la minimisation d'une fonction coût ℓ_0 ou ℓ_1 , est coûteuse en ressources informatiques. La parcimonie est étudiée avec une procédure par séparation-évaluation dans [MWA06]. La méthode de Lasso tient en compte de la parcimonie de la solution et le coût calculatoire dans [Tib96]. Ce coût calculatoire est étudié lors de l'implémentation de cette méthode Lasso dans Matlab [Sjo05], ainsi que lors de son application sur les composantes principales dans [ZHT04]. De plus, une étude concernant le coût de la parcimonie de la solution porte sur la maximisation de la variance des données dans [dEJL07]. Des études sont menées pour optimiser les modèles en imposant une pénalité de la parcimonie dans [BJMO12], d'autres études tiennent en compte une optimisation convexe dans [BJMO11]. Elle est illustrée dans la section d'expérimentations sur des ensembles de données artificielles et réelles.

2.4 Expérimentations

Dans cette section, nous illustrons l'efficacité de nos méthodes proposées, pour deux applications différentes : l'extraction des caractéristiques détaillée dans 1.5.1 et le débruitage décrit dans la section 1.5.2.

2.4.1 Extraction des caractéristiques de signaux ERP

Nous considérons l'extraction des caractéristiques avec une application sur des signaux réels, plus précisément des enregistrements mesurant l'activité du cerveau. Les potentiels évoqués ou *Event-related*

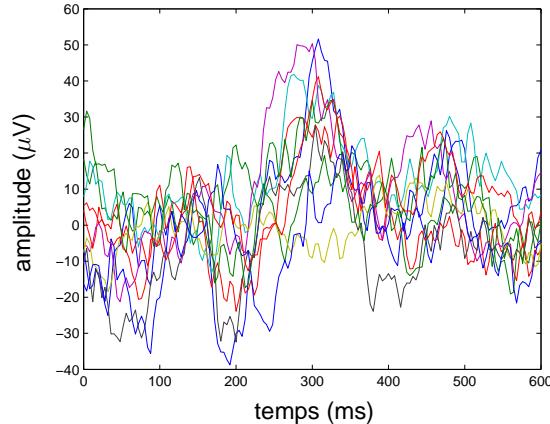


FIGURE 2.3: *Quelques signaux des potentiels évoqués, enregistrés par le canal Cz. La diversité de ces signaux est illustrée, avec quelques uns n'ayant pas de composantes positives au voisinage de 300 ms (voir par exemple —).*

potentials (ERP) en anglais représentent l'activité électrique du cerveau due à une réponse à une stimulation bien spécifique, mesurée avec l'électroencéphalographe (EEG). Il y a un fort consensus sur les composantes d'un enregistrement ERP, indépendamment des sujets ou du type de stimulation. Un tel signal comprend une déviation négative (appelée N200) suivie d'une autre positive (appelée P300), se produisant respectivement à environ 200 ms et 300 ms après le début de la stimulation. Durant l'activité cérébrale, une telle réponse unique n'est pas généralement visible dans ces enregistrements. Pour contourner ce problème, de nombreux essais sont souvent effectués en utilisant la même stimulation. En pratique, on prend la moyenne de ces réponses, ce qui donne un moment du premier ordre des enregistrements ERP. Nous donnons dans cette section une autre statistique tenant compte de la variance de ces signaux, en combinant l'ACP-à-noyaux avec le noyau Gaussien d'une part, et la technique proposée de pré-image d'autre part.

Pour les expériences, nous utilisons des signaux ERP collectés à partir d'expériences réalisées à l'université de Kuopio [oK] et étudiées dans [Tar]; pour plus d'informations, voir aussi [Geo07, Tar04]. La stimulation auditive est composée d'une série de deux signaux de tonalité en alternance, joués aléatoirement avec un temps entre les stimuli (aussi appelé intervalle inter-stimulation ou ISI) d'une seconde. Ces stimuli correspondent soit à une tonalité à la fréquence de 800 Hz ou d'une autre tonalité à 560 Hz, jouées avec un rapport de 85% pour le premier signal et 15% pour le second. Alors que les signaux ERP sont des enregistrements à partir d'un EEG à 64 canaux, seule la ligne médiane du canal central Cz, souvent très fiable, est utilisée pour la détection des potentiels. L'enregistrement capté dans le canal Cz est segmenté en signaux afin de voir la réaction du sujet aux stimulations en utilisant une fenêtre $[0, 600]$ ms, où 0 correspond à l'instance de relance de chaque stimulation. Une telle fenêtre est appropriée pour extraire les deux composantes N200 et P300 de l'ERP. Un ensemble de 87 signaux de longueur 600 ms est recueilli, avec 151 échantillons chacun, comme illustré dans la Figure 2.3 où seulement dix signaux choisis au hasard sont présentés pour afficher la diversité de ces signaux.

Nous appliquons l'ACP-à-noyaux pour extraire le premier axe principal de ces données, dans l'espace fonctionnel associé au noyau Gaussien. L'approche de la pré-image nous a permis de revenir à l'espace

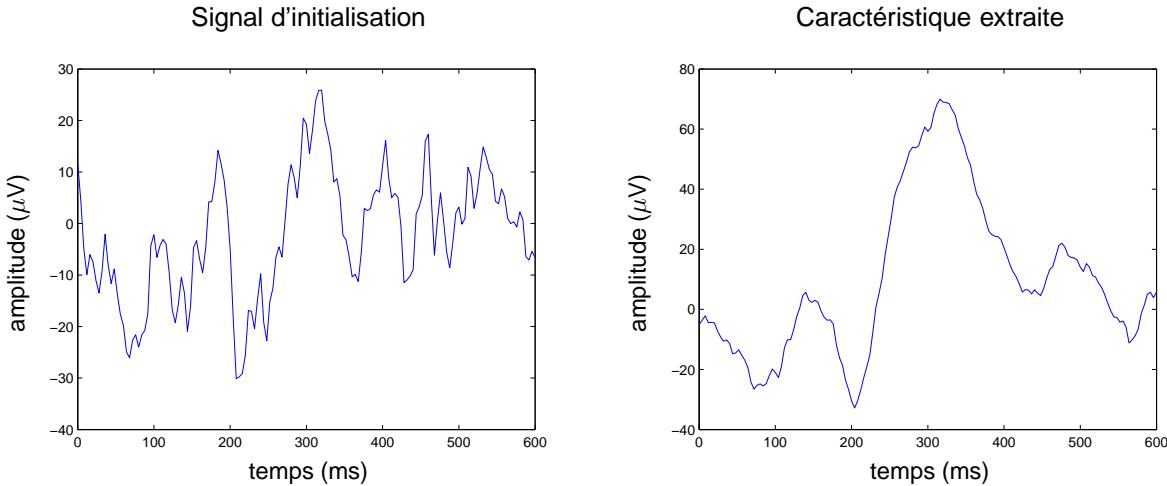


FIGURE 2.4: Extraction de caractéristiques des données des potentiels évoqués, avec l'algorithme initialisé au signal initial (figure à gauche). En évaluant la pré-image du premier axe principal de l'ACP-à-noyaux, nous obtenons la caractéristique extraite (figure à droite).

des observations, qui est, l'espace des signaux étudiés. Ses signaux ont des composantes négatives¹. Nous appliquons la technique de pré-image sous contrainte de non-négativité sur les coefficients. Le noyau Gaussien est utilisé, avec une largeur de bande définie à $\sigma = 500$, et la valeur du pas fixée à $\eta = 0.1$. Nous étudions l'influence de l'initialisation de l'algorithme, selon deux initialisations différentes.

Tout d'abord, l'algorithme est initialisé à des données aléatoires, soit $x(0)$ sans perte de généralité, et le signal d'initialisation est illustré à la Figure 2.4 (figure à gauche). Cette valeur correspond à initialiser le vecteur β à $\beta(0) = (XX^\top)^{-1}Xx(0)$ pour les valeurs non-négatives, et à zéro autrement. En fixant le nombre maximal d'itérations à $t = 100$, nous obtenons la caractéristique illustrée à la Figure 2.4 (figure à droite). Il est facile de retrouver les deux composantes importantes de l'ERP, et qui sont les ondes N200 et P300.

Afin d'étudier l'évolution des coefficients à chaque itération, et par analogie à la pratique où la moyenne est souvent considérée, nous considérons le cas d'initialisation où tous les coefficients sont égaux, c'est-à-dire, $\beta_k = 1/n$ pour tout $k = 1, 2, \dots, n$. Ce qui correspond à la moyenne des données, où le résultat provient d'une contribution uniforme de toutes les données disponibles. L'évolution de la distribution de ces coefficients au cours des cinq premières itérations est donnée par les histogrammes de la Figure 2.5. Ces résultats montrent que l'algorithme proposé, défini par l'expression (2.7), aboutit à des représentations parcimonieuses, avec le niveau de parcimonie qui augmente à chaque itération. La caractéristique qui en résulte est donnée dans la Figure 2.6, ce qui montre à la fois les composantes N200 et P300, même avec très peu d'itérations. En comparant ces résultats, à la moyenne de dix signaux choisis aléatoirement dans la Figure 2.7 (figure à gauche), et à celle de tous les signaux de la Figure 2.7 (figure à droite), nous pouvons voir que la méthode proposée ne nécessite qu'un petit nombre

1. Les mesures de l'activité du cerveau sont toujours positives. Cependant, les praticiens calibrent ces mesures, résultant en des signaux de moyenne nulle.

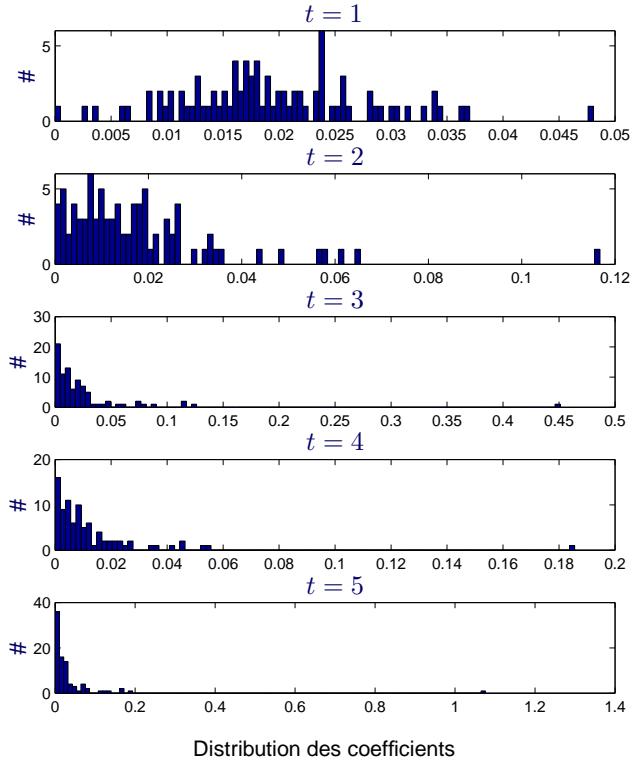


FIGURE 2.5: Distribution des coefficients du modèle de la première itération (première figure) jusqu'à la cinquième itération (dernière figure). Ces figures illustrent l'évolution des coefficients vers une distribution parcimonieuse.

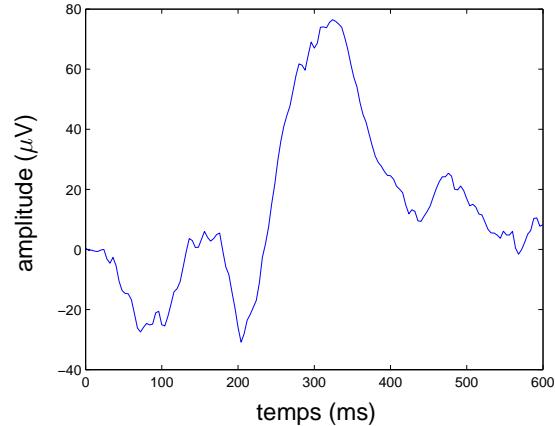


FIGURE 2.6: Extraction des caractéristiques des données des potentiels évoqués, obtenue à l'itération $t = 5$ dans la Figure 2.5 (dernière figure). L'algorithme est initialisé à une contribution uniforme de tous les signaux présents.

de signaux pour en extraire les N200 et P300, à l'opposé des praticiens, il est indispensable de prendre tous les signaux pour extraire ces composantes.

2.4.2 Débruitage de données et des images

Dans cette partie, nous étudions le débruitage sur des données artificielles et sur des chiffres manuscrits réels pris de la base de donnée MNIST. Différentes comparaisons sont faites avec d'autres

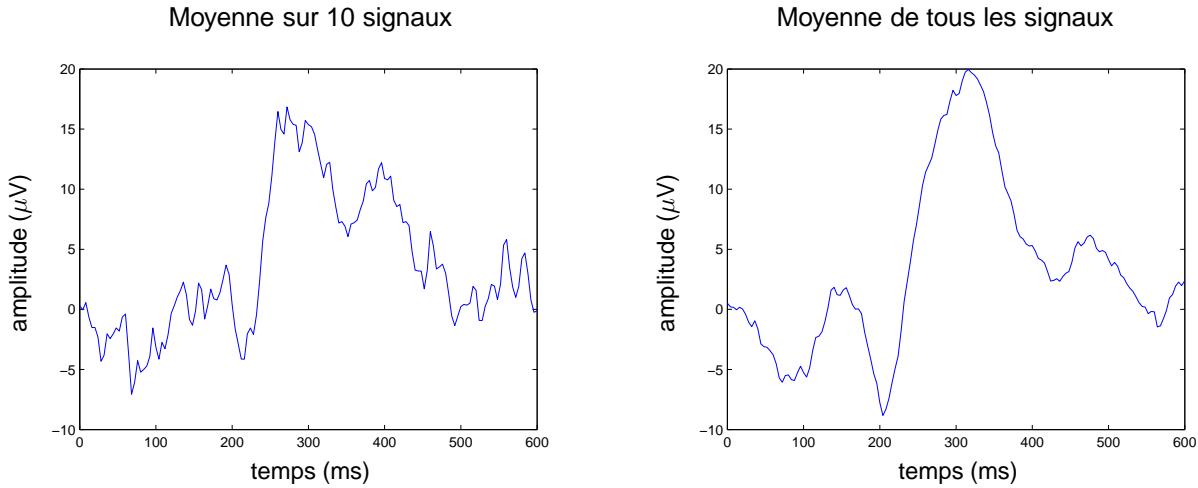


FIGURE 2.7: La moyenne de 10 signaux choisis aléatoirement (figure à gauche) et la moyenne de tous les signaux des potentiels évoqués (figure à droite).

méthodes.

2.4.2.1 Données artificielles

Commençons tout d'abord avec l'ensemble de données artificielles. Afin d'illustrer les résultats, nous considérons un espace à deux dimensions, et nous appliquons la technique de débruitage sur trois distributions de formes différentes : banane, anneau et cadre. L'ensemble de données formant une banane est défini par une parabole de coordonnées $(x, x^2 + \xi)$, où x est uniformément réparti sur l'intervalle $[0.5, 2.5]$ de l'axe des abscisses, et ξ est une variable aléatoire qui suit la loi normale de moyenne nulle et de déviation standard $\nu = 0.2$. L'ensemble de données formant un anneau est défini par un cercle ayant un rayon de 0.9, corrompu par un bruit de distribution uniforme sur $[-\nu, \nu]$, où $\nu = 0.4$. L'ensemble de données formant un cadre est défini par quatre lignes, chacune de longueur 2. Les données aléatoires sont uniformément distribuées sur ces lignes et corrompues par un bruit uniformément distribué sur $[-\nu, \nu]$, avec $\nu = 0.2$. Pour chaque distribution, un ensemble de n échantillons, illustré dans la Figure 2.10 (première ligne), est généré pour la phase d'apprentissage des m vecteurs propres. Pour la distribution sous forme de banane, ayant une forme quadratique, nous choisissons $m = 2$, cependant pour les autres formes qui sont plus compliquées, nous prenons $m = 4$. Un autre ensemble de N échantillons est généré suivant les mêmes distributions, illustré par des très petits points bleus dans la Figure 2.10. Les valeurs des paramètres utilisés pour chaque un des trois ensembles sont données dans le Tableau 2.2. Il est clair que ces formes non-linéaires ne peuvent pas être proprement débruitées en utilisant des approches linéaires, comme l'ACP classique.

Nous comparons l'approche de la pré-image non-négative avec d'autres techniques sans contraintes, y compris la technique du point-fixe défini par (1.12) et l'estimation de la pré-image régularisée décrite dans 1.7.6 et proposée dans [AH11]. Pour appliquer le débruitage par pré-image avec contrainte de non-

TABLE 2.2: Valeurs des paramètres pour les trois distributions

	Distributions			
	banane	anneau	cadre	
Paramètre du bruit	ν	0.2	0.4	0.2
Nombre de données d'apprentissage	n	800	500	550
Nombre de vecteurs propres	m	2	4	4
Largeur du noyau Gaussien	σ	0.7	0.8	0.5
Nombre de données à débruiter	N	200	100	510
Valeur du pas	η	0.3	0.3	0.3
Nombre maximal d'itérations	t_{\max}	20	20	20

négativité, nous opérons une translation des échantillons définissant ces distributions vers le quadrant positif, comme illustré dans la Figure 2.10 (première ligne). Pour tous ces algorithmes, la version bruitée des données a été utilisée pour l'initialisation, *i.e.*, $x(t)$ pour $t = 0$, donnée par des très petits points bleus dans la Figure 2.10. Avec un nombre maximal d'itérations fixé à 20 pour les algorithmes itératifs, les échantillons débruités obtenus par ces techniques de pré-image sont représentés par des points rouges. Les trajectoires obtenues lors de chaque itération sont représentées par des traits verts (excepté la méthode d'estimation de la pré-image régularisée qui n'est pas itérative). Comme nous pouvons voir suivant la taille de ces traits, l'algorithme du point-fixe (deuxième ligne) présente une convergence plus lente par rapport à l'approche proposée (dernière ligne). Ceci est principalement dû à l'utilisation du pas η , ayant ici une valeur fixe de $\eta = 0.3$ pour les trois ensembles de données. Il est à noter que les résultats obtenus à partir de la technique de l'estimation de la pré-image régularisée peuvent aboutir à des minima locaux.

Nous passons maintenant à l'approche pour laquelle la contrainte de non-négativité est appliquée sur les coefficients du modèle. Les ensembles de données sont utilisés directement, et aucune restriction n'est nécessaire dans ce cas, donc, aucune translation n'est réalisée comme auparavant. Nous comparons trois types de noyaux : le noyau Gaussien avec une largeur de bande $\sigma = 0.7$ comme dans le cas précédent, le noyau polynomial quadratique où q vaut 2 et c vaut 1, et le noyau exponentiel avec $\sigma = 1$ (voir le Tableau 1.1). Pour tous ces noyaux, la valeur initiale du vecteur β , étant donné une observation bruitée x_0 a été fixée à la solution de $x_0 = \mathbf{X}^\top \beta$ en ne retenant que les coefficients non-négatifs. En d'autres termes, en utilisant une pseudo-inverse avec

$$\beta(0) = (\mathbf{X} \mathbf{X}^\top)^{-1} \mathbf{X} x_0,$$

pour les valeurs non-négatives ; dans les autres cas, elle est fixée à zéro. Même avec une seule itération ($t = 1$) et un petit pas valant $\eta = 0.1$, l'algorithme proposé a abouti à une pré-image résultante qui est bien débruite, comme le montre la Figure 2.8 (première ligne). En fixant le nombre maximal d'itérations à $t = 100$, les trois noyaux ont donné des résultats comparables reflétant la forme de la banane, comme le montre la Figure 2.8 (dernière ligne). Ce résultat est en opposition avec les résultats précédents ob-

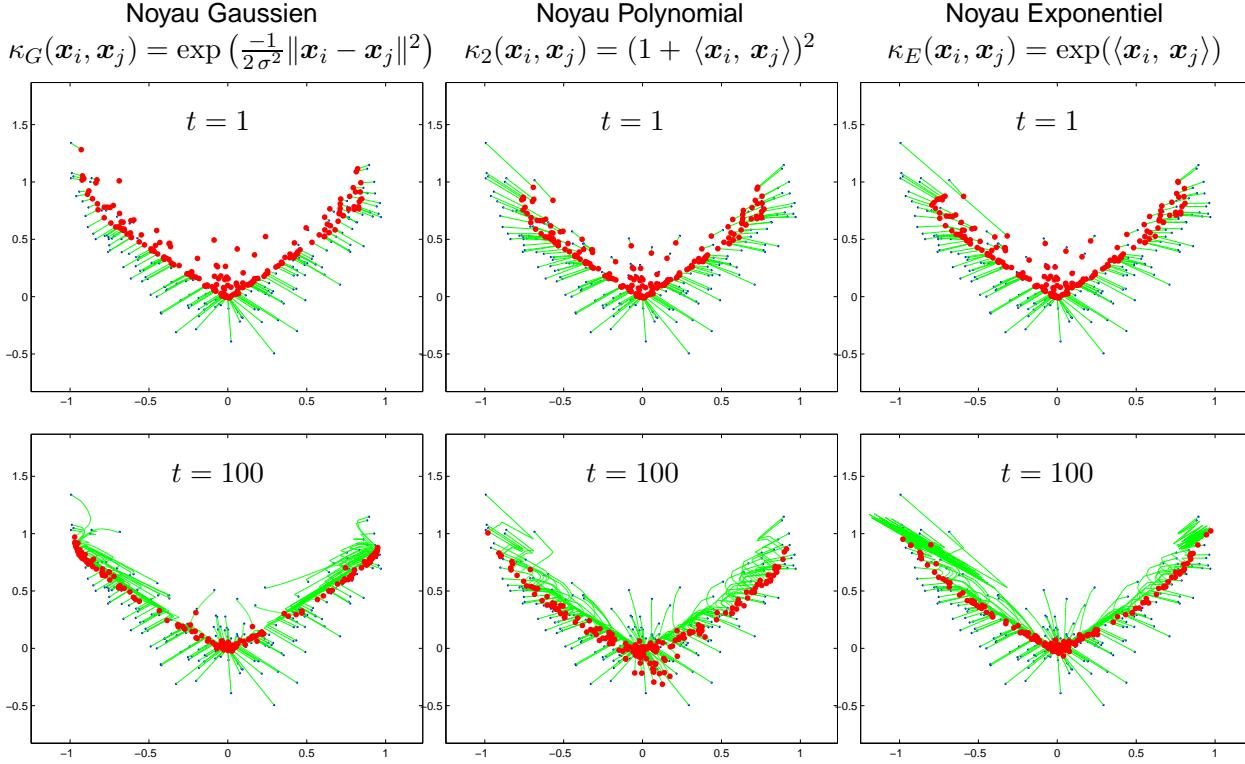


FIGURE 2.8: Débruitage, avec des contraintes sur les coefficients du modèle, de données qui suivent la distribution sous forme de banane, pour une seule itération (première ligne) et après $t = 100$ itérations (deuxième ligne). Trois noyaux différents sont comparés : Gaussien (colonne gauche), polynomial (colonne au milieu), et exponentiel (colonne droite).

servés dans [KT03], où Kwok *et al.* réclament que seulement le noyau Gaussien peut être pré-imagé. Cependant, en appliquant une contrainte de non-négativité des coefficients sur la solution, nous pouvons voir que des noyaux autre que le noyau Gaussien aboutissent à des résultats pertinents.

Passons maintenant à l'analyse des coefficients du modèle, et la parcimonie de la solution. Pour ce faire, nous considérons la distribution des coefficients $\beta_1, \beta_2, \dots, \beta_n$ pour chacun des $N = 200$ échantillons bruités. La Figure 2.9 montre l'histogramme d'une telle distribution, où chacune des couleurs correspond à un échantillon débruité. Alors que nous représentons ici les résultats d'une seule itération, des résultats similaires sont obtenus pour un grand nombre d'itérations. Ces résultats illustrent le fait que les coefficients sont non-négatifs comme prévu, compris entre 0 et 0.018. En outre, la plupart d'entre eux sont proches de zéro, à savoir inférieure à 0.002. C'est la propriété de la parcimonie, bien établie et souvent exigée par une grande classe d'algorithmes dans la communauté d'apprentissage statistique.

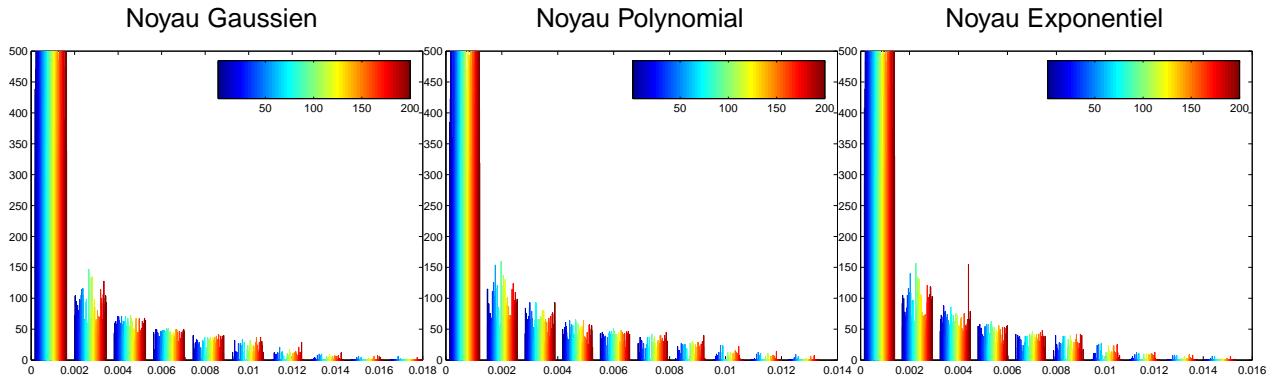


FIGURE 2.9: Distribution des coefficients du modèle pour chacune des 200 données bruitées de la base de données banane, après une seule itération de notre approche, correspondant aux résultats donnés dans la Figure 2.8 (première ligne). Toutes les données débruitées (chacune présentée par une couleur de la barre de couleur) jouissent de la propriété de la parcimonie, avec un grand nombre de valeurs proches de zéro.

2.4.2.2 Données réelles : Chiffres manuscrits

Nous étudions le débruitage des images dans, sur des chiffres manuscrits réels, pris de la base de données MNIST [LC]. MNIST regroupe des données compilées de la base de *National Institute for Standards and Technology* NIST. Elle se compose de 60 000 échantillons pour l'apprentissage et 10 000 pour le test, où chaque échantillon est une image. Nous étudions l'ensemble des chiffres “0”. Chaque image est de taille 28×28 pixels, codés en niveau de gris avec des valeurs comprises entre 0 et 255. Chaque image est écrite sous la forme d'un vecteur de 784 composantes. Les images sont bruitées en ajoutant un bruit de type sel-et-poivre de densité de 0.1. Une collection de 500 images est exploitée pour l'apprentissage de l'ACP-à-noyaux.

Afin d'assurer le débruitage des images sous la contrainte de la non-négativité de la pré-image, comme défini par (2.1), nous utilisons le noyau Gaussien avec une largeur de bande fixée à $\sigma = 500$, pour toutes les techniques de pré-image. Pour construire le sous-espace optimal, 50 vecteurs propres ont été retenus. Une autre collection de dix images, illustrées dans la Figure 2.11 (1^{ère} ligne), bruitées par le même type de bruit (2^{ème} ligne) est utilisée pour le débruitage. La pertinence de la méthode proposée est démontrée pour le débruitage d'images, et est comparée à différentes techniques : la descente du gradient, celle itérative du point-fixe [MSS⁺99], celle de l'échelle multi-dimensionnelle [KT03], l'estimation de la pré-image régularisée [AH11], et la pré-image pénalisée [ZLY10]. Comme nous pouvons le voir, la méthode proposée présente le meilleur débruitage parmi toutes les autres. Nous évaluons l'erreur absolue moyenne (MAE pour *mean absolute error*) pour chacune des images débruitées en utilisant les

TABLE 2.3: L'erreur absolue moyenne pour chaque technique évaluée sur chacune des images débruitées où la densité du bruit vaut 0.1.

Numéro de l'image	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Point-fixe	52,75	13,55	21,77	13,04	10,45	50,95	9,57	54,67	11,02	57,23
Technique MDS	31,31	30,49	25,15	20,83	27,89	25,87	24,92	24,86	24,04	25,85
Pré-image pénalisée	34,70	42,99	40,86	38,70	41,75	34,18	29,15	37,60	35,51	36,62
Pré-image régularisée	176,43	13,55	144,77	145,07	42,00	284,03	17,23	236,62	54,63	20,85
Notre méthode	21,29	13,55	19,46	145,07	42,00	16,18	71,59	126,24	54,63	17,90

TABLE 2.4: L'erreur absolue moyenne pour chaque technique évaluée sur chacune des images débruitées où la densité du bruit vaut 0.25.

Numéro de l'image	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Point-fixe	34,26	76,81	28,99	30,76	31,79	70,89	54,25	67,97	54,55	67,05
Technique MDS	47,56	47,73	41,34	44,13	63,94	44,81	52,43	43,91	58,12	47,39
Pré-image pénalisée	53,11	66,21	43,41	63,68	65,27	57,65	52,76	54,67	46,38	63,97
Pré-image régularisée	34,26	213,43	28,99	30,76	31,79	175,21	124,61	314,11	148,77	156,04
Notre méthode	34,26	30,93	28,99	30,76	31,79	33,25	32,83	30,80	35,77	32,10

différentes techniques de pré-image avec

$$MAE = \frac{1}{28 \times 28} \sum_{i=1}^{28} \sum_{j=1}^{28} |\mathbf{x}_{i,j}^* - \mathbf{x}_{i,j}|,$$

où $\mathbf{x}_{i,j}^*$ est le (i, j) ème élément de l'image \mathbf{x}^* évaluée avec une technique de pré-image et $\mathbf{x}_{i,j}$ est le (i, j) ème élément de l'image correspondante initiale sans bruit. Les Tableaux 2.3, 2.4 et 2.5 présentent les MAE pour chaque image en utilisant les différentes techniques de pré-image où la densité du bruit vaut 0.1, 0.25 et 0.5 respectivement. De même, nous évaluons le rapport signal sur bruit (*PSNR* pour *Peak Signal-to-Noise Ratio*) pour chacune des images débruitées en utilisant les différentes techniques de pré-image avec

$$PSNR = 10 \times \log_{10} \left(\frac{255^2}{\frac{1}{28 \times 28} \sum_{i=1}^{28} \sum_{j=1}^{28} |\mathbf{x}_{i,j}^* - \mathbf{x}_{i,j}|^2} \right),$$

où $\mathbf{x}_{i,j}^*$ est le (i, j) ème élément de l'image \mathbf{x}^* évaluée avec une technique de pré-image et $\mathbf{x}_{i,j}$ est le (i, j) ème élément de l'image correspondante initiale sans bruit. Le Tableau 2.6 montre le PSNR pour chacune des images évaluées à l'aide des différentes techniques de pré-image où la densité du bruit vaut 0.1.

TABLE 2.5: L'erreur absolue moyenne pour chaque technique évaluée sur chacune des images débruitées où la densité du bruit vaut 0.5.

Numéro de l'image	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Point-fixe	77,8	93,51	72,51	60,92	72,86	80,57	77,25	90,47	75,78	60,29
Technique MDS	75,37	73,82	78,55	88,09	65,35	69,75	116,35	71,39	70,86	87,98
Pré-image pénalisée	82,51	91,71	75,82	75,90	86,30	83,39	80,85	87,02	83,15	88,68
Pré-image régularisée	142,89	272,87	62,22	60,92	158,47	58,71	109,55	63	403,75	60,29
Notre méthode	57,61	67,83	62,22	60,92	63,04	58,71	71,82	63	62,37	60,29

TABLE 2.6: Le rapport signal sur bruit pour chaque technique évaluée pour chacune des images débruitées où la densité du bruit vaut 0.1.

Numéro de l'image	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
Point-fixe	12,91	12,89	7,45	8,94	7,61	7,37	13,44	14,21	9,23	7,99
Technique MDS	12,30	11,48	14,83	13,37	14,78	13,90	13,95	12,26	12,84	14,72
Pré-image pénalisée	10,50	7,80	8,10	8,15	7,89	9,78	8,80	9,57	9,09	9,71
Pré-image régularisée	14,04	12,95	-0,43	15,12	15,10	16,60	14,00	13,10	3,86	2,68
Notre méthode	15,27	12,97	14,04	8,61	15,16	6,71	14,22	14,52	8,26	12,80

2.5 Conclusion

Intéressée par la forme et la caractéristique, et plus spécifiquement la physiologie des observations, certaines applications nécessitent des conditions, en particulier, des conditions de non-négativités de la solution. Pour ce faire, nous avons introduit, dans ce chapitre, une nouvelle approche pour la résolution du problème de pré-image sous contrainte de non-négativité. Deux types de contraintes ont été examinés. Le premier type de contraintes s'applique sur la pré-image elle-même, où nous avons élaboré deux approches pour une meilleure convergence de l'algorithme : le pas fixe, et le pas variable. La pré-image s'écrivant sous la forme d'une combinaison linéaire des données, nous proposons alors un second type de contraintes tenant en compte les coefficients de cette combinaison linéaire. Cette contrainte induit souvent la parcimonie. La puissance de l'approche proposée a été illustrée sur trois cas d'applications à savoir : l'extraction de caractéristiques des signaux ERP, le débruitage des données artificielles et le débruitage des images (en l'occurrence des chiffres manuscrits). Notre approche a été comparée avec succès aux méthodes de pré-image telles la descente du gradient, celle du point-fixe, la méthode régularisée et celle pénalisée. Nous avons aussi montré empiriquement son intérêt dans la parcimonie de la solution.

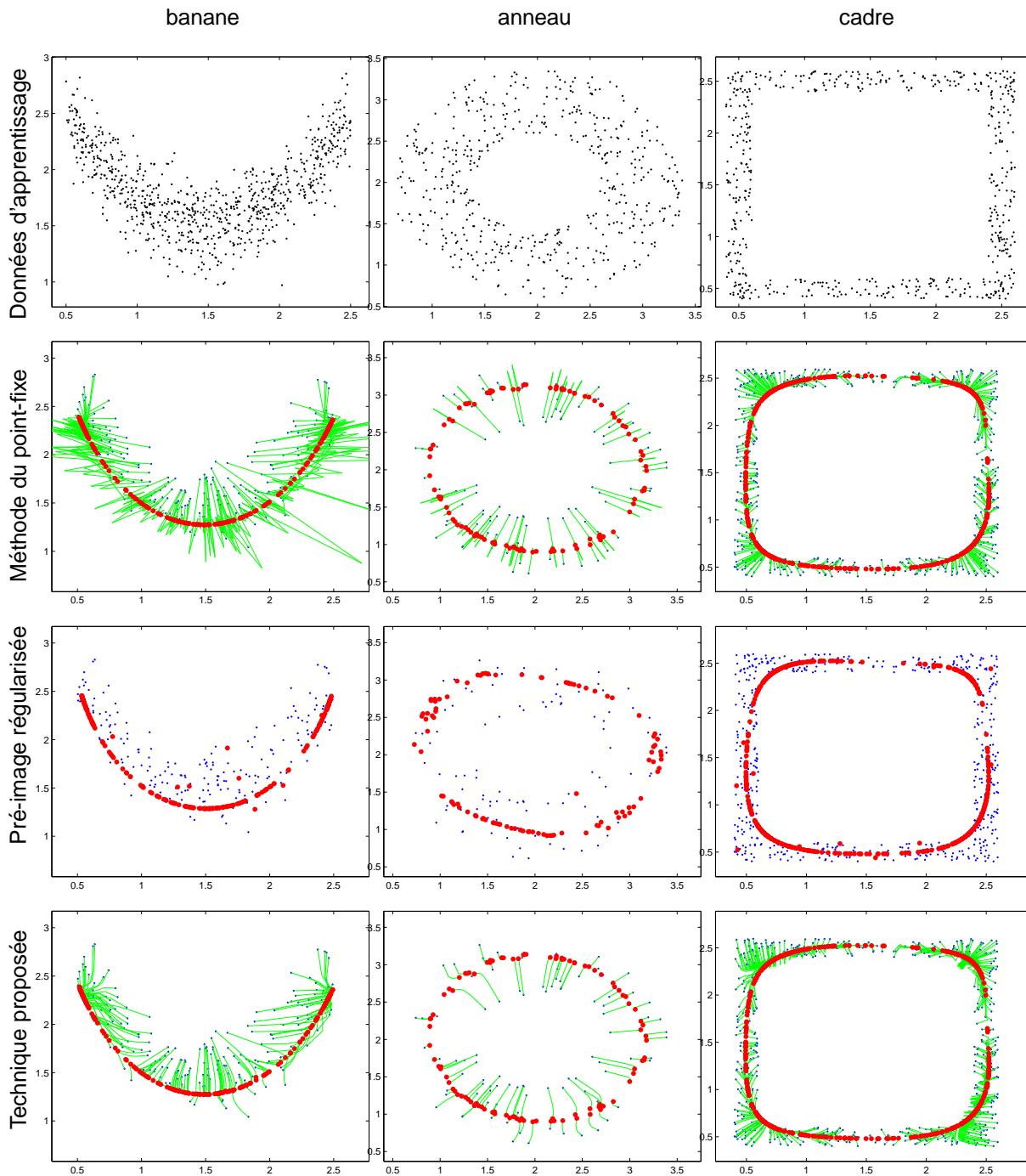


FIGURE 2.10: Débruitage des données artificielles, pour les trois formes : banane (colonne gauche), anneau (colonne au milieu) et cadre (colonne droite). Un ensemble de données d'apprentissage (\blacktriangledown dans la première ligne) est utilisé pour construire le sous-espace pertinent, en utilisant l'ACP-à-noyaux avec le noyau Gaussien. Un autre ensemble de données (désignés par \bullet) est débruité (en \bullet) par les méthodes suivantes : celle du point-fixe (deuxième ligne), celle de l'estimation de la pré-image régularisée (troisième ligne) et celle proposée (dernière ligne). L'évolution de la solution des méthodes itératives pour les 20 itérations est donnée par les traits (illustrés par $\textcolor{green}{—}$).

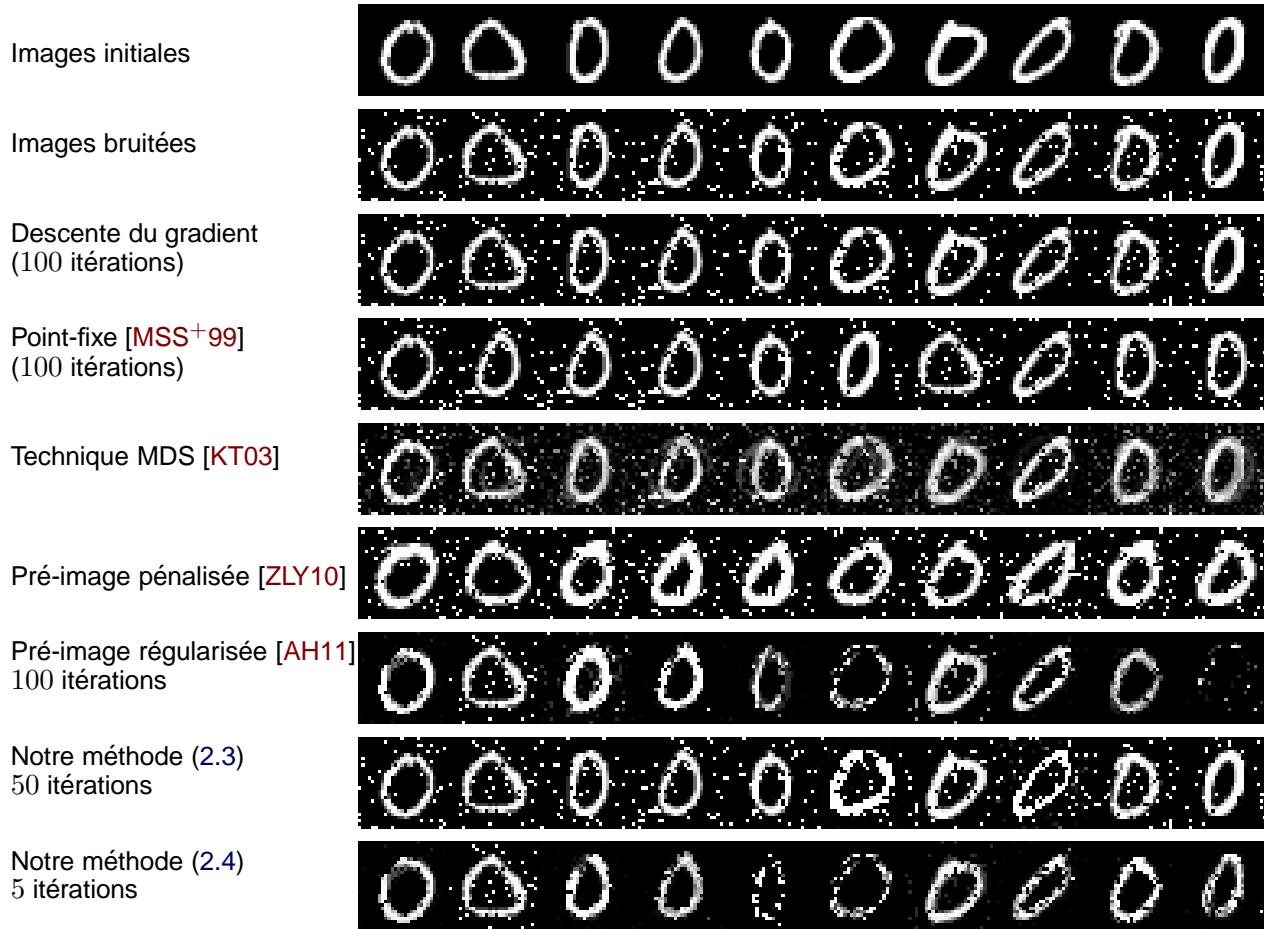


FIGURE 2.11: Un ensemble de 10 images de chiffre “0” (première ligne) corrompues par un bruit sel-et-poivre de densité 0.1 (deuxième ligne). Les résultats de la pré-image sont illustrés en utilisant la descente du gradient (troisième ligne), l'algorithme itératif du point-fixe [MSS⁺99] (quatrième ligne), la méthode de l'échelle multidimensionnelle ou MDS [KT03] (cinquième ligne), la méthode d'apprentissage de la pré-image pénalisée [ZLY10] (sixième ligne), la technique de l'estimation de la pré-image pénalisée [AH11] (septième ligne), la méthode de la pré-image non-négative avec un processus itératif (2.3) (huitième ligne), et celle avec le pas adéquat de la pré-image non-négative (2.4) (dernière ligne).

Modèles autorégressifs-à-noyaux : technique de prédition

Sommaire

3.1 Introduction	55
3.2 Séries temporelles	57
3.2.1 Processus stochastique	58
3.2.2 Propriétés des séries temporelles	58
3.3 Prédiction des séries temporelles avec un modèle autorégressif	59
3.3.1 Méthode des moindres carrés	60
3.3.2 Équations de Yule-Walker	60
3.4 Modèle autorégressif à noyaux pour la prédiction des séries temporelles	61
3.4.1 Modèle autorégressif dans l'espace fonctionnel	62
3.4.2 Le problème de la pré-image comme technique de prédiction	65
3.5 Modèles autorégressifs-à-noyaux sans pré-image	67
3.5.1 Modèle autorégressif sur les valeurs du noyau	68
3.5.2 Un modèle autorégressif hybride	69
3.5.3 Lien entre le modèle AR hybride et les modèles proposés auparavant	70
3.6 Expérimentations	70
3.6.1 Comparaison les techniques prédictives non-linéaires	72
3.6.2 Comparaison entre les différentes techniques de pré-image	74
3.6.3 Comparaison entre les différentes techniques proposées	74
3.7 Conclusion	77

3.1 Introduction

Les techniques liées à la prédiction des séries temporelles constituent un outil d'aide à la décision de première importance. Pour donner une idée sur l'ampleur de la recherche en matière de méthodes de prédiction, nous pouvons citer des applications financières [Tsa05], économétriques [BT98, LBB⁺⁰⁴],

gestionnaires [BT10], statistiques [Ful96], sans oublier les applications météorologiques. De même, ces méthodes sont utilisées en contrôle de processus [MWM83], et en traitement du signal comme les signaux biomédicaux [ZIP06]. Pour ce dernier cas, différentes études ont été réalisées ; nous pouvons en citer le traitement par décomposition empirique [KCV11], la méthode de l'entropie maximale [Lin79], et l'analyse en ligne [GS90].

Le développement des techniques de prédiction a été longtemps basé sur des progrès réalisés en statistique et en probabilités. Les premiers modèles statistiques de prédiction sont sous forme de modèles autorégressifs. Le modèle autorégressif (AR), ou prédictif linéaire, est omniprésent en sciences et technologie, avec un rôle essentiel dans l'analyse des séries temporelles dans des applications allant de la prédiction financière, à l'analyse météorologique. Pour le traitement de la parole par exemple, afin de maintenir une conversation téléphonique, pour chaque 20 millisecondes, le téléphone portable modélise la parole selon un modèle prédictif linéaire facilitant, ainsi la transmission de données [DMK09]. Le modèle prédictif linéaire décrit un échantillon en l'exprimant sous forme d'une combinaison linéaire d'un certain nombre d'échantillons précédents. Afin d'évaluer cette combinaison linéaire, les paramètres qui la décrivent sont estimés sur une série d'échantillons disponibles, avant de l'étendre à la prédiction des futurs échantillons. Ce modèle autorégressif est facile à implémenter grâce à l'algèbre linéaire. Cependant, il est conçu pour les systèmes linéaires.

Les mathématiques sous-jacentes qui maîtrisent le modèle autorégressif sont les équations de Yule-Walker [Yul27, SS89]. Depuis ce temps, la communauté scientifique s'est investie sans cesse afin de maîtriser ces équations pour la prédiction linéaire [CG85]. Les équations de Yule-Walker sont le bloc de construction du modèle linéaire AR, reliant ainsi les paramètres du modèle à la fonction de covariance du processus. Les paramètres du modèle sont donc estimés à partir des covariances de la série temporelle. La prédiction peut être considérée en appliquant le modèle prédictif qui en résulte. Toutefois, l'hypothèse de linéarité est souvent insuffisante pour expliquer les phénomènes non-linéaires. Une première tentative pour établir des équations de Yule-Walker non-linéaires est donnée dans [CB96] avec un modèle non-linéaire d'ordre élevé.

Nous proposons dans ce chapitre un modèle autorégressif non-linéaire pour la modélisation et la prédiction de séries temporelles, en utilisant les méthodes à noyaux et la résolution du problème de la pré-image. Nous dérivons des modèles prédictifs non-linéaires, d'une part par la mise en œuvre de la méthode des moindres carrés dans l'espace fonctionnel, d'autre part en tirant pleinement avantage des équations de Yule-Walker. Cette dernière dérivation conduit à l'estimation des paramètres du modèle en utilisant l'espérance des noyaux. Il est à noter que l'idée de l'espérance des noyaux a montré son efficacité très récemment dans d'autres domaines applicatifs, voir par exemple [AGSJ11, AG11].

La modélisation non-linéaire et la prédiction n'ont pas encore tiré pleinement du profit des progrès récents dans le domaine de l'apprentissage statistique, même si plusieurs essais ont été faits pour développer des techniques non-linéaires pour l'analyse des séries temporelles, telles que la régression à vecteurs de support [Vap98], le filtre de Kalman à noyau [RDB05] et la prédiction en ligne avec les noyaux [RBH09]. Très peu de tentatives ont été faites pour aborder le modèle AR non-linéaire en apprentissage.

Un premier travail dans cette direction, présenté par Kumar *et al.* [KJ07] propose un modèle AR dans l'espace fonctionnel, cependant, sans la capacité de prédire les échantillons futurs. Trois modèles sont à l'étude dans ce chapitre. Le premier modèle est basé sur l'idée sous-jacente des méthodes à noyaux, à savoir la transformation des données. En appliquant un modèle AR sur les échantillons transformés, la prédiction est définie dans l'espace RKHS. Pour interpréter l'échantillon prédit, il est nécessaire de faire le retour inverse à l'espace des observations, à savoir l'espace des échantillons, par la résolution du problème de la pré-image. Ensuite, nous proposons de contourner le problème de la pré-image, en dérivant deux autres modèles. Dans le deuxième modèle, nous considérons le modèle AR sur les valeurs obtenues en considérant le noyau choisi. Cette considération conduit à une résolution plus grossière du problème. Dans le troisième modèle, nous proposons une formulation hybride, comme un compromis entre les modèles précédents, à savoir entre le modèle itératif, affiné, et le modèle direct, général, évalué sur le noyau.

Dans ce chapitre, nous présentons les séries temporelles, en définissant quelques propriétés importantes utilisées pour l'analyse. Ensuite, le modèle autorégressif est défini en utilisant la méthode des moindres carrés, ainsi qu'à partir des équations de Yule-Walker. Une extension de ce modèle pour l'analyse des séries temporelles issues de systèmes non-linéaires est proposée à l'aide des méthodes à noyaux. Les trois techniques AR non-linéaires décrites ci-dessus sont détaillées dans ce chapitre. Finalement, la pertinence de la technique autorégressive-à-noyaux est illustrée en l'appliquant à des séries temporelles unidimensionnelles et multidimensionnelles [MOG97, Wan]. Commençons tout d'abord par une introduction sur les séries temporelles.

3.2 Séries temporelles

Une série temporelle est une suite de valeurs numériques représentant l'évolution d'une quantité spécifique au cours du temps. De telles suites de valeurs peuvent être exprimées mathématiquement afin d'en analyser le comportement [Mad08], généralement pour comprendre leur évolution passée et souvent pour en prévoir leur comportement futur. Une telle transposition mathématique utilise le plus souvent des concepts de probabilités et de statistique.

Définition 3.1. (*Série temporelle*) *Une série temporelle, $\{x_t, t = 1, 2, \dots, n\}$, est une suite finie d'observations réelles d'un phénomène donné x , indexées par une date t .*

La suite d'observations correspondant à une même variable décrit une série temporelle. La série est caractérisée par sa variation en fonction du temps. L'instant auquel l'observation est prise correspond à une information importante pour décrire le système. Nous pouvons calculer les statistiques descriptives usuelles : moyenne, variance, coefficients d'aplatissement et d'asymétrie. La Figure 3.1 montre trois séries temporelles, à gauche deux séries des indices du prix de l'once d'or en dollars et en euros, entre les années 1998 et 2011, et à droite une série d'un enregistrement d'électrocardiogramme.

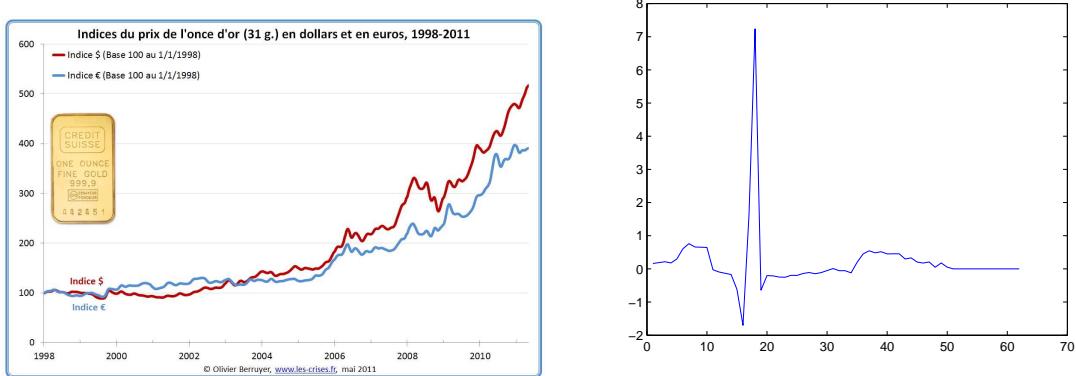


FIGURE 3.1: Exemples de séries temporelles statistique à gauche et électrocardiogramme à droite

3.2.1 Processus stochastique

La première étape de l'analyse d'une série temporelle est la sélection d'un modèle mathématique approprié pour les données. Il est naturel de supposer que chaque observation x_t est une réalisation d'une certaine variable aléatoire [BD09]. Nous avons besoin de définir précisément ce qu'on entend par processus stochastique et ses réalisations.

Définition 3.2. (*Processus stochastique*) *Un processus stochastique (ou aléatoire) est une famille de variables aléatoires $\{x_t, t = 1, 2, \dots, n\}$ définies sur le même espace de probabilité.*

Remarque 3.1. *Dans l'analyse des séries temporelles, l'ensemble $\{t = 1, 2, \dots, n\}$ est un ensemble représentant le temps, très souvent $\{0, \pm 1, \pm 2, \dots\}$, $\{1, 2, 3, \dots\}$. Les processus stochastiques pour lesquels l'ensemble représentant le temps n'est pas un sous-ensemble de \mathbb{R} ont également une importance, comme pour le cas des processus stochastiques géophysiques.*

Nous passons maintenant à définir quelques propriétés des séries temporelles. Nous parlerons de la stationnarité et de l'ergodicité.

3.2.2 Propriétés des séries temporelles

Lors du travail avec un nombre fini de variables aléatoires, il est souvent utile d'évaluer la matrice de covariance pour mieux comprendre la dépendance des données. Pour une série temporelle $\{x_t, t = 1, 2, \dots, n\}$, l'idée de la matrice de covariance est étendue à des collections infinies de variables aléatoires. La fonction d'autocovariance nous fournit cette extension nécessaire pour une série temporelle. Nous désignons par $\mathbb{E}[\cdot]$ l'espérance. Il est à noter que toutes les espérances dans ce document sont prises sur le temps t .

Définition 3.3. (*Fonction d'autocovariance*) Si $\{x_t, t = 1, 2, \dots, n\}$ est un processus de variance finie, alors sa fonction d'autocovariance $\gamma_x(\cdot, \cdot)$ est définie par

$$\gamma_x(r, s) = \text{Cov}(x_r, x_s) = \mathbb{E}[(x_r - \mathbb{E}[x_r])(x_s - \mathbb{E}[x_s])], \quad \text{pour } r, s \in \{1, 2, \dots, n\}.$$

Définition 3.4. (*Stationnarité*) Le processus $\{x_t, t = 1, 2, \dots, n\}$ de moyenne constante, est dit stationnaire si

- $\mathbb{E}[x_t]^2 < \infty$ pour tout $t \in \{1, 2, \dots, n\}$,
- $\gamma_x(r, s) = \gamma_x(r + t, s + t)$ pour tout $r, s, t \in \{1, 2, \dots, n\}$.

Remarque 3.2. La stationnarité comme définie ci-dessus est souvent connue sous le nom de stationnarité au sens large, ou stationnarité de second degré. Dans la pratique, le terme stationnarité fait référence à la Définition 3.4.

Une autre notion de stationnarité, importante et fréquemment utilisée, la stationnarité au sens strict, est donnée par la définition suivante.

Définition 3.5. (*Stationnarité au sens strict*) Le processus $\{x_t, t = 1, 2, \dots, n\}$ est dit un processus stationnaire au sens strict si les distributions conjointes de $(x_{t_1}, \dots, x_{t_k})$ et $(x_{t_1+h}, \dots, x_{t_k+h})$ sont les mêmes pour tous les $t_1, \dots, t_k, h \in \{1, 2, \dots, n\}$.

La stationnarité au sens strict signifie intuitivement que les réalisations de la série temporelle, sur deux intervalles de temps différents, doivent présenter des caractéristiques statistiques similaires.

Définition 3.6. (*Ergodicité*) Un processus stochastique $\{x_t, t = 1, 2, \dots, n\}$ est dit ergodique si toutes les moyennes temporelles existent et ont même valeur indépendamment de l'instant choisi.

Cette notion d'ergodicité est très importante du fait que pratiquement, pour évaluer les moyennes statistiques, l'on ne dispose généralement que d'un échantillon sur lequel on estime une moyenne temporelle. Cette définition n'a de valeur que si le processus stochastique étudié est stationnaire et ergodique.

3.3 Prédiction des séries temporelles avec un modèle autorégressif

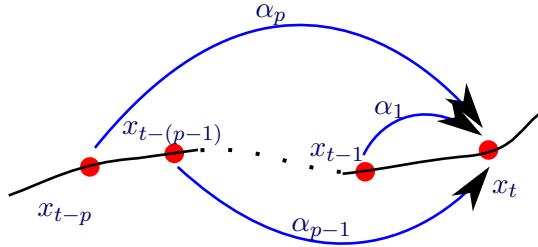
Dans cette partie, nous décrivons le modèle autorégressif. Il est utilisé pour la prédiction et la modélisation des séries temporelles. Soit la série temporelle x_1, x_2, \dots, x_n .

Définition 3.7. (*Processus autorégressif*) Un processus est dit autorégressif (AR) d'ordre p s'il existe $\alpha_1, \dots, \alpha_p \in \mathbb{R}$ tel que

$$\begin{aligned} x_t &= \alpha_1 x_{t-1} + \alpha_2 x_{t-2} + \dots + \alpha_p x_{t-p} + \varepsilon_t \\ &= \sum_{j=1}^p \alpha_j x_{t-j} + \varepsilon_t \end{aligned} \tag{3.1}$$

où les constantes α_k sont les paramètres du modèle, et ε_t est un bruit souvent supposé blanc Gaussien.

FIGURE 3.2: Schéma illustratif du modèle AR, où x_t est défini par une combinaison linéaire de p échantillons précédents $x_{t-1}, x_{t-2}, \dots, x_{t-p}$, avec les paramètres $\alpha_1, \alpha_2, \dots, \alpha_p$ comme étant ses coefficients.



Le processus autorégressif est donc caractérisé par son ordre p , et les paramètres $\alpha_1, \alpha_2, \dots, \alpha_p$ qui décrivent la combinaison linéaire. Il est basé sur la prédiction d'un échantillon à partir d'un certain nombre d'échantillons de son passé, suivant une simple combinaison linéaire [Tsa05]. Bien que son principe est simple, le modèle AR est largement utilisé en modélisation et prédiction de séries temporelles. La Figure 3.2 illustre l'idée du modèle autorégressif.

Différentes méthodes ont été proposées dans la littérature pour le calcul des paramètres du modèle. Nous citons les équations de Yule-Walker qui seront étudiées dans le chapitre suivant, la technique de Levinson-Durbin, la méthode des moindres carrés...

3.3.1 Méthode des moindres carrés

Dans cette section, nous détaillons la méthode des moindres carrés. Cette méthode cherche à minimiser l'erreur quadratique de prédiction entre la vraie valeur de la série et la valeur prédictive, à savoir

$$\sum_{t=p+1}^n \left(x_t - \sum_{j=1}^p \alpha_j x_{t-j} \right)^2.$$

Pour aboutir aux valeurs optimales des α_k recherchées, nous procédons par le calcul de la dérivée de cette expression par rapport à $\alpha_1, \alpha_2, \dots, \alpha_p$. En définissant le vecteur $\boldsymbol{\alpha} = [\alpha_1 \alpha_2 \cdots \alpha_p]^\top$ des p paramètres à estimer, ces paramètres sont estimés selon

$$\boldsymbol{\alpha} = \left(\sum_{t=p+1}^n \mathbf{x}_t \mathbf{x}_t^\top \right)^{-1} \sum_{t=p+1}^n \mathbf{x}_t x_t,$$

où \mathbf{x}_t comprend les p échantillons précédents, à savoir $\mathbf{x}_t = [x_{t-1} \ x_{t-2} \ \dots \ x_{t-p}]^\top$.

3.3.2 Équations de Yule-Walker

Nous passons maintenant à détailler les équations de Yule-Walker utilisées pour l'estimation des coefficients $\boldsymbol{\alpha}$. Les paramètres $\alpha_1, \alpha_2, \dots, \alpha_p$ sont directement liés à la fonction de covariance du processus. Nous déterminons alors ces paramètres à partir de la fonction d'autocovariance. C'est l'essence même des équations de Yule-Walker, comme illustré ci-dessous.

Soit μ l'espérance des x_t , c'est-à-dire

$$\mu = \mathbb{E}[x_t].$$

En appliquant l'espérance sur les deux membres de l'équation (3.1), nous avons alors l'expression de l'espérance du bruit, selon $(1 - \sum_{j=1}^p \alpha_j)\mu = \mathbb{E}[\varepsilon_t]$. Pour tout décalage positif τ , nous évaluons la fonction d'autocovariance de chaque série temporelle. Soit $r(\cdot)$ la contrepartie empirique de la fonction d'autocorrelation de la série temporelle, alors $r(\tau) = \sum_{j=1}^p \alpha_j r(\tau - j)$, pour tout décalage τ . Puisque la fonction d'autocorrelation est paire, *i.e.*, $r(-\tau) = r(\tau)$, nous obtenons sous forme matricielle les équations de Yule-Walker

$$\mathbf{r} = \mathbf{R}\boldsymbol{\alpha},$$

où \mathbf{r} est un vecteur regroupant les p fonctions d'autocorrelation empiriques pour tout décalage τ entre 1 et p , à savoir, $\mathbf{r} = [r(1) \ r(2) \ \cdots \ r(p)]^\top$, et \mathbf{R} représente une matrice des fonctions d'autocorrelation empiriques pour les décalages τ entre 0 et $p - 1$, donnée par

$$\mathbf{R} = \begin{bmatrix} r(0) & r(1) & \dots & r(p-1) \\ r(1) & r(0) & \dots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \dots & r(0) \end{bmatrix}.$$

En supposant que la matrice symétrique \mathbf{R} de taille $p \times p$ est inversible, les coefficients $\boldsymbol{\alpha}$ sont estimés par le produit entre l'inversion de la matrice \mathbf{R} et le vecteur \mathbf{r} , selon $\boldsymbol{\alpha} = \mathbf{R}^{-1}\mathbf{r}$.

Après avoir déterminé les paramètres $\alpha_1, \alpha_2, \dots, \alpha_p$, le modèle AR d'ordre p permet de prédire directement un futur échantillon, selon l'équation (3.1) pour $t = n + 1$ (et au delà). Bien que cette technique réussisse à prédire les futurs échantillons issus de systèmes linéaires, elle n'est pas adaptée aux systèmes non-linéaires.

3.4 Modèle autorégressif à noyaux pour la prédiction des séries temporelles

Nous passons maintenant à étendre le concept du modèle autorégressif pour des données des systèmes non-linéaires. Dans le même esprit et par le biais des méthodes à noyaux, nous proposons d'étendre le modèle AR à une approche non-linéaire dans un RKHS, en appliquant à chaque échantillon une transformation non-linéaire.

3.4.1 Modèle autorégressif dans l'espace fonctionnel

Considérons une fonction non-linéaire $\Phi(\cdot)$ de l'espace des observations \mathcal{X} à l'espace RKHS qui, à chaque x_t , fait correspondre son image $\Phi(x_t)$. Le modèle AR décrit dans le RKHS est alors donné par

$$\Phi(x_t) = \sum_{j=1}^p \alpha_j \Phi(x_{t-j}) + \varepsilon_t^\Phi, \quad (3.2)$$

où ε_t^Φ représente un bruit dans l'espace fonctionnel. Soit $\varphi_t = [\Phi(x_{t-1}) \ \Phi(x_{t-2}) \ \cdots \ \Phi(x_{t-p})]$ le vecteur regroupant les transformées, par la fonction $\Phi(\cdot)$, des p échantillons précédents à x_t . Nous écrivons le modèle non-linéaire sous la forme matricielle

$$\Phi(x_t) = \varphi_t \alpha.$$

Nous étudions deux méthodes pour estimer les coefficients α , d'une part la minimisation de la distance quadratique dans l'espace fonctionnel, et d'autre part les équations de Yule-Walker dans cet espace fonctionnel.

3.4.1.1 Minimisation de la distance quadratique dans l'espace fonctionnel

En se basant sur la méthode des moindres carrés présentée dans le paragraphe 3.3.1, nous minimisons l'erreur quadratique moyenne, dans l'espace fonctionnel, entre la fonction prédictive $\sum_{j=1}^p \alpha_j \Phi(x_{t-j})$ et la vraie fonction de l'image $\Phi(x_t)$. Ainsi le critère cité ci-dessus sera-t-il défini par :

$$\min_{\alpha} \sum_{t=p+1}^n \left\| \Phi(x_t) - \sum_{j=1}^p \alpha_j \Phi(x_{t-j}) \right\|_{\mathcal{H}}^2,$$

où $\|\cdot\|_{\mathcal{H}}$ désigne la norme dans l'espace en question. En utilisant l'écriture matricielle, nous obtenons $\sum_{t=p+1}^n (\langle \varphi_t \alpha, \varphi_t \alpha \rangle_{\mathcal{H}} - 2 \langle \varphi_t \alpha, \Phi(x_t) \rangle_{\mathcal{H}} + \langle \Phi(x_t), \Phi(x_t) \rangle_{\mathcal{H}})$. En dérivant l'expression de l'erreur quadratique par rapport à α , et en annulant sa dérivée, nous aboutissons aux valeurs optimales des paramètres avec

$$\alpha = \left(\sum_{t=p+1}^n \langle \varphi_t, \varphi_t \rangle \right)^{-1} \sum_{t=p+1}^n \langle \varphi_t, \Phi(x_t) \rangle_{\mathcal{H}}.$$

Il est clair que cette expression n'implique que le produit scalaire entre les couples d'images par la fonction non-linéaire $\Phi(\cdot)$ des données, définis par $\Phi(x_1), \Phi(x_2), \dots, \Phi(x_n)$, ce qui permet de l'évaluer en utilisant simplement une fonction noyau $\kappa(x_t, x_i) = \langle \Phi(x_t), \Phi(x_i) \rangle_{\mathcal{H}}$.

3.4.1.2 Équations de Yule-Walker dans l'espace fonctionnel

Nous passons maintenant à une autre technique pour l'estimation des coefficients α dans l'espace fonctionnel. Cette technique est basée sur les équations de Yule-Walker dans le RKHS, et tient compte

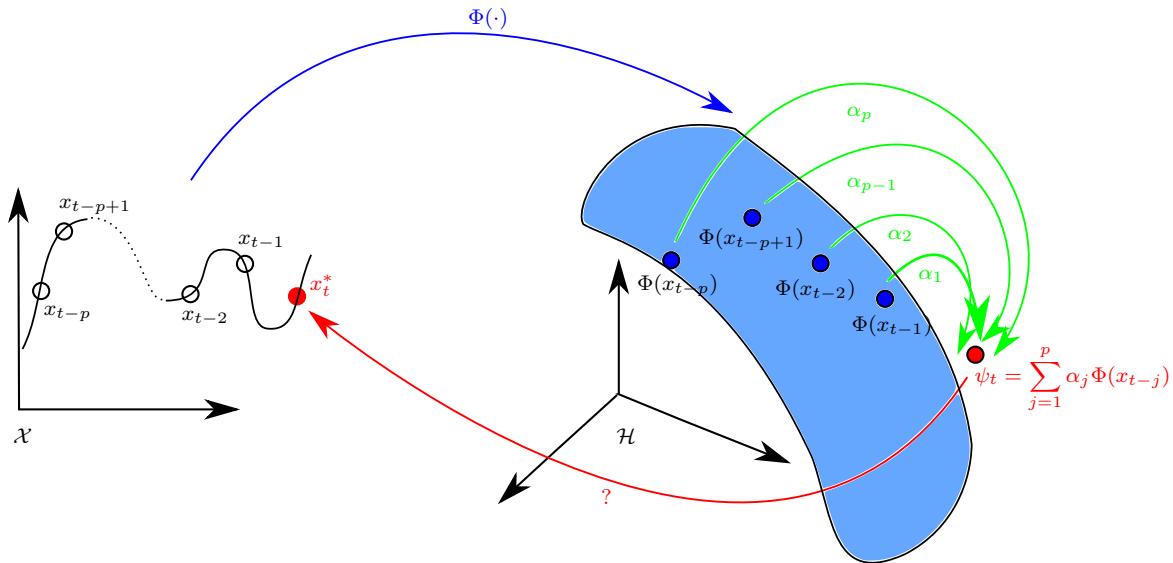


FIGURE 3.3: Schéma illustrant le modèle AR-à-noyaux : les échantillons sont transformés de l'espace des données \mathcal{X} à l'espace fonctionnel \mathcal{H} , où le modèle AR est appliqué sur les échantillons transformés. Une fois que la fonction ψ_t est estimée, il est nécessaire de faire le retour inverse à l'espace des données \mathcal{X} , afin de trouver la prédiction x_t^* . Une technique de pré-image est alors nécessaire pour prédire les nouveaux échantillons dans l'espace initial.

de l'idée décrite dans le paragraphe 3.3.2. Bien que les échantillons x_t soient supposés de moyenne nulle dans l'espace des observations, ce n'est pas le cas pour les images $\Phi(x_t)$ dans l'espace fonctionnel.

Soit μ_Φ l'espérance des fonctions $\Phi(x_t)$ dans l'espace fonctionnel, à savoir

$$\mu_\Phi = \mathbb{E}[\Phi(x_t)].$$

En calculant l'espérance des deux membres de l'équation (3.5), nous obtenons une expression de l'espérance de l'erreur donnée par $(1 - \sum_{j=1}^p \alpha_j)\mu_\Phi = \mathbb{E}[\varepsilon_t^\Phi]$, sous l'hypothèse de stationnarité du processus. D'un autre côté, nous définissons les fonctions centrées dans l'espace fonctionnel par

$$\begin{aligned}\Phi(x_t) - \mu_\Phi &= \sum_{j=1}^p \alpha_j \Phi(x_{t-j}) + \varepsilon_t^\Phi - \mu_\Phi \\ &= \sum_{j=1}^p \alpha_j (\Phi(x_{t-j}) - \mu_\Phi) + \varepsilon_t^\Phi - \left(1 - \sum_{j=1}^p \alpha_j\right) \mu_\Phi.\end{aligned}$$

En couplant ces résultats, et en considérant le produit scalaire dans l'espace fonctionnel entre les deux membres de l'équation ci-dessus d'une part, et le terme $(\Phi(x_{t-\tau}) - \mu_\Phi)$ d'autre part, pour un décalage

positif τ , nous aboutissons à

$$\langle \Phi(x_t) - \mu_\Phi, \Phi(x_{t-\tau}) - \mu_\Phi \rangle = \langle \varepsilon_t^\Phi - \mathbb{E}[\varepsilon_t^\Phi], \Phi(x_{t-\tau}) - \mu_\Phi \rangle + \sum_{j=1}^p \alpha_j \langle \Phi(x_{t-j}) - \mu_\Phi, \Phi(x_{t-\tau}) - \mu_\Phi \rangle. \quad (3.3)$$

Par analogie avec le modèle AR linéaire, nous supposons que le bruit ε_t^Φ et la fonction $\Phi(x_{t-\tau})$ sont non corrélés pour tout décalage positif τ . Par suite, en considérant l'espérance de l'expression (3.3) et sous l'hypothèse de la stationnarité de la séquence, nous avons pour tout décalage τ supérieur ou égal à 1

$$\mathbb{E}[\kappa_c(x_t, x_{t-\tau})] = \sum_{j=1}^p \alpha_j \mathbb{E}[\kappa_c(x_{t-j}, x_{t-\tau})], \quad (3.4)$$

où $\kappa_c(\cdot, \cdot)$ est la *version centrée* du noyau $\kappa(\cdot, \cdot)$, définie par le produit scalaire des fonctions centrées, comme suit

$$\kappa_c(x_i, x_j) = \langle \Phi(x_i) - \mu_\Phi, \Phi(x_j) - \mu_\Phi \rangle.$$

Finalement, nous considérons toutes les valeurs possibles du décalage, et nous écrivons l'expression de l'équation (3.4) sous forme matricielle, avec

$$\mathbf{r}_{\kappa_c} = \mathbf{R}_{\kappa_c} \boldsymbol{\alpha},$$

où \mathbf{r}_{κ_c} regroupe les p valeurs des espérances des noyaux centrés pour les décalages entre 1 et p , à savoir,

$$\mathbf{r}_{\kappa_c} = \begin{bmatrix} \mathbb{E}[\kappa_c(x_t, x_{t-1})] & \mathbb{E}[\kappa_c(x_t, x_{t-2})] & \cdots & \mathbb{E}[\kappa_c(x_t, x_{t-p})] \end{bmatrix}^\top,$$

et \mathbf{R}_{κ_c} est la matrice définie par les espérances des noyaux, selon

$$\mathbf{R}_{\kappa_c} = \begin{bmatrix} \mathbb{E}[\kappa_c(x_t, x_t)] & \mathbb{E}[\kappa_c(x_t, x_{t-1})] & \cdots & \mathbb{E}[\kappa_c(x_t, x_{t-p+1})] \\ \mathbb{E}[\kappa_c(x_t, x_{t-1})] & \mathbb{E}[\kappa_c(x_t, x_t)] & \cdots & \mathbb{E}[\kappa_c(x_t, x_{t-p+2})] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[\kappa_c(x_t, x_{t-p+1})] & \mathbb{E}[\kappa_c(x_t, x_{t-p+2})] & \cdots & \mathbb{E}[\kappa_c(x_t, x_t)] \end{bmatrix}.$$

Le vecteur des coefficients $\boldsymbol{\alpha}$, obtenu en inversant cette matrice, est donné par

$$\boldsymbol{\alpha} = \mathbf{R}_{\kappa_c}^{-1} \mathbf{r}_{\kappa_c}.$$

En pratique, les espérances sont estimées sur un ensemble de n échantillons disponibles [CMR12]. La version centrée du noyau est évaluée à partir de

$$\kappa_c(x_i, x_j) = \kappa(x_i, x_j) - \frac{1}{n} \sum_{k=1}^n \kappa(x_i, x_k) - \frac{1}{n} \sum_{k=1}^n \kappa(x_j, x_k) + \frac{1}{n^2} \sum_{k,k'=1}^n \kappa(x_{k'}, x_k).$$

3.4.2 Le problème de la pré-image comme technique de prédiction

Une fois que les observations ont été transformées vers le RKHS, et les paramètres déterminés, nous pouvons alors prédire un élément à partir de son passé, avec

$$\psi_t = \sum_{j=1}^p \alpha_j \Phi(x_{t-j}). \quad (3.5)$$

Comme illustré dans la Figure 3.3, cet élément appartient à l'espace engendré par les images des p précédents échantillons, une interprétation est nécessaire dans le domaine des échantillons. Alors que la fonction $\Phi(\cdot)$ permet de passer de ce domaine au RKHS, la fonction inverse n'existe pas dans le cas général. C'est le problème de la *pré-image* en méthodes d'apprentissage à noyaux décrit dans la section 1.6, qui consiste à déterminer l'échantillon x_t^* dont l'image $\Phi(x^*)$ est la plus proche possible de ψ_t . En utilisant la norme définie dans le RKHS, le problème d'optimisation est donné par

$$x_t^* = \arg \min_x \frac{1}{2} \left\| \sum_{j=1}^p \alpha_j \Phi(x_{t-j}) - \Phi(x) \right\|_{\mathcal{H}}^2, \quad (3.6)$$

où les α_k sont les paramètres déjà estimés.

Le problème d'optimisation s'écrit sous forme générale avec

$$x_t^* = \arg \min_x J_t(x),$$

où $J_t(x)$ représente la fonction coût décrite par le développement de la norme de l'équation (3.6), elle est définie par

$$J_t(x) = - \sum_{j=1}^p \alpha_j \kappa(x_{t-j}, x) + \frac{1}{2} \kappa(x, x).$$

Dans cette expression, le terme indépendant de x , donné par $\frac{1}{2} \sum_{k=1}^p \sum_{j=1}^p \alpha_k \alpha_j \kappa(x_{t-k}, x_{t-j})$, est éliminé puisque sa dérivée par rapport à x s'annule.

Afin de résoudre ce problème, différentes techniques sont présentes dans la littérature. Une possibilité est d'étudier le gradient de la fonction coût $J_t(x)$ par rapport à x . A l'optimum, le gradient par rapport à x s'annule, à savoir $\partial J_t(x) / \partial x = 0$. Le gradient est alors défini par :

$$\frac{\partial J_t(x)}{\partial x} = - \sum_{j=1}^p \alpha_j \frac{\partial \kappa(x_{t-j}, x)}{\partial x} + \frac{1}{2} \frac{\partial \kappa(x, x)}{\partial x}. \quad (3.7)$$

C'est la forme générale pour tout type de noyau. Cette expression peut être simplifiée pour la classe des noyaux radiaux, comme le noyau Gaussien. Dans ce cas, la définition du noyau entraîne l'annulation de la dérivée du terme $\kappa(x, x)$ par rapport à x , et le gradient est donné par le premier terme de (3.7),

selon

$$\frac{\partial J_t(x)}{\partial x} = -\frac{1}{\sigma^2} \sum_{j=1}^p \alpha_j \kappa(x_{t-j}, x) (x_{t-j} - x).$$

En annulant ce gradient à l'optimum x_t^* , nous obtenons l'expression itérative du point fixe

$$x_t^* = \frac{\sum_{j=1}^p \alpha_j \kappa(x_{t-j}, x_t^*) x_{t-j}}{\sum_{k=1}^p \alpha_k \kappa(x_{t-k}, x_t^*)}.$$

Ce résultat peut être interprété comme étant un modèle AR, à l'instar de (3.1), puisque nous avons la forme $x_t^* = \sum_{j=1}^p \beta_j x_{t-j}$, cependant les paramètres ne sont plus des *constantes*, avec

$$\beta_j = \left(\sum_{k=1}^p \alpha_k \kappa(x_{t-k}, x_t^*) \right)^{-1} \alpha_j \kappa(x_{t-j}, x_t^*).$$

Pour le cas bien spécifique du noyau Gaussien, l'expression itérative est donnée par

$$x_t^* = \frac{\sum_{j=1}^p \alpha_j \exp(-\|x_{t-j} - x_t^*\|^2/2\sigma^2) x_{t-j}}{\sum_{k=1}^p \alpha_k \exp(-\|x_{t-k} - x_t^*\|^2/2\sigma^2)}.$$

Pour le cas de noyau polynomial défini par $\kappa_q(x_i, x_j) = (\langle x_i, x_j \rangle + c)^q$, où q est un entier positif, et c est un paramètre non négatif, la fonction coût $J_t(x)$ est donnée par

$$J_t(x) = - \sum_{j=1}^p \alpha_j (\langle x_{t-j}, x \rangle + c)^q + \frac{1}{2} (\langle x, x \rangle + c)^q,$$

menant à l'expression itérative du point-fixe décrite par

$$x_t^* = \frac{-\sum_{j=1}^p \alpha_j (\langle x_{t-j}, x_t^* \rangle + c)^{q-1} x_{t-j}}{(\langle x_t^*, x_t^* \rangle + c)^{q-1}}.$$

En utilisant le Théorème 1.3, pour le cas du modèle autorégressif ; nous pouvons alors écrire la pré-image comme étant une combinaison linéaire des données disponibles, à savoir $x_t^* = \sum_{j=1}^p \delta_j^* x_{t-j}$. Nous prouvons cette déclaration pour les noyaux radiaux et projectifs.

Corollaire 3.1. *Pour un modèle AR-à-noyaux d'ordre p , toute pré-image x_t^* s'écrit sous la forme d'une combinaison linéaire des p derniers échantillons, à savoir*

$$x_t^* = \sum_{j=1}^p \delta_j^* x_{t-j}$$

pour certains poids δ_j^ .*

Démonstration. Premièrement, nous étudions la classe des noyaux radiaux, définie par l'expression

(1.3). Dans de pareils cas, le terme $\partial\kappa(x, x)/\partial x$ s'annule. Le gradient à l'optimum s'écrit comme

$$\sum_{j=1}^p \alpha_j \frac{\partial\kappa(x_{t-j}, x_t^*)}{\partial x_t^*} = 0,$$

où le premier membre est donné par

$$\sum_{j=1}^p \alpha_j \frac{\partial\kappa(x_{t-j}, x_t^*)}{\partial x_t^*} = \sum_{j=1}^p \alpha_j \frac{\partial g(\|x_{t-j} - x_t^*\|^2)}{\partial(\|x_{t-j} - x_t^*\|^2)} 2(x_t^* - x_{t-j}).$$

Par conséquence, le résultat final s'exprime comme

$$x_t^* = \sum_{j=1}^p \alpha_j \frac{g^{(1)}(\|x_{t-j} - x_t^*\|^2)}{\sum_{k=1}^p \alpha_k g^{(1)}(\|x_{t-k} - x_t^*\|^2)} x_{t-j},$$

donc de la forme $x_t^* = \sum_{j=1}^p \delta_j^* x_{t-j}$.

Passons maintenant à l'étude du cas des noyaux projectifs, de la forme (1.2). Dans ce cas, le gradient est donné à l'optimum par

$$\sum_{j=1}^p \alpha_j \frac{\partial\kappa(x_{t-j}, x_t^*)}{\partial x_t^*} = \frac{1}{2} \frac{\partial\kappa(x_t^*, x_t^*)}{\partial x_t^*}.$$

Nous évaluons le premier membre ainsi que le second respectivement en utilisant la forme du noyau, et nous combinons les deux expressions pour aboutir à

$$x_t^* = \sum_{j=1}^p \alpha_j \frac{f^{(1)}(\langle x_{t-j}, x_t^* \rangle)}{f^{(1)}(\langle x_t^*, x_t^* \rangle)} x_{t-j},$$

qui est aussi de la forme $x_t^* = \sum_{j=1}^p \delta_j^* x_{t-j}$. □

3.5 Modèles autorégressifs-à-noyaux sans pré-image

Le modèle AR-à-noyaux proposé auparavant nécessite la résolution du problème mal-posé de la pré-image, pour chaque échantillon prédit. Dans cette section, afin de surmonter ce problème, nous étudions l'estimation d'une fonction où la sortie désirée est la vraie valeur x_t , et l'entrée est un vecteur \mathbf{x}_t représentant les p échantillons précédents avec $\mathbf{x}_t = [x_{t-1} \ x_{t-2} \ \cdots \ x_{t-p}]^\top$. Il s'agit alors d'une fonction à valeur réelle définie sur \mathcal{X}^p . Nous considérons alors des noyaux multivariés, définis sur $\mathcal{X}^p \times \mathcal{X}^p$. En tenant compte de cette écriture, nous proposons deux modèles non-linéaires différents, et dérivons les équations de Yule-Walker correspondantes.

Il est important de noter que pour le noyau Gaussien, cette approche correspond à considérer le

noyau

$$\kappa(\mathbf{x}_{t-j}, \mathbf{x}_t) = \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_{t-j} - \mathbf{x}_t\|^2\right),$$

où la distance entre les deux vecteurs \mathbf{x}_{t-j} et \mathbf{x}_t est donné par $\|\mathbf{x}_{t-j} - \mathbf{x}_t\|^2 = \sum_{k=1}^p (x_{t-j-k} - x_{t-k})^2$.

Il est facile de voir de cette écriture mène à l'expression suivante du noyau Gaussien

$$\kappa(\mathbf{x}_{t-j}, \mathbf{x}_t) = \prod_{k=1}^p \kappa(x_{t-j-k}, x_{t-k}),$$

donc reliant le noyau multivarié au noyau univarié utilisé dans les deux précédentes sections de ce chapitre.

3.5.1 Modèle autorégressif sur les valeurs du noyau

Par opposition à la méthode ci-dessus, où nous appliquons le modèle autorégressif aux images des observations par la fonction $\Phi(\cdot)$ dans l'espace fonctionnel, nous considérons ici un modèle autorégressif sur les valeurs du noyau. Le modèle proposé est défini par la prédiction de l'échantillon x_t , avec

$$\varphi(\mathbf{x}_t) = \sum_{j=1}^p \beta_j \kappa(\mathbf{x}_{t-j}, \mathbf{x}_t) + \varepsilon_t, \quad (3.8)$$

où \mathbf{x}_t est un vecteur représentant les p échantillons précédents avec $\mathbf{x}_t = [x_{t-1} \ x_{t-2} \ \cdots \ x_{t-p}]^\top$, et $\varphi(\cdot)$ est la fonction définissant la prédiction des échantillons futurs tel que $\varphi(\mathbf{x}_t) = x_t$.

Soit μ_x l'espérance de toutes les valeurs transformées $\varphi(\mathbf{x}_t)$, et μ_j l'espérance de chaque noyau $\kappa(\mathbf{x}_{t-j}, \mathbf{x}_t)$, $j = 1, 2, \dots, p$, à savoir

$$\begin{aligned} \mu_x &= \mathbb{E}[\varphi(\mathbf{x}_t)], \\ \mu_j &= \mathbb{E}[\kappa(\mathbf{x}_t, \mathbf{x}_{t-j})]. \end{aligned}$$

En suivant les développements donnés dans la Section 3.4.1.2, et par analogie avec l'expression (3.3) où le produit scalaire se fait sur les échantillons centrés, nous obtenons alors

$$\mathbb{E}[\langle \varphi(\mathbf{x}_t) - \mu_x, \varphi(\mathbf{x}_{t-\tau}) - \mu_x \rangle] = \sum_{j=1}^p \beta_j \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-j}) - \mu_j, \varphi(\mathbf{x}_{t-\tau}) - \mu_x \rangle], \quad (3.9)$$

pour tout décalage $\tau = 1, 2, \dots, p$. En combinant toutes les valeurs possibles du décalage, nous avons la forme matricielle

$$\mathbf{r}_\kappa = \mathbf{R}_\kappa \boldsymbol{\beta},$$

où \mathbf{r}_κ , défini par l'espérance du produit scalaire entre les valeurs évaluées pour tout décalage, est donné

par

$$\mathbf{r}_\kappa = \left[\mathbb{E}[\langle \varphi(\mathbf{x}_t) - \mu_{\mathbf{x}}, \varphi(\mathbf{x}_{t-1}) - \mu_{\mathbf{x}} \rangle], \dots, \mathbb{E}[\langle \varphi(\mathbf{x}_t) - \mu_{\mathbf{x}}, \varphi(\mathbf{x}_{t-p}) - \mu_{\mathbf{x}} \rangle] \right]^\top \quad (3.10)$$

et \mathbf{R}_κ est la matrice décrite par l'espérance du produit scalaire entre les noyaux avec

$$\mathbf{R}_\kappa = \begin{bmatrix} \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-1}) - \mu_1, \varphi(\mathbf{x}_{t-1}) - \mu_{\mathbf{x}} \rangle] \dots \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-p}) - \mu_p, \varphi(\mathbf{x}_{t-1}) - \mu_{\mathbf{x}} \rangle] \\ \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-1}) - \mu_1, \varphi(\mathbf{x}_{t-2}) - \mu_{\mathbf{x}} \rangle] \dots \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-p}) - \mu_p, \varphi(\mathbf{x}_{t-2}) - \mu_{\mathbf{x}} \rangle] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-1}) - \mu_1, \varphi(\mathbf{x}_{t-p}) - \mu_{\mathbf{x}} \rangle] \dots \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-p}) - \mu_p, \varphi(\mathbf{x}_{t-p}) - \mu_{\mathbf{x}} \rangle] \end{bmatrix}.$$

Le vecteur des coefficients $\boldsymbol{\beta}$ est obtenu en inversant la matrice \mathbf{R}_κ , avec

$$\boldsymbol{\beta} = \mathbf{R}_\kappa^{-1} \mathbf{r}_\kappa.$$

Une fois que les paramètres $\boldsymbol{\beta}$ du modèle sont déterminés, nous pouvons directement prédire tout échantillon futur, en utilisant $x_t^* = \varphi(\mathbf{x}_t)$, avec

$$x_t^* = \sum_{j=1}^p \beta_j \kappa(\mathbf{x}_{t-j}, \mathbf{x}_t). \quad (3.11)$$

3.5.2 Un modèle autorégressif hybride

Dans cette section, nous étudions un autre modèle AR, en estimant une fonction $\varphi(\cdot)$, telle que $\varphi : \mathcal{X}^p \rightarrow \mathcal{X}$. Le modèle proposé est défini par

$$\varphi(\mathbf{x}_t) = \sum_{j=1}^p \beta_j \kappa(\mathbf{x}_{t-j}, \mathbf{x}_t) x_{t-j} + \varepsilon_t, \quad (3.12)$$

de façon que l'image d'un échantillon \mathbf{x}_t par cette fonction $\varphi(\cdot)$ donne la valeur prédite, à savoir $\varphi(\mathbf{x}_t) = x_t^*$, qui n'est autre que la valeur de l'échantillon prédit dans l'espace des observations \mathcal{X} , et \mathbf{x}_t est un vecteur représentant les p échantillons précédents avec $\mathbf{x}_t = [x_{t-1} \ x_{t-2} \ \dots \ x_{t-p}]^\top$. La Section 3.5.3 détaille la motivation pour proposer ce modèle.

Soient μ_{xj} l'espérance du produit entre le noyau $\kappa(\mathbf{x}_{t-j}, \mathbf{x}_t)$ et l'échantillon x_{t-j} pour tout $j = 1, 2, \dots, p$, à savoir

$$\mu_{xj} = \mathbb{E}[\kappa(\mathbf{x}_t, \mathbf{x}_{t-j}) x_{t-j}],$$

et $\mu_{\mathbf{x}}$ l'espérance de toutes les observations transformées $\varphi(\mathbf{x}_t)$, précisément $\mu_{\mathbf{x}} = \mathbb{E}[\varphi(\mathbf{x}_t)]$. Par

conséquent, et par analogie à ce qui est donné auparavant, nous obtenons

$$\mathbb{E}[\langle \varphi(\mathbf{x}_t) - \mu_{\mathbf{x}}, \varphi(\mathbf{x}_{t-\tau}) - \mu_{\mathbf{x}} \rangle] = \sum_{j=1}^p \beta_j \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-j}) x_{t-j} - \mu_{xj}, \varphi(\mathbf{x}_{t-\tau}) - \mu_{\mathbf{x}} \rangle]. \quad (3.13)$$

En considérant tous les décalages possibles, nous obtenons la forme matricielle $\mathbf{r}_\kappa = \mathbf{R}_{\kappa x} \boldsymbol{\beta}$, où \mathbf{r}_κ est défini par (3.10), et $\mathbf{R}_{\kappa x}$ est la matrice décrite par les espérances du produit entre les noyaux et les échantillons pour tout décalage possible, selon

$$\mathbf{R}_{\kappa x} = \begin{bmatrix} \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-1}) x_{t-1} - \mu_{x1}, \varphi(\mathbf{x}_{t-1}) - \mu_{\mathbf{x}} \rangle] & \dots & \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-p}) x_{t-p} - \mu_{xp}, \varphi(\mathbf{x}_{t-1}) - \mu_{\mathbf{x}} \rangle] \\ \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-1}) x_{t-1} - \mu_{x1}, \varphi(\mathbf{x}_{t-2}) - \mu_{\mathbf{x}} \rangle] & \dots & \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-p}) x_{t-p} - \mu_{xp}, \varphi(\mathbf{x}_{t-2}) - \mu_{\mathbf{x}} \rangle] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-1}) x_{t-1} - \mu_{x1}, \varphi(\mathbf{x}_{t-p}) - \mu_{\mathbf{x}} \rangle] & \dots & \mathbb{E}[\langle \kappa(\mathbf{x}_t, \mathbf{x}_{t-p}) x_{t-p} - \mu_{xp}, \varphi(\mathbf{x}_{t-p}) - \mu_{\mathbf{x}} \rangle] \end{bmatrix}.$$

Le vecteur des coefficients $\boldsymbol{\beta}$ est obtenu en inversant la matrice \mathbf{R}_κ , avec $\boldsymbol{\beta} = \mathbf{R}_{\kappa x}^{-1} \mathbf{r}_\kappa$.

Une fois que les paramètres $\boldsymbol{\beta}$ sont évalués, le modèle autorégressif hybride permet de prédire un échantillon futur en appliquant $x_t^* = \varphi(\mathbf{x}_t)$, à savoir

$$x_t^* = \sum_{j=1}^p \beta_j \kappa(\mathbf{x}_{t-j}, \mathbf{x}_t) x_{t-j}. \quad (3.14)$$

Nous passons maintenant à donner la motivation qui nous a permis de proposer ce modèle.

3.5.3 Lien entre le modèle AR hybride et les modèles proposés auparavant

La principale motivation du modèle AR hybride est sa relation avec les deux modèles proposés auparavant, comme décrit dans cette section. Tout d'abord, nous considérons le modèle AR-à-noyaux, comme défini dans la Section 3.4.1.2. Le Corollaire 3.1 montre que l'échantillon prédit à l'instant t prend la forme

$$x_t^* = \sum_{j=1}^p \delta_j^* x_{t-j}.$$

En limitant la solution à $\delta_j^* = \beta_j \kappa(\mathbf{x}_{t-j}, \mathbf{x}_t)$, nous obtenons le modèle AR hybride défini par (3.14). Ensuite, nous considérons le modèle AR sur les valeurs du noyau, décrit dans la Section 3.5.1, à savoir $x_t^* = \sum_{j=1}^p \beta_j \kappa(\mathbf{x}_{t-j}, \mathbf{x}_t)$. En limitant β_j à la forme $\beta_j x_{t-j}$, donc dépendant linéairement du j ème précédent échantillon, nous retrouvons le modèle AR hybride.

3.6 Expérimentations

Nous passons maintenant aux expérimentations afin d'étudier la pertinence de la méthode proposée. Dans un premier temps, la méthode proposée est comparée à des méthodes non-linéaires connues

dans la littérature : le perceptron multicouche, la régression à vecteurs de support [MOG97] et le filtre de Kalman non-linéaire [RDB05]. Dans un second temps, une comparaison entre les différentes techniques de pré-image est faite pour la méthode proposée : la descente du gradient, la méthode du point-fixe, la technique multidimensionnelle, et l'approche conforme. Dans un troisième temps, une comparaison entre les différents modèles AR-à-noyaux proposés est réalisée. A cette fin, nous prenons pour les deux premières expérimentations quatre séries temporelles classiques, et les différents paramètres de ces séries sont choisis comme dans [RDB05] afin de faire la comparaison avec les autres techniques, et pour la troisième expérimentation, nous avons choisi un signal électrocardiogramme (ECG) de [GAG⁺13, GAG^{+b}]. Les séries sont définies par [Wan] :

- La série temporelle *Laser* provient des intensités des impulsions d'un laser NH_3 . La série est chaotique et possède un attracteur de faible dimension. Cette série fait partie d'un ensemble ayant fait l'objet d'une compétition organisée en 1992 par l'institut de Sante Fe à New Mexico [HAW89].
- La série temporelle *Mackey-Glass* fournit un modèle de l'évolution de la production des cellules sanguines [Cas89]. Elle est définie par l'équation différentielle suivante

$$\frac{dx(t)}{dt} = -0.1 x(t) + \frac{0.2 x(t - \tau)}{1 + x(t - \tau)^{10}}.$$

Pour les valeurs de τ supérieures à 16.8, cette série temporelle possède un attracteur étrange et les trajectoires montrent un comportement chaotique fortement non-linéaire [Cas89]. Pour $\tau = 30$, la série temporelle est désignée par MG_{30} .

- La série temporelle *Ikeda map* réfère à une série bidimensionnelle de la dynamique du laser. A partir d'un point initial, $\mathbf{x}(0) = [x_1(0) \ x_2(0)]^\top$, elle est définie par

$$\begin{cases} \omega(t) = c1 - c3/(1 + x_1^2(t) + x_2^2(t)) \\ x_1(t+1) = r + c2(x_1(t) \cos \omega(t) + x_2(t) \sin \omega(t)) \\ x_2(t+1) = c2(x_1(t) \sin \omega(t) + x_2(t) \cos \omega(t)) \end{cases}$$

où $c1, c2, c3$ et r sont des constantes ; dans notre cas, nous considérons $c1 = 0.4, c2 = 0.84, c3 = 6.0, r = 1.0$ et $\mathbf{x}(0) = [1 \ 0.001]^\top$.

- La série temporelle *attracteur Lorenz* est la solution du système défini par les équations différentielles suivantes :

$$\begin{cases} \frac{dx(t)}{dt} = -a x(t) + a y(t) \\ \frac{dy(t)}{dt} = -x(t) z(t) + r x(t) - y(t) \\ \frac{dz(t)}{dt} = +x(t) y(t) - b z(t) \end{cases}$$

Les constantes sont fixées à $a = 10, r = 28$ et $b = 8/3$.

- Le signal ECG représente des enregistrements ECG des sujets référencés au Laboratoire d'Arythmie où les sujets inclus dans cette base de données ont été connus comme ayant aucune arythmie

significative [GAG^{+b}, GAG⁺¹³]. Les signaux ECG sont considérés comme stationnaires, car ils sont pris durant une 1 minute.

- Le signal EMG représente des enregistrements d'électromyogrammes (EMG) ayant un problème de neuropathie [Kim] où les données sont prises en insérant une aiguille sur le muscle antérieur de la jambe pour chaque patient.

Les deux séries *Laser* et MG_{30} sont unidimensionnelles, tandis que *Ikeda* et *Lorenz* sont deux séries chaotiques multidimensionnelles. Nous commençons tout d'abord par la comparaison avec d'autres techniques de modélisation et prédiction.

3.6.1 Comparaison les techniques prédictives non-linéaires

Dans cette partie, nous prenons en compte une comparaison entre différentes techniques non-linéaires de prédiction. Deux expériences ont été réalisées sur chacune de ces séries. Dans un premier temps, nous prenons $n = 300$ échantillons pour l'apprentissage, l'estimation de la largeur de bande du noyau Gaussien σ , et l'ordre optimal p . Lors des simulations, l'ordre optimal est choisi parmi les valeurs $\{1, 2, \dots, 10\}$, et nous obtenons $p = 3$ pour l'attracteur de *Lorenz* et $p = 6$ pour toutes les autres séries. Un ensemble formé des 300 échantillons suivants est utilisé pour évaluer les performances du modèle. Le critère de la comparaison est l'erreur quadratique moyenne de prédiction (en anglais *mean square error* (MSE)) qui est estimée sur les échantillons pris pour le test de performance :

$$err = \frac{1}{n} \sum_{i=n+1}^{2n} \|x_t^* - x_t\|^2,$$

où x_t^* est la valeur prédite à l'instant t , x_t est la vraie valeur de la série au même instant, et n est le nombre d'échantillons pris pour chaque expérience.

Le Tableau 3.1 comporte les valeurs de la largeur de bande σ choisies pour chacune des séries temporelles et utilisées pour chaque expérience. Afin de comparer les résultats obtenus avec d'autres méthodes de prédiction non-linéaires, nous considérons les méthodes du perceptron multicouche, du régresseur à vecteurs supports et du filtre de Kalman non-linéaire. Pour ces deux dernières méthodes à noyaux, le noyau Gaussien a été utilisé. La non-linéarité utilisée pour le perceptron multicouche est implémentée par une fonction `tanh`, et le code utilisé pour implémenter le perceptron multicouche et les autres techniques de filtrage est pris de la boîte à outil ReBEL de R. van der Merwe [WdM10]. L'apprentissage des séries temporelles multidimensionnelles avec des régresseurs à vecteurs supports est fait en décomposant le problème d'apprentissage en autant de problèmes que de dimension des séries temporelles. Pour toutes les expérimentations, la “régression ridge à noyau” avec un noyau Gaussien dont sa largeur de bande est déterminée par rapport à la performance mesurée sur l'ensemble utilisé pour l'apprentissage [RDB05] est utilisée. Une comparaison avec ces différentes méthodes est donnée dans le Tableau 3.2.

Dans un second temps, un autre ensemble comprenant seulement $n = 50$ échantillons est

TABLE 3.1: Valeurs de la largeur de bande σ du noyau Gaussien pour chaque série temporelle pour les deux expériences.

	Laser	MG_{30}	Ikeda	Lorenz
1 ^{ere} expérience : $n = 300$	0.3	0.015	0.0046	0.04
2 ^{nde} expérience : $n = 50$	0.4	0.035	0.0255	0.035

TABLE 3.2: Erreur quadratique moyenne pour différentes méthodes de prédiction et plusieurs séries temporelles.

	Laser	MG_{30}	Ikeda	Lorenz
perceptron multicouche	1.4326	0.0461	0.00071	0.2837
régression à vecteurs de support	0.2595	0.0313	0.00081	0.1811
filtre de Kalman non-linéaire	0.2325	0.0307	0.00077	0.3133
modèle AR-à-noyaux (1 ^{ere} expérience : $n = 300$)	0.0702	0.0008	0.00088	0.1792
modèle AR-à-noyaux (2 ^{nde} expérience : $n = 50$)	0.1813	0.0049	0.0053	0.2625

considéré. La même démarche est réalisée lors de l'apprentissage et les 50 échantillons suivants sont utilisés pour tester l'efficacité de la méthode proposée. Même en prenant des séries courtes, avec seulement 50 échantillons, l'approche proposée présente aussi de bons résultats. Cependant, elle peut éventuellement présenter des résultats moins bonnes que d'autres méthodes comme dans le cas des données *Ikeda*. Même pour des séries ayant un nombre d'échantillons élevé, notre approche donne de meilleurs résultats, et l'amélioration des performances est considérable. Nous notons aussi sa simplicité algorithmique, propriété héritée du modèle linéaire, où l'estimation des paramètres du modèle ne nécessite qu'une simple inversion de matrice de taille $p \times p$. Les autres méthodes non-linéaires souffrent d'une charge calculatoire conséquente, comme c'est le cas de la régression à vecteurs supports qui nécessite la résolution d'un problème d'optimisation quadratique avec contraintes.

Une autre expérience est faite pour faire une comparaison avec le modèle dynamique à noyau proposé dans [RdB03]. À cette fin, nous réalisons la même configuration décrite dans [RdB03]. Chaque série de taille T est notée par $x_{1:T}$. Nous considérons une dimension d et un pas k tel que les vecteurs $x_t = (x_t, x_{t-k}, \dots, x_{t-(d-1)k})^\top$ sont utilisés. Dans un premier cas, pour la série MG_{17} , nous construisons des vecteurs de dimension $d = 6$ et un pas de $k = 6$. En utilisant cette configuration, nous obtenons six séries indépendantes. Pour l'apprentissage, nous utilisons les 100 premiers points de S_1 , cependant les 100 premiers points de S_2 servent à choisir les hyperparamètres. L'erreur de prédiction est mesurée en fonction des points dans l'intervalle 201 à 300 de la série S_1 . Dans un second cas, pour la série *Laser*, la dimension est fixée à $d = 3$ et le pas à $k = 1$. La série est divisée comme suit. Les 100 premiers points sont utilisés pour l'apprentissage cependant les points entre 201 et 300 sont utilisés pour la sélection des hyperparamètres. L'erreur est calculée sur points entre 101 et 200. Le Tableau 3.3 présente l'erreur quadratique moyenne pour les méthodes de régression à vecteurs supports, le modèle

TABLE 3.3: Erreur quadratique moyenne pour différentes méthodes de prédiction pour les séries temporelles Laser et MG_{17} .

	Laser	MG_{17}
régression à vecteurs de support	15.81	0.0812
modèle dynamique à noyau	13.96	0.0859
modèle AR à noyaux	3.37	0.0465

dynamique à noyau et le modèle AR à noyaux proposé pour les deux séries *Laser* et MG_{17} . Le modèle proposé présente la meilleure MSE pour les deux séries en considération.

Passons maintenant à une comparaison entre les différentes techniques de pré-image, en considérant la méthode autorégressive proposée.

3.6.2 Comparaison entre les différentes techniques de pré-image

Dans cette partie, une comparaison entre les techniques de pré-image pour la prédiction en utilisant un modèle autorégressif-à-noyaux est faite. Pour cette étude comparative, nous utilisons $n = 300$ échantillons des séries temporelles lors de l'apprentissage des paramètres. Nous estimons l'erreur de prédiction quadratique sur les 300 échantillons suivants. Lors de l'apprentissage, les paramètres du modèle et son ordre p sont estimés, ainsi que la largeur de bande du noyau Gaussien σ , le degré q avec $c = 1$ pour le noyau polynomial, et le pas η . Les valeurs optimales pour la largeur de bande σ et le pas η sont estimées en utilisant une recherche sur une grille de $\{2^{-12}, 2^{-11}, \dots, 2^{11}, 2^{12}\}$. Pour le noyau polynomial, la valeur optimale du paramètre q est choisie de l'ensemble $\{1, 2, 3, 4, 5, 6\}$. Pour chacune des techniques étudiées, les valeurs optimales de ces paramètres sont utilisées. Pour les deux méthodes itératives, le nombre maximal d'itérations est fixé à 100 itérations. Le temps de calcul est estimé en utilisant Matlab 2009 fonctionnant sur un ordinateur dual-core PC Pentium 3.4 GHz et de mémoire vive (RAM) de 1.00 GO.

Les résultats sont donnés dans le Tableau 3.4 pour le noyau Gaussien et le Tableau 3.5 pour le noyau polynomial. Il est évident que, indépendamment du type de noyau utilisé, les techniques itératives de pré-image telles que la descente du gradient et la méthode itérative du point-fixe exigent plus de temps de calcul que l'échelle multidimensionnelle MDS et l'approche conforme. Pour l'erreur quadratique moyenne, excepté le cas des données *Laser* avec le noyau Gaussien, les deux méthodes itérative du point-fixe et approche conforme présentent essentiellement l'erreur MSE la plus petite. Il est clair qu'en tenant compte du temps de calcul, l'approche conforme est le meilleur compromis.

3.6.3 Comparaison entre les différentes techniques proposées

Chaque série temporelle est décomposée en deux parties. Les premiers $n = 300$ échantillons sont utilisés lors de l'apprentissage, pour l'estimation de la valeur de l'ordre p optimal choisi dans

TABLE 3.4: Comparaison entre plusieurs techniques de pré-image appliquées sur différents types de séries temporelles, en utilisant un noyau Gaussien.

		<i>Laser</i>	<i>MG₃₀</i>	<i>Ikeda</i>	<i>Lorenz</i>
Gradient	σ	2^{-10}	2^{-6}	2^{-3}	2^3
	MSE	876.1293	0.0832	0.7187	150.0145
	temps (s)	3.0736	3.0953	6.0193	9.4399
Pt-fixe	σ	2^2	2	2^{10}	2^6
	MSE	16.5673	0.0162	0.5194	0.00035
	temps (s)	5.6392	8.2743	12.8256	34.2385
MDS	σ	2^{10}	2^{-10}	2^{10}	2^2
	MSE	11.5991	0.083	0.5825	99.2851
	temps (s)	0.1002	0.1608	0.1976	0.2735
Conforme	σ	2^5	2	2^3	2^2
	η	2^{-10}	2^{-10}	2^{-10}	2^{-10}
	MSE	17.1484	0.0166	0.5201	0.1079
	temps (s)	1.0172	0.9916	1.9249	2.3637

TABLE 3.5: Comparaison entre plusieurs techniques de pré-image, en utilisant le noyau polynomial.

		<i>Laser</i>	<i>MG₃₀</i>	<i>Ikeda</i>	<i>Lorenz</i>
Gradient	q	2	5	6	2
	η	2^{-10}	2^{-2}	2^{-12}	2^{-11}
	MSE	876.1293	0.1000	0.7187	339.3405
	temps (s)	1.9851	2.0025	3.7099	5.5710
Pt-fixe	q	5	2	2	5
	MSE	16.0169	0.0161	0.5246	0.007
	temps (s)	7.2244	8.1867	19.0268	23.6776
Conforme	q	2	2	2	2
	η	2^{-9}	2^{-10}	2^{-9}	2^{-10}
	MSE	18.6591	0.0160	0.5171	0.0025
	temps (s)	0.5113	0.4877	0.9250	1.2632

$p \in \{1, 2, \dots, 5\}$, les coefficients de l'expansion AR, et la largeur de bande σ du noyau Gaussien. Les 300 échantillons suivants sont utilisés pour évaluer la pertinence du modèle résultant, en considérant le calcul de l'erreur quadratique moyenne de prédiction.

Le Tableau 3.6 regroupe l'erreur quadratique moyenne de prédiction et le temps de calcul mesuré sur un Intel Core 2, avec une vitesse de 2.40 GHz et une mémoire vive de 1.00 GO, pour chacun des trois modèles AR non-linéaires proposés, ainsi que le modèle AR linéaire. Pour la technique de pré-image, le critère d'arrêt est donné par une borne inférieure sur la tolérance, qui vaut 10^{-6} , tout en limitant le

TABLE 3.6: Temps de calcul estimé et erreur quadratique moyenne de prédiction entre les valeurs prédictes et les vraies valeurs.

		MG_{30}	Lorenz	ECG	EMG-Neuro
Modèle AR linéaire (avec (3.1))	MSE	0.0655	0.2907	0.0332	0.1397
	Time (s)	0.0107	0.0539	0.0126	0.0133
Modèle AR-à-noyaux (pré-image) (avec (3.4))	MSE	0.00001	0.1498	0.0006	0.0001
	Time (s)	5.3201	10.5241	7.8921	9.0315
Modèle AR sur les valeurs du noyau (avec (3.9))	MSE	3.7466	7276.8	4.8224	2.8418
	Time (s)	0.0619	0.1975	0.0632	7.3417
Modèle AR hybride (avec (3.13))	MSE	0.0623	0.0213	0.0290	0.0061
	Time (s)	0.0861	0.6091	0.1834	0.1280

TABLE 3.7: L'erreur quadratique moyenne de prédiction pour différentes techniques non-linéaires de prédiction comparées aux méthodes proposées.

	MG_{30}	Lorenz
Perceptron multicouche	0.0461	0.2837
Régression à support vecteur	0.0313	0.1811
Filtre de Kalman non-linéaire	0.0307	0.3133
Modèle AR-à-noyaux (pré-image)	0.00001	0.1498
Modèle AR sur les valeurs du noyau	3.7466	7276.8
Modèle AR hybride	0.0623	0.0213

nombre maximal d'itérations à 50. Il est évident que le modèle AR-à-noyaux avec une technique de pré-image présente la meilleure précision. Cependant, une telle mise au point nécessite d'importantes ressources de calcul. Le modèle AR évalué sur les valeurs obtenues en calculant le noyau considéré n'est pas approprié pour ces séries temporelles. Le modèle AR hybride présente un bon compromis entre la précision et la complexité de calcul, et surpassé le modèle AR linéaire. Les résultats obtenus pour les trois séries temporelles sont illustrés dans la Figure 3.4.

Les configurations expérimentales sont les mêmes que celles utilisées dans [RDB05] pour les deux séries temporelles *Lorenz* et MG_{30} . Ces configurations nous permettent de faire une étude comparative avec les différentes techniques non-linéaires données par le papier cité auparavant, comme le perceptron multicouche, le régresseur à vecteurs supports, et le filtre de Kalman non-linéaire. Le Tableau 3.6.3 montre l'erreur quadratique moyenne évaluée sur chacune des séries temporelles en utilisant les différentes techniques non-linéaires de prédiction. Comme nous pouvons le voir, le modèle AR-à-noyaux avec une technique de pré-image présente la meilleure erreur quadratique pour la série MG_{30} , et le modèle AR hybride présente la meilleure erreur quadratique pour la série temporelle *attracteur de Lorenz*. Un modèle approprié est attribué à chaque série temporelle en fonction de la nature de ses données.

3.7 Conclusion

Dans ce chapitre, nous avons présenté une méthode autorégressive pour l'analyse et la prédition des séries temporelles. Deux méthodes ont été détaillées afin de déterminer les divers paramètres nécessaires pour l'implémentation de cette méthode. Nous avons montré qu'à l'aide des méthodes à noyaux, nous pouvons étendre l'usage de la technique AR pour les systèmes non-linéaires. Différents modèles AR-à-noyaux ont été alors proposés, dont une nécessitant la résolution du problème de la pré-image comme un schéma prédictif. Finalement, la méthode développée dans ce chapitre a été comparée avec succès à la méthode AR linéaire et aux méthodes du perceptron multicouche, du régresseur à vecteurs supports et du filtre de Kalman non-linéaire. A cette fin plusieurs types de séries temporelles ont été utilisés tel que des séries temporelles unidimensionnelles et chaotiques multidimensionnelles. Dans un premier temps, une comparaison avec d'autres techniques non-linéaires de prédition montre l'efficacité de la méthode proposée même si un nombre d'échantillons réduit est utilisé lors de l'apprentissage. Dans un second temps, une comparaison entre les différentes techniques de pré-image montre que l'approche conforme est la technique de pré-image qui présente un compromis entre la précision et le temps de calcul. Finalement, une comparaison entre les différentes techniques proposées ainsi qu'avec le modèle AR linéaire montre que le modèle AR-à-noyaux avec pré-image présente la meilleure prédition, au prix d'importantes ressources calculatoires. De plus, le modèle AR basé sur les valeurs obtenues en évaluant le noyau en considération n'était pas approprié pour les séries temporelles en question. De même, le modèle AR hybride est un bon compromis entre l'efficacité de prédition et le temps de calcul.

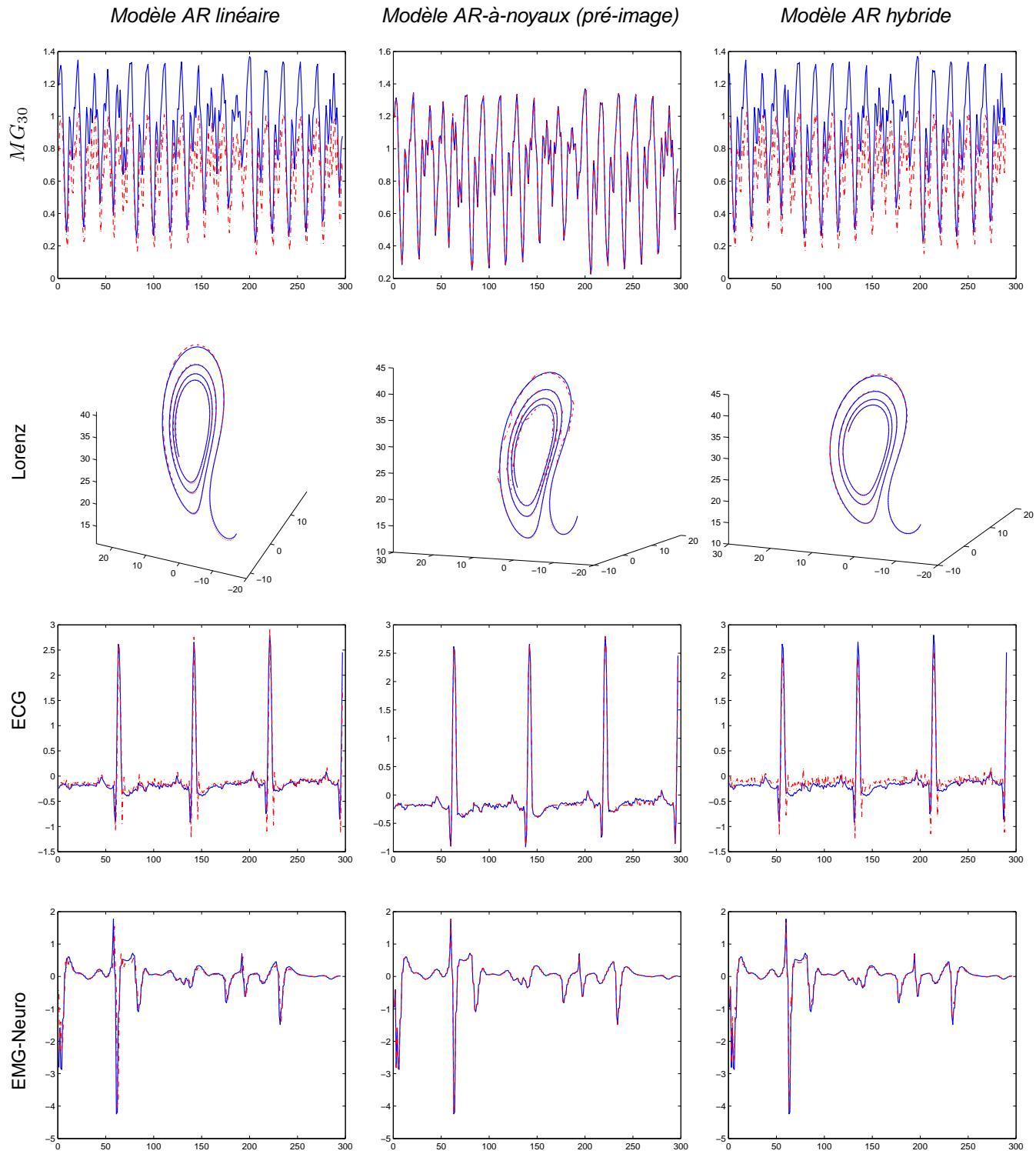


FIGURE 3.4: Visualisation des résultats obtenus lors de la prédiction des quatre séries temporelles en utilisant le modèle AR linéaire (première colonne), le modèle AR-à-noyaux avec pré-image (seconde colonne) et le modèle hybride (dernière colonne). Les séries temporelles prédites sont données par des lignes pointillées rouges, tandis que les lignes bleues définissent les données originales.

Étude de cas : les classifications binaire et multi-classes avec les méthodes à noyaux

Sommaire

4.1	Introduction	79
4.2	Classification binaire par SVM	80
4.3	Classification multi-classes	82
4.3.1	Un contre tous	82
4.3.2	Un contre un	83
4.4	Carte d'auto-organisation	84
4.5	Expérimentations	87
4.5.1	Critères d'évaluation de la classification	87
4.5.2	Classification binaire	88
4.5.3	Classification multi-classes	91
4.5.4	Carte d'auto-organisation	93
4.6	Conclusion	96

4.1 Introduction

L'apprentissage est l'acquisition de connaissances et de compétences permettant la synthèse d'information. Un algorithme d'apprentissage permet le passage d'un espace des exemples à un espace dit des hypothèses. Pour un ensemble de paramètres en entrée, l'apprentissage fournit un ensemble de résultats en sortie. Par exemple, par apprentissage, les personnes saines sont distinguées des personnes malades. Nous parlons alors de classification ou catégorisation. La classification est une opération de structuration qui vise à regrouper les données ayant des propriétés similaires. Chaque regroupement est dit une classe. Différentes techniques sont présentes pour la classification. La méthode la plus connue est celle des machines à vecteurs support.

L'idée des méthodes à noyaux est de plus en plus répandue suite à l'usage de ces machines à vecteurs support (*SVM* pour *Support Vector Machines*). Initialement introduites par Vapnik [Vap95] dans

le cadre de la théorie statistique de l'apprentissage, les SVM sont une méthode de classification binaire par apprentissage supervisé [BGV92a, CV95]. Cette méthode permet alors de discriminer les données par des algorithmes de traitement non-linéaires, *i.e.* linéaires dans un espace approprié. Depuis leur parution, les SVM sont utilisées notamment pour la régression [CV95], la multi-classification [WW99], la détection de nouveautés [SPST⁺01], et l'estimation de sorties multiples [EW02, PCCVSO⁺02]. Plusieurs domaines du traitement du signal ont bénéficié de l'application de ces algorithmes. Nous pouvons en citer la détection de visage [OFG97], la détection d'images tatouées (watermarking) [TW06], et même l'identification d'un locuteur [WC00] et la reconnaissance de texte [Joa00]. L'idée essentielle consiste à avoir recours à des espaces de Hilbert pour la discrimination des données. Décrise pour la première fois dans les années 1960 [VL63], cette idée consiste à déterminer l'hyperplan séparateur à marge maximale. Puisque le problème d'optimisation à résoudre est (convexe) quadratique, et ne souffre donc pas d'optima locaux, contrairement aux réseaux de neurones, cette méthode s'avère en outre particulièrement bien adaptée aux données de très grande dimension, telles que les images par exemple.

Un autre outil pour la classification est la carte d'auto-organisation. Initialement introduite par la carte de Kohonen [Koh82], la carte d'auto-organisation est présentée récemment en statistique comme une généralisation qui introduit une notion de voisinage entre les différentes classes [KSH01]. En d'autres-termes, la carte organise les classes selon leur proximité. Dans un premier temps, cette carte sera présentée comme un support graphique d'analyse du résultat de la classification en mettant en avant la variété des représentations possibles. Certaines sont plutôt performantes pour synthétiser un résultat, d'autres proposent une représentation des données de séries temporelles adaptée à leurs caractéristiques. Dans un second temps, cette carte est utilisée comme la représentation d'une surface susceptible de regrouper le nuage de points.

Dans ce chapitre, nous présentons une étude de cas portant sur la discrimination. Pour ce faire, nous étudions deux méthodes pour la classification des données en vue de la classification de signaux électrocardiogrammes. La première, basée sur l'apprentissage supervisé, est les machines à vecteurs support. Nous détaillons la classification binaire initialement traitée par les machines à vecteurs supports. La mise en œuvre de ces dernières pour une tâche de classification multi-classes est étudiée avec deux stratégies, qui sont : “un contre tous” et “un contre un”. La seconde, basée sur l'apprentissage non-supervisé, est la carte d'auto-organisation. Nous étudions alors l'apprentissage de la carte afin de faire une classification multi-classes. Les performances de ces techniques sont illustrées sur des signaux ECG pris de deux bases de données [BKS, GAG⁺a] pour détecter les signaux venant de personnes saines des signaux de personnes présentant une certaine arythmie.

4.2 Classification binaire par SVM

La classification est une action de discriminer les données d'entrée par classes ou par catégories. Cette structuration vise à organiser les données en des classes homogènes afin de faciliter l'analyse des informations. Une méthode bien connue pour la classification est les machines à vecteurs supports.

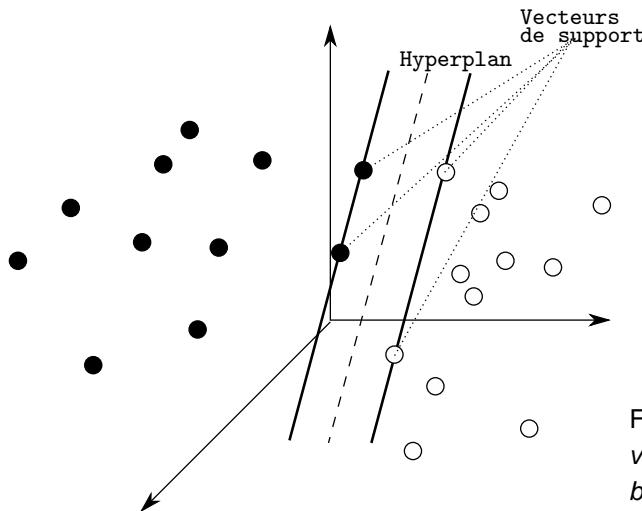
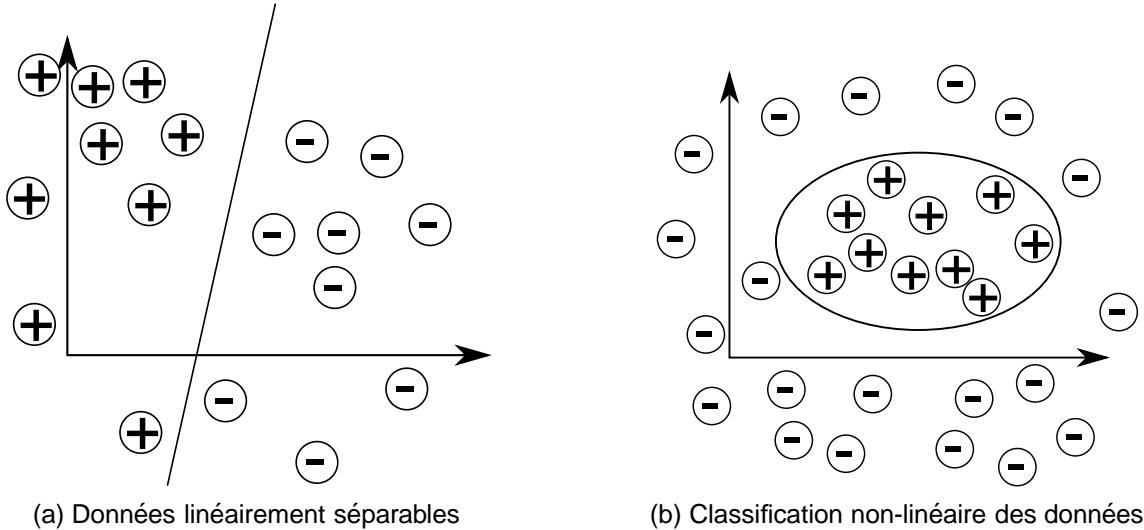


FIGURE 4.1: Schéma illustrant l'hyperplan avec les vecteurs de support permettant la classification binaire.

La théorie d'apprentissage statistique de Vapnik est la base de nouvelles méthodes d'apprentissage. En 1995, Cortes et Vapnik ont introduit les machines à vecteurs supports [CV95]. Voir aussi [BL07]. Les SVM sont utilisées pour la classification non-linéaire des données. Elles reposent sur deux propriétés qui expliquent leur succès : la première est celle de la marge maximale qui représente la distance entre la frontière de séparation et les échantillons les plus proches appelés vecteurs supports, et la deuxième est l'utilisation de fonction noyau permettant de transformer les données à un espace approprié à la séparation linéaire.

Les SVM font partie des méthodes d'apprentissage supervisé. Les échantillons d'apprentissage sont représentés par un ensemble de paires entrée/sortie où la sortie est une étiquette binaire pour une discrimination à deux classes. Le but est de construire une fonction à partir de ces exemples d'apprentissage qui peut prédire les sorties pour des entrées n'appartenant pas à l'ensemble des données d'apprentissage. Les entrées peuvent être des descriptions d'objets et les sorties sont les classes de ces objets données en entrées [Bur98]. Pour deux classes d'exemples donnés, le but de SVM est de trouver un classifieur séparant les données en maximisant la distance entre ces deux classes. Dans la plupart des problèmes réels, il n'y a pas de séparation linéaire possible entre les données. La mise en œuvre d'une transformation non-linéaire, par l'usage de noyau (voir chapitre 1), permet de contourner le problème. Pour les SVM, ce classifieur dans l'espace transformé est un classifieur linéaire appelé hyperplan. La Figure 4.1 montre l'hyperplan pour la classification entre deux classes. Les points les plus proches, qui seuls définissent l'hyperplan, sont appelés vecteurs supports. Plusieurs hyperplans permettent une séparation valide, mais les SVM considèrent l'hyperplan dont la distance aux exemples d'apprentissage est maximale. Cette distance est la “marge”. La Figure 4.2 montre deux exemples de classification binaire, la première dont les données sont linéairement séparables et la seconde non-linéairement séparables.

En SVM, le paramètre C dit de régularisation détermine le compromis entre la fraction de données d'apprentissages mal classées et la régularité de la solution. De plus, à part ce paramètre, le choix du noyau et de ses paramètres est crucial. Nous rappelons par exemple que le noyau Gaussien dépend

FIGURE 4.2: Exemples de classification binaire, linéaire et non-linéaire dans \mathbb{R}^2 .

de sa largeur de bande. Voir le Tableau 1.1 pour les expressions des noyaux les plus utilisés. Pour une application donnée, il est difficile de déterminer à l'avance quel type de noyau ou quels paramètres nous donnent les meilleurs résultats. Notre objectif est d'optimiser les performances de classification. A cet effet, la validation croisée à k -plis est utilisée afin de tenter cet objectif. La validation croisée à k -plis est utilisée pour évaluer les classificateurs SVM étant donné les paramètres d'un certain noyau ainsi que celui de régularisation. Elle consiste à partitionner équitablement les données d'apprentissage dans k plis, où $k - 1$ plis sont utilisés à chaque fois pour l'apprentissage, et le pli qui reste, nommé ensemble de validation, est utilisé pour les tests. De cette façon, toutes les données participent à la validation.

4.3 Classification multi-classes

En SVM, bien que les hyperplans séparateurs de marge maximale sont souvent développés pour les problèmes de discrimination binaire, il est nécessaire de les adapter pour traiter des problèmes multi-classes. L'idée est simplement de transformer le problème de classification de ℓ classes en plusieurs classificateurs binaires. Il existe deux stratégies de décomposition, “un contre tous” (One-Against-All OAA) et “un contre un” (One-Against-One OAO). Considérons un problème de ℓ classes, où nous avons n échantillons d'apprentissage, l'entrée est un ensemble $\{(x_1, y_1), \dots, (x_n, y_n)\}$ de vecteurs d'apprentissage $x_i \in \mathcal{X}$ et les étiquettes correspondantes $y_i \in \{1, 2, \dots, \ell\}$.

4.3.1 Un-contre-tous

Cette stratégie “un contre tous” (en anglais One-Against-All (OAA)), la plus simple et la plus ancienne stratégie de décomposition, a été introduite par Vapnik en 1995 [Vap95]. Cette approche utilise une architecture parallèle de ℓ classificateurs, un pour chaque classe. La formulation initiale de la méthode

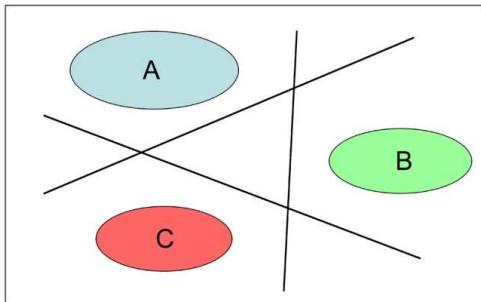


FIGURE 4.3: Schéma représentant les frontières binaires des régions OAA pour un problème fondamental.

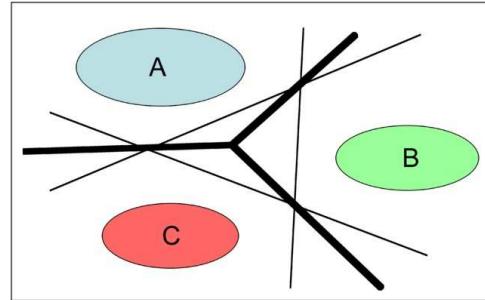


FIGURE 4.4: Schéma représentant les frontières continues des régions OAA pour un problème fondamental.

un contre tous déclare qu'une donnée serait classée dans une certaine classe si et seulement si le classifieur de la classe associée l'a acceptée et les classificateurs de toutes les autres classes l'ont rejetée. Alors que pour les classes de précision étroitement groupées, cette approche laisse des régions d'ambiguïté pour lesquelles plus d'une classe les acceptent ou toutes les classes les rejettent. La Figure 4.3 illustre cette formulation.

Une amélioration des performances de l'OAA a été proposée par Vapnik en 1998 [Vap98]. La solution la plus simple pour résoudre un SVM multi-classes est de le décomposer en un ensemble de sous-problèmes binaires et construire des SVM indépendants pour chacun d'eux. Cette stratégie, appelée “un contre tous” consiste en la construction d'un nombre de SVM égal au nombre de classes. Chaque SVM est ensuite entraînée pour séparer les données d'une classe étiquetée 1, de celles de toutes les autres classes qui sont étiquetées -1. Ainsi, chaque SVM est associée à une classe et sa sortie avant seuillage appartient à la classe. La règle de décision est l'application du principe “winner takes all”, elle est donc généralement utilisée pour répartir les données inconnues à la classe correspondant au classifieur avec la plus grande valeur de sortie [PC07, FHL08, MCS06]. La Figure 4.4 illustre cette idée.

4.3.2 Un contre un

Une autre stratégie de décomposition est “un contre un” (en anglais One-Against-One (OAO)), également connue sous le nom “couplage par paires”, “toutes les paires” ou “round robin” [MCS06]. Cette stratégie consiste en la construction d'un classifieur pour chaque paire de classes, c'est-à-dire $\ell(\ell - 1)/2$ classificateurs binaires pour un problème à ℓ classes. Chaque classificateur est entraîné pour séparer les données d'une classe de celles d'une autre classe. En combinant les règles de décision des différents (sous)-classificateurs, la règle de décision finale utilisée est généralement la méthode du vote majoritaire appelée “max-wins voting”. En d'autres termes, chaque classificateur vote pour une classe et l'échantillon étudié est finalement associé à la classe recevant le plus de votes [PC07, FHL08]. D'autres méthodes de combinaison de règles de décision comprennent l'utilisation de graphes de décision pour déterminer la classe sélectionnée de manière similaire à des tournois à élimination directe [Bur98]. La

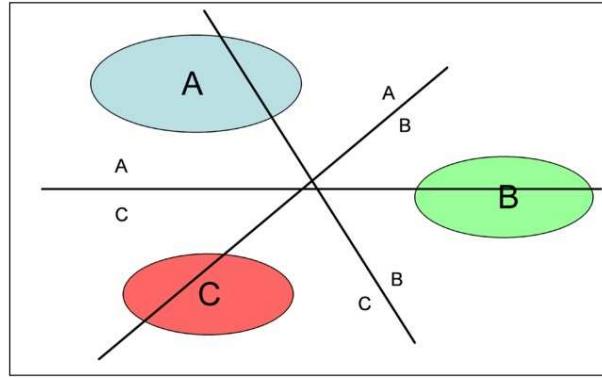


FIGURE 4.5: Schéma des frontières décisionnelles, deux à deux, pour la méthode de décomposition OAO dans le cas de trois classes.

Figure 4.5 illustre cette stratégie.

Les auteurs de [HL02] montrent que la stratégie “un contre un” a une meilleure précision que la stratégie “un contre tous”, mais dans toutes les comparaisons, le taux de précision reste inférieur à 2%. Même si la différence de précision est faible, il existe un argument plus important en faveur de la stratégie “un contre un”, qui n'est autre que le temps nécessaire pour l'apprentissage. Pour cette méthode, le temps est de 2 à 6 fois plus rapide que pour la stratégie “un contre tous”. Cette condition est due au nombre de données beaucoup plus élevé de chaque classifieur binaire de cette dernière stratégie.

4.4 Carte d'auto-organisation

Une carte d'auto-organisation (SOM pour *Self-Organizing Map*) est un type de réseaux de neurones artificiels qui est entraîné en utilisant l'apprentissage non supervisé pour produire une faible dimension (typiquement deux dimensions) pour la représentation discrétisée de l'espace d'entrée des échantillons d'apprentissage. Elle a été conçue comme une alternative aux réseaux de neurones traditionnels. Elle est utilisée pour des tâches similaires à celles des réseaux de neurones, citons par exemple : la reconnaissance des formes, la robotique, le contrôle de processus et même le traitement de l'information sémantique. La ségrégation spatiale des différentes réponses et de leurs organisations dans les résultats des sous-ensembles produisent un degré élevé d'efficacité dans les opérations typiques de réseaux de neurones. Les cartes d'auto-organisation diffèrent des autres réseaux de neurones artificiels dans le sens où elles utilisent une fonction de voisinage afin de préserver les propriétés topologiques de l'espace des observations. Elles sont considérées comme un outil d'analyse des données et de prise de décisions pour le pré-traitement et de sélection des algorithmes de classification. Les résultats formés par les SOM sont plus orientés vers l'utilisateur permettant une forte interaction avec l'utilisateur pour différentes tâches.

L'idée d'une carte d'auto-organisation a été initialement introduite par Kohonen [Koh82]. Mais ce n'est que très récemment qu'elles sont utilisées pour résoudre des problèmes de grande dimension et non-linéaires telles que l'extraction de caractéristiques et la classification des images et des

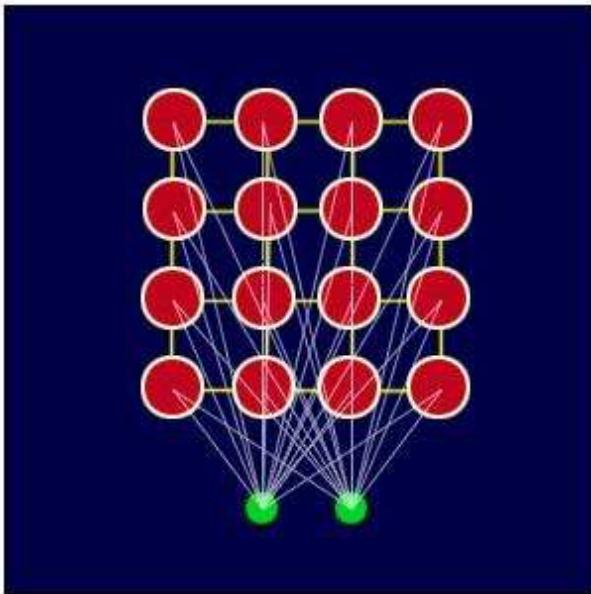


FIGURE 4.6: Carte simple de Kohonen.

modèles acoustiques, la commande adaptative de robots, la démodulation en transmission de signaux de télécommunication, ainsi que pour l'organisation de collections de documents très volumineux [KSH01]. La SOM est actuellement utilisée comme l'un des outils des réseaux de neurones génériques pour la visualisation de la structure des données à dimension élevée [Gac11]. Une carte d'auto-organisation se compose d'éléments appelés nœuds ou neurones. Un vecteur de pondération de la même dimension que les vecteurs de données d'entrées est associé à chaque nœud, qui possède une position bien précise dans la carte. La disposition normale des nœuds est un espacement régulier dans un réseau hexagonal ou rectangulaire, illustrée dans la Figure 4.6.

L'objectif de la SOM est de visualiser des données de dimensions élevées dans un espace de faible dimension, généralement placé dans un plan de deux ou trois dimensions. Pour que cette visualisation ait un sens, une exigence ultime est que cette représentation des données de grande dimension doit conserver les propriétés topologiques de l'ensemble de données. Cette implication signifie que deux données qui sont proches l'un de l'autre dans l'espace à haute dimension doivent préserver cette similarité (ou ressemblance) lors de leur représentation sur la carte. Comme la plupart des réseaux de neurones artificiels, la SOM fonctionne en deux modes : l'apprentissage et la correspondance. Durant l'apprentissage, la carte est construite à partir des échantillons d'entrées. Il s'agit d'un processus compétitif, également appelé quantification vectorielle. La procédure de la mise en place d'une entrée de l'espace des données sur la carte est de trouver le premier nœud ayant le vecteur le plus proche de celui avec le poids pris de l'espace des données. Une fois que l'apprentissage est réalisé, la correspondance classifie automatiquement une nouvelle entrée dans la classe appropriée.

Phase d'apprentissage

L'objectif de l'apprentissage de la carte d'auto-organisation est de provoquer les différentes parties du réseau pour répondre de manière similaire à certains modèles d'entrées. Chaque nœud possède une position spécifique topologique (une coordonnée dans le treillis) et contient un vecteur de coefficients de pondération de la même dimension que les vecteurs d'entrée. En d'autres termes, chaque nœud contient alors un vecteur de pondération w , de même dimension que les données d'entrée. À partir d'une distribution initiale de poids aléatoires, et sur plusieurs itérations, la SOM établit finalement une carte avec des zones stables. Chaque zone est effectivement un classifieur de caractéristiques, de sorte que la sortie graphique devient un type de carte de caractéristiques de l'espace des observations. Tous les nouveaux vecteurs d'entrée présentés au réseau stimuleront les nœuds dans la zone des vecteurs de poids similaires. L'algorithme pour l'apprentissage de la carte est donné dans l'algorithme 4.1. Lors de la pondération du nœud, celui gagnant est communément connu sous le nom de l'unité correspondant le mieux (*Best Matching Unit ou BMU*). À partir du BMU, le rayon de son voisinage est alors calculé. Il s'agit d'une valeur initialement importante, typiquement réglée au "rayon" de la grille, mais qui diminue à chaque itération. Plus un nœud est proche du BMU, plus son poids se modifie.

```

Initialization : Poids de chaque nœud  $\leftarrow$  valeur aléatoire ;
pour  $t \leftarrow 1$  à  $nbr\ iteration$  faire
    Choix au hasard d'un vecteur dans l'ensemble de données d'apprentissage ;
    Présentation du vecteur au réseau ;
    Examination du nœud pour calculer une pondération la plus proche du vecteur d'entrée ;
    Recherche de tous les nœuds trouvés dans le rayon de voisinage de BMU ;
    Ajustement de chaque poids du nœud voisin pour les rendre similaires au vecteur d'entrée ;
fin
```

Algorithme 4.1: Algorithme d'apprentissage de la carte d'auto-organisation.

Identification du BMU

Pour déterminer l'unité correspondant le mieux, la méthode consiste à parcourir tous les nœuds et calculer la distance Euclidienne entre le vecteur poids de chaque nœud et l'entrée utilisée. Nous désignons par w_k le vecteur poids du $k^{\text{ième}}$ nœud. Soit x_i l'échantillon sélectionné aléatoirement à l'itération t courante. Le nœud avec le vecteur poids le plus proche de l'échantillon courant est identifié comme étant la BMU, en minimisant la distance Euclidienne, selon

$$\min_k \|w_k - x_i\|^2.$$

Le vecteur poids w_k de chaque nœud k est alors ajusté s'il est dans le voisinage de la BMU, en l'adaptant à

$$w_k + \eta_t \kappa_G(w_{BMU}, w_k)(x_i - w_k),$$

où η_t est le taux d'apprentissage qui diminue à chaque itération selon l'équation suivante

$$\eta_t = \eta_0 \exp\left(\frac{-t}{\lambda}\right),$$

où λ est une constante du temps. Dans cette expression le noyau Gausien $\kappa_G(\mathbf{w}_{BMU}, \mathbf{w}_k)$ représente la quantité d'influence que la distance du $k^{\text{ième}}$ nœud à la BMU a sur son apprentissage, suivant l'équation

$$\kappa_G(\mathbf{w}_{BMU}, \mathbf{w}_k) = \exp\left(\frac{-\|\mathbf{w}_{BMU} - \mathbf{w}_k\|^2}{2\sigma_t^2(t)}\right)$$

Une caractéristique unique de l'algorithme d'apprentissage de Kohonen est que la zone du voisinage se rétrécit à chaque itération. Cette caractéristique est accomplie en réduisant le rayon du voisinage avec le temps. Pour ce faire, une fonction décroissante exponentielle est utilisée pour la largeur de bande du noyau Gaussien ci-dessus, selon :

$$\sigma_t = \sigma_0 \exp\left(\frac{-t}{\lambda}\right),$$

où σ_0 indique la largeur du treillis à l'itération initiale, et en pratique $\lambda = \frac{\text{nombre total d'itérations}}{\log(o_0)}$. Le voisinage se rétrécit au cours des itérations, pour ne contenir que la BMU.

4.5 Expérimentations

Dans cette section, nous présentons l'efficacité de la classification avec des méthodes à noyaux. Pour ce faire, nous considérons un enregistrement d'électrocardiogramme (ECG), pris de [GAG⁺a, MM01]. Cet enregistrement est découpé en n segments, chacun comportant un battement commençant avant l'onde P [CLS⁺07]. Il est important de noter que l'ECG est un signal irrégulier qui n'a pas de période constante de sorte que le nombre d'échantillons des différents battements n'est pas unique. Pour contourner ce problème, nous complétons les segments courts avec la moyenne des derniers échantillons du segment [CCSK04].

4.5.1 Critères d'évaluation de la classification

Après un test, l'une des quatre conclusions est tirée : TP qui représente un vrai positif (*True Positive*), FP pour faux positif (*False Positive*), FN désigne faux négatif (*False Negative*) et TN pour vrai négatif (*True Negative*). Si un battement anormal est classifié comme anormal, alors on dit que le battement est classé TP. Tout battement normal qui est classifié comme un battement anormal par erreur donne alors un résultat FN. Finalement, tout battement normal classifié comme normal produit un TN. Alors, TN et FP représentent des personnes saines et TP et FN désignent des personnes malades.

Quatre critères d'évaluation sont utilisés dans la littérature pour évaluer la performance de chacune des techniques de classification et d'extraction de caractéristiques : la spécificité, sensibilité, précision et prédictivité positive. La spécificité est le rapport de personnes que le test considère sain parmi tous ceux

TABLE 4.1: Les valeurs optimales des paramètres pour chacune des fonctions noyaux.

Noyau	C	paramètre du noyau
Linéaire	40	-
Polynomial	1	$q = 10$
Gaussien	100	$\sigma = 0.8$

qui sont vraiment en bonne santé. Une spécificité élevée indique qu'il y a une erreur de classification mineure et moins de gens en bonne santé sont considérés comme malades. L'erreur produite par la classification erronée est appelée erreur de première espèce (erreur de type I). La sensibilité est le rapport de personnes que le test considère comme malades parmi tous ceux qui sont vraiment malades. Une sensibilité élevée indique qu'il y a une erreur de classification mineure et moins de malades sont considérés comme sains. L'erreur produite par la classification erronée est appelée erreur de second type (erreur de type II). L'erreur de type II est plus dangereuse que l'erreur de type I puisque dans ce cas une personne malade est considérée en bonne santé. Son traitement sera ignoré conduisant à de graves problèmes. La précision est définie comme le rapport de battements classés correctement par rapport au nombre total de battements. La prédictivité positive est le pourcentage des personnes ayant un test positif et qui sont malades. Les erreurs de classification SVM peuvent souvent se produire à proximité des frontières de décision, où les classes sont proches les unes des autres. Ces critères sont définis comme suit :

1. Précision = $\frac{TP+TN}{TP+TN+FP+FN} * 100$
2. Sensibilité = $\frac{TP}{TP+FN} * 100$
3. Prédictivité Positive = $\frac{TP}{TP+FP} * 100$
4. Spécificité = $\frac{TN}{TN+FP} * 100$

4.5.2 Classification binaire

Dans cette partie, la classification binaire est prise en compte. Deux expérimentations sont faites. La première tient compte des observations initiales tandis que la deuxième est appliquée sur les signaux après avoir réduit leur dimension.

4.5.2.1 SVM binaire appliquée sur les observations initiales

Les fonctions noyaux utilisées sont, linéaire, polynomial et Gaussien. Le Tableau 4.1 résume les valeurs optimales des paramètres utilisés. Comme nous pouvons voir dans le Tableau 4.2, les résultats obtenus avec le classifieur Gaussien sont mieux que ceux obtenus avec les classificateurs à noyaux linéaire ou polynomial. Cette expérience révèle la supériorité de la classification SVM basée sur le noyau Gaussien par rapport à d'autres techniques lors d'un traitement sur les données initiales transformées dans

TABLE 4.2: Performance de la classification SVM binaire.

Méthode	Précision	Sensibilité	Spécificité	Prédicтивité Positive
SVM linéaire	54	33	40	40
SVM polynomial	63	50	80	75
SVM Gaussian	68	58	80	77

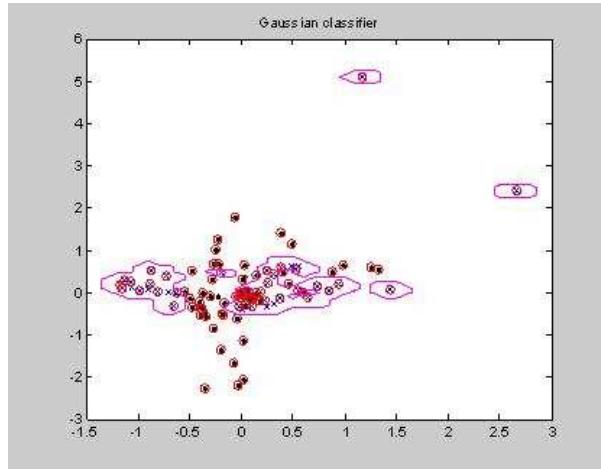


FIGURE 4.7: Résultats de l'application du classifieur Gaussian.

un espace fonctionnel de grande dimension. Les résultats de l'application du classifieur Gaussian sur les données sont présentés dans la Figure 4.7.

4.5.2.2 Classification basée sur la réduction des caractéristiques

Nous passons maintenant à l'application de la classification après avoir réduit les caractéristiques du signal ECG.

Réduction des caractéristiques avec l'ACP

L'ACP détermine les axes principaux qui rendent compte du maximum de variance des données. Pour ce faire, les vecteurs propres ayant la plus grande variance sont retenus, voir la Section 1.4.1. La Figure 4.8 illustre les résultats de l'application de la classification SVM binaire combinée avec l'ACP.

L'algorithme 4.2 est appliqué pour la classification binaire avec l'analyse en composantes principales pour lequel nous avons choisi le noyau Gaussian puisqu'il s'est révélé dans l'expérience précédente qu'il est la fonction noyau la plus adéquate pour la classification des ECG. L'estimation des paramètres du classifieur est faite via la méthode de la validation croisée à 5-plis comme illustré dans la Figure 4.8.

Réduction des caractéristiques avec l'ACP-à-noyaux

Passons maintenant à la réduction des caractéristiques avec l'ACP-à-noyaux. Comme le signal ECG a

Application de l'ACP sur les données d'apprentissage ;
 Choix du classifieur SVM basé sur le noyau Gaussien ;
 Apprentissage de ce classifieur sur les composantes principales ;
 Estimation des paramètres du noyau Gaussien avec la méthode de la validation croisée à 5-plis;
 Application de l'ACP sur les données du test;
 Adoption du modèle d'apprentissage pour la classification des échantillons du test;

Algorithme 4.2: Algorithme de l'ACP-SVM.

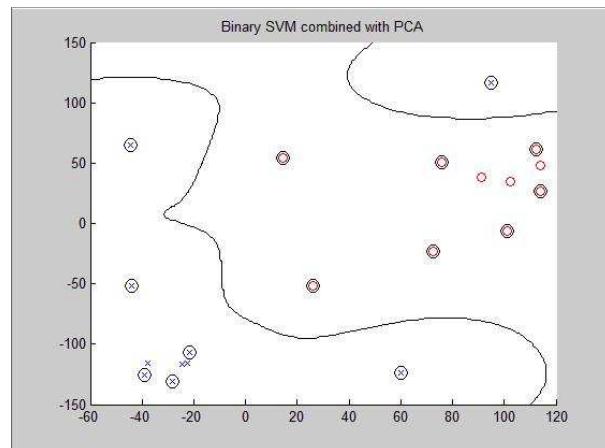


FIGURE 4.8: Résultats de l'application de la classification SVM binaire combinée avec l'ACP.

beaucoup de caractéristiques non-linéaires et puisque l'ACP extrait uniquement celles linéaires, nous avons adopté l'ACP-à-noyaux pour extraire les composantes non-linéaires. L'algorithme ACP-à-noyaux SVM est le même que ACP-SVM avec une seule différence, c'est que nous avons appliqué l'ACP-à-noyaux comme méthode d'extraction de caractéristiques.

Pour trouver le nombre optimal de composantes principales extraites, nous exécutons le même algorithme chaque fois pour une nouvelle autre dimension. Cette dimension réduite varie de un au nombre de variables en augmentant une composante à chaque étape. La valeur qui génère l'erreur minimale sur l'ensemble de validation est adoptée. Les meilleurs résultats obtenus dans SVM en utilisant l'entrée d'origine, sans extraction de caractéristiques, et en utilisant l'ACP avec les paramètres ($C = 40$ et $\sigma = 10$ pour les SVM) et l'ACP-à-noyaux avec les paramètres ($C = 10$ et $\sigma = 0,5$ pour les SVM) pour l'extraction des caractéristiques sont donnés dans le Tableau 4.3. Ce Tableau montre que la classification SVM couplée à l'extraction de caractéristiques à l'aide de l'ACP ou de l'ACP-à-noyaux présente une précision supérieure à la classification SVM sans extraction de caractéristiques. De plus, la précision de l'ACP-à-noyaux est meilleure que l'ACP classique, et elle est due à des composantes non-linéaires extraites par la méthode à noyaux.

La Figure 4.9 illustre les résultats de l'application de la classification SVM binaire combinée avec l'ACP-à-noyaux. Nous pouvons donc conclure que pour la classification binaire des signaux ECG, la méthode la plus appropriée est de combiner les SVM avec l'ACP-à-noyaux.

TABLE 4.3: Performance de la classification SVM binaire combinée avec l'ACP et l'ACP-à-noyaux.

Méthode	Précision	Sensibilité	Spécificité	Prédicтивité Positive
SVM-Gaussien	68	58	80	77
ACP-SVM	85	80	88	90
ACP-à-noyaux-SVM	95	100	90	90

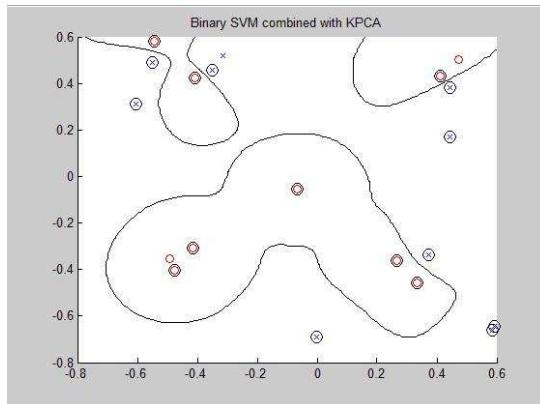


FIGURE 4.9: Résultats de l'application de la classification SVM binaire combinée avec l'ACP-à-noyaux.

4.5.3 Classification multi-classes

Nous passons maintenant à la classification multi-classes des signaux ECG. Pour ce faire, nous avons appliqué les classifieurs SVM multi-classes mentionnés dans la Section 4.3 sur les signaux ECG. Par conséquent, le signal lui-même est divisé en 3 classes. Dans notre étude, nous avons examiné trois différentes classes : une classe pour le cas normal et deux classes pour les deux différents cas anormaux, contraction ventriculaire prématûrée (*Premature Ventricular Contraction PVC*) et bloc de branche gauche (*Left Bundle Branch Block LBBB*). Un PVC est une extrasystole impliquant les ventricules du cœur, produisant parfois des palpitations accompagnantes, cependant un bloc de branche gauche est l'échec de l'impulsion cardiaque de se propager vers le bas de la branche gauche, résultant en une activation précoce de la partie droite du septum et un myocarde ventriculaire droit.

Notre objectif est d'optimiser les performances des classifieurs et d'atteindre un pourcentage de bonne classification plus élevé. Les entrées de l'algorithme sont les données d'apprentissage, le paramètre σ du noyau Gaussien et le paramètre de régularisation C . Nous avons testé différentes valeurs des paramètres avec la recherche de la meilleure solution en appliquant la validation croisée à 5-plis. Pour la classification binaire, nous avons montré que le noyau Gaussien présente de meilleurs résultats que les autres noyaux lorsqu'il s'agit de données initiales projetées dans un espace fonctionnel, en considérant les valeurs retrouvées pour la précision, la sensibilité, la spécificité et la prédictivité positive. Ainsi, dans la classification multi-classes, nous avons également examiné le noyau Gaussien. Après avoir appliqué la méthode de validation croisée à 5-plis pour le noyau Gaussien, nous avons trouvé que pour le paramètre de régularisation C de valeur égale à 100 et une largeur de bande σ égale à 1, la précision est maximale.

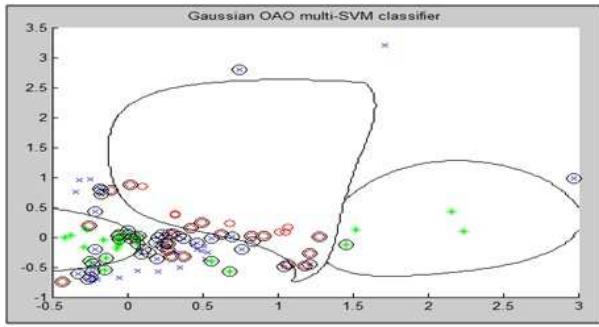


FIGURE 4.10: Classifieur SVM avec un noyau Gaussien en utilisant la stratégie OAO.

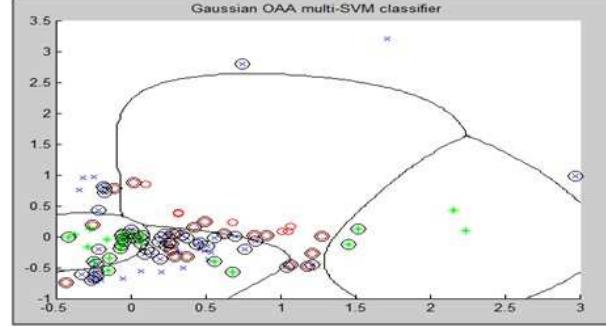


FIGURE 4.11: Classifieur SVM avec un noyau Gaussien en utilisant la stratégie OAA.

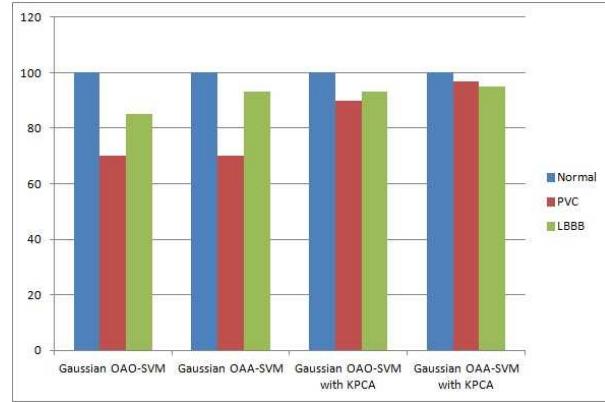


FIGURE 4.12: Performances des classifieurs SVM avec un noyau Gaussien pour les stratégies OAO et OAA, ainsi que les performances lorsque les stratégies sont combinées à l'ACP-à-noyaux.

Lors de la validation, nous avons utilisé 10 signaux normaux, 10 signaux anormaux présentant de PVC et 14 signaux avec l'anomalie LBBB. Nous avons appliqué la classification multi-classes pour les signaux ECG dans deux cas : sans extraction de caractéristiques et après extraction de caractéristiques. Les Figures 4.10 et 4.11 montrent la classification multi-classes avec un noyau Gaussien pour les stratégies de décomposition citées auparavant qui sont le “un-contre-tous” (OAA) et le “un-contre-un” (OAO).

Classification multi-classes SVM après extraction des caractéristiques

Nous étudions la combinaison de l'ACP-à-noyaux avec les SVM. Après avoir appliqué la méthode de validation croisée à 5-plis pour le noyau Gaussien, nous avons obtenu que pour le paramètre de régularisation C égal à 100 et une largeur de bande σ égale à 0.1, la précision est maximale. Les battements normaux présentent une précision de 100% dans les deux stratégies OAO et OAA en combinaison avec l'ACP-à-noyaux, les battements de cœur avec PVC ont une précision de 90% dans OAO et 97.34% pour OAA, et les battements de cœur avec LBBB ont une précision 92.65% pour OAO et 94.85% pour la stratégie de OAA. Ces résultats illustrent la supériorité de l'ACP-à-noyaux avec OAA lorsqu'il s'agit d'une classification multi-classes. La Figure 4.12 représente le pourcentage de signaux ECG correctement classés dans chaque catégorie avec et sans extraction de caractéristiques. Les batte-

TABLE 4.4: Liste faisant correspondre chaque couleur à son arythmie correspondante pour les signaux ayant réduit leur dimension.

Couleur	Arythmie
Rouge	Bloc de branche gauche
Vert	Normal
Bleu	Pacemaker
Cyan	Hypertrophie
Magenta	Bloc de branche droite
Noir	Cardiomyopathie
Orange	Valvulopathie cardiaque
Jaune	Infarctus du myocarde
Vert foncé	Extrasystole auriculaire

ments normaux sont entièrement séparés des autres battements de cœur anormaux avec une précision de 100%, les battements de cœur avec PVC ont une précision de 70% dans les deux stratégies OAO et OAA, et les battements de cœur avec LBBB présentent une précision de 85.71% pour OAO et 92.85% pour la stratégie de OAA.

4.5.4 Carte d'auto-organisation

Passons maintenant à une classification multi-classes en utilisant la carte d'auto-organisation. Deux expérimentations sont réalisées. La première tient compte de la carte d'auto-organisation appliquée directement sur les signaux ECG, tandis que la seconde implique l'utilisation de l'ACP-à-noyaux sur les signaux avant de les faire entrer dans l'algorithme de construction de la carte.

4.5.4.1 SOM appliquée sur les signaux ECG

Les signaux ECG sont pris de deux bases de données : Massachusetts Institute of Technology [GAG^{+a}] et Physikalisch-Technische Bundesanstalt [BKS]. L'ensemble de données est une matrice de dimensions 63×42 composé de neuf classes : Normal (N), bloc de branche gauche (*Left Bundle Branch Block LBBB*), bloc de branche droite (*Right Bundle Branch Block RBBB*), pacemaker (*Pace Beat PB*), infarctus du myocarde (*Myocardial Infarction MI*), hypertrophie (Hyp), cardiomyopathie (Card), valvulopathie cardiaque (*Valvular Heart Disease Valv*) et extrasystole auriculaire (*Atrial Premature Beat Atr*). Le Tableau 4.4 montre la correspondance couleur dans la SOM à une arythmie cardiaque bien donnée. Afin de bien choisir la dimension et le nombre d'itérations les plus appropriés pour cette application, différents tests ont été réalisés. Le Tableau 4.5 illustre ces différentes possibilités. À partir de ce Tableau, nous avons choisi 40×40 comme dimension pour la carte et le nombre d'itérations est fixé à 20 000. L'erreur d'apprentissage est alors 0.000625%.

La Figure 4.13 montre le positionnement des maladies représentées chacune par une couleur. Grâce à tous les tests réalisés, nous avons remarqué que les nœuds bleus qui représentent les signaux de

TABLE 4.5: Valeurs de l'erreur d'apprentissage pour différentes dimensions et différents nombres d'itérations.

Dimension ($o \times o$)	Nombre d'itérations	Erreur d'apprentissage (%)
30	30000	0.0022
30	40000	0.0010
40	10000	0.0133
40	20000	0.000625
40	30000	0.0000138
40	40000	0.00000068

pacemaker sont toujours entourés par les nœuds rouges qui représentent les signaux avec un bloc de branche gauche et les nœuds de couleur magenta indiquant les signaux avec un bloc de branche droite. Il s'agit d'une mise en place correcte, puisque le pacemaker peut être placé soit sur le côté droit soit sur le côté gauche de la poitrine. En d'autres termes, s'il est placé sur le côté gauche de la poitrine, il provoque un signal RBBB, et un signal LBBB s'il est placé sur le côté droit de la poitrine. Tandis que s'il est centré sur la poitrine, nous obtenons un signal de battement rythmé avec pacemaker. La Figure 4.13 illustre les résultats obtenus lorsque les entrées sont les signaux ECG introduits suivant la liste ci-dessous :

1. Extrasystole auriculaire
2. Pacemaker
3. Bloc de branche droite
4. Bloc de branche gauche
5. Hypertrophie
6. Infarctus du myocarde
7. Normal
8. Valvulopathie cardiaque
9. Cardiomyopathie

4.5.4.2 ACP-à-noyaux et SOM appliquées sur les signaux ECG

Dans cette partie, nous avons appliqué l'ACP-à-noyaux afin de réduire la dimension des signaux ECG. Une fois que la dimension est réduit, la carte d'auto-organisation est utilisée sur les signaux obtenus. La correspondance couleur et arythmie est donnée par le Tableau 4.4.

Plusieurs tests ont été réalisés afin de choisir la dimension et le nombre d'itérations adéquats pour être utilisés sur les signaux à dimension réduite. Le Tableau 4.7 montre ces différents tests. Nous avons alors choisi de prendre une dimension de 25×25 pour la carte et le nombre d'itérations fixé à 50 000. L'erreur d'apprentissage est donnée par la moyenne de la distance qui sépare chaque nœud k de son voisinage, pour ce cas, elle vaut 0.000625%. La carte ainsi formée est donnée par la Figure 4.15. Pour

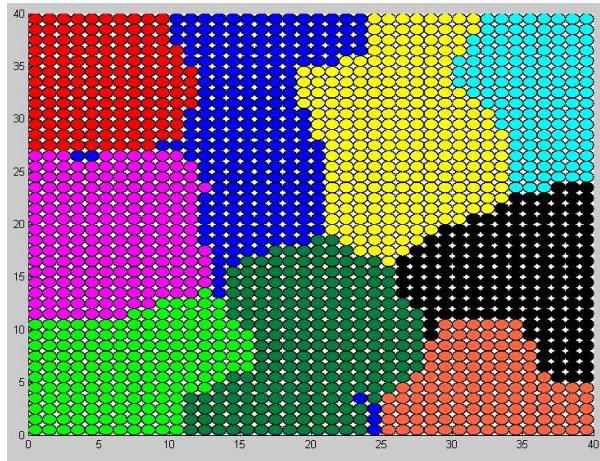


FIGURE 4.13: Carte SOM pour les données d'apprentissage

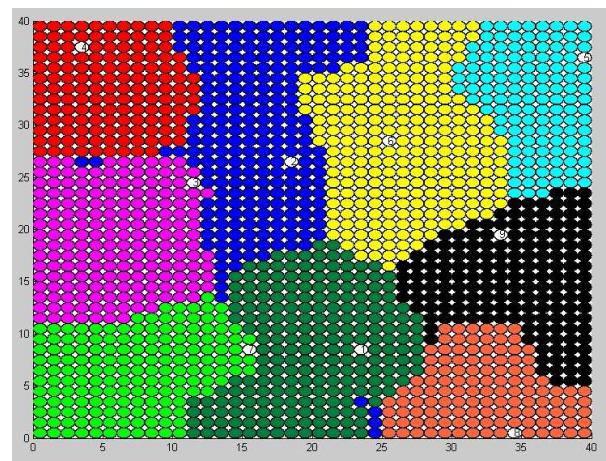


FIGURE 4.14: Résultats du test appliquée sur la carte SOM des ECG

TABLE 4.6: Valeurs de l'erreur d'apprentissage pour différentes dimensions et différents nombres d'itérations pour les signaux ayant réduit leur dimension.

Dimension ($o \times o$)	Nombre d'itérations	Erreur d'apprentissage (%)
20	20000	0.2250
20	25000	0.1345
20	40000	0.0572
25	30000	0.00140
25	50000	0.00063
30	20000	0.000053
30	40000	0.00004

valider cette approche, nous avons testé la carte en choisissant différents signaux ECG. La Figure 4.16 montre les résultats du test appliquée suivant la liste ci-après :

1. Bloc de branche droite
2. Normal
3. Extrasystole auriculaire
4. Bloc de branche gauche
5. Infarctus du myocarde
6. Hypertrophie
7. Cardiomyopathie
8. Pacemaker
9. Valvulopathie cardiaque

La SOM combinée à l'ACP-à-noyaux permet la plus grande précision même sur un ensemble de 100 signaux utilisés pour le test. Ces signaux sont répartis en différentes catégories, sains ou anormaux, où

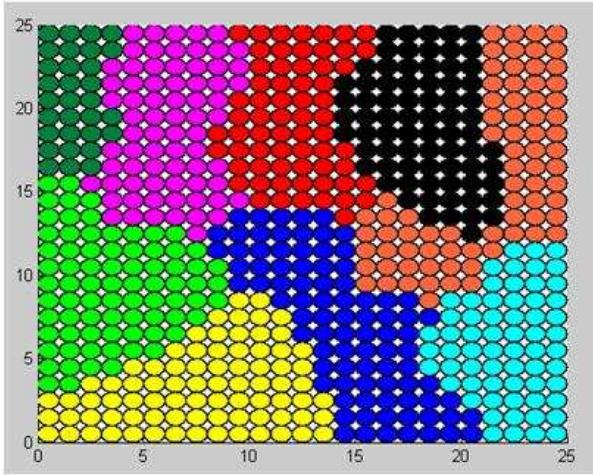


FIGURE 4.15: Carte SOM pour les données d'apprentissage après réduction de dimension.

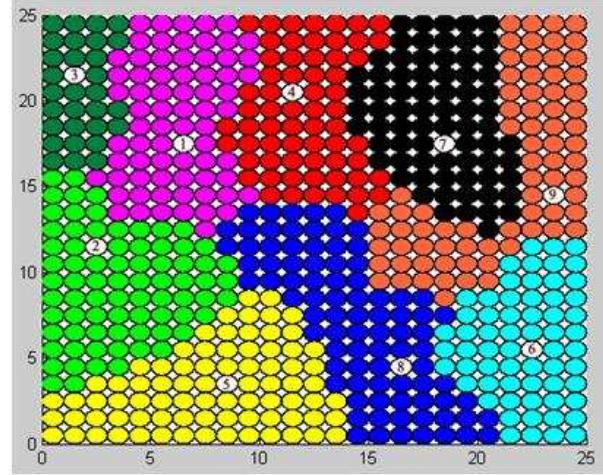


FIGURE 4.16: Résultats du test appliqués sur la carte SOM après réduction de dimension.

TABLE 4.7: Résultats de la classification des signaux avec le taux de précision.

Classe du signal	Correct	Faux	Place erronée	Total	Précision
Bloc de branche gauche	15	0	—	15	100%
Normal	16	0	—	16	100%
Pacemaker	10	0	—	10	100%
Infarctus du myocarde	10	0	—	10	100%
Hypertrophie	10	0	—	10	100%
Cardiomyopathie	15	0	—	15	100%
Bloc de branche droite	4	0	—	4	100%
Valvulopathie cardiaque	10	0	—	10	100%
Extrasystole auriculaire	10	0	—	10	100%

les anomalies sont les huit cas cités auparavant. Le Tableau 4.7 regroupe les résultats de la classification des signaux tout en indiquant le pourcentage de la précision.

Le Tableau 4.8 regroupe une synthèse générale des résultats obtenus lors de cette étude de cas, en appliquant les classifications binaire et multi-classes. Comme nous pouvons le voir, le noyau Gaussien est le plus utilisé dans de telles études. Le noyau Gaussien est plus adapté pour l'étude de signaux en tenant compte de sa largeur de bande. Lorsque nous appliquons une réduction de caractéristiques, nous obtenons de meilleurs résultats. Cette synthèse est valable dans le cas de classification binaire ainsi que dans le cas de classification multi-classes.

4.6 Conclusion

Dans ce chapitre, nous avons montré la puissance des méthodes à noyaux dans le domaine de la classification des signaux biomédicaux. Deux techniques de classification ont été couplées aux méthodes

TABLE 4.8: Synthèse générale de l'étude de cas sur la classification.

Classification	Noyau et spécification	Précision pre (%)
Binaire	Gaussien	$pre = 68$
Binaire avec réduction	Gaussien avec ACP-à-noyaux	$pre = 95$
Multi-classes	Gaussien avec stratégie Un-Contre-Tous	$70 < pre < 100$
Multi-classes avec extraction	Gaussien avec stratégie Un-Contre-Tous combiné à l'ACP-à-noyaux	$95 < pre < 100$
Carte d'auto-organisation	Carte d'auto-organisation avec l'ACP-à-noyaux	$pre > 97$

à noyaux : les machines à vecteurs supports (SVM) et la carte d'auto-organisation (SOM). Les SVM sont une méthode de classification basée sur l'apprentissage supervisé tandis que la SOM est basée sur l'apprentissage non-supervisé. Nous avons introduit pour les SVM les classifications binaire et celle multi-classes avec deux stratégies “un-contre-tous” (OAA) et “un-contre-un” (OAO), ainsi que l'algorithme de la SOM pour une classification multi-classes. Les expérimentations pour séparer les signaux des personnes saines de celles présentant des anomalies montrent que les SVM avec réduction des caractéristiques présentent une précision meilleure que le cas sans réduction, plus spécifiquement les SVM avec l'ACP-à-noyaux classent les données mieux que le cas avec l'ACP linéaire. De plus pour une classification multi-classes des arythmies cardiaques, la stratégie OAA est meilleure que celle OAO, et la précision de ces deux techniques augmente une fois elles sont couplées à l'ACP-à-noyaux. Finalement, après avoir réduit la dimension des signaux ECG de l'espace des observations, la SOM présente une classification parfaite de ces données avec une cartographie de dimensions réduites par rapport au cas sans réduction de dimensions.

Conclusion générale et perspectives

Durant les deux dernières décennies, nous avons assisté à une prolifération des méthodes à noyaux grâce à la diversité des traitements non-linéaires qu'elles autorisent avec un faible coût calculatoire. Un élément fondamental, assurant le succès de ces méthodes, est l'*astuce du noyau*. Le principe clé réside dans l'interprétation d'un noyau défini positif comme un produit scalaire dans un espace transformé. Ainsi un tel noyau assure-t-il le passage des données de l'espace des observations à l'espace dit de Hilbert à noyau reproduisant, sans la nécessité d'exhiber la fonction de transformation non-linéaire associée. L'objectif de ce mémoire est de proposer dans ce cadre des méthodes pour l'extraction des caractéristiques, l'analyse de séries temporelles et la classification.

Dans un premier temps, nous avons présenté la théorie des noyaux reproduisants tout en décrivant leurs caractéristiques, et en précisant les deux éléments fondamentaux des méthodes à noyaux qui sont l'*astuce du noyau* et le théorème de représentation. L'Analyse en Composantes Principales (ACP)-à-noyaux a été introduite après avoir décrit l'ACP classique qui échoue au traitement des données non-linéaires. Ensuite, l'ACP-à-noyaux a été présentée pour la reconnaissance des formes, en vue de l'extraction des caractéristiques et le débruitage. Pour ces applications, le retour inverse de l'espace fonctionnel à l'espace des observations est exigé. Pour ce faire, nous avons défini ce problème, dit de la *pré-image*. L'état de l'art sur les différentes techniques de résolution de ce problème a été effectué, et nous avons proposé une écriture de la pré-image sous la forme d'une combinaison des données disponibles.

Dans un second temps, il nous a paru indispensable de rajouter des contraintes qui sont liées à la physiologie des données, en particulier la contrainte de non-négativité. Nous avons alors proposé la résolution du problème de pré-image sous contraintes de non-négativité. Pour ce faire, deux approches sont introduites. La première impose la non-négativité de la pré-image elle-même, tandis que la seconde l'impose sur l'additivité des contributions qui sont les coefficients d'un modèle décrivant une solution de la pré-image. Cette dernière approche illustre le principe de parcimonie des résultats. L'approche proposée a été appliquée pour l'extraction des composantes de signaux électroencéphalogrammes, et le débruitage de données artificielles et de chiffres manuscrits. Le débruitage des données artificielles ainsi que des chiffres manuscrits a été réalisé en utilisant différentes techniques de pré-image. Les résultats de la comparaison de cette approche avec d'autres techniques de pré-image telles la descente du gradient, celle du point-fixe, la méthode régularisée et celle pénalisée, ont montré la pertinence de la méthode proposée.

Dans un troisième temps, nous avons considéré les méthodes à noyaux pour l'analyse et la prédiction de séries temporelles. Les séries temporelles sont des séquences de valeurs qui varient avec le temps. Nous pouvons en citer les données financières, économétriques, gestionnaires et statistiques et les bio-signaux dont les signaux électroencéphalogrammes et électrocardiogrammes. Pour ce faire, le modèle prédictif linéaire dit autorégressif (AR) a été introduit, tout en proposant son extension pour le cas non-linéaire. Le modèle AR est défini soit par la méthode des moindres carrés, soit les équations de Yule-

Walker qui sont les mathématiques sous-jacentes maîtrisant un tel modèle. Alors, nous avons élaboré trois modèles AR-à-noyaux. Le premier nécessite une technique de résolution du problème de pré-image pour la prédiction. Il tient compte de la définition des méthodes à noyaux, en d'autres termes la transformation vers un espace fonctionnel où l'estimation de l'échantillon futur est réalisée. Une extension de la méthode des moindres carrés est alors présentée dans l'espace fonctionnel, et une extension des équations de Yule-Walker est réalisée dans cet espace afin de définir le modèle AR-à-noyaux en question. Une fois, cette estimation est faite, il s'avère nécessaire de faire le retour inverse pour prédire l'échantillon dans l'espace des observations. Le second modèle utilise directement les valeurs obtenues en évaluant le noyau considéré afin de prédire les échantillons futurs, tandis que le dernier est un modèle hybride en lien avec les deux modèles précédents. Les séries temporelles utilisées lors des expérimentations sont d'une part unidimensionnelles telles que les séquences *MG₃₀* et *Laser* et les signaux électroencéphalogrammes et d'autre part chaotiques multidimensionnelles telles que les séquences *Ikeda* et *Lorenz*. Les différentes expérimentations ont montré la pertinence de la méthode proposée. L'efficacité des modèles proposés est montrée en les comparant au modèle AR linéaire, au perceptron multicouche, au régresseur à vecteurs supports et au filtre de Kalman non-linéaire. La première approche proposée donne des meilleurs résultats que les autres techniques au détriment du temps de calcul. Le modèle hybride est un bon compromis entre temps de calcul et précision de la prédiction. Il donne des résultats moins bons que la première approche proposée mais ses performances restent au-delà de celles des méthodes traditionnelles.

Dans un dernier temps, la classification binaire et multi-classes ont été examinées. À cette fin, deux méthodes sont utilisées. La première utilise les machines à vecteurs supports (SVM), en se basant sur l'apprentissage supervisé. Elles sont principalement conçues pour une classification binaire, et sont adaptées pour une classification multi-classes selon deux stratégies de décompositions différentes, "un contre-un" et "un contre-tous". Cette méthode a été appliquée pour la discrimination des signaux électrocardiogrammes pour distinguer les personnes saines des personnes présentant des anomalies cardiaques, et pour différencier entre plusieurs problèmes cardiaques en utilisant ces deux stratégies. Les résultats ont montré que pour une classification binaire, les SVM avec un noyau Gaussien après réduction des caractéristiques donnent une précision plus élevée par rapport à une classification sur les données sans réduction des caractéristiques. De plus, si la réduction est réalisée par l'ACP-à-noyaux, la précision est meilleure et elle atteint 95%. De même, pour la classification multi-classes, les SVM après extraction des caractéristiques classifient mieux les différentes anomalies. La seconde méthode utilise la carte d'auto-organisation (SOM). Contrairement aux SVM, la SOM est basée sur l'apprentissage non-supervisé. Elle est utilisée en vue d'une classification multi-classes sans et avec réduction de la dimension des signaux. Les résultats ont montré que la SOM avec l'ACP-à-noyaux pour la réduction de dimension présentent une précision presque parfaite.

Perspectives

Notre travail ne s'achève pas à la fin de ce manuscrit. Un certain nombre de pistes méritent encore d'être explorées :

Méthodes à noyau multiple. Nous avons appliqué les méthodes à noyau en choisissant une fonction non-linéaire bien donnée. Une extension intéressante serait d'appliquer l'idée des méthodes à noyau multiple, impliquant la sommation de différents noyaux. Cette sommation est équivalente à la concaténation des espaces fonctionnels. Il fallait alors voir la régularisation de ces noyaux multiples imposant ainsi une certaine parcimonie.

La sélection des noyaux et des paramètres. Une perspective de travail intéressante serait la sélection du noyau à partir des données, ainsi que le choix des hyperparamètres. Un plan d'étude consistera à considérer conjointement l'apprentissage de la statistique et l'apprentissage du noyau, c'est-à-dire en incluant tous les paramètres dans les variables du problème. Bien qu'aucun algorithme de résolution global n'existe, il ne s'agit pas d'une impossibilité théorique : les méthodes actuelles de résolution de problèmes sont simplement incompatibles avec la nature non-linéaire des fonctions à noyaux. Pour la classification binaire, nous avons utilisé trois types de noyaux : le noyau Gaussien, le noyau polynomial et le noyau exponentiel. Suite à ce choix, nous avons pris pour la classification multi-classes, le noyau Gaussien pour discriminer les signaux ECG. Nous pensons que ce sujet mérite d'être approfondi afin de mieux exploiter le potentiel des méthodes à noyaux. De plus, le choix du paramètre de régularisation C est pris selon les données d'entrée. Des études sont nécessaires pour le développement de procédures qui permettent l'optimisation de ce paramètre.

La factorisation non-négative. Il serait intéressant d'implémenter une méthode analogue à l'Analyse en Composantes Principales, cependant en préservant la positivité des résultats. Nous parlons alors de la factorisation non-négative (*NonNegative Matrix Factorization soit NMF*). Il serait alors avantageux de comparer les résultats obtenus par une méthode classique de la pré-image de facteurs obtenus par NMF avec les résultats des méthodes proposées.

Les contraintes de boîtes et la parcimonie. Nous avons considéré la contrainte de non-négativité de la pré-image ou des coefficients du modèle définissant la pré-image. Une extension vers les contraintes de boîtes mérite d'être étudiée. Ces contraintes de boîtes, en anglais *box-constraints*, imposent des conditions non seulement sur une borne mais sur les bornes, d'où vient le nom de boîte. À cette fin, les limites supérieures et inférieures doivent être satisfaites, comme pour le traitement des images en niveaux de gris. Les contraintes de non-négativité des coefficients du modèle définissant la pré-image ont abouti à un effet secondaire qui est la parcimonie du résultat. Une étude intéressante serait de lancer des investigations plus poussées vers des critères de parcimonie lors de l'application des méthodes à noyaux en reconnaissance des formes.

Le modèle autorégressif. Le modèle autorégressif dépend de l'ordre, et des coefficients le définissant. Différents critères, comme le critère d'information d'Akaike ou le critère d'information Bayesien et la fonction partielle d'autocorrelation, sont présents dans la littérature pour estimer cet ordre. Il serait souhaitable d'élaborer ces critères dans l'espace fonctionnel. Un développement pourrait être envisagé pour des

techniques d'estimation des paramètres, telles que la méthode de Levinson-Durbin. En outre, nous envisageons l'utilisation des méthodes de processus Gaussiens où la fonction de covariance est la fonction noyau. Il serait aussi souhaitable d'étendre l'approche proposée pour le modèle autorégressif à moyenne mobile (*ARMA pour autoregressive moving average*).

Annexe

Dans cet annexe, nous dérivons les expressions, associées aux deux classes de noyaux, du gradient de (1.10) par rapport à \mathbf{x} , à savoir

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i \nabla_{\mathbf{x}} \kappa(\mathbf{x}_i, \mathbf{x}) + \frac{1}{2} \nabla_{\mathbf{x}} \kappa(\mathbf{x}, \mathbf{x}). \quad (\text{A.1})$$

ainsi que les expressions de la méthode itérative du point fixe.

Pour cela, nous rappelons l'expression du gradient d'une composition de deux fonctions. Soient h_2 une fonction à valeurs réelles définie sur \mathbb{R} et h_1 une fonction à valeurs réelles définie sur \mathcal{X} (nous nous intéressons en particulier à $h_1(\mathbf{x}) = \langle \mathbf{x}_i, \mathbf{x} \rangle$ ou encore $h_1(\mathbf{x}) = \|\mathbf{x}_i - \mathbf{x}\|^2$, voir plus bas). Sous condition de différentiabilité de la fonction h_2 en $h_1(\mathbf{x})$, nous avons alors

$$\nabla_{\mathbf{x}} (h_2 \circ h_1) \mathbf{x} = h_2^{(1)}(h_1(\mathbf{x})) \nabla_{\mathbf{x}} h_1(\mathbf{x}),$$

où $h_2^{(1)}(\zeta)$ désigne la première dérivée de la fonction h_2 par rapport à ζ , c'est-à-dire

$$h_2^{(1)}(\zeta) = \frac{\partial h_2(\zeta)}{\partial \zeta}.$$

A.1 Noyaux projectifs

Les noyaux projectifs sont de la forme $\kappa(\mathbf{x}_i, \mathbf{x}_j) = f(\langle \mathbf{x}_i, \mathbf{x}_j \rangle)$. Dans ce cas,

$$\nabla_{\mathbf{x}} f(\langle \mathbf{x}_i, \mathbf{x} \rangle) = f^{(1)}(\langle \mathbf{x}_i, \mathbf{x} \rangle) \nabla_{\mathbf{x}} (\langle \mathbf{x}_i, \mathbf{x} \rangle)$$

avec le gradient $\nabla_{\mathbf{x}} (\langle \mathbf{x}_i, \mathbf{x} \rangle)$ est donné par le vecteur dont la $k^{\text{ième}}$ composante est

$$\frac{\partial \langle \mathbf{x}_i, \mathbf{x} \rangle}{\partial [\mathbf{x}]_k} = \sum_{j=1}^{\dim(\mathcal{X})} [\mathbf{x}_i]_j \frac{[\mathbf{x}]_j}{\partial [\mathbf{x}]_k} = \sum_{j=1}^{\dim(\mathcal{X})} [\mathbf{x}_i]_j \delta_{jk} = [\mathbf{x}_i]_k,$$

où δ_{jk} désigne le symbole de Kronecker. En d'autres termes, $\nabla_{\mathbf{x}} (\langle \mathbf{x}_i, \mathbf{x} \rangle) = \mathbf{x}_i$, soit

$$\nabla_{\mathbf{x}} f(\langle \mathbf{x}_i, \mathbf{x} \rangle) = f^{(1)}(\langle \mathbf{x}_i, \mathbf{x} \rangle) \mathbf{x}_i.$$

D'autre part, il est facile de montrer que

$$\nabla_{\mathbf{x}} f(\langle \mathbf{x}, \mathbf{x} \rangle) = 2 f^{(1)}(\langle \mathbf{x}, \mathbf{x} \rangle) \mathbf{x},$$

puisque

$$\frac{\partial \langle \mathbf{x}, \mathbf{x} \rangle}{\partial [\mathbf{x}]_k} = \frac{\partial \|\mathbf{x}\|^2}{\partial [\mathbf{x}]_k} = 2 [\mathbf{x}]_k.$$

Le gradient (A.1) s'écrit alors

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i f^{(1)}(\langle \mathbf{x}_i, \mathbf{x} \rangle) \mathbf{x}_i + f^{(1)}(\langle \mathbf{x}, \mathbf{x} \rangle) \mathbf{x}.$$

A l'optimum \mathbf{x}^* , c'est-à-dire lorsque l'expression ci-dessus s'annule, nous obtenons

$$\sum_{i=1}^n \gamma_i f^{(1)}(\langle \mathbf{x}_i, \mathbf{x}^* \rangle) \mathbf{x}_i = f^{(1)}(\langle \mathbf{x}^*, \mathbf{x}^* \rangle) \mathbf{x}^*,$$

ou encore sous la forme d'expression du point fixe :

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

Le noyau polynomial est un cas particulier de noyaux projectifs, avec

$$\kappa_q(\mathbf{x}_i, \mathbf{x}_j) = (c + \langle \mathbf{x}_i, \mathbf{x}_j \rangle)^q.$$

Nous avons alors

$$\nabla_{\mathbf{x}} \kappa_q(\mathbf{x}_i, \mathbf{x}) = \nabla_{\mathbf{x}} (c + \langle \mathbf{x}_i, \mathbf{x} \rangle)^q = q (c + \langle \mathbf{x}_i, \mathbf{x} \rangle)^{q-1} \mathbf{x}_i = q \kappa_{q-1}(\mathbf{x}_i, \mathbf{x}) \mathbf{x}_i,$$

ce qui permet de déterminer l'expression du point fixe

$$\mathbf{x}^* = \frac{\sum_{i=1}^n \gamma_i \kappa_{q-1}(\mathbf{x}_i, \mathbf{x}^*) \mathbf{x}_i}{\kappa_{q-1}(\mathbf{x}^*, \mathbf{x}^*)}.$$

A.2 Noyaux radiaux

Ne dépendant que de la distance, les noyaux radiaux sont de la forme $\kappa(\mathbf{x}_i, \mathbf{x}_j) = g(\|\mathbf{x}_i - \mathbf{x}_j\|^2)$.

Nous avons alors

$$\nabla_{\mathbf{x}} g(\|\mathbf{x}_i - \mathbf{x}\|^2) = g^{(1)}(\|\mathbf{x}_i - \mathbf{x}\|^2) \nabla_{\mathbf{x}} (\|\mathbf{x}_i - \mathbf{x}\|^2).$$

Dans ce cas, il est facile de montrer que la $k^{\text{ième}}$ composante du gradient $\nabla_{\mathbf{x}}(\|\mathbf{x}_i - \mathbf{x}\|^2)$ est :

$$\frac{\partial(\|\mathbf{x}_i - \mathbf{x}\|^2)}{\partial[\mathbf{x}]_k} = -2([\mathbf{x}_i]_k - [\mathbf{x}]_k),$$

ce qui permet d'écrire $\nabla_{\mathbf{x}}(\|\mathbf{x}_i - \mathbf{x}\|^2) = -2(\mathbf{x}_i - \mathbf{x})$, et par conséquent

$$\nabla_{\mathbf{x}} g(\|\mathbf{x}_i - \mathbf{x}\|^2) = -2 g^{(1)}(\|\mathbf{x}_i - \mathbf{x}\|^2) (\mathbf{x}_i - \mathbf{x}).$$

Le gradient (A.1) s'écrit alors

$$\nabla_{\mathbf{x}} J(\mathbf{x}) = - \sum_{i=1}^n \gamma_i g^{(1)}(\|\mathbf{x}_i - \mathbf{x}\|^2) (\mathbf{x}_i - \mathbf{x}).$$

Cette dernière s'annulant à l'optimum \mathbf{x}^* , nous obtenons l'expression du point fixe

$$\mathbf{x}^* = \frac{\sum_{i=1}^n \gamma_i g^{(1)}(\|\mathbf{x}_i - \mathbf{x}\|^2) \mathbf{x}_i}{\sum_{i=1}^n \gamma_i g^{(1)}(\|\mathbf{x}_i - \mathbf{x}\|^2)}.$$

Un exemple de noyaux radiaux est le noyau Gaussien, avec $\kappa_G(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{-1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$.

Nous avons alors :

$$\mathbf{x}^* = \frac{\sum_{i=1}^n \gamma_i \kappa_G(\mathbf{x}_i, \mathbf{x}^*) \mathbf{x}_i}{\sum_{i=1}^n \gamma_i \kappa_G(\mathbf{x}_i, \mathbf{x}^*)}.$$

Bibliographie

- [ABR64] M. A. Aizerman, E. A. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control*, number 25, pages 821–837, 1964. 12, 17
- [AG11] H. S. Anderson and M. R. Gupta. Expected kernel for missing features in support vector machines. In *IEEE Workshop on Statistical Signal Processing*, Nice, France, 28-30 June 2011. 56
- [AGSJ11] H. S. Anderson, M. R. Gupta, E. Swanson, and K. Jamieson. Channel-robust classifiers. *IEEE Transactions on Signal Processing*, 59(4) :1421–1434, 2011. 56
- [AH09] T. J. Abrahamsen and L. K. Hansen. Input space regularization stabilizes pre-images for kernel PCA de-noising. In *IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Grenoble, France, 2009. 27, 31
- [AH11] T. J. Abrahamsen and L. K. Hansen. Regularized pre-image estimation for kernel pca de-noising : Input space regularization and sparse reconstruction. *Journal of Signal Processing Systems*, 65(3) :403–412, 2011. 47, 50, 54
- [Aro50] N. Aronszajn. Theory of reproducing kernels. *Trans. Amer. Math. Soc.*, 68 :337–404, 1950. 14, 15, 16
- [Bak05] G. Bakir. *Extension to kernel dependency estimation with applications to robotics*. PhD thesis, Technical University, Berlin, Germany, November 2005. 28
- [BD09] P.J. Brockwell and R.A. Davis. *Time Series : Theory and Methods*. Springer Series in Statistics. Springer, 2009. 58
- [BGV92a] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM. 80
- [BGV92b] B. E. Boser, L. M. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual ACM workshop on computational learning theory*, pages 144–152. ACM Press, 1992. 12
- [BJMO11] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Structured sparsity through convex optimization. *CoRR*, 2011. 43
- [BJMO12] F. R. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1) :1–106, 2012. 43
- [BKS] R. Bousseljot, D. Kreiseler, and A. Schnabel. The PTB diagnostic ECG database. <http://physionet.org/physiobank/database/ptbdb/>. 80, 93

- [BL07] L. Bottou and C.-J. Lin. Support vector machine solvers. In *Large-Scale Kernel Machines*, Neural Information Processing Series. MIT Press, 2007. [81](#)
- [BSW07] G. Bakir, B. Schölkopf, and J. Weston. On the pre-image problem in kernel methods. In *Eds. G. Camps-Valls and J.L. Rojo-Alvarez and M. Martínez-Ramon, Kernel Methods in Bioengineering, Signal and Image Processing*, pages 284–302. Hershey, PA : Idea Group publishing, 2007. [28](#)
- [BT98] R. Bourbonnais and M. Terraza. *Analyse des séries temporelles en économie*. Économie (Paris). Presses Universitaires de France, 1998. [55](#)
- [BT10] R. Bourbonnais and M. Terraza. *Analyse des séries temporelles : Applications à l'économie et à la gestion*. Éco sup. Manuel et exercices corrigés. Dunod, 2010. [56](#)
- [Bur98] C.J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2 :121–167, June 1998. [81, 83](#)
- [Bur99] Christopher J. C. Burges. Geometry and invariance in kernel based methods. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in kernel methods*, pages 89–116, Cambridge, MA, USA, 1999. MIT Press. [16, 17](#)
- [BWS04] G.H. Bakir, J. Weston, and B. Schölkopf. Learning to find pre-images. In *Advances in Neural Information Processing Systems*, pages 449–456. MIT Press, 2004. [28](#)
- [Cas89] M. Casdagli. Nonlinear prediction of chaotic time series. *Physica D : Nonlinear Phenomena*, 35(3) :335 – 356, 1989. [71](#)
- [CB96] R. Cusani and E. Baccarelli. Parameter identification of frequency-selective noisy fast-fading rayleigh digital channels via nonlinear Yule-Walker-like equations. In *Proc. of European Conference on Signal Processing (EUSIPCO)*. Eurasip, 1996. [56](#)
- [CCC00] Trevor F. Cox, Michael A. A. Cox, and Trevor F. Cox. *Multidimensional Scaling, Second Edition*. Chapman & Hall/CRC, sep 2000. [28](#)
- [CCSK04] H.H. Chou, Y.J. Chen, Y.C. Shiao, and T.S. Kuo. A high performance compression algorithm for ECG with irregular periods. In *Proc. IEEE International Workshop on Biomedical Circuits and Systems*, 2004. [87](#)
- [CG85] M. Chevalier and Y. Grenier. Autoregressive models with time-dependent log area ratios. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, volume 10, pages 1049 – 1052, apr 1985. [56](#)
- [CLS⁺07] F. Castells, P. Laguna, L. Sörnmo, A. Bollmann, and J.M. Roig. Principal component analysis in ECG signal processing. *EURASIP J. Appl. Signal Process.*, 2007 :98–98, January 2007. [87](#)
- [CMR12] C. Cortes, M. Mohri, and A. Rostamizadeh. Algorithms for learning kernels based on centered alignment. *Journal of Machine Learning Research*, 13 :795–828, 2012. [64](#)

- [CRH⁺10a] J. Chen, C. Richard, P. Honeine, H. Lantéri, and C. Theys. System identification under non-negativity constraints. In *Proc. of European Conference on Signal Processing (EUSIPCO)*, Aalborg, Denmark, 2010. Eurasip. 35
- [CRH⁺10b] J. Chen, C. Richard, P. Honeine, H. Snoussi, H. Lantéri, and C. Theys. Techniques d'apprentissage non-linéaires en ligne avec contraintes de positivité. In *Actes de la VI^{ème} Conférence Internationale Francophone d'Automatique*, Nancy, France, 2 - 4 Juin 2010. 35
- [CRHB10] J. Chen, C. Richard, P. Honeine, and J. C. M. Bermudez. Non-negative distributed regression for data inference in wireless sensor networks. In *Proc. of the 44th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove (CA), USA, 2010. 35
- [CS02] F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the American Mathematical Society*, 39 :1–49, 2002. 16
- [CV95] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3) :273–297, 1995. 80, 81
- [CW08] E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2) :21–30, March 2008. 43
- [dEJL07] A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation of sparse PCA using semidefinite programming. *SIAM Review*, 49(3), 2007. 43
- [DMK09] T. Dutoit, N. Moreau, and P. Kroon. How is speech processed in a cell phone conversation ? In *Eds. T. Dutoit, F. Marques, Applied Signal Processing*. Springer, 2009. 56
- [ESK07] P. Etyngier, F. Ségonne, and R. Keriven. Shape priors using manifold learning techniques. In *in Proc. 11th IEEE International Conference on Computer Vision, Rio de Janeiro*, Rio de Janeiro, Brazil, October 2007. 29
- [EW02] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS-01)*, pages 681–687, 2002. 80
- [FHL08] J.H. Fu, C.H. Huang, and S.L. Lee. A multi-class svm classification system based on methods of self-learning and error filtering. In *Department of Computer Science and Information Engineering National Chung Cheng University Chiayi 62107*, Taiwan, Republic of China, 2008. 83
- [Ful96] Wayne A. Fuller. *Introduction to Statistical Time Series (Wiley Series in Probability and Statistics)*. Wiley-Interscience, April 1996. 56
- [Gac11] A. Gacek. Preprocessing and analysis of ecg signals - a self-organizing maps approach. *Expert Systems with Applications*, 38(7) :9008–9013, 2011. 85

- [GAG^{+a}] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. The MIT-BIH arrhythmia database. <http://physionet.org/physiobank/database/mitdb/>. 80, 87, 93
- [GAG^{+b}] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. The MIT-BIH normal sinus rhythm database. <http://physionet.org/physiobank/database/nsrdb/>. 71, 72
- [GAG⁺¹³] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet : Components of a new research resource for complex physiologic signals. *Circulation*, 101(23) :e215–e220, 2000 (June 13). Circulation Electronic Pages : <http://circ.ahajournals.org/cgi/content/full/101/23/e215> PMID :1085218 ; doi : 10.1161/01.CIR.101.23.e215. 71, 72
- [Geo07] S. D. Georgiadis. *State-Space Modeling and Bayesian Methods for Evoked Potential Estimation*. PhD thesis, Department of Applied Physics, University of Kuopio, Finland, May 2007. 44
- [GS90] K. Gordon and A. F. M. Smith. Modeling and monitoring biomedical times series. *Journal of the American Statistical Association*, 85(410) :pp. 328–337, 1990. 56
- [Han10] H. Han. Nonnegative principal component analysis for mass spectral serum profiles and biomarker discovery. *BMC Bioinformatics*, 2010. 36
- [HAW89] U. Hübner, N. B. Abraham, and C. O. Weiss. Dimensions and entropies of chaotic intensity pulsations in a single-mode far-infrared nh3 laser. *Phys Rev A*, 40(11) :6354–6365, 1989. 71
- [HL02] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13 :415–425, 2002. 84
- [HO00] A. Hyvärinen and E. Oja. Independent component analysis : algorithms and applications. *Neural Netw.*, 13(4-5) :411–430, 2000. 36
- [HR09] P. Honeine and C. Richard. Solving the pre-image problem in kernel machines : a direct method. In *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, Grenoble, France, September 2009. 30
- [HR10] P. Honeine and C. Richard. A closed-form solution for the pre-image problem in kernel-based machines. *Journal of Signal Processing Systems*, April 2010. 30
- [HR11] P. Honeine and C. Richard. The pre-image problem in kernel-based machine learning. *IEEE Signal Processing Magazine, special issue on “dimensionality reduction via subspace and manifold learnin”*, 28 (2) :77–88, March 2011. 25
- [Joa00] T. Joachims. *The Maximum Margin Approach to Learning Text Classifiers : Methods, Theory, and Algorithms*. PhD thesis, Universität Dortmund, Informatik, LS VIII, 2000. 80

- [KCV11] A. Karagiannis, P. Constantinou, and D. Vouyioukas. Biomedical time series processing and analysis methods : The case of empirical mode decomposition. In *Advanced Biomedical Engineering, Gaetano D. Gargiulo, Co-editor : Alistair McEwan (Ed.)*. 2011. 56
- [KFH⁺12] M. Kallas, C. Francis, P. Honeine, H. Amoud, and C. Richard. Modeling electrocardiogram using Yule-Walker equations and kernel machines. In *19th International Conference on Telecommunications*, Jounieh, Lebanon, 23-25 April 2012. 9
- [KFK⁺12] M. Kallas, C. Francis, L. Kanaan, D. Merheb, P. Honeine, and H. Amoud. Multi-class SVM classification combined with kernel PCA feature extraction of ECG signals. In *19th International Conference on Telecommunications*, Jounieh, Lebanon, 23-25 April 2012. 9
- [KHAF11] M. Kallas, P. Honeine, H. Amoud, and C. Francis. Sur le problème de la pré-image en reconnaissance des formes avec contraintes de non-négativité. In *23ème édition du Colloque GRETSI*, Bordeaux, France, 5-8 Septembre 2011. 9
- [KHFA11] M. Kallas, P. Honeine, C. Francis, and H. Amoud. Kernel autoregressive models. a comparative study of pre-image techniques. In *25th IEEE Workshop on Signal Processing Systems SiPS'2011*, Beirut, Lebanon, 4-7 Octobre 2011. 9
- [KHFA12] M. Kallas, P. Honeine, C. Francis, and H. Amoud. Kernel autoregressive models using Yule-Walker equations. *Elsevier, Signal Processing*, Soumis en Juillet 2012. 9
- [KHR⁺10] M. Kallas, P. Honeine, C. Richard, H. Amoud, and C. Francis. Nonlinear feature extraction using kernel principal component analysis with non-negative pre-image. In *Proc. 32nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Buenos Aires, Argentina, 31 Aug. - 4 Sept. 2010. 9, 41
- [KHR⁺11a] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Kernel-based autoregressive modeling with a pre-image technique. In *16th IEEE Workshop on Statistical Signal Processing*, Nice, France, 28-30 June 2011. 9
- [KHR⁺11b] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Modèle autorégressif non-linéaire à noyau : une première approche. In *23ème édition du Colloque GRETSI*, Bordeaux, France, 5-8 Septembre 2011. 9
- [KHR⁺11c] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Non-negative pre-image in machine learning for pattern recognition. In *19th European Signal Processing Conference*, Barcelona, Spain, 29 August - 2 September 2011. 9
- [KHR⁺12a] M. Kallas, P. Honeine, C. Richard, H. Amoud, and C. Francis. Non-negativity constraints on the pre-image for pattern recognition with kernel machines. *Elsevier, Pattern Recognition*, Soumis en Avril 2012. 9

- [KHR⁺12b] M. Kallas, P. Honeine, C. Richard, C. Francis, and H. Amoud. Prediction of time series using Yule-Walker equations with kernels. In *37th IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012. 9
- [Kim] J. Kimura. Examples of electromyograms. <http://physionet.org/physiobank/database/emgdb/>. 72
- [KJ07] R. Kumar and C. V. Jawahar. Kernel approach to autoregressive modeling. In *The 13th National Conference on Communications (NCC)*, Kanpur, India, January 2007. 57
- [KMK⁺11] L. Kanaan, D. Merheb, M. Kallas, C. Francis, H. Amoud, and P. Honeine. PCA and KPCA of ECG signals with binary SVM classification. In *25th IEEE Workshop on Signal Processing Systems SiPS'2011*, Beirut, Lebanon, 4-7 Octobre 2011. 9
- [Koh82] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1) :59–69, January 1982. 80, 84
- [KSH01] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001. 80, 85
- [KT03] J. T. Kwok and I. W. Tsang. The pre-image problem in kernel methods. In T. Fawcett and N. Mishra, editors, *Proc. 20th International Conference on Machine Learning (ICML)*, pages 408–415, Washington, August 2003. AAAI Press. 27, 29, 49, 50, 54
- [KT04] J. T. Kwok and I. W. Tsang. The pre-image problem in kernel methods. *IEEE Trans. on Neural Networks*, 15(6) :1517–1525, November 2004. 29
- [KW71] G. S. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, 33(1) :82–95, 1971. 12, 18
- [LBB⁺04] H. Lütkepohl, J. Breitung, R. Brüggemann, H. Herwartz, T. Teräsvirta, R. Tschernig, and M. Krätzig. *Applied Time Series Econometrics*. Cambridge University Press, August 2004. 55
- [LC] Y. LeCun and C. Cortes. The MNIST database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>. 50
- [Lin79] D. A. Linkens. Maximum entropy analysis of short time-series biomedical rhythms. *Journal of Interdisciplinary Cycle Research*, 10(2) :145–163, 1979. 56
- [LRCA01] H. Lantéri, M. Roche, O. Cuevas, and C. Aime. A general method to devise maximum-likelihood signal restoration multiplicative algorithms with non-negativity constraints. *Signal Processing*, 81(5) :945–974, May 2001. 35, 39
- [LTBM09] H. Lantéri, C. Theys, F. Benvenuto, and D. Mary. Méthode algorithmique de minimisation de fonctions d'écart entre champs de données. application à la reconstruction d'images astrophysiques. In *Colloque GRETSI'2009*, Dijon, France, 8-11 septembre 2009. 40

- [Mad08] H. Madsen. *Time Series Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, 2008. 57
- [MCS06] J. Milgram, M. Cheriet, and R. Sabourin. One Against One or One Against All : Which One is Better for Handwriting Recognition with SVMs ? In G. Lorette, editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France), October 2006. Université de Rennes 1, Suvisoft. 83
- [MM01] G B Moody and R G Mark. The impact of the mit-bih arrhythmia database. *IEEE Eng Med Biol Mag*, 20(3) :45–50, 2001. 87
- [MOG97] S. Mukherjee, E. Osuna, and F. Girosi. Nonlinear prediction of chaotic time series using support vector machines. In *IEEE Workshop on Neural Networks for Signal Processing VII*, pages 511–519. IEEE Press, 1997. 57, 71
- [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. 27, 50, 54
- [MWA06] B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA : Exact and greedy algorithms. In *Advances in Neural Information Processing Systems*, pages 915–922. MIT Press, 2006. 36, 43
- [MWM83] S.G. Makridakis, S.C. Wheelwright, and V.E. McGee. *Forecasting : Methods and Applications*. Wiley series in management. Wiley, 1983. 56
- [OFG97] E. Osuna, R. Freund, and F. Girosi. Training support vector machines : an application to face detection. In *Conference on Computer Vision and Pattern Recognition*, pages 130–136, 1997. 80
- [oK] University of Kuopio. Erp signal. <http://www.uku.fi/>. 44
- [OP03] E. Oja and M. Plumley. Blind separation of positive sources using non-negative pca. In *In 4th International Symposium on Independent Component Analysis and Blind Signal Separation*, pages 11–16, 2003. 36
- [PC07] A. Pronobis and B. Caputo. Confidence-based cue integration for visual place recognition. In *in Proc. IROS07 Centre for Autonomous Systems Royal Institute of Technology*, SE-100 44 Stockholm, Sweden, 2007. 83
- [PCCVSO⁺02] F. Pérez-Cruz, G. Camps-Valls, E. Soria-Olivas, J. J. Pérez-Ruixo, A. R. Figueiras-Vidal, and A. Artés-Rodríguez. Multi-dimensional function approximation and regression estimation. In *Proceedings of the International Conference on Artificial Neural Networks*, ICANN '02, pages 757–762, London, UK, UK, 2002. Springer-Verlag. 80
- [PG84] R. Prost and R. Goutte. Discrete constrained iterative deconvolution algorithms with optimized rate of convergence. *Signal Processing*, 7(3) :209–230, December 1984. 35

- [RBH09] C. Richard, J. C. M. Bermudez, and P. Honeine. Online prediction of time series data with kernels. *IEEE Transactions on Signal Processing*, 57(3) :1058–1067, 2009. [56](#)
- [RdB03] L. Ralaivola and F. d’Alche Buc. Dynamical modeling with kernels for nonlinear time series prediction. In *Advances in Neural Information Processing Systems 16*, page 2004. MIT Press, 2003. [73](#)
- [RDB05] L. Ralaivola and F. D’alche-Buc. Time series filtering, smoothing and learning using the kernel kalman filter. In *Proc. IEEE International Joint Conference on Neural Networks*, volume 3, pages 1449–1454, 2005. [56, 71, 72, 76](#)
- [RGTO0] R. Rosipal, M. Girolami, and L. J. Trejo. Kernel PCA for feature extraction and de-noising in non-linear regression. *Neural Computing and Applications*, 10 :231–243, 2000. [18](#)
- [RS00] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290 :2323–2326, December 2000. [29](#)
- [SB08] C. D. Sigg and J. M. Buhmann. Expectation-maximization for sparse and non-negative PCA. In *25th International Conference on Machine Learning (ICML)*. ACM, 2008. [36](#)
- [SBS98] B. Schölkopf, C.J.C. Burges, and A.J. Smola. *Advances in Kernel Methods : Support Vector Learning*. MIT Press, 1998. [12](#)
- [SHS01] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *Proc. 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT ’01/EuroCOLT ’01, pages 416–426, London, UK, 2001. Springer-Verlag. [17, 18](#)
- [Sjo05] K. Sjøstrand. Matlab implementation of LASSO, LARS, the elastic net and SPCA, jun 2005. Version 2.0. [43](#)
- [SPST⁺01] B. Schölkopf, J. C. Platt, J. C. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Comput.*, 13(7) :1443–1471, 2001. [80](#)
- [SS89] R. A. Stine and P. Shaman. A fixed point charactérisation for bias of autoregressive estimators. *Annals of statistics*, 17(3) :1275–1284, 1989. [56](#)
- [SSM98a] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10 :1299–1319, July 1998. [12](#)
- [SSM98b] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, 10(5) :1299–1319, 1998. [18](#)
- [SSO92] D.L. Snyder, T.J. Schulz, and J.A. O’Sullivan. Deblurring subject to nonnegativity constraints. *IEEE Transactions on Signal Processing*, 40 :1143 – 1150, May 1992. [35](#)
- [STC04] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004. [12](#)

- [Tar] M. Tarvainen. Medical signal analysis. <http://venda.uku.fi/opiskelu/kurssit/LSA/>. 44
- [Tar04] M. P. Tarvainen. *Estimation Methods for Nonstationary Biosignals*. PhD thesis, Department of Applied Physics, University of Kuopio, Finland, June 2004. 44
- [Tho83] G. Thomas. A positive optimal deconvolution procedure. *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 8 :651 – 654, April 1983. 35
- [Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58 :267–288, 1996. 43
- [TS91] G. Thomas and N. Souilah. Utilisation des multiplicateurs de lagrange pour la restauration d'image avec contraintes. *Colloques sur le Traitement du Signal et des Images*, 1991. 35
- [Tsa05] Ruey S. Tsay. *Analysis of Financial Time Series (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2nd edition, August 2005. 55, 60
- [TW06] P. Then and Y. C. Wang. Support vector machine as digital image watermark detector. In *Proceeding of Society of Photo-Optical Instrumentation Engineers*, volume 6064, 2006. 80
- [Vap95] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995. 79, 82
- [Vap98] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, September 1998. 12, 43, 56, 83
- [VL63] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 1963. 80
- [Wah90] G. Wahba. *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990. 12, 17
- [Wan] E. A. Wan. Time series data. <http://www.bme.ogi.edu/~ericwan/data.html>. 57, 71
- [WC00] V. Wan and W. M. Campbell. Support vector machines for speaker verification and identification. In *Proceeding of IEEE International Workshop on Neural Networks for Signal Processing*, volume 2, pages 775–784, 2000. 80
- [WdM10] E. A. Wan and R. Van der Merwe. ReBEL : Recursive bayesian estimation library. <http://icoregon.technologypublisher.com/technology/5019>, April 2010. 72
- [Wil01] C. K.I. Williams. On a connection between kernel pca and metric multidimensional scaling. In *Advances in Neural Information Processing Systems 13*, pages 675–681. MIT Press, 2001. 29

- [WW99] J. Weston and C. Watkins. Support vector machines for multi-class pattern recognition. In *7th European Symposium on Artificial Neural Networks*, pages 219–224, Bruges, Belgium, April 21-23 1999. 80
- [Yul27] Udny G. Yule. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London Series. Series A, Containing papers of a Mathematical or Physical Character*, 226 :267–298, 1927. 56
- [YV07] Y. Yamanishi and J.-P. Vert. Kernel matrix regression. Technical report, 2007. 30
- [ZHT04] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15 :2006, 2004. 43
- [ZIP06] S. L. Zeger, R. Irizarry, and R. D. Peng. On time series analysis of public health and biomedical data. *Annual review of public health*, 27(1) :57–79, 2006. 56
- [ZL06] W.-S. Zheng and J.-H. Lai. Regularized locality preserving learning of pre-image problem in kernel principal component analysis. *Pattern Recognition, International Conference on*, 2 :456–459, 2006. 28
- [ZLY10] W.S. Zheng, J.H. Lai, and P. C. Yuen. Penalized preimage learning in kernel principal component analysis. *IEEE Transaction Neural Networks*, 21 :551–570, April 2010. 28, 31, 50, 54
- [ZS07] R. Zass and A. Shashua. Nonnegative Sparse PCA. In *Neural Information Processing Systems*, 2007. 36