N^o d'ordre : H509

HABILITATION A DIRIGER DES RECHERCHES

Présentée à

L'Université des Sciences et Technologies de Lille

par

Manuel DAVY

OUTILS DE DÉCISION POUR L'ACCÈS À L'INFORMATION AUDIO
MÉTHODES DE MONTE CARLO ET MÉTHODES À NOYAUX

Soutenue le 20 juin 2006 devant le jury d'examen composé de :

Jean-Marc Geib, Professeur, USTL	Président
Stéphane Canu, Professeur, INSA de Rouen	Rapporteur
Pierre Del Moral, Professeur, Université de Nice	Rapporteur
Patrick Flandrin, Directeur de Recherches, CNRS	Rapporteur
Christian Doncarli, Professeur, Ecole Centrale de Nantes	Examinateur
Arnaud Doucet, Professeur, University of British Columbia, Canada	Examinateur
Philippe Vanheeghe, Professeur, Ecole Centrale de Lille	Examinateur

Habilitation à diriger des recherches préparée sous la direction de Monsieur le Professeur Philippe Vanheeghe au Laboratoire d'Automatique, de Génie Informatique et Signal (LAGIS).

Table des matières

Chapitre 1

Introduction

Ce document propose une synthèse de mes travaux de recherche menés depuis la thèse, soutenue en septembre 2000. Le cœur de mon activité concerne les *signaux*, essentiellement temporels. Cependant, comme tente de le montrer ce manuscrit, les méthodes développées sont adaptables à de nombreux contextes, et peuvent plus généralement être qualifiées de «méthodes d'analyse de données». Le contexte méthodologique est très largement lié aux statistiques (autant pour les modèles que pour les algorithmes) et les applications concernent essentiellement les signaux audio ainsi que, dans une moindre mesure, les systèmes multi-capteurs. De fait, ces activités de recherche ont été largement déterminées par le projet de recherche approuvé par le CNRS lors de mon recrutement en 2001. Avant d'en rappeler les points les plus marquants dans la section 1.2, l'organisation du manuscrit est détaillée ci-dessous.

1.1 Organisation du manuscrit

Ce document est structuré de la façon suivante. Dans le **chapitre 2**, je présente mon cursus, différentes activités liées à mon métier de chercheur, et ma liste de publications. Le **chapitre 3** aborde, sous la forme d'un état de l'art, les deux cadres théoriques et méthodologiques où se situent mes contributions. Il s'agit, d'une part, des méthodes statistiques bayésiennes – aujourd'hui indissociables des méthodes de Monte Carlo, souvent indispensables à leur mise en œuvre. D'autre part, il s'agit des méthodes d'apprentissage statistique à noyaux, et l'une de leurs instanciations les plus connues, les algorithmes à vecteurs support (*Support Vector Machines* – SVM).

Unifier des contributions concernant des méthodes et des applications variées m'a été difficile. J'ai choisi de présenter, dans un premier temps, les contributions selon le point de vue méthodologique et théorique (**chapitre 4**). Dans un second temps, elles sont évoquées sous l'angle des applications (**chapitre 5**). Ce choix entraîne inévitablement des redondances, qui me semblent cependant limitées. Enfin, dans le **chapitre 6**, je propose un point de vue unifié sur l'ensemble des contributions, et propose quelques conclusions et perspectives de recherche.

En annexe, le lecteur trouvera une sélection des publications les plus significatives dans les divers domaines abordés.

1.2 Projet de recherche établi en 2001

Cette section rappelle les objectifs de recherche proposés au CNRS lors de mon recrutement en 2001, et qui ont orienté mes activités depuis. L'extrait choisi constitue l'introduction de ce projet. Comme elle suffit à poser le cadre général du travail, je n'ai pas inclus le projet au complet.

(Extrait du projet de recherche présenté en vue du recrutement au CNRS en 2001. Les propositions qui y sont formulées ont été écrites à l'époque, aussi certain sujets sont évoqués au « futur », alors qu'ils appartiennent aujourd'hui au passé.)

1.2.1 Cadre du travail

L'Internet a créé de nouvelles possibilités d'accès à l'information, notamment par les bibliothèques numériques, la diffusion de l'information, le commerce ou l'éducation personnalisée. La force de l'Internet tient à sa nature distribuée, ce qui permet à chaque individu de contribuer localement et à sa façon à la mise à disposition d'informations. Mais cette nature distribuée est aussi une faiblesse, dans le sens où il n'y a pas d'organisation des données. Aussi, l'accès à des informations précises est-il rendu très difficile par l'ampleur de la base de données et l'absence d'une structure de recherche (d'indexation) des données.

Dans les bases de données plus anciennes (bibliothèques, administrations, archives), la solution a consisté à indexer les données avant leur stockage. La recherche s'effectue alors par mots clés. Ce type de solution est en fait inadapté à l'Internet, en raison de sa taille et de l'action non coordonnée de ses acteurs. En outre, quelques mots clés ne permettent pas une recherche naturelle.

Dans cette variété impressionnante de données, mon intérêt se porte sur les signaux audio qui constituent une catégorie bien particulière. Ils se différencient d'informations de type texte par la difficulté d'accès au contenu puisqu'un outil de recherche syntaxique permet de retrouver facilement des mots clés dans un texte. On comprend aisément que cette technique n'est pas transposable en l'état à des sons, dont l'information n'est pas accessible directement. Les enregistrements sonores se différencient également des image fixes – dont le contenu peut être appréhendé d'un seul coup d'œil – par la dimension temporelle (il faut plusieurs secondes d'écoute pour interpréter un signal sonore). En outre, chaque pixel composant une image tend à apporter de l'information sur un seul objet (par exemple, il correspond soit à un personnage au premier plan, soit à l'arrière plan). Dans un signal audio, un échantillon temporel porte de l'information sur tous les objets constitutifs du son à un instant donné (les différents instruments d'une musique de fond, mais aussi la voix d'un narrateur par exemple). Les signaux audio se différencient enfin des séquences vidéo qui peuvent être résumées par une série d'images clés sur la base desquelles sera faite une recherche de données. Ces spécificités font que les outils de traitement des média texte, image et vidéo, aujourd'hui assez avancés, ne sont pas directement transposables au médium audio. Le Traitement du Son constitue une discipline à part entière qui participe au même effort de recherches que le Traitement de l'Image et le Traitement de la Vidéo, et avec lesquelles elle interagit naturellement.

Les méthodes que je propose de développer, et que je détaillerai dans ce qui suit, permettront à l'internaute de retrouver des motifs audio qui auront été directement extraits des

données. Dans un objectif proche, et comme une étape vers l'application à l'échelle de l'Internet, je proposerai des algorithmes de recherche applicables à des bases de données plus organisées que l'Internet, telles les archives des stations de radio, de télévision, de cinéma ou encore la recherche de mélodies, de cris d'animaux, etc. Ces applications d'envergure importante – néanmoins plus modeste que l'Internet – sont nombreuses et présentent des enjeux scientifiques et industriels de premier plan[1].

1.2.2 Programme des recherches – Aspects méthodologiques

J'expose à présent quelques aspects méthodologiques qui me permettront de mener à bien l'extraction d'informations contenues dans des signaux audio. Je vais plus précisement décrire les problématiques concernant les méthodes à développer, sous la forme d'un programme des recherches :

- à moyen terme, je développerai des modèles ou des méthodes permettant de segmenter le signal en plages comportant de la musique, du bruit, de la parole, ou les trois simultanément. Je proposerai des algorithmes pour séparer ces différentes sources (éventuellement à partir d'enregistrements monophoniques), afin de les traiter par des algorithmes spécifiques préexistants (reconnaissance de la parole, musique) ou à développer (bruits divers, musique). Ces méthodes prendront en compte l'*a priori* fort qu'est la nature harmonique de beaucoup de phénomènes sonores, ou encore l'aspect rythmique de la musique. A moyen terme également, je participerai à la diffusion et au développement des méthodes de simulation telles que les MCMC et les filtres particulaires (Davy *et al.* (2000b,c)) et à l'adaptation de leur utilisation à des problèmes formels généraux, comme nous l'avons fait par exemple pour la classification optimale de signaux à modèle non-linéaire non-gaussien (Davy *et al.* (2000a)). Par ailleurs, j'aurai le souci de privilégier les techniques permettant une mise en œuvre en temps-réel, avec par exemple les filtres particulaires ;
- à court terme, il s'agira donc d'utiliser les modèles préexistants (adaptés à des signaux audio simples), de les généraliser en s'autorisant des hypothèses éventuellement non-gaussiennes, non-linéaires et non-stationnaires : cela est rendu possible par les techniques émergentes évoquées précédemment. Par exemple, on peut mettre en œuvre une segmentation dépendante du signal (et non plus systématique, toutes les 100ms) à l'aide de méthodes telles que dans (Laurent et Doncarli (1998)), ou la détection-estimation de composantes harmoniques, voir Vermaak *et al.* (2000).

Face à des signaux réels aussi complexes que les enregistrements sonores, une attitude méthodologique ouverte est incontournable. On ne peut définitivement prendre parti pour une classe de méthodes, par exemple non-paramétriques, contre d'autres, par exemple paramétriques. C'est pourquoi j'ai travaillé au cours de la thèse (Davy (2000)) dans un cadre ouvert, et je poursuivrai dans cet esprit, avec l'utilisation d'outils novateurs à la fois paramétriques et non-paramétriques. Je décris maintenant, avec davantage de détails, les méthodes émergentes non-paramétriques puis paramétriques que j'emploierai.

[1] L'intérêt porté par les industriels et les scientifiques au Multimédia (au sens large) est illustré par le succès des groupes normatifs comme MPEG4 et ses successeurs.

Les outils non-paramétriques : une efficacité sous contrôle

Le difficile problème de la reconnaissance des événements audio trouve des solutions avec les méthodes non-paramétriques dévolues aux signaux non-stationnaires : par exemple, les représentations temps-fréquence ou les techniques à base d'ondelettes. Pouvoir se passer d'un modèle rend possible une décision face aux signaux à structure complexe tels qu'un enregistrement musical de Jazz où il y a à la fois présence de parole chantée, de structures harmoniques dues à la guitare, et impulsionnelles,caractéristiques de la batterie.

Dans des recherches précédentes, il a été montré (Davy *et al.* (2001b); Laurent et Doncarli (1998)) que l'utilisation de représentations temps-fréquence et de distances permet de classer ou segmenter efficacement des signaux audio (il s'agissait d'enregistrements musicaux de piano, de signaux de contrôle d'enceintes acoustiques ou de reconnaissance de locuteurs). Il est également apparu que l'efficacité était fortement dépendante de la paramétrisation de la représentation temps-fréquence, laquelle devait être optimisée en fonction du problème de décision à traiter. Je prolonge actuellement ces travaux dans le cadre d'un Projet Jeunes Chercheurs du GdR-PRC ISIS. Un autre prolongement de ces travaux est la voie «optimisation d'ondelettes pour la classification», où l'on recherche non pas les meilleures bases de paquets d'ondelettes (Saito et Coifman (1994)), mais la meilleure ondelette pour la classification, selon la même méthodologie que celle utilisée dans (Davy *et al.* (2001b)) (ce thème est aujourd'hui en cours d'études préliminaires à l'IRCCyN). Un des enjeux de la recherche future sera en outre le prétraitement des signaux – c'est-à-dire leur conditionnement, certaines méthodes nécessitant des signaux de même longueur, ou un recalage temps-fréquence précis, etc. – aujourd'hui effectué au cas par cas avec plus ou moins de succès.

Bien que complexe, la structure des signaux audio laisse toutefois apparaître des caractéristiques spécifiques. Par exemple, de nombreux phénomènes vibratoires provoquent des signaux de type harmonique (le son du violoncelle, les phonèmes voisés, le ronflement d'un moteur de Ford Mustang) caractérisés par des structures filaires dans le plan temps-fréquence, se prêtant à une modélisation. Nous avons développé un algorithme semi-paramétrique de suivi de ces trajectoires spectrales sur la base d'une représentation ARCAP glissante (Cottereau *et al.* (2000); Davy *et al.* (1998)). Cette idée mêlant techniques paramétriques et non-paramétriques s'est révélée très efficace et devra être étendue dans un domaine paramétrique général, notamment pour le traitement des structures harmoniques présentes dans les sons.

Les outils paramétriques : l'optimalité à portée

Lorsque les phénomènes physiques à l'origine d'un signal sont compris, il est parfois intéressant de déterminer un modèle mathématique du signal. Se positionnant dans un cadre statistique, on peut alors mettre en œuvre des outils de décision, estimation, filtrage, etc. optimaux par rapport au modèle statistique utilisé. Malheureusement, les techniques numériques ont longtemps imposé de se restreindre au cas 'linéaire-gaussien', dont un exemple est le filtre de Kalman, et l'on ne savait pas résoudre des problèmes généraux non-linéaires et non-gaussiens. Depuis moins d'une dizaine d'années, les méthodes de simulation développées par la communauté des chercheurs en Statistiques Appliquées et Traitement du Signal ont ouvert un champ d'investigations gigantesque. Il est maintenant possible de mettre en œuvre des modèles très généraux, sur lesquels les recherches commencent à peine dans le champ du

Traitement du Signal, et en particulier dans le cadre de signaux audio.

Les méthodes MCMC que j'ai utilisées pour la classification permettent une utilisation optimale (mais hors ligne) des techniques bayésiennes. La mise au point d'algorithmes de simulation, et plus précisément l'adéquation entre l'algorithme et le modèle simulé, sont des domaines de recherche très ouverts où nous souhaitons continuer de faire avancer les connaissances pour les besoins du traitement de l'audio. Leur version 'en ligne', les filtres particulaires (Gordon *et al.* (1993)), seront également utilisés pour des algorithmes de traitement audio en temps-réel.

1.2.3 En résumé

Dans ce projet, je propose de développer les nouvelles méthodes du Traitement du Signal pour extraire, reconnaître et classer l'information contenue dans les signaux audio. Comme je l'ai fait jusqu'à aujourd'hui, je tirerai parti des qualités respectives des nouvelles méthodes paramétriques et non-paramétriques. Le fil rouge de ces recherches sera le Traitement de l'Audio, mais les résultats théoriques engendrés seront présentés dans un cadre général.

(Fin de l'extrait du projet de recherche présenté en vue du recrutement)

Chapitre 2

Curriculum vitæ

Dans ce chapitre, j'expose mes activités liées à la recherche, à l'enseignement, à la valorisation et à l'administration. La section 2.1 présente mon cursus, tandis que les sections suivantes détaillent les activités, en privilégiant la période post-thèse septembre 2000–avril 2006.

2.1 Cursus

Déc. 2005-	**Chargé de Recherches** de première classe, LAGIS, UMR 8146
Fév. 2004-Nov. 2005	**Chargé de Recherches** de deuxième classe, LAGIS, UMR 8146
Sep. 2001-Fév. 2004	**Chargé de Recherches** de deuxième classe, IRCCyN, UMR CNRS 6597
Sep. 2001-Juin 2002	**Embedded Researcher** à l'Université de Cambridge (complément de formation) Projet : *Outils de décision pour l'accès à l'information audio*
2000-2001	**Research Associate** au *Signal Processing Group*, Université de Cambridge, GB Titre : *Audio information retrieval in Multimedia databases, MOUMIR European Project*
1997-2000	**Doctorat** de l'Université de Nantes - Spécialité AIA (Traitement du Signal) à l'IRCCyN, bourse BDI CNRS Soutenance le 18 septembre 2000, devant un jury composé de G. Allengrin, D. Barba, C. Doncarli (directeur de thèse), P. Duvaut (rapporteur), P. Flandrin (rapporteur), F. Hlawatsch, N. Martin, J.Y. Tourneret – Mention *très honorable avec félicitations*
1995-1996	**D.E.A.** Automatique et Informatique Appliquée, filière Traitement du Signal, sous la direction de M. Guglielmi (mention *très bien*), Ecole Centrale de Nantes Sujet : *Synthèse du mouvement brownien fractionnaire par la décomposition de Choleski*

1993-1996 Cycle d'**études d'ingénieur**, option *Automatique*, Ecole Centrale de Nantes.

2.2 Formation doctorale et enseignements

Au cours des quatre dernières années, mes activités liées à la formation doctoral comprennent des encadrements de thèse, de stages de recherche de DEA/Mastère, et des cours en Mastère. En outre, j'ai eu l'occasion d'intervenir en école d'ingénieur, pour des cours ou des encadrements de projets.

2.2.1 Thèse encadrée soutenue

- **F. Desobry**

Sujet : Détection de ruptures à l'aide de noyaux

Directeur de thèse : C. Doncarli (IRCCyN, Ecole Centrale de Nantes)

Implication : co-encadrement – 90%

Publications : Deux revues internationales (Desobry *et al.* (2005a); Davy *et al.* (2006b)), deux conférences internationales (Desobry et Davy (2003b, 2004)) et une conférence nationale (Desobry et Davy (2003a)).

Durée de la thèse : 3 ans

Soutenance : le 23 septembre 2004 devant le jury composé de S. Canu (INSA de Rouen–Rapporteur), M. Davy (CNRS/LAGIS), C. Doncarli (Ecole Centrale de Nantes/ IRCCyN), P. Flandrin (CNRS/ENS Lyon–Président), B. Torrésani (Université de Provence–Rapporteur) et B. Schölkopf (Max Planck Institut, Tübingen, Allemagne).

Devenir du doctorant : Actuellement en post-doctorat à l'Université de Cambridge (Royaume Uni)

2.2.2 Thèses en cours

- **C. Dubois**

Sujet : Détection d'effets sonores dans l'environnement urbain

Directeur de thèse : G. Hégron (CERMA, Ecole d'architecture de Nantes, Ecole des Mines de Nantes)

Implication : co-encadrement – 90%

Publications : Une revue internationale soumise à *IEEE trans. on Audio Speech and Language processing*, deux conférences internationales (Dubois *et al.* (2005); Dubois et Davy (2005a)) et une conférence nationale (Dubois et Davy (2005c)).

Début de la thèse : Novembre 2003

Soutenance prévue : Octobre 2006.

Mobilité : C. Dubois a choisi de continuer sa thèse à Lille en octobre 2004, dans les locaux du LAGIS.

- **B. Fergani**

  Sujet : Segmentation en locuteurs de signaux de parole par méthodes à noyau et temps-fréquence

  Directeur de thèse : P. Vanheeghe (LAGIS, Ecole Centrale de Lille)

  Implication : co-encadrement – 66%

  Publications : Une conférences internationale (Fergani *et al.* (2006c)) et une conférence nationale (Fergani *et al.* (2006a)).

  Début de la thèse : Novembre 2003 sous la direction de I. Magrin-Chagnolleau et H. Paugam-Moisy, et depuis Octobre 2004

  Soutenance prévue : Octobre 2007.

  Activité professionnelle : B. Fergani est Professeur chargé de cours à l'Université USTHB d'Alger.

2.2.3 Participation à d'autres jurys de thèses

En plus du jury de thèse de F. Desobry, j'ai participé aux jurys de thèses suivants :

- **H. Cottereau** (IRCCyN, Nantes) en tant qu'examinateur (thèse dirigée par C. Doncarli)
- **V. Guigue** (PSI, Rouen) le 12 décembre 2005, en tant qu'examinateur (thèse dirigée par S. Canu)

2.2.4 Encadrement de stages de DEA/Mastère de Recherche

A Nantes et à Lille, j'ai participé à l'encadrement de six stages de DEA et de Mastère de recherche :

- **N. Leroy** sur la *mise en œuvre de méthodes MCMC pour la détection de défauts d'enceintes acoustiques*, 80% avec C. Doncarli (IRCCyN)
- **G. Ménard** sur la *détection d'ondes gravitationnelles*, 50% avec E. Chassande-Mottin (Observatoire de Nice)
- **M. Dallier** sur la *fusion de sonnées GPS - centrale inertielle*, 33 % avec Thalès Avionics Valence
- **S. Ridard** sur la *détection d'ondes gravitationnelles*, 33%, avec A. Doucet (Université de Cambridge, Royaume Uni)
- **B. Parent** sur la *recherche de conformations stables de protéines*, (participation sur les aspects Monte Carlo, avec D. Horvath de l'Institut de biologie de Lille)
- **L. Katekondji** sur la *recherche de conformations stables de protéines*, 50 % avec D. Horvath (Institut de biologie de Lille)
- **M. Loth** sur les *Méthodes à noyaux pour l'apprentissage par renforcement*, 33% avec R. Coulom (LIFL) et Ph. Preux (LIFL)

2.2.5 Cours de DEA/ Mastère de recherche

Depuis 2004, j'ai eu l'occasion d'enseigner à des élèves de Mastère de Recherche et de troisième année d'écoles d'ingénieur.

Intitulé	Année(s)	Lieu
Filtrage particulaire (4h de cours) (4h de Cours et 8h de TP)	 2004, 2006 2005	 Mastère Automatique et informatique appliquée, Ecole Centrale de Nantes
Data Mining (15h de cours)	2005	Mastère de Génie Industriel, Ecole Centrale de Lille
Méthodes à noyaux (3h de cours)	2005	Mastère d'Informatique, Université des Sciences et Technologies de Lille

2.2.6 Projets Elèves Ingénieurs

L'IRCCyN et le LAGIS sont situés sur les campus d'écoles d'ingénieurs, et j'ai été souvent amené à intervenir dans leurs projets ou stages de recherche afin de donner des avis scientifiques, ou proposer des solutions techniques (Ecole Centrale de Nantes et Ecole Centrale de Lille). Le volume horaire concerné est de 50 à 100 heures par an.

2.3 Responsabilités collectives et administration de la recherche

Au cours de ces quatre dernières années, j'ai eu plueisurs types de responsabilités dans l'animation, l'administration et la gestion de la recherche. D'une part, l'organisation d'écoles d'été, de sessions spéciales, de journées. Ensuite, la proposition et la gestion d'actions de recherche, et la participation à des réseaux.

2.3.1 Organisation des écoles d'été MLSS'03 et MLSS'04

Depuis octobre 2002 jusqu'à septembre 2004, j'ai participé à l'organisation de deux école d'été internationales (*Machine Learning Summer School* 2003 et 2004) sur le thème de l'apprentissage automatique (*Machine Learning*). Ces écoles sont destinées aux doctorants du monde entier, ainsi qu'aux chercheurs du monde universitaire et industriel. dans chacune de ces écoles d'été, plus de 20 nationalités étaient représentées, et les participants français étaient minoritaires.

La première a eu lieu à Tübingen, en Allemagne en août 2003, et la seconde dans l'Ile de Berder, France, en septembre 2004. Les autres organisateurs étaient O. Bousquet, F. Desobry, B. Schölkopf et A. Smola. Il est toujours possible de consulter les pages Internet de ces écoles l'adresse `http://www.mlss.cc`.

Plus précisément, ma participation à l'organisation à porté sur les tâches suivantes :
- Choix, invitation et prise en charge des professeurs
- Sélection des candidatures de participants
- Conception de l'emploi du temps, des activités, de l'organisation locale
- Enseignement des Méthodes de Monte Carlo (travaux pratiques 3×4h en 2003 et 2004) et du Traitement du Signal (8h de cours en 2004)

En outre, pour l'école d'été MLSS'04 dans l'Ile de Berder, j'ai géré l'ensemble de l'organisation locale. Je coordonne actuellement l'organisation d'une nouvelle école d'été qui aura lieu en France en 2008.

2.3.2 Organisation de sessions spéciales et d'un atelier

Les quatre dernières années, j'ai pu organiser les sessions spéciales et ateliers présentés dans le tableau suivant. Cela à consisté à rédiger des propositions aux organisateurs des conférences concernées, et à concevoir le programme scientifique.

Intitulé	**Année**	**Objet**
Conférence IEEE ICASSP'03	2003	Session spéciale *Kernel methods in Signal Processing*
Conférence GRETSI'03	2003	Session spéciale *Méthodes à noyaux en Traitement du signal*
Conférence ICML'06	2006	*Workshop* intitulé *Kernel methods and Reinforcement Learning* (avec Ph. Preux, R. Munos et C. Szepesvari)

2.3.3 Actions ponctuelles

En plus de ces actions d'animation de la communauté scientifique, j'ai participé à la création et la gestion d'actions des actions de recherche suivantes :

- de septembre 2002 à septembre 2004 : animation de l'**Action Spécifique 66** *Méthodes à Vecteurs Support et Méthodes à Noyaux* d'un budget de 60 kEuros. C'est dans le cadre de cette AS que j'ai organisé les deux sessions spéciales de conférences et les deux écoles d'été.
- En octobre 2002 et octobre 2003 : participation à la **rédaction du projet de réseau européen InRiCo** (*Information Retrieval in Context* – programme *Marie Curie Research Training Network*), en tant que responsable d'un *workpackage*. Ce projet a été évalué deux fois, a passé les seuils mais n'a pas été financé (notes 79.6/100 et 78.5/100)
- En mai 2003 : **expertise de projets scientifiques du 6ème PCRD** à la Commission Européenne à Bruxelles, pendant une semaine. En mai 2005 et mars 2006 : **expertises à un et deux ans** de l'un des projets qui a été financé en 2003.
- En avril 2004 : coordination de la **rédaction d'un projet d'école d'été européenne** *Marie Curie small conferences and training courses*. Ce projet a obtenu la note de 90/100, a passé tous les seuils mais n'a pas été financé.
- En mai 2005 : animation d'une **table ronde** sur la structuration de la recherche en apprentissage (30 mai 2005) pendant la conférence CAP tenue à Nice, qui a donné lieu à la **création de la FERA** (*Fédération des Équipes de Recherche en Apprentissage*). Dans le cadre de la FERA, nous avons organisé une journée FERA-GdR ISIS (en collaboration avec Michèle Sébag et Stéphane Canu), le 21 octobre 2005.
- En mars 2006 : co-rédaction d'un **projet ANR blanc** intitulé *KERNSIG*, portant sur l'utilisation des méthodes à noyaux en Traitement du Signal (avec O. Cappé – LTCI/ENST, S. Canu – PSI/INSA de Rouen, C. Richard LM2S/UTT).
- En avril 2006 : création de l'**équipe INRIA-FUTURS** «**SequeL**», avec R. Coulom (LIFL/Université de Lille III), E. Duflos (LAGIS/Ecole Centrale de Lille), R. Munos (CEMAP/École Polytechnique), Ph. Preux (LIFL/Université de Lille III), Ph. Vanheeghe (LAGIS/Ecole Centrale de Lille).

Depuis quatre ans, j'ai **expertisé des articles scientifiques** (une quinzaine par an) pour les revues *IEEE Trans. on Signal Processing*, *IEEE Signal Processing Letters*, *IEEE*

Trans. on Speech, Audio and Language Processing, *Speech Communications*, *Signal Processing*, *Traitement du Signal*, *IEEE Trans. on Circuits and Systems I*, *Control Engineering Practice*, etc.

2.4 Collaborations et partenariats académiques

Mes collaborations académiques sont de plusieurs types. D'une part, les collaborations « institutionnalisées », par exemple dans le cadre du GdR ISIS (Directeur J.-P. Cocquerez), de l'Action Spécifique 66 *Méthodes à Vecteurs Support et Méthodes à Noyaux*, de l'Action spécifique 67 *Méthodes particulaires* (responsables O. Cappé – LTCI/ENST et F. Legland – IRISA) et de l'équipe projet multi-laboratoire *Sémantique des Objets Sonores* (responsable D. Dubois – LAM). D'autre part, j'ai entretenu les collaborations de recherche détaillées ci-dessous.

En France (Les collaborations internes au laboratoire ne figurent pas dans la liste) :

- **S. Canu** de l'INSA de Rouen, sur les méthodes à noyaux en traitement du Signal (Davy *et al.* (2006a))
- **Ph. Preux, R. Coulom** du LIFL/Université de Lille III, et **R. Munos** du CEMAP/Ecole Polytechnique sur l'apprentissage par renforcement (une publication en cours de rédaction, et montage de l'équipe INRIA-FUTURS SequeL)

A l'étranger :

- **S. Godsill** de l'Université de Cambridge, pour l'analyse bayésienne de la musique harmonique (Godsill et Davy (2002); Davy et Godsill (2002c,b,a); Godsill et Davy (2003); Davy et Godsill (2003); Godsill et Davy (2005); Davy *et al.* (2006c))
- **A. Doucet** de l'Université de Colombie Britannique à Vancouver (Canada), sur les méthodes de Monte Carlo séquentielles. Certain de ces travaux ont également été menés avec C. Andrieu de l'Université de Bristol (Royaume Uni). Les publications suivantes sont issues de ces collaborations : Andrieu *et al.* (2001a, 2002a); Doucet *et al.* (2002); Andrieu *et al.* (2002b); Davy et Doucet (2003); Davy *et al.* (2003); Andrieu *et al.* (2003); Johansen *et al.* (2006); Caron *et al.* (2006a).
- **P.J. Wolfe** de l'Université Harvard à Boston (USA), pour les méthodes bayésiennes et l'analyse de l'environnement sonore. Cette collaboration en démarrage à donné lieu à la publication Davy et Wolfe (2005).
- **A. Klapuri** de l'Université de Tampere (Finlande), pour l'analyse automatique de la musique dans le cadre du livre Klapuri et Davy (2006).

2.5 Transfert technologique, relations industrielles et valorisation

Le tableau présenté ci-dessous résume les contrats de recherche industrielle dont j'ai négocié les modalités et que je gère d'un point de vue scientifique et financier.

Période	Montant	Intitulé
2003-2007	43 kEuros	Détection de défauts d'enceintes acoustiques, avec *Neutrik Test Instruments* (Liechtenstein), leader européen du matériel de test audio
2005-2006	56 kEuros	Amélioration des procédures de reconnaissance automatique de la parole, par des méthodes d'apprentissage à noyaux, avec l'équipe *reconnaissance de la parole* de France Telecom R&D, Lannion.
2005-2006	5 kEuros	Analyse de données industrielles, avec la petite société lilloise Innov'Process.

2.6 Mobilité

La période octobre 2000 – avril 2006 a donné lieu à deux mobilités importantes. Mon affectation initiale était l'IRCCyN, UMR 6597, Nantes.

- d'octobre 2000 à septembre 2001, l'Université de Cambridge m'a accueilli sur un poste de *research associate*, financé par le projet Européen MOUMIR (*Multimedia Information Retrieval*), sous la responsabilité de Simon Godsill. Le groupe d'accueil était le *Signal Processing Group* du *Department of Engineering*, dirigé à l'époque par le professeur Peter Rayner.
- d'octobre 2001 à juin 2002, j'ai poursuivi mon séjour comme chercheur invité à l'Université de Cambridge, Grande-Bretagne (dans le groupe *Signal Processing* du *dept. of Engineering*). Cette mobilité a permis d'encore renforcer mes connaissances dans les méthodes de Monte Carlo, et de finir plusieurs travaux entamés dans le domaine des signaux audio.
- En février 2004 : changement d'affectation, pour raisons familiales. Ma nouvelle affectation est le LAGIS, UMR 8146 à Villeneuve d'Ascq.

Par ailleurs, la période octobre 2000 – octobre 2005 a donné lieu à plusieurs missions de deux à trois semaines en 2002 (Université de Melbourne, Australie), 2003 (Max Plank Institut, Tübingen, Allemagne) et 2004 (Ile de Berder, France).

2.7 Publications

Les publications sont triées par catégories, et dans chaque catégorie, par ordre chronologique. Celles qui sont suivies d'un numéro de page sont annexées au document, à la page indiquée.

2.7.1 Livre édité

1. Klapuri et Davy (2006) – voir page 89 : «Signal Processing Methods for Music Transcription», A. Klapuri et M. Davy éditeurs, Springer, New-York, USA, avril 2006.

2.7.2 Chapitres de livres

1. Davy (2003) : M. Davy, «Classification», in «Décision dans le Plan Temps-Fréquence», Nadine Martin et Christian Doncarli éditeurs, traité IC2, Hermès, Paris, France, 2004.

2. Davy (2005a) : M. Davy, «Application de techniques temps-fréquence aux signaux sonores : reconnaissance et diagnostic», in «Temps-fréquence : concepts et outils», F. Hlawatsch et F. Auger éditeurs, traité IC2, Hermès, Paris, France, 2005.
3. Davy (2006a) : M. Davy, «An introduction to Statistical Signal Processing and Spectrum Estimation» in «Signal Processing Methods for Music Transcription», A. Klapuri et M. Davy éditeurs, Springer, New-York, USA, 2006.
4. Davy (2006b) : M. Davy, «Multiple F0 Frequency Estimation Based on Generative Models» in «Signal Processing Methods for Music Transcription», A. Klapuri et M. Davy éditeurs, Springer, New-York, USA, 2006.
5. Herrera-Boyer *et al.* (2006) : P. Herrera-Boyer, A. Klapuri et M. Davy, «Automatic Classification of Pitched Musical Instrument Sounds» in «Signal Processing Methods for Music Transcription», A. Klapuri et M. Davy éditeurs, Springer, New-York, USA, 2006.

2.7.3 Articles de journaux internationaux à comités de lecture

1. Davy *et al.* (2001b) : M. Davy, C. Doncarli et G.F. Boudreaux-Bartels, «Improved Optimization of Time-Frequency Based Classifiers», *IEEE Signal Processing Letters*, Vol. 8, No 2, pp. 52–57, février 2001 (6 pages).
2. Davy *et al.* (2002a) : M. Davy, C. Doncarli et J.Y. Tourneret, «Classification of Chirp Signals using Hierarchical Bayesian Learning and MCMC Methods», *IEEE Trans. on Signal Processing*, No 2, Vol. 50, pp. 377–388, février 2002 (12 pages).
3. Davy et Doncarli (2002) : M. Davy et C. Doncarli, «A New Nonstationary Test Procedure for Improved Loudspeaker Fault Detection», Journal of the Audio Engineers Society, No 6, Vol. 50, pp. 458-469, juin 2002 (12 pages).
4. Davy et Godsill (2002a) : M. Davy et S. Godsill, «Bayesian Harmonic Models for Musical Signal Analysis», *Bayesian Statistics 7*, Oxford University Press, pp. 105–118, 2002 (14 pages).
5. Davy *et al.* (2002b) – voir page 77 : M. Davy, A. Gretton, A. Doucet et P.W.J. Rayner, «Optimised Support Vector Machines for Nonstationary Signal Classification», *IEEE Signal Processing Letters*, Vol. 9, No 12, pp. 442-445, Décembre 2002 (4 pages).
6. Davy et Doucet (2003) : M. Davy and A. Doucet, «Copulas : A New Insight Into Positive Time-Frequency Distributions», *IEEE Signal Processing Letters*, Vol. 10, No 7, pp. 215-218, juillet 2003 (4 pages).
7. Andrieu *et al.* (2003) – voir page 79 : C. Andrieu, M. Davy et A. Doucet, «Efficient Particle Filtering for Jump Markov Systems», *IEEE Trans. on Signal Processing*, Vol. 51, No 7, pp. 1762-1769, juillet 2003 (8 pages).
8. Desobry *et al.* (2005a) – voir page 83 : F. Desobry, M. Davy et C. Doncarli, «An Online Kernel Change Detection Algorithm», *IEEE Trans. on Signal Processing*, Vol. 53, No. 8 (partie 2), pp. 2961-2974, août 2005 (14 pages).
9. Davy *et al.* (2006c) – voir page 85 : M. Davy, S. Godsill et J. Idier, «Bayesian Analysis of Polyphonic Western Tonal Music», *Journal of the Acoustical Society of America*, Volume 119, Numéro 4, pp. 2498-2517, avril 2006 (20 pages).

10. Potin *et al.* (2006) : D. Potin, P. Vanheeghe, E. Duflos et M. Davy, «An Abrupt Change Detection Algorithm for Buried Landmines Localization», *IEEE Trans. on Geoscience and Remote Sensing*, , Volume 44, Numéro 2, pp. 260–272, février 2006 (13 pages).
11. Davy *et al.* (2006b) : M. Davy, F. Desobry, A. Gretton et C. Doncarli, «An Online Support Vector Machine for Abnormal Events Detection», *Signal Processing*, Volume 86, Issue 8, pages 2009-2025, août 2006 (17 pages).
12. Dobigeon *et al.* (2006a) (à paraître) : N. Dobigeon, J.Y. Tourneret et M. Davy, «Joint Segmentation of Piecewise Constant Autoregressive Processes by Using a Hierarchical Model and a Bayesian Sampling Approach», *IEEE Trans. on Signal Processing*, 2006.
13. Caron *et al.* (2006c) (à paraître) : F. Caron, M. Davy, E. Duflos et P. Vanheeghe, «Particle Filtering for Multisensor Data Fusion with Switching Observation Models. Application to Land Vehicle Positioning», *IEEE Trans. on Signal Processing*, 2006.
14. Dubois et Davy (2005b) (à paraître) : C. Dubois et M. Davy, «Joint Detection and Tracking of Time-Varying Harmonic Components : an Online Bayesian Framework.», *IEEE Trans. on Speech and Audio Processing*, 2006.

Articles en cours d'expertise

15. Caron *et al.* (2006b) (en première lecture) : F. Caron, M. Davy, A. Doucet, E. Duflos et Ph. Vanheeghe, «Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures», soumis à *IEEE Trans. on Signal Processing* en avril 2006.
16. Fergani *et al.* (2006b) (en première lecture) : B. Fergani, M. Davy et A. Houacine, «Speaker Segmentation Using One Class Support Vector Machine», soumis à *Speech Communications* en juin 2006.

2.7.4 Conférences internationales avec actes à comité de lecture

1998 – 2000

1. Davy *et al.* (1998) : M. Davy, B. Leprettre, C. Doncarli et N. Martin, «Tracking of Spectral Lines in an ARCAP Time-Frequency Representation», EUSIPCO-98, 9th European Signal Processing Conference, Ile de Rhodes, Grèce, septembre 1998.
2. Davy et Doncarli (1998) : M. Davy et C. Doncarli, «Optimal Kernels of Time-Frequency Representations for Signal Classification», IEEE Int. Symp. on TFTS, Pittsburgh, USA, octobre 1998.
3. Cottereau *et al.* (2000) : H. Cottereau, M. Davy, C. Doncarli et N. Martin, «Using ARCAP Time-Frequency Representations for Decisions», EUSIPCO-2000, 10th European Signal Processing Conference, Tampere, Finlande, septembre 2000.
4. Davy *et al.* (2000a) : M. Davy, C. Doncarli, et J.Y. Tourneret, «Supervised Classification using MCMC Methods», IEEE ICASSP 2000, Istanbul, juin 2000.

2001

5. Davy *et al.* (2001a) : M. Davy, H. Cottereau et C. Doncarli, «Loudspeaker Fault Detection Using Time-Frequency Representations», IEEE ICASSP 2001, Salt Lake City, mai 2001.
6. Andrieu *et al.* (2001a) : C. Andrieu, M. Davy et A. Doucet, «Improved Auxiliary Particle Filtering : Applications to Time-Varying Spectral Analysis», IEEE SSP 2001, Singapour, août 2001.

7. Gretton *et al.* (2001) : A. Gretton, M. Davy, A. Doucet et P.W. Rayner, «Nonstationary Signal Classification Using Support Vector Machines», IEEE SSP 2001, Singapour, août 2001.

2002

8. Davy et Godsill (2002b) : M. Davy et S. Godsill, «Detection of Abrupt Spectral Changes using Support Vector Machines. An application to audio signal segmentation», IEEE ICASSP 2002, Orlando, USA, mai 2002.
9. Godsill et Davy (2002) : S. Godsill and M. Davy, «Bayesian harmonic models for musical pitch estimation and analysis», IEEE ICASSP 2002, Orlando, USA, mai 2002.
10. Doncarli *et al.* (2002) : C. Doncarli, M. Davy and J.Y. Tourneret, «Hierarchical Bayesian Classification of Chirp Signals», IEEE ICASSP 2002, Orlando, USA, mai 2002.
11. Andrieu *et al.* (2002a) : C. Andrieu, M. Davy and A. Doucet, «Efficient Particle Filtering for Jump Markov Systems», IEEE ICASSP 2002, Orlando, USA, mai 2002.
12. Andrieu *et al.* (2002b) : C. Andrieu, M. Davy and A. Doucet, «A Particle Filtering Technique for Jump Markov Systems», EUSIPCO-02, 11th European Signal Processing Conference, Toulouse, France, septembre 2002.
13. Doucet *et al.* (2002) : A. Doucet, V. Ba-Ngu, C. Andrieu et M. Davy, «Particle Filtering for Multi-Target Tracking and Sensor Management», Proceedings of the Fifth International Conference on Information Fusion, juillet 2002, Annapolis, USA.

2003

14. Cottereau *et al.* (2003) : H. Cottereau, J.M. Piasco, C. Doncarli et M. Davy, «Two approaches for the estimation of time-varying amplitude multichirp signals», IEEE ICASSP 2003, Hong-Kong, Chine, avril 2003.
15. Desobry et Davy (2003b) : F. Desobry et M. Davy, «Support Vector-Based Online Detection of Abrupt Changes», IEEE ICASSP 2003, Hong-Kong, Chine, avril 2003
16. Godsill et Davy (2003) : S. Godsill et M. Davy, «Bayesian harmonic models for musical pitch analysis», 54th Session of the International Statistical Institute (ISI), août 2003, Berlin, Allemagne.

2004

17. Davy et Idier (2004) : M. Davy et J. Idier, «Fast MCMC computations for the estimation of sparse processes from noisy observations», ICASSP 2004, Montréal, Canada, mai 2004.
18. Desobry et Davy (2004) – voir page 81 : F. Desobry et M. Davy, «Dissimilarity measures in feature space», ICASSP 2004, Montréal, Canada, mai 2004.

2005

19. Desobry *et al.* (2005b) : F. Desobry, M. Davy et W.J. Fitzgerald, «A Class of Kernels for Sets of Vectors», ESANN 2005, Bruges, Belgique, avril 2005.
20. Dubois et Davy (2005a) : C. Dubois et M. Davy, «Harmonic Tracking Using Sequential Monte Carlo», IEEE SSP 2005, Bordeaux, France, juillet 2005.
21. Davy (2005b) : M. Davy, «Bayesian separation of harmonic sources», ASA-Joint Statistical Meeting, Minneapolis, USA, août 2005.

22. Dubois *et al.* (2005) : C. Dubois, M. Davy et J. Idier, «Tracking of Time-Frequency Components Using Particle Filtering», ICASSP 2005, Philadelphie, USA, mars 2005.
23. Godsill et Davy (2005) : S. Godsill et M. Davy, «Bayesian Computational Models For Inharmonicity In Musical Instruments», WASPAA 2005, Mohonk, USA, octobre 2005.

2006

24. Johansen *et al.* (2006) : A. Johansen, A. Doucet et M. Davy, «Maximum Likelihood Parameter Estimation for Latent Variable Models using Sequential Monte Carlo», ICASSP 2006, Toulouse, France, mai 2006.
25. Davy *et al.* (2006a) – voir page 87 : M. Davy, F. Desobry et S. Canu, «Estimation of minimum measure sets in reproducing kernel Hilbert spaces and applications», ICASSP 2006, Toulouse, France, mai 2006.
26. Dobigeon *et al.* (2006b) : N. Dobigeon, J.Y. Tourneret et M. Davy, «Joint Segmentation of Piecewise Constant Autoregressive Processes by Using a Hierarchical Model and a Bayesian Sampling Approach», ICASSP 2006, Toulouse, France, mai 2006.
27. Caron *et al.* (2006a) (à paraître) : F. Caron, M. Davy, A. Doucet, E. Duflos et P. Vanheeghe, «Bayesian Inference for Dynamic Models with Dirichlet Process Mixtures», Fusion 2006, Florence, Italie, juin 2006.
28. Fergani *et al.* (2006c) (à paraître) : B. Fergani, M. Davy et A. Houacine, «Unsupervised Speaker Indexing Using One-Class Support Vector Machines», EUSIPCO 2006, Florence, Italie, Septembre 2006.

Articles en cours d'expertise

29. E. Jackson, M. Davy, A. Doucet et W. fitzgerald, «Unsupervised Functional Classification using a Dirichlet Process Mixture of Gaussian Processes», soumis à *NIPS 2006* en juin 2006.
30. A. Rakotomamonjy et M. Davy, «One-class SVM regularization path and Multiclass Application», soumis à *NIPS 2006* en juin 2006.
31. F. Desobry, M. Davy et W. fitzgerald, «Density kernels on unordered sets», soumis à *NIPS 2006* en juin 2006.

2.7.5 Conférences nationales avec actes à comités de lecture

1. Davy et Doncarli (1999) : M. Davy et C. Doncarli, «Distances et critères de contraste dans le plan temps-fréquence», 17ème Colloque GRETSI, Vannes, France, septembre 1999, pp. 287-290.
2. Desobry et Davy (2003a) : F. Desobry et M. Davy, «Détection de ruptures en ligne par estimateur SVM de support de densité», 19ème Colloque GRETSI, Paris, France, septembre 2003.
3. Davy *et al.* (2003) : M. Davy, P. Del Moral, et A. Doucet, «Méthodes de Monte Carlo Séquentielles pour l'Analyse Spectrale Bayésienne», 19ème Colloque GRETSI, Paris, France, septembre 2003.
4. Davy et Wolfe (2005) : M. Davy et P. Wolfe, «Une Méthode à Noyaux pour l'Approximation Parcimonieuse des Représentations Temps-Fréquence Bilinéaires», 20ème Colloque GRETSI, Louvain-la-Neuve, Belgique, septembre 2005.

5. Dubois et Davy (2005c) : C. Dubois et M. Davy, «Suivi de Trajectoires Temps-Fréquence par Filtrage Particulaire», 20ème Colloque GRETSI, Louvain-la-Neuve, Belgique, Septembre 2005.
6. Caron *et al.* (2005) : F. Caron, M. Davy, E. Duflos et P. Vanheeghe, «Fusion de Capteurs Potentiellement Défaillants par Filtrage Particulaire», 20ème Colloque GRETSI, Louvain-la-Neuve, Belgique, septembre 2005.
7. Fergani *et al.* (2006a) (à paraître) : B. Fergani, M. Davy et A. Houacine, «Application des Machines à Vecteurs Support Mono-Classe à L'indexation en Locuteurs de Documents Audio», Journées d'étude de la parole, Dinard, France, juin 2006.
8. Loth *et al.* (2006) (à paraître) : M. Loth, R. Coulom, M. Davy et P. Preux, "Least Angle Temporal Difference Learning : LATD(lambda)", Journ es Francophones Planification, D cision, Apprentissage, Toulouse, France, Mai 2006.

2.7.6 Conférences avec actes sans comité de lecture

1. Martin *et al.* (1998) : N. Martin, C. Doncarli, et M. Davy, «Méthodes de décision dans un contexte de signaux non stationnaires», Third SFM-IMEKO-SFA International Conference on Acoustic and Vibratory Surveillance Methods and Diagnostic Techniques, Senlis, France, 13-15 octobre 1998
2. Davy et Godsill (2003) : M. Davy et S. Godsill, «Bayesian Harmonic Models for Musical Signal Analysis», Cambridge Music Processing Colloquium, 28 mars 2003, Cambridge, Grande-Bretagne.

2.7.7 Séminaires et tutoriaux

1. M. Davy et P. Gonçalvès, «Optimisation des noyaux de représentations temps-fréquence», Journée Tutoriale du GT2, GdR-PRC ISIS, 28 avril 1998.
2. M. Davy, «Distributions Conjointes des Signaux», INRIA, Séminaire du projet Fractales, 8 juin 1998.
3. M. Davy, C. Doncarli et J.Y. Tourneret, «Méthodes MCMC pour l'apprentissage bayésien», Journée Tutoriale du CNES, 11 mai 2000.
4. M. Davy, C. Doncarli et J.Y. Tourneret, «Bayesian learning using MCMC methods», séminaire du *Signal Processing Group*, Université de Cambridge, juin 2000.
5. M. Davy, C. Doncarli et J.Y. Tourneret, «Méthodes MCMC pour l'apprentissage bayésien», réunion plénière du GT2, GdR ISIS, 5 décembre 2000.
6. M. Davy et P.J. Wolfe, «Improved Audio Feature Extraction», *in* Probabilistic Models in Vision and Signal Processing Meeting, British Machine Vision Association and Society for Pattern Recognition, Londres, 9 mai 2001.
7. M. Davy, «Audio Features for Classification, Segmentation, Indexation and Retrieval of Audio», réunion du projet MOUMIR, Lisbonne, 30 mars 2001.
8. M. Davy, «Filtrage non-linéaire et non-gaussien : filtres particulaires», séminaire de l'IRCCyN, 6 septembre 2001.
9. M. Davy, C. Andrieu et A. Doucet, «Advanced Particle Filtering Techniques for Jump Markov Systems», Journée de la *Royal Statistical Society*, Bristol, 19 octobre 2001.

10. M. Davy et S. Godsill, «Audio information retrieval at Cambridge University — Audio signals segmentation», réunion du projet MOUMIR, Dublin, 15 mars 2002.
11. M. Davy, C. Andrieu et A. Doucet, «Advanced Particle Filtering Techniques for Jump Markov Systems», Journée Steven Kay, GdR ISIS, Paris, 20 mars 2002.
12. M. Davy, C. Andrieu et A. Doucet, «An efficient particle filtering technique for jump Markov systems», Séminaire interne, Université de Cambridge, 2 mai 2002.
13. M. Davy, «Une Introduction aux Méthodes à Vecteurs Support et autres Algorithmes à Noyaux», séminaire à l'assemblée Générale du GdR ISIS, Dourdan, 21 mars 2003.
14. M. Davy, «Une Introduction aux Méthodes à Vecteurs Support et autres Algorithmes à Noyaux», séminaire de l'Institut de Recherche en Informatique, Toulouse, 11 décembre 2003.
15. M. Davy, «A Bayesian Harmonic model for polyphonic music analysis», séminaire du *Signal Processing Group*, Université de Cambridge, Cambridge, Grande-Bretagne, 29 juin 2004.
16. M. Davy, «A Bayesian Harmonic model for polyphonic music analysis», séminaire du *Audio Processing Group*, Institut de Traitement du Signal, Université de Tampere, Tampere, Finlande, 4 juillet 2004.
17. M. Davy, «A Bayesian Harmonic model for polyphonic music analysis», 1st Hanse Workshop on HEARing Research, Hanse Institut for Advanced Study, Delmenhorst, Allemagne, 22-24 août 2004.
18. M. Davy, «Un modèle Autorégressif à coefficients variables. Modèle Bayésien et algorithme», Séminaire de Probabilités et Statistiques, Laboratoire Paul Painlevé, Villeneuve d'Ascq, 10 novembre 2004.
19. M. Davy, «Filtrage non-linéaire et non-gaussien : filtres particulaires», Séminaire du GRAPPA, Villeneuve d'Ascq, 10 mars 2005.
20. M. Davy, «An introduction to Support Vector Machines and Other Kernel Algorithms», Tutorial, 9th International conference on Engineering Applications of Neural Networks (EANN'05), Lille, août 2005.

2.7.8 Rapports techniques (non publiés par ailleurs)

1. M. Davy et S.J. Godsill, «Audio information retrieval : A bibliographical study», rapport technique CUED/F-INFENG/TR.429, Université de Cambridge, février 2002.
2. M. Davy et S.J. Godsill, «Bayesian Harmonic Models for musical pitch estimation and analysis», rapport technique CUED/F-INFENG/TR.431, Université de Cambridge, novembre 2002.

2.7.9 Autres

– En Novembre 2003, le Journal de la région des Pays-de-la-Loire (numéro 78) a consacré un article à l'aide qu'elle apporte aux jeunes chercheurs. A cette occasion, j'ai été interviewé, et la transcription a été publiée dans un encadré en page 2.
– Le journal du CNRS m'a consacré un article de la rubrique « jeune chercheur », paru en juin 2006 (numéro 197, page 17).

Chapitre 3

Cadre théorique et méthodologique

Dans ce chapitre, je présente le cadre théorique et méthodologique dans lequel s'inscrivent la majorité de mes contributions. Je ne prétends pas à l'exhaustivité, aussi me suis-je concentré sur l'essentiel. Ici, deux cadres méthodologies et théoriques sont abordés. Le premier concerne l'inférence bayésienne et les méthodes de Monte Carlo (section 3.1). Le second aborde les méthodes à noyaux (section 3.2) utilisées pour l'apprentissage automatique.

3.1 Méthodes de Monte Carlo

La théorie bayésienne propose un cadre général d'inférence statistique. Le concept central est celui de distribution *a posteriori*, qui provient du produit de la vraisemblance par une distribution *a priori*. Dans la suite, $\boldsymbol{\theta} \in \Theta$ désigne le paramètre inconnu d'un modèle décrivant des données (ou observations) notées collectivement X. A partir de la distribution *a posteriori* $\mathsf{p}(\boldsymbol{\theta}|\mathsf{X})$, telle que[1]

$$\mathsf{p}(\boldsymbol{\theta}|\mathsf{X}) \;\propto\; \mathsf{p}(\mathsf{X}|\boldsymbol{\theta})\mathsf{p}(\boldsymbol{\theta}), \tag{3.1}$$

(où $\propto$ signifie « proportionnel à ») il est possible de construire des estimateurs. Les plus populaires sont :

- le *maximum a posteriori* (MAP). La valeur estimée est celle pour laquelle la distribution *a posteriori* est maximale soit, dans le cas où celle-ci admet une densité

$$\widehat{\boldsymbol{\theta}}_{\text{MAP}} \;=\; \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmax}}\; \mathsf{p}(\boldsymbol{\theta}|\mathsf{X}) \tag{3.2}$$

- le minimum d'erreur quadratique (*Minimum mean square error – MMSE*). L'estimateur est obtenu comme l'espérance a posteriori

$$\widehat{\boldsymbol{\theta}}_{\text{MMSE}} \;=\; \mathbb{E}_{\mathsf{p}(\boldsymbol{\theta}|\mathsf{X})}[\boldsymbol{\theta}] \tag{3.3}$$

Bien qu'ils soient définis simplement, ces estimateurs sont souvent difficiles à calculer numériquement. Par exemple, l'estimateur MMSE requiert le calcul de l'intégrale suivante :

$$\widehat{\boldsymbol{\theta}}_{\text{MMSE}} \;=\; \int_{\Theta} \boldsymbol{\theta}\mathsf{p}(\boldsymbol{\theta}|\mathsf{X})\, d\boldsymbol{\theta} \tag{3.4}$$

[1]Par souci de simplification des notations, nous supposons ici que toutes les distributions admettent une densité par rapport à, par exemple, la mesure de Lebesgue. On peut lever cette hypothèse aisément.

Dans la plupart des problèmes d'inférence provenant de données réelles, la dimension de l'espace Θ peut être très grande (parfois supérieure à 1000). Lorsque le calcul analytique de (3.4) ne peut être effectué, une technique d'intégration numérique est inévitable. Par exemple, considérons le calcul numérique de l'intégrale :

$$I\left[\mathsf{h}\right] \;=\; \int_{\Theta} \mathsf{h}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{3.5}$$

où $\pi(\boldsymbol{\theta})$ dénote une densité quelconque de $\boldsymbol{\theta}$, par exemple la densité *a posteriori* $\mathsf{p}(\boldsymbol{\theta}|\mathsf{X})$. Quand la dimension de Θ est petite (typiquement, inférieure à 3), $I\left[\mathsf{h}\right]$ peut être calculée numériquement à la Riemann en utilisant une grille régulière (sous réserve que Θ soit un compact). Par exemple, si $\boldsymbol{\theta}$ est mono-dimensionnel, et $\Theta = [0, 1]$, l'intégration à la Riemann consiste à calculer :

$$I\left[\mathsf{h}\right] \;\approx\; \frac{1}{N}\sum_{i=1}^{N} \mathsf{h}(i/N)\pi(i/N) \tag{3.6}$$

Cette méthode atteint vite sa limite, car le nombre de points de la grille augmente exponentiellement avec la dimension de l'espace Θ, notée d_Θ : en supposant que la grille comporte 100 points dans chaque dimension, et que $d_\Theta = 50$, le nombre total de points de la grille est $100^{50} = 10^{100}$: cette méthode de calcul est inapplicable.

Une autre technique de calcul numérique peut toutefois être mise en œuvre. Supposons que des échantillons aléatoires $\widetilde{\boldsymbol{\theta}}^{(i)}$, $i = 1, \ldots, N$ sont disponibles, chacun d'entre eux étant distribué selon $\pi(\boldsymbol{\theta})$ (ce qui est noté $\widetilde{\boldsymbol{\theta}}^{(i)} \sim \pi(\boldsymbol{\theta})$ dans la suite). Alors, par la loi des grands nombres,

$$\widehat{I}_N\left[\mathsf{h}\right] \;=\; \frac{1}{N}\sum_{i=1}^{N} \mathsf{h}\big(\widetilde{\boldsymbol{\theta}}^{(i)}\big) \;\approx\; I\left[\mathsf{h}\right] \tag{3.7}$$

Les échantillons aléatoires $\widetilde{\boldsymbol{\theta}}^{(i)}$, $i = 1, \ldots, N$ sont généralement appelés *échantillons de Monte Carlo* et $\widehat{I}_N\left[\mathsf{h}\right]$ l'*estimation Monte Carlo* de $I\left[\mathsf{h}\right]$. D'un point de vue formel, l'approximation de Monte Carlo consiste à remplacer la vraie distribution $\pi(d\boldsymbol{\theta})$ par la distribution empirique

$$\mathsf{P}_N(d\boldsymbol{\theta}) \;=\; \sum_{i=1}^{N} \frac{1}{N}\delta_{\widetilde{\boldsymbol{\theta}}^{(i)}}(d\boldsymbol{\theta}) \tag{3.8}$$

où $\delta_a(du)$ est la distribution de Dirac en a, pour la variable u. Ainsi, l'approximation de Monte Carlo consiste à remplacer une espérance par rapport à $\pi(\cdot)$ par une espérance par rapport à $\mathsf{P}_N(\cdot)$. En effet,

$$\widehat{I}_N\left[\mathsf{h}\right] \;=\; \mathbb{E}_{\mathsf{P}_N(\cdot)}\big[\mathsf{h}(\cdot)\big] \tag{3.9}$$

L'estimée $\widehat{I}_N\left[\mathsf{h}\right]$ est non-biaisée pour tout N, c'est-à-dire $\mathbb{E}_{\pi(\widetilde{\boldsymbol{\theta}}^{(1)},\ldots,\widetilde{\boldsymbol{\theta}}^{(N)})}\big(\widehat{I}_N\left[\mathsf{h}\right]\big) = I\left[\mathsf{h}\right]$, et consistante. Finalement, l'*erreur quadratique empirique* $\widehat{\sigma}^2_N\left[\mathsf{h}\right]$ fournit une indication fiable de la variance de $\widehat{I}_N\left[\mathsf{h}\right]$ (dans le sens où, asymptotiquement quand $N \to \infty$, $\widehat{\sigma}^2_N\left[\mathsf{h}\right]$ tend vers la vraie variance d'estimation) :

$$\widehat{\sigma}^2_N\left[\mathsf{h}\right] \;=\; \frac{1}{N}\sum_{i=1}^{N} \Big[\mathsf{h}\big(\widetilde{\boldsymbol{\theta}}^{(i)}\big) - \widehat{I}_N\left[\mathsf{h}\right]\Big]^2 \tag{3.10}$$

Les méthodes de Monte Carlo peuvent en outre être utilisées pour calculer d'autres types d'estimées, y compris en dehors du cadre bayésien. Par exemple, les méthodes d'*optimisation*

de Monte Carlo peuvent être appliquées pour calculer des estimées par maximum de vraisemblance ou par MAP, voir Robert et Casella (2000); Del Moral et Doucet (2003).

Jusqu'à présent, nous avons supposé que les échantillons de Monte Carlo sont disponibles. La véritable difficulté réside en fait dans la génération des échantillons de Monte Carlo. Lorsque $\pi(\boldsymbol{\theta})$ est une densité standard (par exemple uniforme ou gaussienne), il est possible d'utiliser directement des générateurs de variables aléatoires. Dans les autres cas, typiquement quand $\pi(\boldsymbol{\theta})$ est une densité *a posteriori* en grande dimension, de constante de normalisation inconnue, le problème devient difficile. Plusieurs techniques ont été proposées dans le but de générer des échantillons à partir d'une distribution de probabilité donnée. Nous présentons ici les méthodes de Monte Carlo par Chaînes de Markov (MCMC) et l'échantillonnage préférentiel (*importance sampling*). Le lecteur pourra se reporter au travail de Robert et Casella (2000) pour une présentation complète de ces méthodes..

3.1.1 Méthodes de Monte Carlo par Chaînes de Markov (MCMC)

Le principe des algorithmes MCMC est le suivant : étant donnée une densité $\pi(\boldsymbol{\theta})$ à échantillonner, une chaîne d'échantillons est générée itérativement, voir l'algorithme 3.1 ci-dessous. La chaîne est (statistiquement) entièrement déterminée par la distribution de l'échantillon initial $\widetilde{\boldsymbol{\theta}}^{(0)}$, notée $\pi_0(\boldsymbol{\theta})$ et du *noyau de Markov* $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')$, qui est une distribution par rapport à $\boldsymbol{\theta}$ (à $\boldsymbol{\theta}'$ fixé). Sous réserve que $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ vérifie certaines propriétés (Robert et Casella (2000)), la distribution de chaque échantillon $\widetilde{\boldsymbol{\theta}}^{(i)}$ converge vers la distribution cible $\pi(\boldsymbol{\theta})$ lorsque i augmente. En particulier, le noyau doit être construit de telle façon que $\pi(\boldsymbol{\theta})$ est la distribution invariante de $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')$, c'est-à-dire,

$$\int_{\Theta} \mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')\pi(\boldsymbol{\theta}')\, d\boldsymbol{\theta}' \;=\; \pi(\boldsymbol{\theta}) \quad \text{pour tout} \quad \boldsymbol{\theta} \in \Theta \tag{3.11}$$

Algorithme 3.1: Algorithme MCMC générique

- Initialisation : échantillonner $\widetilde{\boldsymbol{\theta}}^{(0)} \;\sim\; \pi_0(\boldsymbol{\theta})$.
- itérations : pour $i = 1, 2, \ldots, N$, faire
  - Echantillonner $\widetilde{\boldsymbol{\theta}}^{(i)} \;\sim \mathcal{K}(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}^{(i-1)})$

Les algorithmes MCMC fournissent une série d'échantillons qui sont distribués asymptotiquement selon $\pi(\boldsymbol{\theta})$. En pratique, il est nécessaire de générer des « échantillons de chauffage » avant d'atteindre la convergence. Les échantillons de chauffage ne sont pas conservés dans le calcul d'estimées par Monte Carlo à l'équation (3.7). La figure 3.1 montre des réalisations typiques de chaînes de Markov, incluant les itérations de chauffage : les courbes montrent des fluctuations amples au début (pendant la phase de chauffage) puis les fluctuations sont de plus faible amplitude, autour de la valeur la plus probable selon $\pi(\boldsymbol{\theta})$.

Il peut être délicat de construire de toutes pièces un noyau de Markov fournissant des échantillons correctement distribués. Par bonheur, deux algorithmes simples permettent de construire facilement de tels noyaux : l'*échantillonneur de Gibbs*, basé sur de l'échantillonnage selon des distributions conditionnelles et l'*algorithme de Metropolis-Hastings (MH)*, basé sur des acceptations/rejets d'échantillons.

L'échantillonneur de Gibbs

Supposons que l'on souhaite générer des échantillons de Monte Carlo selon $\mathsf{p}(\boldsymbol{\theta})$, avec $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_{d_\Theta}]^\mathsf{T}$, où d_Θ est la dimension de l'espace Θ. Supposons en outre que l'on peut échantillonner aisément selon chacune des distributions conditionnelles $\mathsf{p}(\theta_1|\theta_2, \ldots, \theta_{d_\Theta})$, $\mathsf{p}(\theta_2|\theta_1, \theta_3, \ldots, \theta_{d_\Theta}), \ldots, \mathsf{p}(\theta_{d_\Theta}|\theta_1, \ldots, \theta_{d_\Theta-1})$. Typiquement, cette situation se rencontre lorsque certaines distributions sont gaussiennes tandis que d'autres sont des distributions gamma. L'échantillonneur de Gibbs consiste à générer une composante (ou un bloc de composantes) de $\boldsymbol{\theta}$ à la fois, à partir de la distribution *a posteriori* conditionnelle, voir l'algorithme 3.2.

Algorithme 3.2: L'échantillonneur de Gibbs

- Initialisation : échantillonner $\widetilde{\boldsymbol{\theta}}^{(0)} \sim \pi_0(\boldsymbol{\theta})$.
- itérations : pour $i = 1, 2, \ldots, N$, et pour $j = 1, \ldots, d_\Theta$, faire
  - Echantillonner $\widetilde{\theta}_1^{(i)} \sim \mathsf{p}(\theta_1|\widetilde{\theta}_2^{(i-1)}, \ldots, \widetilde{\theta}_{d_\Theta}^{(i-1)})$
  - Echantillonner $\widetilde{\theta}_2^{(i)} \sim \mathsf{p}(\theta_2|\widetilde{\theta}_1^{(i)}, \widetilde{\theta}_3^{(i-1)}, \ldots, \widetilde{\theta}_{d_\Theta}^{(i-1)})$
  - ...
  - Echantillonner $\widetilde{\theta}_{d_\Theta}^{(i)} \sim \mathsf{p}(\theta_{d_\Theta}|\widetilde{\theta}_1^{(i)}, \ldots, \widetilde{\theta}_{d_\Theta-1}^{(i)})$

Dans l'algorithme 3.2 ci-dessus, chaque θ_j, $j = 1, \ldots, d_\Theta$ peut aussi représenter un groupe de variables. L'échantillonneur de Gibbs est simple, cependant il nécessite de pouvoir échantillonner directement selon les distributions conditionnelles. Cela n'étant pas toujours possible, on peut utiliser à la place l'algorithme de MH.

L'algorithme de Metropolis-Hastings

En plus de la distribution cible $\pi(\boldsymbol{\theta})$, l'algorithme MH requiert une *distribution de proposition* $\mathsf{q}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ à partir de laquelle on peut générer des échantillons directement. Elle doit vérifier $\mathsf{q}(\boldsymbol{\theta}|\boldsymbol{\theta}') \neq 0$ quand $\pi(\boldsymbol{\theta}) \neq 0$. Le principe de l'algorithme de Metropolis-Hastings est présenté ci-dessous

---

Algorithme 3.3: l'algorithme de Metropolis-Hastings

- Initialisation : échantillonner $\widetilde{\boldsymbol{\theta}}^{(0)} \sim \pi_0(\boldsymbol{\theta})$.
- itérations : pour $i = 1, 2, \ldots, N$ faire
  - Echantillonner une valeur candidate $\boldsymbol{\theta}^\star$ pour le paramètre $\boldsymbol{\theta}^\star \sim q(\boldsymbol{\theta}|\widetilde{\boldsymbol{\theta}}^{(i-1)})$
  - Calculer
$$\alpha_{\mathsf{MH}}(\boldsymbol{\theta}^\star, \widetilde{\boldsymbol{\theta}}^{(i-1)}) = \min\left[1, \frac{\pi(\boldsymbol{\theta}^\star)}{\pi(\widetilde{\boldsymbol{\theta}}^{(i-1)})} \frac{q(\widetilde{\boldsymbol{\theta}}^{(i-1)}|\boldsymbol{\theta}^\star)}{q(\boldsymbol{\theta}^\star|\widetilde{\boldsymbol{\theta}}^{(i-1)})}\right] \tag{3.12}$$
  - Avec la probabilité $\alpha_{\mathsf{MH}}(\boldsymbol{\theta}^\star, \widetilde{\boldsymbol{\theta}}^{(i-1)})$, accepter le candidat, c'est-à-dire faire $\widetilde{\boldsymbol{\theta}}^{(i)} \leftarrow \boldsymbol{\theta}^\star$
  - Sinon (c'est-à-dire avec la probabilité $1 - \alpha_{\mathsf{MH}}(\boldsymbol{\theta}^\star, \widetilde{\boldsymbol{\theta}}^{(i-1)})$, rejeter le candidat en faisant $\widetilde{\boldsymbol{\theta}}^{(i)} \leftarrow \widetilde{\boldsymbol{\theta}}^{(i-1)}$

Dans l'algorithme 3.3, il est clair que le noyau $\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ résultat de l'algorithme de MH est déterminé par la distribution de proposition $\mathsf{q}(\boldsymbol{\theta}|\boldsymbol{\theta}')$. Une remarque importante est que $\pi(\boldsymbol{\theta})$ apparaît sous la forme d'un rapport : cette distribution peut n'être connue qu'à une constante

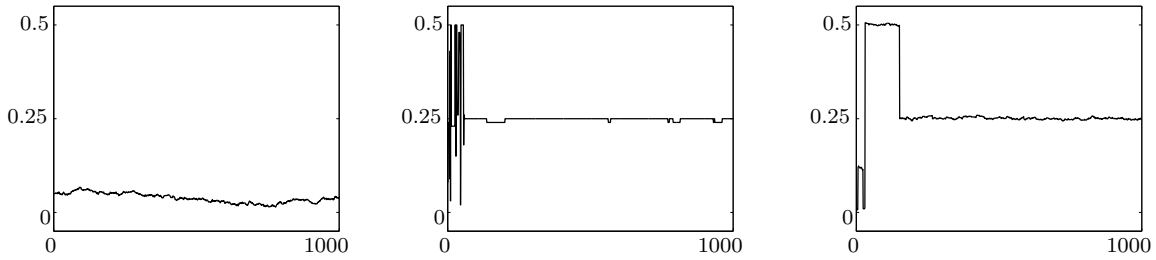
multiplicative près (ce qui est le plus souvent le cas lorsque $\pi(\boldsymbol{\theta})$ est une distribution *a posteriori*). Différents choix de la distribution de proposition conduisent à différents algorithmes MH. Trois cas importants sont présentés ci-dessous.

- Choisir $\mathsf{q}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathcal{N}\big(\boldsymbol{\theta}; \boldsymbol{\theta}', \boldsymbol{\Sigma}\big)$ où $\mathcal{N}\big(a; b, c\big)$ désigne la loi gaussienne pour la variable a, de moyenne b et de covariance c. Cela correspond à une *marche aléatoire gaussienne*, qui est symétrique au sens où $\mathsf{q}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathsf{q}(\boldsymbol{\theta}'|\boldsymbol{\theta})$, et le rapport de distributions (3.12) ne dépend pas de q. Ce type de distribution de proposition est souvent appelé *locale* car elle fait évoluer la chaîne localement autour du dernier échantillon accepté. Ces chaînes ont le mérite de ne pas rester au même point trop longtemps : la proportion de candidats acceptés peut généralement être réglée autour de 50% en ajustant la matrice de covariance $\boldsymbol{\Sigma}$. Toutefois, les distributions de proposition locales ne peuvent pas explorer les différentes régions de l'espace Θ très rapidement.
- Choisir $\mathsf{q}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathsf{q}(\boldsymbol{\theta})$, indépendant de $\boldsymbol{\theta}'$. Cette distribution de proposition est qualifiée d'*indépendante*, ou encore *globale*. Contrairement à la version locale, elle ne considère pas le dernier échantillon de la chaîne pour construire le candidat. Cela facilite en principe de grands sauts d'une région de Θ à une autre, et facilite la convergence. Cependant, elle est aussi à l'origine de points stationnaires de la chaîne, dans lesquels tout nouveau candidat est rejeté pendant de nombreuses itérations successives (i.e., $\widetilde{\boldsymbol{\theta}}^{(i)} = \widetilde{\boldsymbol{\theta}}^{(i+1)} = \ldots = \widetilde{\boldsymbol{\theta}}^{(i+I)}$ où I est typiquement de l'ordre de grandeur de 100, 1000 voire 10000). En pratique, il est crucial de construire de telles distributions de proposition en incorporant des heuristiques favorisant les régions de Θ où $\pi(\boldsymbol{\theta})$ a des chances d'être grande.
- Pour chaque composante θ_j ($j = 1, \ldots, d_\Theta$, d_Θ étant la dimension de Θ) utiliser une loi de proposition $\mathsf{q}(\theta_j|\theta'_j)$, sans faire évoluer les autres composantes. D'une manière globale, cette distribution de proposition s'écrit $\mathsf{q}_j(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathsf{q}(\theta_j|\theta'_j) \prod_{i=1, i\neq j}^{d_\Theta} \delta_{\theta'_i}(\theta_i)$ pour tout $j = 1, \ldots, d_\Theta$, où $\delta_u(v)$ est la fonction delta de Dirac. Ce type de distribution de proposition conduit à un algorithme appelé *une variable à la fois* car il ne met à jour qu'une composante de $\boldsymbol{\theta}$ à chaque itération, en utilisant une proposition locale ou globale. Une extension simple est de considérer des blocs de variables à mettre à jour, plutôt que des variables scalaires.

La figure 3.1 présente le tracé de chaînes de Markov générées à partir des distributions de proposition précédentes. Ces noyaux MH peuvent être mélangés, permettant ainsi de construire d'autres noyaux admissibles : étant donnée une famille de noyaux de Markov $\{\mathcal{K}_j(\boldsymbol{\theta}|\boldsymbol{\theta}'), j = 1, \ldots, J\}$ ayant la même distribution invariante $\pi(\boldsymbol{\theta})$, et des coefficients positifs $\{\beta_j, j = 1, \ldots, J\}$ tels que $\sum_{j=1}^{J} \beta_j = 1$, alors le noyau suivant

$$\mathcal{K}(\boldsymbol{\theta}|\boldsymbol{\theta}') \;=\; \sum_{j=1}^{J} \beta_j \mathcal{K}_j(\boldsymbol{\theta}|\boldsymbol{\theta}') \tag{3.13}$$

admet également $\pi(\boldsymbol{\theta})$ comme distribution invariante. En pratique, ce noyau est mis en œuvre comme suit : une variable discrète u est choisie dans $\{1, 2, \ldots, J\}$ avec la probabilité $\mathsf{P}(u = j) = \beta_j$, et l'on échantillonne selon le noyau MH $\mathcal{K}_j(\boldsymbol{\theta}|\boldsymbol{\theta}')$. Cela permet par exemple de construire des algorithmes de type *une variable à la fois*, dans lequel chaque composante est mise à jour de façon locale ou globale.



Loi de proposition locale | Loi de proposition globale | Loi de proposition mélangée

FIG. 3.1 – Chaînes de Markov typiques produites par différents noyaux de Markov, c'est-à-dire pour différentes distributions de proposition. Le paramètre échantillonné est une fréquence dont la « vraie »valeur est 0.25. a) La loi de proposition est une marche aléatoire gaussienne (locale) b) la loi de proposition est indépendante de l'itération précédente (globale) c) La loi de proposition résulte du mélange des lois de proposition locale et globale. La loi locale ne réussit pas à explorer rapidement l'espace des fréquences, tandis que la loi globale conduit à des blocages et la chaîne n'évolue presque plus. La loi obtenue par mélange local/global permet à la fois d'explorer rapidement l'espace et de l'explorer localement autour d'une valeur probable (un *mode* de la distribution *a posteriori*).

Sauts réversibles

Dans ce paragraphe, nous abandonnons l'hypothèse que la distribution cible admet une densité. Cela couvre les cas où, par exemple, l'espace Θ est discret, les cas où il est composite discret/continu et les cas où il a une structure emboîtée. La distribution cible est notée $\pi(d\boldsymbol{\theta})$. Dans le cas de l'algorithme de Metropolis-Hastings, le ratio (3.12) s'écrit sous la forme d'une dérivée de Radon-Nikodym,

$$\alpha_{\mathrm{MH}}(\boldsymbol{\theta}, \boldsymbol{\theta}') \;=\; \min\left[1, \frac{\pi(d\boldsymbol{\theta})}{\pi(d\boldsymbol{\theta}')} \frac{\mathcal{Q}(d\boldsymbol{\theta}'|\boldsymbol{\theta})}{\mathcal{Q}(d\boldsymbol{\theta}|\boldsymbol{\theta}')}\right] \tag{3.14}$$

où $\mathcal{Q}$ désigne le noyau de Markov de proposition sélectionné. Cette expression permet d'écrire par exemple les algorithmes à sauts réversibles (Richardson et Green (1997); Andrieu *et al.* (2001b)), dans lesquels $\boldsymbol{\theta}$ et $\boldsymbol{\theta}'$ peuvent ne pas avoir la même dimension. Plus précisément, l'algorithme consiste à mettre en correspondance l'espace dans lequel se trouve $\boldsymbol{\theta}$ avec celui où se trouve $\boldsymbol{\theta}'$ par une transformation réversible (cela peut requérir d'étendre l'espace de $\boldsymbol{\theta}$ et/ou l'espace de $\boldsymbol{\theta}'$ à des espaces de même structure, voir Andrieu *et al.* (2001b)).

3.1.2 Echantillonnage préférentiel et échantillonnage préférentiel séquentiel

Une autre technique importante de calcul d'intégrales par Monte Carlo est l'échantillonnage préférentiel (*importance sampling* en anglais). Soit $\pi(\boldsymbol{\theta})$ une distribution de probabilité[2], par exemple une distribution *a posteriori*. Soit $\mathsf{q}(\boldsymbol{\theta})$ une autre distribution (dite *distribution d'importance*) telle que $\mathsf{q}(\boldsymbol{\theta}) \neq 0$ quand $\pi(\boldsymbol{\theta}) \neq 0$. Par ailleurs, $\mathsf{q}(\boldsymbol{\theta})$ est choisie de telle sorte qu'il

[2]Ici encore, pour simplifier les notations, nous supposons qu'il s'agit d'une densité.

est possible d'échantillonner directement selon elle. Alors, (3.5) peut être écrite

$$I\left[\mathsf{h}\right] \;=\; \int_{\Theta} \mathsf{h}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,d\boldsymbol{\theta} \;=\; \int_{\Theta} \mathsf{h}(\boldsymbol{\theta})\frac{\pi(\boldsymbol{\theta})}{\mathsf{q}(\boldsymbol{\theta})}\mathsf{q}(\boldsymbol{\theta})\,d\boldsymbol{\theta} \tag{3.15}$$

Supposons maintenant qu'un ensemble d'échantillons de Monte Carlo $\widetilde{\boldsymbol{\theta}}^{(i)}$, $i = 1, \ldots, N$ sont générés selon $\mathsf{q}(\boldsymbol{\theta})$. L'approximation de Monte Carlo est

$$\widehat{I}_N\left[\mathsf{h}\right] \;=\; \sum_{i=1}^{N} \widetilde{\omega}^{(i)}\mathsf{h}\big(\widetilde{\boldsymbol{\theta}}^{(i)}\big) \;\approx\; I\left[\mathsf{h}\right]\;, \qquad \text{avec } \widetilde{\omega}^{(i)} = \frac{\pi(\widetilde{\boldsymbol{\theta}}^{(i)})}{\mathsf{q}(\widetilde{\boldsymbol{\theta}}^{(i)})} \tag{3.16}$$

où $\widetilde{\omega}^{(i)}$ est le *poids d'importance* de l'échantillonnage $\widetilde{\boldsymbol{\theta}}^{(i)}$ qui est destiné à corriger l'écart entre $\mathsf{q}(\boldsymbol{\theta})$ et $\pi(\boldsymbol{\theta})$. En d'autres termes, l'échantillonnage préférentiel consiste à remplacer la distribution $\pi(\boldsymbol{\theta})$ par la distribution empirique

$$\mathsf{P}_N(d\boldsymbol{\theta}) \;=\; \sum_{i=1}^{N} \widetilde{\omega}^{(i)}\delta_{\widetilde{\boldsymbol{\theta}}^{(i)}}(d\boldsymbol{\theta}) \tag{3.17}$$

Un remarque importante est que l'estimateur (3.16) dépend fortement de la distribution d'importance $\mathsf{q}(\boldsymbol{\theta})$ choisie. On peut démontrer facilement que, pour un N donné, la distribution d'importance qui minimise la variance de l'estimateur (3.16) est $\mathsf{q}(\boldsymbol{\theta}) \propto |\mathsf{h}(\boldsymbol{\theta})|\pi(\boldsymbol{\theta})$ (Robert et Casella (2000)). Toutefois, cette dernière ne peut généralement pas être utilisée car il faut pouvoir échantillonner directement selon elle. En outre, la constante de normalisation peut être difficile à calculer (il s'agit de $I\left[\mathsf{h}\right]$ elle-même quand h est une fonction positive). Ce résultat fournit néanmoins des indications pour choisir $\mathsf{q}(\boldsymbol{\theta})$: elle doit être aussi proche que possible de la distribution d'importance optimale.

Filtrage particulaire

Une application importance de l'échantillonnage préférentiel est le filtrage particulaire, voir le livre de Doucet *et al.* (2001) pour présentation générale. Nous considérons ici un système dynamique comprenant un *paramètre d'état caché* $\boldsymbol{\theta}_n$ et des *observations* $\mathbf{x}_n$ pour $n = 1, 2, \ldots$ Ce système est caractérisé par les équations suivantes (voir aussi la représentation graphique figure 3.2).

$$\boldsymbol{\theta}_n \;=\; f_n\big[\boldsymbol{\theta}_{n-1}\big] + \mathbf{v}_n \qquad \text{(équation de transition)} \tag{3.18}$$

$$\mathbf{x}_n \;=\; g_n\big[\boldsymbol{\theta}_n\big] + \boldsymbol{\epsilon}_n \qquad \text{(équation d'observation)} \tag{3.19}$$

En supposant que $\mathbf{v}_n$ et $\boldsymbol{\epsilon}_n$ sont indépendants et identiquement distribués (i.i.d.) au cours des itérations, et indépendants entre eux, de distributions connues, on peut exprimer (3.18)-(3.19) sous la forme d'une distribution *a priori* $\mathsf{p}(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1})$ et d'une vraisemblance $\mathsf{p}(\mathbf{x}_n|\boldsymbol{\theta}_n)$. En définissant en outre la distribution initiale $\mathsf{p}_0(\boldsymbol{\theta}_0)$, un modèle bayésien séquentiel est alors complètement caractérisé. On rencontre de tels modèles dans de nombreux problèmes : par exemple, $\boldsymbol{\theta}_n$ peut être un vecteur comprenant la position, la vitesse et l'accélération d'un avion au temps n, et $\mathbf{x}_n$ est l'observation délivrée par un radar. Dans ce problème, $\boldsymbol{\theta}_n$ n'est pas observé directement : il doit être estimé.

L'objectif du filtrage bayésien est de réaliser l'estimation séquentielle de la trajectoire du paramètre d'état entre les instants 0 et n, noté $\boldsymbol{\theta}_{0:n} = \{\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n\}$ dont la distribution *a*

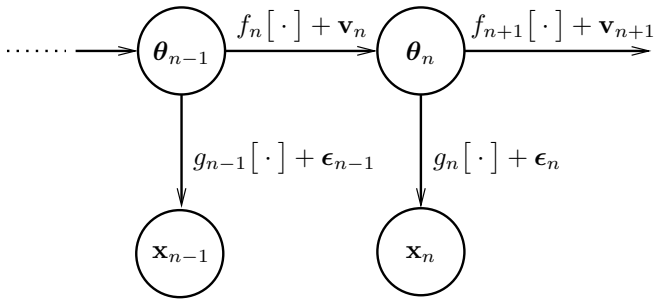


FIG. 3.2 – Illustration graphique du modèle dynamique (3.18)–(3.19). Dans ce modèle, seules les $\mathbf{x}_n$ sont observés. Le paramètre d'état $\boldsymbol{\theta}_n$ est caché pour $n = 1, 2, \ldots$. Le paramètre d'état $\boldsymbol{\theta}_n$ suit un modèle markovien, d'où le nom de *modèle de Markov caché* (*Hidden Markov Model* en anglais).

posteriori est $\mathsf{p}(\boldsymbol{\theta}_{0:n}|\mathbf{x}_{1:n})$ à chaque instant n. Dans le cas particulier où les fonctions $f_n\big[\,\cdot\,\big]$ et $g_n\big[\,\cdot\,\big]$ sont linéaires et où les bruits $\mathbf{v}_n$ et $\boldsymbol{\epsilon}_n$ sont gaussiens, une solution analytique existe : la distribution *a posteriori* est gaussienne, sa moyenne et sa matrice de covariance sont fournies par le filtre de Kalman, voir Anderson et Moore (1979). Dans le cas général, l'estimation de $\boldsymbol{\theta}_{0:n}$ repose sur des intégrales non calculables analytiquement (comme les estimées MMSE (3.3)) à chaque instant n. Le filtrage particulaire fournit une approximation de Monte Carlo par échantillonnage préférentiel séquentiel, où $\mathsf{p}(\boldsymbol{\theta}_{0:n-1}|\mathbf{x}_{1:n-1})$ joue le rôle de $\pi(\boldsymbol{\theta})$ dans (3.15). Pour écrire l'échantillonnage sous forme séquentielle, nous exprimons d'abord la distribution *a posteriori* $\mathsf{p}(\boldsymbol{\theta}_{0:n}|\mathbf{x}_{1:n})$ à l'instant n en fonction de $\mathsf{p}(\boldsymbol{\theta}_{0:n-1}|\mathbf{x}_{1:n-1})$, à l'instant $n-1$:

$$\mathsf{p}(\boldsymbol{\theta}_{0:n}|\mathbf{x}_{1:n}) \;=\; \mathsf{p}(\boldsymbol{\theta}_{0:n-1}|\mathbf{x}_{1:n-1}) \frac{\mathsf{p}(\boldsymbol{\theta}_n|\boldsymbol{\theta}_{n-1})\mathsf{p}(\mathbf{x}_n|\boldsymbol{\theta}_n)}{\mathsf{p}(\mathbf{x}_n|\mathbf{x}_{1:n-1})} \tag{3.20}$$

où $\mathsf{p}(\mathbf{x}_n|\mathbf{x}_{1:n-1})$ est un terme de normalisation qu'il n'est pas nécessaire de calculer. Ensuite, la distribution d'importance est choisie de forme séquentielle

$$\mathbf{q}_n(\boldsymbol{\theta}_{0:n}) \;=\; \mathbf{q}_{n-1}(\boldsymbol{\theta}_{0:n-1})\mathsf{q}_n(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n - 1) \;=\; \mathbf{q}_0(\boldsymbol{\theta}_0) \prod_{l=1}^{n} \mathsf{q}_l(\boldsymbol{\theta}_l|\boldsymbol{\theta}_{l-1}) \tag{3.21}$$

Il est maintenant possible d'échantillonner l'état à l'instant n en utilisant $\mathsf{q}_n(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n - 1)$ et de calculer séquentiellement le poids sous la forme $\mathsf{p}(\boldsymbol{\theta}_{0:n}|\mathbf{x}_{1:n})/\mathbf{q}_n(\boldsymbol{\theta}_{0:n})$. Ces éléments permettent d'écrire l'algorithme de filtrage particulaire présenté ci-dessous.

Algorithme 3.4: Algorithme de filtrage particulaire

- Initialisation : Pour chaque particule $i = 1, \ldots, N$, échantillonner indépendamment $\widetilde{\boldsymbol{\theta}}_0^{(i)} \sim \mathbf{q}_0(\boldsymbol{\theta}_0)$ et calculer les poids initiaux $\widetilde{\omega}_0^{(i)} = \mathsf{p}_0(\widetilde{\boldsymbol{\theta}}_0^{(i)})/\mathbf{q}_0(\widetilde{\boldsymbol{\theta}}_0^{(i)})$
- Itérations : pour $n = 1, 2, \ldots, N$ faire

  % Les trajectoires des particules sont mises à jour en utilisant la distribution d'importance séquentielle
  - Pour $i = 1, \ldots, N$, échantillonner le nouvel état au temps n pour la particule i,

$$\widetilde{\boldsymbol{\theta}}_n^{(i)} \;\sim\; \mathsf{q}_n(\boldsymbol{\theta}_n|\widetilde{\boldsymbol{\theta}}_{n-1}^{(i)}) \tag{3.22}$$

  % Les poids sont calculés et normalisés

– Pour $i = 1, \ldots, N$, calculer les poids d'importance séquentiels $\widetilde{\omega}_n^{(i)}$ comme suit

$$\widetilde{\omega}_n^{(i)} \propto \widetilde{\omega}_{n-1}^{(i)} \frac{\mathsf{p}(\widetilde{\boldsymbol{\theta}}_n^{(i)}|\widetilde{\boldsymbol{\theta}}_{n-1}^{(i)})\mathsf{p}(\mathbf{x}_n|\boldsymbol{\theta}_n)}{\mathsf{q}_n(\widetilde{\boldsymbol{\theta}}_n^{(i)}|\widetilde{\boldsymbol{\theta}}_{n-1}^{(i)})} \tag{3.23}$$

– Calculer la constante de normalisation des poids $W_n = \sum_{i=1}^N \widetilde{\omega}_n^{(i)}$ (cela normalise les poids, et évite d'avoir à calculer le terme $\mathsf{p}(\mathbf{x}_n|\mathbf{x}_{1:n-1})$ dans (3.15))
– Pour les particules $i = 1, \ldots, N$, faire $\widetilde{\omega}_n^{(i)} \leftarrow \widetilde{\omega}_n^{(i)}/W_n$
% L'état courant est estimé
– Estimer l'état à partir des particules, par Monte Carlo. Par exemple, $\widehat{\boldsymbol{\theta}_{n\,\mathrm{MMSE}}} = \sum_{i=1}^N \widetilde{\omega}_n^{(i)} \widetilde{\boldsymbol{\theta}}_n^{(i)}$
% Les particules sont rééchantillonnées
– Calculer l'indicateur d'efficacité $N_{\mathsf{eff}} = \left[\sum_{i=1}^N (\widetilde{\omega}_n^{(i)})^2\right]^{-1}$
– Si $N_{\mathsf{eff}} \leq N_{\mathsf{Threshold}}$, alors rééchantillonner les particules : dupliquer les particules de poids importants, et supprimer les particules de poids faible. Faire $\widetilde{\omega}_n^{(i)} \leftarrow 1/N$ pour $i = 1, \ldots, N$.

A chaque itération n, l'algorithme 3.4 produit des échantillons de Monte Carlo $\widetilde{\boldsymbol{\theta}}_{0:n}^{(i)}$, $i = 1, \ldots, N$ (appelés *particules* dans ce contexte) et des poids $\widetilde{\omega}_n^{(i)}$ qui approchent la distribution *a posteriori* $\mathsf{p}(\boldsymbol{\theta}_{0:n}|\mathbf{x}_{1:n})$. Cependant, cette stratégie simple produit des échantillons qui dégénèrent : après quelques itérations, la plupart des poids sont proches de zéro, tandis que quelques-uns seulement sont significativement non nuls. Ce phénomène peut être mesuré par N_{eff}. Si cette quantité devient trop faible, la dégénérescence est effective et les particules sont rééchantillonnées : les particules de poids faible sont supprimées, et remplacées par de nouvelles particules, copiées à partir des particules de poids important. Cela peut être mis en œuvre en choisissant aléatoirement le nombre de particules filles, selon une distribution de probabilité proportionnelle au poids des particules initiales. Un algorithme effectuant le rééchantillonnage est celui du *stratified resampling* (Kitagawa (1996)).

Différentes distributions d'importance conduisent à différents algorithmes. Un choix simple consiste à prendre $\mathsf{q}_n(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n - 1) = \mathsf{p}(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n - 1)$. L'algorithme correspondant est simple, mais la variance des estimateurs de Monte Carlo ainsi construits est généralement grande. La distribution d'importance qui minimise la variance des poids (et ainsi, la variance des estimateurs) prend en compte la nouvelle observation. Elle s'exprime $\mathsf{q}_n(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n - 1) = \mathsf{p}(\boldsymbol{\theta}_n|\mathbf{x}_n, \boldsymbol{\theta}_{n-1})$, mais il est souvent impossible d'en générer directement des échantillons. Cependant, elle peut être approchée localement par des gaussiennes : $\mathsf{q}_n(\boldsymbol{\theta}_n|\boldsymbol{\theta}_n - 1) = \mathcal{N}\big(\boldsymbol{\theta}_n; \boldsymbol{\theta}_{n|n}, \boldsymbol{\Sigma}_{n|n}\big)$ où $\boldsymbol{\theta}_{n|n}$ et $\boldsymbol{\Sigma}_{n|n}$ sont l'état et la matrice de covariance estimés par le filtre de Kalman étendu ou par le filtre de Kalman sans parfum, voir Julier et Uhlmann (2004).

3.2 Méthodes à noyaux

Les méthodes statistiques bayésiennes présentées dans la section précédente sont toutes construites sur des *modèles génératifs* des données, c'est-à-dire que la connaissance des paramètres du modèle permet de générer les données (au sens statistique du terme).

Par contraste, les méthodes à base de *modèles discriminatifs* visent à apprendre une fonction de décision (pour la classification ou la détection par exemple). D'un point de vue méthodologique, les approches discriminatives se situent dans un cadre formel différent des méthodes bayésiennes génératives. Ici, nous nous intéressons aux méthodes visant à produire un modèle discriminatif dans un espace fonctionnel donné *a priori*, sous forme non

paramétrique. Par soucis de brièveté, je ne présente ici que les méthodes à noyaux reproduisant défini-positifs (générant des espaces de Hilbert à noyau reproduisant) – des extensions à d'autres espaces à noyaux ont été étudiées récemment (Ong *et al.* (2004); Guigue *et al.* (2005)).

3.2.1 Contexte

Dans bien des situations, des données sont collectées et l'on cherche à caractériser statistiquement le système ou le processus qui les a générées. Plus précisément, les données observées peuvent être associées à une quantité qui peut être un nombre entier (classification), un réel (régression), un vecteur (régression multivariée) ou aucune quantité (détection de nouveauté) – dans cette section, par souci de simplicité, seules ces situations sont étudiées. En pratique, les données collectées peuvent être des vecteurs dans $\mathbb{R}^{d_{\mathbf{x}}}$, des images, des séries temporelles, des séquences d'ADN ou des chaînes de caractères. Dans la suite, l'espace des données est noté $\mathcal{X}$. L'objet de tous ces problèmes est d'*apprendre*, à partir des observations, le processus qui associe une variable entière, réelle ou vectorielle à chaque donnée. L'*Apprentissage Automatique* est la discipline scientifique qui étudie ces problématiques et propose des algorithmes d'apprentissage.

Ces problèmes d'apprentissage peuvent être posés de façon plus précise. Tout d'abord, nous supposons qu'un ensemble d'observations sont disponibles sous la forme de couples $(\mathsf{X},\mathsf{Y}) = \{(\mathbf{x}_1,\mathsf{y}_1),\ldots,(\mathbf{x}_m,\mathsf{y}_m)\}$ où $\mathbf{x}_i \in \mathcal{X}$ $(i=1,\ldots,m)$ sont les données et les $\mathsf{y}_i \in \mathcal{Y}$ $(i=1,\ldots,m)$ sont les *étiquettes* (ou *labels*) associés. Ces observations sont dites d'*apprentissage*. Des exemples de tels couples sont $\mathcal{X}=\mathbb{R}^{d_{\mathbf{x}}}$ et $\mathcal{Y}=\{-1,1\}$ (classification binaire), $\mathcal{Y}=\{1,2,\ldots,\}$ (classification multiclasses), $\mathcal{Y}=\{0,1\}$ (détection de nouveauté–dans ce cas, les labels Y ne sont pas fournis avec les données d'apprentissage X) et $\mathcal{Y}=\mathbb{R}$ (régression). Le but de l'apprentissage automatique est de prédire correctement le label $\mathsf{y}\in\mathcal{Y}$ d'une nouvelle donnée $\mathbf{x}\in\mathcal{X}$ qui n'appartient pas à l'ensemble d'apprentissage (X,Y). Cette prédiction doit être construite en utilisant uniquement l'information issue de $\mathbf{x}$ et de (X,Y) – en plus des hypothèses *a priori*.

3.2.2 Fonctions de coût et risque

Comme indiqué ci-dessus, l'apprentissage consiste à apprendre une fonction $\mathsf{F}_{(\mathsf{X},\mathsf{Y})} : \mathcal{X} \rightarrow \mathcal{Y}$ qui associe le label y à une donnée $\mathbf{x}$. La notation $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$ met en évidence le fait que cette fonction est apprise sur la base de l'ensemble d'apprentissage (X,Y). Il va de soit que cette fonction doit faire le moins d'erreur possible, à la fois sur l'ensemble d'apprentissage, mais aussi sur tout nouvelle donnée à traiter. La «gravité» d'une erreur est déterminée par la *fonction de coût*.

Fonctions de coût

La fonction $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$ n'a d'intérêt pratique que si le risque qu'elle commette des erreurs est faible. Soit F une fonction quelconque qui associe le label y à une donnée $\mathbf{x}$. Une erreur est commise lorsque $\mathsf{F}(\mathbf{x})=\mathsf{y}'$ alors que le «vrai» label de $\mathbf{x}$ est y, avec $\mathsf{y}'\neq\mathsf{y}$. Dans certains cas pratiques comme la régression, il est simplement impossible de ne pas faire d'erreur, mais certaines erreurs sont plus importantes que d'autres. Aussi mesure-t-on la gravité des erreurs pour concevoir $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$; c'est le rôle des fonction de coût. Soit $\mathsf{c}(\mathbf{x},\mathsf{y};\mathsf{F}(\mathbf{x}))$ une fonction de coût mesurant la gravité de l'erreur commise par F en $\mathbf{x}$, dont le vrai label est y. Un exemple

standard de fonction de coût est le coût 0-1 $\mathsf{c}\,(\mathbf{x}, \mathsf{y}; \mathsf{F}(\mathbf{x})) = \delta_{\mathsf{y}}\big(\mathsf{F}(\mathbf{x})\big)$ qui associe un coût nul aux prédictions $\mathsf{F}(\mathbf{x})$ correctes, et un coût unité aux erreurs, quelle que soit leur gravité. En classification bi-classe par exemple, on fait souvent l'hypothèse que F est à valeurs dans $\mathbb{R}$ (et pas simplement dans $\mathcal{Y} = \{-1, 1\}$), et l'on choisit $\mathsf{y} = \text{signe}\,(\mathsf{F}(\mathbf{x}))$. Cela rend possible l'utilisation du coût quadratique, du coût «charnière», etc. (voir figure 3.3). Nous notons que les fonctions de coût sont positives et, de préférence, convexes.

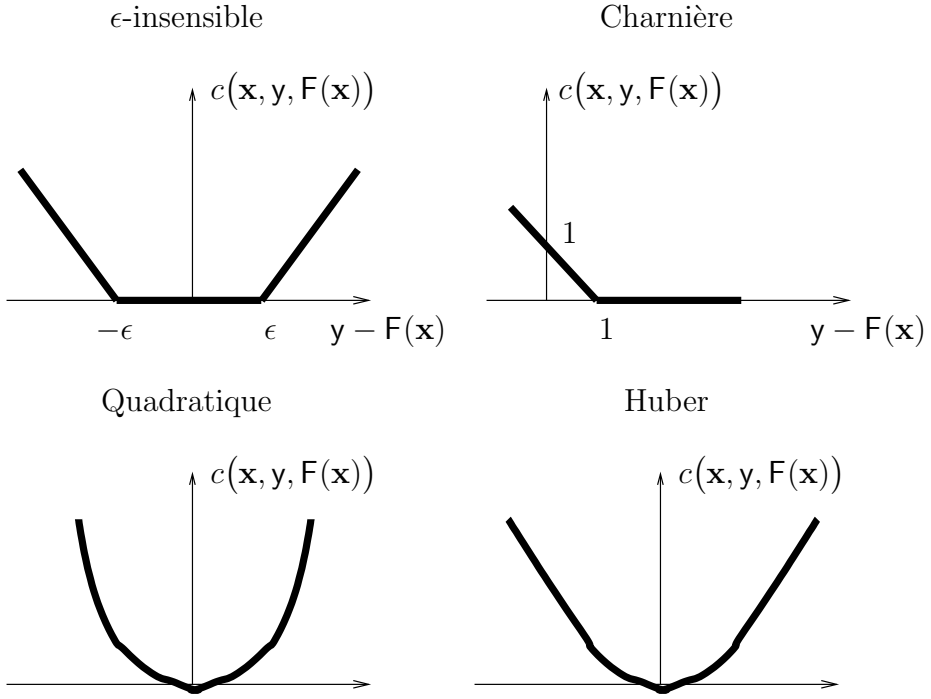


FIG. 3.3 – Fonctions de coût utilisées fréquemment pour l'apprentissage automatique.

Les fonctions de coût permettent de définir précisément la notion de *précision* d'une fonction apprise en un point $\mathbf{x}$ de label y. Toutefois, elles ne permettent pas de connaître la précision *globale* pour chaque couple (donnée,label) que l'on peut rencontrer.

Risques

Une façon simple de définir la précision globale d'une fonction F pour un processus de génération de données (X, Y) est le *risque* $\mathsf{R}\,[\mathsf{F}]$ (qui caractérise F) comme l'espérance mathématique du coût, c'est-à-dire

$$\mathsf{R}\,[\mathsf{F}] \;=\; \mathbb{E}_{\mathsf{p}(\mathbf{x},\mathsf{y})}\big[\,\mathsf{c}\,(\mathbf{x}, \mathsf{y}; \mathsf{F}(\mathbf{x}))\,\big] \;=\; \int_{\mathcal{X}\times\mathcal{Y}} \mathsf{c}\,(\mathbf{x}, \mathsf{y}; \mathsf{F}(\mathbf{x}))\, \mathsf{P}(d\mathbf{x}, d\mathsf{y}) \tag{3.24}$$

où $\mathsf{P}(d\mathbf{x}, d\mathsf{y})$ est la loi de probabilité conjointe des données et de leurs labels. Il est clair que cette distribution est inconnue en pratique, aussi le risque ne peut-il pas être calculé. Supposons cependant que $\mathsf{R}\,[\mathsf{F}]$ puisse effectivement être calculé pour toute fonction F, il serait alors possible de trouver la fonction F recherchée : c'est celle qui minimise le risque. Il convient de remarquer que cette approche est évidemment inapplicable, mais elle fournit la

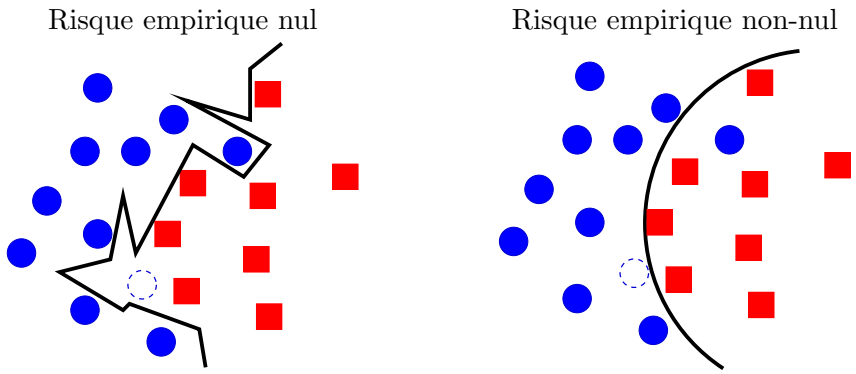


FIG. 3.4 – Dans un problème de classification à deux classes (ici, la classe des ronds et la classe des carrés), minimiser le risque empirique conduit à des solutions triviales. A gauche : la fonction de classification (ligne noire) a un risque empirique nul (aucune erreur de classification des ronds et carrés d'apprentissage) mais elle classe mal la nouvelle donnée (rond en pointillés). A droite : la fonction de classification en forme d'arc de cercle n'a pas un risque empirique nul, mais sa simplicité assure de meilleurs performances de généralisation : elle classe correctement la nouvelle donnée.

base d'une méthode de recherche de la fonction F. Dans ce but, une première remarque est que le risque peut être estimé sur la base de l'ensemble d'apprentissage, en remplaçant dans l'équation (3.24) $\mathsf{P}(d\mathbf{x}, d\mathsf{y})$ par la distribution empirique

$$\mathsf{P}_m^{\text{emp}}(d\mathbf{x}, d\mathsf{y}) \;=\; \frac{1}{m}\sum_{i=1}^{m} \delta_{\mathbf{x}_i,\mathsf{y}_i}(d\mathbf{x}, d\mathsf{y}) \tag{3.25}$$

Ainsi, le risque empirique s'écrit

$$\mathsf{R}^{\text{emp}}_{(\mathsf{X},\mathsf{Y})}\left[\mathsf{F}\right] \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathsf{c}\left(\mathbf{x}_i, \mathsf{y}_i; \mathsf{F}(\mathbf{x}_i)\right) \tag{3.26}$$

et il s'agit d'un estimateur du vrai $\mathsf{R}\left[\mathsf{F}\right]$, basé sur l'information fournie par l'ensemble d'apprentissage.

Un des problèmes clé est que $\mathsf{R}^{\text{emp}}_{(\mathsf{X},\mathsf{Y})}\left[\mathsf{F}\right]$ est un mauvais estimateur de $\mathsf{R}\left[\mathsf{F}\right]$ dans le cas général, et il existe presque toujours des fonctions F telles que $\mathsf{R}^{\text{emp}}_{(\mathsf{X},\mathsf{Y})}\left[\mathsf{F}\right] = 0$, dont une majorité de fonctions triviales. Les fonctions assurant un risque empirique nul ont souvent une mauvaise capacité de prédiction, le label associé à une donnée ne faisant pas partie de l'ensemble d'apprentissage. En outre, si le risque empirique tend effectivement vers le vrai risque lorsque m tend vers l'infini, cette convergence n'est généralement pas monotone, et est très lente. La convergence du risque empirique vers le vrai risque (quelle que soit F) dépend de la classe de fonctions dans laquelle on recherche F. Dans la plupart des cas, si le risque empirique tend effectivement vers le vrai risque (consistance), il faut avoir un ensemble d'apprentissage très vaste pour que l'écart entre risque empirique et vrai risque soit faible *in fine*.

Régularisation

Nous l'avons vu, minimiser le risque empirique conduit à de nombreuses solutions sans intérêt $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$. Cependant, ce problème peut être résolu (au moins partiellement) en minimisant un risque qui inclut le risque empirique $\mathsf{R}^{\text{emp}}_{(\mathsf{X},\mathsf{Y})}[\mathsf{F}]$ et un nouveau terme $\widetilde{\Omega}\left(\mathsf{F}\right)$ qui pénalise les solutions indésirables. Nous définissons le *risque régularisé* par

$$\mathsf{R}^{\text{reg}}[\mathsf{F}] \;=\; \mathsf{R}^{\text{emp}}_{(\mathsf{X},\mathsf{Y})}[\mathsf{F}] + \lambda\widetilde{\Omega}\left(\mathsf{F}\right) \tag{3.27}$$

où λ est le paramètre de régularisation. Au sens de ce risque, la fonction de classification optimale $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$ vérifie

$$\mathsf{F}_{(\mathsf{X},\mathsf{Y})} \;=\; \underset{\mathsf{F}\in\mathcal{F}}{\operatorname{argmin}}\; \mathsf{R}^{\text{reg}}[\mathsf{F}] \tag{3.28}$$

où la fonction $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$ est recherchée dans l'espace de fonctions $\mathcal{F}$. Le terme de pénalisation $\widetilde{\Omega}\left(\mathsf{F}\right)$ a un rôle central, et il doit être choisi de telle sorte que $\mathsf{R}^{\text{reg}}[\mathsf{F}]$ converge rapidement vers le vrai risque lorsque m tend vers l'infini. Selon Vapnik (1995), $\widetilde{\Omega}\left(\mathsf{F}\right)$ doit pénaliser la complexité de F, car une fonction simple de risque empirique faible a de meilleurs performances de généralisation qu'une fonction plus complexe qui colle parfaitement aux données d'apprentissage (Hastie *et al.*, 2001, Chap. 7). Le compromis entre *adéquation aux données d'apprentissage* et *complexité* (c'est-à-dire le compromis biais/variance) est réglé par le paramètre λ.

Le problème régularisé de l'équation (3.27) conduit à de bonnes solutions dans la mesure où l'espace de recherche $\mathcal{F}$ contient des fonctions de faible risque. La section suivante présente les espaces de Hilbert à noyaux reproduisants (*Reproducing Kernel Hilbert Space* –RKHS), utilisés dans beaucoup de méthodes récentes, voir Berlinet et Thomas-Agnan (2004).

3.2.3 Espaces de Hilbert à Noyau reproduisant

Dans le but d'introduire les espaces de Hilbert à noyaux reproduisants, nous rappelons plusieurs définitions, notamment celle d'un produit scalaire et celle d'un espace complet. Pour commencer, soit un espace vectoriel $\mathcal{H}$ de fonctions de $\mathcal{X}$ dans $\mathbb{R}$. Par ailleurs, nous supposons que cet espace est muni d'un produit scalaire.

Définition 3.1: Un produit scalaire sur $\mathcal{H}$ est une opération notée $\langle \mathsf{f}, \mathsf{f}'\rangle_{\mathcal{H}}$ qui associe un réel à chaque paire $(\mathsf{f}, \mathsf{f}')$ d'éléments de $\mathcal{H}$, et qui vérifie :

- Pour tout $(\mathsf{f}, \mathsf{f}') \in \mathcal{H}\times\mathcal{H}$, $\langle \mathsf{f}, \mathsf{f}'\rangle_{\mathcal{H}} = \langle \mathsf{f}', \mathsf{f}\rangle_{\mathcal{H}}$ (symétrique) ;
- Pour tout $a \in \mathbb{R}$, et pour tous $(\mathsf{f}, \mathsf{f}', \mathsf{f}'') \in \mathcal{H}^3$, $\langle \mathsf{f} + a\mathsf{f}'', \mathsf{f}'\rangle_{\mathcal{H}} = \langle \mathsf{f}, \mathsf{f}'\rangle_{\mathcal{H}} + a\langle \mathsf{f}'', \mathsf{f}'\rangle_{\mathcal{H}}$ (linéaire) ;
- Pour tout $\mathsf{f} \in \mathcal{H}$, $\langle \mathsf{f}, \mathsf{f}\rangle_{\mathcal{H}} \geq 0$ (positif) ;
- $\langle \mathsf{f}, \mathsf{f}\rangle_{\mathcal{H}} = 0$ implique $\mathsf{f} = 0$ (défini). □

A partir d'un produit scalaire, il est possible de définir une norme pour "mesurer" f, notée $\|\mathsf{f}\|_{\mathcal{H}}$ et définie par $\|\mathsf{f}\|^2_{\mathcal{H}} = \langle \mathsf{f}, \mathsf{f}\rangle_{\mathcal{H}}$. Nous rappelons maintenant la définition d'un espace complet.

Définition 3.2: Un espace complet $\mathcal{H}$ est tel que toute suite f_n $(n = 1, 2, \ldots)$ d'éléments de $\mathcal{H}$ qui vérifie $\|\mathsf{f}_{n+p} - \mathsf{f}_n\|_{\mathcal{H}} \underset{n\to\infty}{\longrightarrow} 0$ pour $p > 0$ converge vers une limite $\overline{\mathsf{f}}$ appartenant à $\mathcal{H}$. Ici,

$\mathcal{H}$ est *complet pour la norme* $\|\cdot\|_{\mathcal{H}}$ car la condition $\|\mathsf{f}_{n+p} - \mathsf{f}_n\|_{\mathcal{H}} \underset{n\to\infty}{\longrightarrow} 0$ est exprimée pour cette norme. □

Ces éléments permettent de rappeler la définition d'un espace de Hilbert.

Définition 3.3: Un espace de Hilbert $\mathcal{H}$ de fonctions est un espace vectoriel de fonctions muni d'un produit scalaire $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, et qui est complet pour la norme $\|\cdot\|_{\mathcal{H}}$. □

Les espaces de Hilbert à noyau reproduisant sont des espaces de Hilbert avec une propriété additionnelle.

Définition 3.4: Un espace de Hilbert à noyau reproduisant est un espace de Hilbert de $\mathcal{X}$ dans $\mathbb{R}$ tel qu'il existe une *fonction noyau* $\mathsf{k}(\cdot,\cdot)$ de $\mathcal{X}\times\mathcal{X}$ vers les réels telle que
- $\mathsf{k}(\cdot,\cdot)$ est symétrique ;
- $\mathsf{k}(\mathbf{x},\cdot)$ est une fonction de $\mathcal{H}$ pour tout $\mathbf{x}\in\mathcal{X}$ fixé ;
- Pour tout $\mathbf{x}\in\mathcal{X}$ et pour tout $\mathsf{f}\in\mathcal{H}$, $\langle \mathsf{k}(\mathbf{x},\cdot), \mathsf{f}\rangle_{\mathcal{H}} = \mathsf{f}(\mathbf{x})$ (propriété reproduisante). □

Il va de soi que n'importe quelle fonction $\mathsf{k}(\cdot,\cdot)$ de $\mathcal{X}\times\mathcal{X}$ dans $\mathbb{R}$ n'est pas toujours un noyau reproduisant. Selon Aronszajn (1950), un noyau $\mathsf{k}(\cdot,\cdot)$ donne naissance à un RKHS si et seulement si il est *défini positif*.

Définition 3.5: Un noyau est dit *défini positif* (Aronszajn (1950); Mercer (1909)) si, pour tout $m\geq 1$, toute famille $\{\mathbf{x}_1,\ldots,\mathbf{x}_m\}\in\mathcal{X}^m$ et tout ensemble de coefficients $\{a_1,\ldots,a_m\}\in\mathbb{R}^m$,

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i a_j \mathsf{k}(\mathbf{x}_i,\mathbf{x}_j) \;\geq\; 0 \tag{3.29}$$

□

Il existe une bijection entre l'ensemble des noyaux définis positifs et les RKHS. En d'autres termes, il suffit de choisir un noyau défini positif pour implicitement choisir un espace de fonctions. Un noyau très souvent choisi est le noyau gaussien (pour tous $\mathbf{x}$ et $\mathbf{x}'$ dans $\mathcal{X}$, muni de la norme $\|\cdot\|_{\mathcal{X}}$)

$$\mathsf{k}(\mathbf{x},\mathbf{x}') = \exp -\frac{1}{2\sigma^2}\|\mathbf{x}-\mathbf{x}'\|_{\mathcal{X}}^2 \tag{3.30}$$

Une définition équivalente des RKHS est rappelée par Wahba (1990). Avant de la rappeler, nous introduisons la fonctionnelle d'évaluation.

Définition 3.6: On appelle *fonctionnelle d'évaluation* l'application suivante

$$\begin{array}{rccc} \mathsf{L}_{\mathbf{x}} \;:\; & \mathcal{H} & \longmapsto & \mathbb{R} \\ & \mathsf{F}(\cdot) & \longrightarrow & \mathsf{L}_{\mathbf{x}}\mathsf{F} = \mathsf{F}(\mathbf{x}) \end{array} \tag{3.31}$$

□

Alors, un RKHS $\mathcal{H}$ est un espace de Hilbert de fonctions de $\mathcal{X}$ dans $\mathbb{R}$ tel que la fonctionnelle d'évaluation est linéaire et bornée au sens où il existe une constante M telle que

pour tout $\mathbf{x} \in \mathcal{X}$, $|\mathsf{L}_{\mathbf{x}}\mathsf{F}| \leq M\|\mathsf{F}\|_{\mathcal{H}}$. Alors, par le théorème de Riesz appliqué à $\mathsf{L}_{\mathbf{x}}$, il existe un unique noyau $\mathsf{k}(\cdot,\cdot)$ tel que pour tous $\mathbf{x}$ et F,

$$\mathsf{L}_{\mathbf{x}}\mathsf{F} \;=\; \langle \mathsf{k}(\mathbf{x},\cdot), \mathsf{F}(\cdot)\rangle_{\mathcal{H}} \;=\; \mathsf{F}(\mathbf{x}) \tag{3.32}$$

et on peut montrer que le noyau apparaissant ici est défini positif.

Les problèmes de minimisation du risque régularisé (3.28) dans un RKHS où le terme de pénalisation est une fonction monotone croissante de la norme $\|\cdot\|_{\mathcal{H}}$ (i.e. $\widetilde{\Omega}\left(\mathsf{F}\right) = \Omega\left(\|\mathsf{f}\|_{\mathcal{H}}\right)$ où $\Omega\left(\cdot\right)$ est monotone croissante) admettent une solution simple, donnée par le théorème du représentant :

Théorème 3.1: Soit une fonction $\Omega\left(\cdot\right) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ monotone croissante. Alors, la fonction $\mathsf{f}_{(\mathsf{X},\mathsf{Y})}$ qui minimise $\mathsf{R}^{\text{reg}}\left[\mathsf{F}\right] \;=\; \mathsf{R}^{\text{emp}}_{(\mathsf{X},\mathsf{Y})}\left[\mathsf{F}\right] + \lambda\Omega\left(\|\mathsf{f}\|_{\mathcal{H}}\right)$ dans un RKHS $\mathcal{H}$ de noyau $\mathsf{k}(\cdot,\cdot)$, s'écrit[3]

$$\mathsf{f}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x}) \;=\; \sum_{i=1}^{m} \alpha_i \mathsf{k}(\mathbf{x}_i, \mathbf{x}) \tag{3.33}$$

c'est-à-dire que la solution appartient au sous-espace vectoriel engendré par la famille $\{\mathsf{k}(\mathbf{x}_1,\cdot), \ldots, \mathsf{k}(\mathbf{x}_m,\cdot)\}$ □

En résumé, les RKHS jouent un rôle central dans beaucoup de méthodes d'apprentissage statistique. D'abord, ils forment des espaces de fonctions dont la complexité peut être mesurée aisément par la norme $\|\cdot\|_{\mathcal{H}}$. Ensuite, le théorème du représentant assure que la solution du problème régularisé est simple, et que sa complexité est liée au *nombre de données*, et non à leur *dimension*.

Nous présentons maintenant les algorithmes de classification à vecteur support, qui sont un exemple important de ce cadre général.

3.2.4 Classifieurs à vecteurs support

Les classifieurs à vecteurs support (*Support Vector Machines* – SVM Schölkopf et Smola (2002)) sont un cas particulier de minimisation du risque régularisé dans un RKHS. Tout d'abord, nous supposons que le risque régularisé (3.27) est minimisé dans un espace $\mathcal{F}$ tel que

$$\mathcal{F} = \{\mathsf{F}(\cdot) = \mathsf{f}(\cdot) + b \text{ tel que } \mathsf{f} \in \mathcal{H} \text{ et } b \in \mathbb{R}\} \tag{3.34}$$

où $\mathcal{H}$ est un RKHS de noyau $\mathsf{k}(\cdot,\cdot)$ sur $\mathcal{X} \times \mathcal{X}$. Afin de définir l'algorithme d'apprentissage, il nous faut en outre définir l'espace des labels $\mathcal{Y} = \{-1, 1\}$. La fonction de coût intervenant dans le calcul du risque est le coût dit "charnière" $\mathsf{c}_{\text{charnière}}\left(\mathbf{x}, \mathsf{y}; \mathsf{F}(\mathbf{x})\right)$ (voir figure 3.3), et le terme de pénalisation $\widetilde{\Omega}\left(\mathsf{F}\right)$ est la norme carrée induite par le produit scalaire dans $\mathcal{H}$, c'est-à-dire $\widetilde{\Omega}\left(\mathsf{F}\right) = \|\mathsf{f}\|^2_{\mathcal{H}}$. Cela conduit au risque régularisé suivant, à minimiser dans $\mathcal{F}$

$$\mathsf{R}^{\text{reg}}\left[\mathsf{F}\right] \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathsf{c}_{\text{charnière}}\left(\mathbf{x}_i, \mathsf{y}_i; \langle \mathsf{k}(\mathbf{x}_i,\cdot), \mathsf{f}\rangle_{\mathcal{H}} + b\right) + \lambda\|\mathsf{f}\|^2_{\mathcal{H}} \tag{3.35}$$

[3]La version du thèorème présentée ici est en fait simplifiée, et n'inclut pas un terme additionnel qui est nul dans les cas que nous considérons ici.

Il est possible de modifier ce risque régularisé de façon à remplacer le paramètre λ par un nouveau paramètre $\nu \in [0,1]$ qui peut être interprété plus facilement (Schölkopf et Smola (2002)). Minimiser ce risque régularisé modifié revient à résoudre

$$\begin{array}{rl} \text{Minimiser} & \frac{1}{2}\|\mathsf{f}\|^2_{\mathcal{H}} - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i \qquad \text{par rapport à } \mathsf{f}, \rho, \xi_i, b \\ \text{avec} & \mathsf{y}_i \langle \mathsf{k}\big(\mathbf{x}_i, \cdot\big), \mathsf{f}\rangle_{\mathcal{H}} + b \;\geq \rho - \xi_i \text{ pour tout } i = 1, \ldots, m \\ \text{et} & \xi_i \geq 0 \text{ pour tout } i = 1, \ldots, m\,,\ \rho \geq 0 \end{array} \tag{3.36}$$

où les variables *de relâchement* ξ_i ($i = 1, \ldots, m$) correspondent à l'action linéaire de la fonction de pénalisation. Après introduction des multiplicateurs de Lagrange $\alpha_1, \ldots, \alpha_m$, l'optimisation (3.36) devient un problème d'optimisation quadratique sous contraintes linéaires :

$$\begin{array}{l} \text{Maximiser} \quad -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j\mathsf{y}_i\mathsf{y}_j\mathsf{k}\big(\mathbf{x}_i, \mathbf{x}_j\big) \text{ par rapport à } \alpha_i \ (i = 1, \ldots, m) \\ \text{avec } \ 0 \leq \alpha_i \leq 1/m \ (i = 1, \ldots, m)\ , \quad \sum_{i=1}^{m}\alpha_i\mathsf{y}_i = 0 \text{ et } \sum_{i=1}^{m}\alpha_i \geq \nu \end{array} \tag{3.37}$$

Les multiplicateurs de Lagrange α_i, $i = 1, \ldots, m$ qui sont solution de (3.37) permettent d'écrire la fonction de classification $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$ optimale, qui s'écrit pour tout $\mathbf{x} \in \mathcal{X}$

$$\mathsf{F}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x}) \;=\; \text{signe}\ \left[\sum_{i=1}^{m}\mathsf{y}_i\alpha_i\mathsf{k}\big(\mathbf{x}_i, \mathbf{x}\big) + b\right] \tag{3.38}$$

Une propriété cruciale de ce classifieur SVM est sa parcimonie : ν est une borne supérieure sur la proportion de multiplicateurs de Lagrange non-nuls (Schölkopf et Smola (2002)). En d'autres termes, pour $\nu = 0.2$ par exemple, au moins 80% des multiplicateurs de Lagrange $\{\alpha_i, i = 1, \ldots, m\}$ sont nuls. Ainsi, le calcul numérique de $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x})$ ne nécessite l'évaluation de $\mathsf{k}\big(\mathbf{x}_i, \mathbf{x}\big)$ que pour une petite fraction (au plus 20%) des données d'apprentissage $\mathbf{x}_i$, $i = 1, \ldots, m$. Les $\mathbf{x}_i$ avec un multiplicateur de Lagrange non nul sont appelés les *vecteurs support*, notés $\mathbf{x}_{i^\star}$. Dans (3.38), b est calculé comme la moyenne sur tous les vecteurs support $\mathbf{x}_{i^\star}$ de $\{-\sum_{j=1}^{m}\alpha_j\mathsf{y}_j\mathsf{k}\big(\mathbf{x}_j, \mathbf{x}_{i^\star}\big)\}$.

Pour tout $\mathbf{x}$, la propriété reproduisante assure que $\mathsf{F}(\mathbf{x}) \;=\; \text{signe}\ \big[\langle \mathsf{k}\big(\mathbf{x}, \cdot\big), \mathsf{f}\rangle_{\mathcal{H}} + b\big]$, ce qui signifie que classer une donnée $\mathbf{x}$ est une opération affine (à la fonction signe près) en termes d'éléments de $\mathcal{H}$, et une opération non-linéaire en termes d'éléments de $\mathcal{X}$. La figure 3.5 propose une interprétation géométrique des SVMs.

3.2.5 Commentaire bibliographique

Les méthodes à noyau sont connues depuis au moins les années 1960-70 à partir des travaux de Parzen (voir par exemple Parzen (1963), Wahba (voir son livre plus récent, Wahba (1990)) et Kailath (voir la série d'articles publiès après Kailath (1971)) pour ne citer qu'eux. De nombreux algorithmes entrent dans ce formalisme, dont les splines, les estimateurs de densités de Parzen et certaines méthodes d'ondelettes. La « révolution » des algorithmes SVM, au milieu des années 1990, tient à la conjonction de plusieurs résultats : tout d'abord, le fait d'utiliser des RKHS comme espaces de recherche de solutions. Ensuite, des résultats probabilistes basés sur des inégalités de concentration, permettant de définir des bornes supérieures

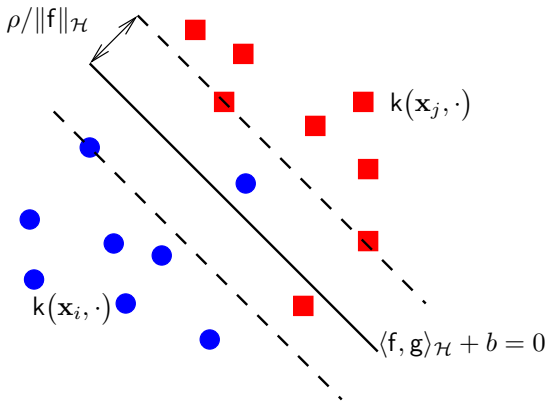


FIG. 3.5 – Interprétation géométrique des ν-SVM à marges douces dans $\mathcal{H}$. Chaque élément de $\mathcal{H}$ est une fonction, représentée ici par un point ou un carré dans un espace de dimension infini (ici, $\mathcal{H}$ est représenté dans le plan). Les ronds représentent les fonctions $\mathsf{k}(\mathbf{x}_i,\cdot)$ de label $\mathsf{y}_i = 1$, et les carrés représentent les fonctions $\mathsf{k}(\mathbf{x}_j,\cdot)$ de label $\mathsf{y}_j = -1$. La fonction de classification f est un vecteur orthogonal à l'hyperplan d'équation $\langle \mathsf{f},\mathsf{g}\rangle_{\mathcal{H}} + b = 0$. Cet hyperplan sépare les données de chacune des classes avec une marge aussi large que possible. La largeur de la marge est $\rho/\|\mathsf{f}\|_{\mathcal{H}}$. Comme il s'agit d'un SVM à marge douce, quelques vecteurs d'apprentissage se trouvent dans la marge.

sur le risque quelle que soit la distribution $\mathsf{P}(d\mathbf{x}, d\mathsf{y})$ des données (résultats dus initialement à Vapnik – voir par exemple Vapnik (1995)). Ces résultats permettent par exemple de relier la largeur de la marge des SVM à une borne supérieure sur le risque de ce classifieur. Enfin, des méthodes d'optimisation performantes, adaptées aux problèmes convexes (quadratiques, sous contraintes linéaires) permettent la résolution efficace, par exemple Bordes *et al.* (2005).

Chapitre 4

Contributions méthodologiques

Dans ce chapitre, je présente mes contributions méthodologiques de façon unifiée, en renvoyant aux publications pour les détails. Elles concernent essentiellement les deux domaines introduits au chapitre précédent : les méthodes de simulation statistique (méthodes de Monte Carlo) sont abordées en section 4.1 et les méthodes d'apprentissage statistique dites « à noyaux » sont évoquées en section 4.2.

4.1 Méthodes bayésiennes et Monte Carlo

Cette section présente mes contributions dans le domaine des modèles bayésiens, et les algorithmes proposés pour leur mise en œuvre. Plus précisément, le paragraphe 4.1.1 présente plusieurs modèles proposés dans le cadre d'applications génériques. Le paragraphe 4.1.2 détaille les contributions dans le domaine des algorithmes dédiés à ces modèles.

4.1.1 Modèles bayésiens à sauts

Les modèles dits «à sauts» correspondent aux cas où le vecteur paramètre comporte une ou plusieurs composantes discrètes «structurelles» qui déterminent par exemple la dimension du vecteur paramètre. D'une manière plus formelle, on considère une famille de modèles $\{\mathcal{M}_r\}$ indexés par la variable discrète r, chacun étant paramétré par $\boldsymbol{\theta}_r$. La variable indicatrice appartient à un espace $\mathcal{R}$ qui est typiquement

- l'ensemble des entiers naturels $\mathcal{R} = \mathbb{N}$. Ce cas se rencontre lorsque r est une variable de comptage (par exemple, le nombre de sinusoïdes formant un signal Andrieu et Doucet (1999), ou le nombre de coefficients AR) ;
- un ensemble de deux entiers naturels $\mathcal{R} = \mathbb{N} \times \mathbb{N}$. Un exemple est celui des modèles AR à coefficients variables, où r comprend le nombre de pôles réels et le nombre de paires de pôles complexes conjugués (Andrieu *et al.* (2003)) ;
- un ensemble de r_1 entiers positifs $(r_2, \ldots, r_{r_1+1})$. Dans ce cas, $\mathcal{R} = \mathbb{N} \times \cdots \times \mathbb{N}$. Par exemple, dans un signal musical harmonique, composé de r_1 notes, chacune ayant r_i harmoniques $(i = 2, \ldots, r_1 + 1)$ (Davy et Godsill (2002a); Davy *et al.* (2006c)).

On parle ici de modèles à sauts parce que l'inférence statistique nécessite de sauter d'un modèle $\mathcal{M}_r$ à un autre modèle $\mathcal{M}_{r'}$. J'ai proposé plusieurs modèles à sauts de ce type, qui sont résumés ci-dessous. Pour les descriptions détaillées, on se reportera aux publications, dont certaines sont annexées à ce manuscrit.

Modèles séquentiels

Un modèle markovien à saut s'écrit sous forme séquentielle comme suit. Tout d'abord, on définit une chaîne de Markov discrète $\{r_n\}_{n=0,1,\ldots}$ sur un espace fini $\mathcal{R}$ dont les probabilités de transition sont

$$\pi_{i,j} = \mathsf{P}\big(r_n = j | r_{n-1} = i\big) \qquad (i,j \in \mathcal{R}) \tag{4.1}$$

et par la distribution initiale notée $\pi_i = \mathsf{P}(r_0 = i)$, $i \in \mathcal{R}$. Etant donnée cette chaîne, on définit une famille de densités de transition du paramètre d'état $\boldsymbol{\theta}_n$

$$\mathsf{p}(\boldsymbol{\theta}_n | \boldsymbol{\theta}_{0:n-1}, r_{0:n}) \;=\; f_{r_{n-1},r_n}(\boldsymbol{\theta}_n | \boldsymbol{\theta}_{n-1}) \tag{4.2}$$

où $\boldsymbol{\theta}_n$ appartient à un espace dont la structure peut être fixée par r_n. Dans les problèmes étudiés, ni r_n ni $\boldsymbol{\theta}_n$ $(t = 1, 2, \ldots)$ ne sont observés directement. On observe par contre la séquence $\mathbf{x}_{1:n}$, qui est liée aux paramètres non observés par une famille de densités d'observation

$$\mathsf{p}(\mathbf{x}_n | \boldsymbol{\theta}_{0:n}, r_{0:n}, \mathbf{x}_{1:n-1}) \;=\; g_{r_n}(\mathbf{x}_n | \boldsymbol{\theta}_n, \mathbf{x}_{1:n-1}) \tag{4.3}$$

Les équations (4.1) à (4.3) définissent un modèle de Markov à sauts sous une forme séquentielle. Ce modèle a été appliqué à trois problèmes avec, pour chacun d'eux, un algorithme dédié.

– Dans le cas des **modèles autorégressifs à coefficients variables** (*Time Varying AutoRegressive* – TVAR) la chaîne r_n $(n = 0, 1, 2, \ldots)$ est à deux dimensions. En effet, un modèle autorégressif mono-dimensionnel à coefficients variables s'écrit

$$x(n) \;=\; a_{n,1}x(n-1) + a_{n,2}x(n-2) + \ldots, + a_{n,p_n}x(n-p_n) + \epsilon(n) \tag{4.4}$$

dont le polynôme caractéristique

$$Q_n(u) \;=\; 1 - a_{n,1}\,u - a_{n,2}\,u^2 - a_{n,p_n}x(n-p_n)u^{p_n} \tag{4.5}$$

admet k_n^r racines réelles et k_n^c racines complexes conjuguées, avec $k_n^r + 2k_n^c = p_n$. Nous avons défini dans ce cas $r_n = \{k_n^r, k_n^c\}$. Le vecteur paramètre $\boldsymbol{\theta}_n$ est composé des k_n^r racines réelles, des k_n^c modules et k_n^c arguments des racines complexes conjuguées. La dimension de $\boldsymbol{\theta}_n$ est donc liée à r_n. L'équation de transition (4.2) est donc définie sur les racines réelles, les modules et arguments (lesquels s'interprètent en termes de fréquences) des racines complexes conjuguées. Dans (Andrieu *et al.* (2002a,b, 2003)), ces équations de transition sont linéaires avec des bruits gaussiens, avec des variances elles-mêmes évolutives. L'équation d'observation est, elle, fortement non linéaire en $\boldsymbol{\theta}_n$, et les observations sont les échantillons du signal.

– Dans le cas de la **poursuite de cibles** (Doucet *et al.* (2002)), r_n est le nombre de cibles, tandis que (4.2) est l'équation de la dynamique de chacune d'elles. L'équation d'observation est liée au type de radar utilisé, et elle est généralement non-linéaire.

– Dans le cas des **modèles de trajectoires harmoniques** (Dubois et Davy (2005a,b,c); Dubois *et al.* (2005)), r_n est le nombre de composantes harmoniques présentes dans le signal à l'instant courant. Une trajectoire harmonique est constituée d'une fréquence fondamentale et de ses harmoniques (en nombre fixé, celles qui se trouveraient au-delà de la fréquence de Nyquist se voient attribuer une amplitude nulle). Ici, les équations de transition portent sur les fréquences et amplitudes des différentes fréquences, ainsi que

sur les variances des bruits de transition. Dans Dubois *et al.* (2005), l'équation d'observation est écrite sur chaque trame du signal (définie par la multiplication du signal par une fenêtre glissante), et elle est linéaire pour une partie des paramètres (les amplitudes). Ce modèle a la particularité d'être linéaire et gaussien conditionnellement aux paramètres de fréquences, de variance et à au nombre r_n de trajectoires harmoniques. Plus précisément, dans Dubois et Davy (2005c); Dubois *et al.* (2005), pour une fenêtre w, une trame $s^{\mathrm{w}}_{n_0}(n)$ localisée en n est définie par

$$s^{\mathrm{w}}_n(i) \;=\; x(i)\mathrm{w}(n-i) \tag{4.6}$$

Notre modèle suppose que chaque trame résulte d'une somme de composantes harmoniques,

$$s^{\mathrm{w}}_n(i) \;=\; \mathrm{w}(n-i)\sum_{j=1}^{r_n}\sum_{m=1}^{M} \alpha^{\mathrm{s}}_{n,j,m}\sin(2\pi k_{n,j,m} i) + \alpha^{\mathrm{c}}_{n,j,m}\cos(2\pi k_{n,j,m} i) + \epsilon(i) \tag{4.7}$$

où $\epsilon(i)$ est gaussien et M est le nombre d'harmoniques par note. L'équation (4.7) définit notre équation d'observation. Les fréquences $k_{n,j,m}$ des harmoniques $m = 1, \ldots, M$ pour une note j donnée sont liées par une relation de proportionnalité, mais on autorise une légère inharmonicité (en d'autres termes, les fréquences des harmoniques sont autorisées à dévier de multiples entiers de la fréquence fondamentale). Dans les publications Dubois et Davy (2005a,b), l'équation d'observation est écrite dans le domaine du spectrogramme, ce qui se révèle moins robuste en pratique.

– Dans le cas de la **détection de capteurs potentiellement défaillants** dans les applications de fusion multicapteurs, r_n est un vecteur booléen dont chaque composante correspond à un capteur. Lorsque la variable $r_n(j)$ du capteur j (pour $j = 1, \ldots, J$) est à 0, le capteur est supposé fonctionner normalement, ce qui correspond au choix de l'équation d'observation (4.3) nominale. Lorsque $r_n(j)$ vaut 1, le capteur est supposé défaillant, et une autre équation d'observation (4.3) est sélectionnée. Cette dernière peut soit modéliser une faible corrélation entre l'observation du capteur j et l'état, soit représenter un type de défaut bien connu et modélisable précisément. Bien sûr, ce modèle s'étend facilement à plus de deux états pour chaque capteur. Le modèle que nous avons mis en œuvre dans Caron *et al.* (2005, 2006c) est en fait un cas dégénéré du modèle (4.1)–(4.3), car la relation entre r_{n-1} et r_n n'est pas directe, mais passe par une variable de fiabilité de chaque capteur (modélisant la probabilité de chacun des modes du capteur), dont l'évolution est markovienne, voir Caron *et al.* (2006c).

Modèles hors-ligne à niveaux multiples

Les modèles présentant un seul niveau sont relativement fréquents dans la littérature. Citons par exemple le cas de sinusoïdes bruitées (Andrieu et Doucet (1999)) et le mélange de gaussiennes (Richardson et Green (1997)). Les modèles présentant plusieurs niveaux de hiérarchie sont plus rares. En fait, le seul exemple hormis le nôtre semble être celui de Punskaya *et al.* (2002). Nous avons proposé un modèle à deux niveaux, adapté aux enregistrements musicaux comportant plusieurs instruments, et supposé segmenté en plages stationnaires (c'est-à-dire en notes). Sur un segment, le modèle suivant est défini :

$$x(n) \;=\; \sum_{i=0}^{I} \mathrm{w}(n - i\Delta_n) \sum_{j=1}^{J}\sum_{m=1}^{M_j} \left\{\alpha^{\mathrm{s}}_{j,m,i}\sin(2\pi k_{j,m} n) + \alpha^{\mathrm{c}}_{k,m,i}\cos(2\pi k_{j,m} n)\right\} + \epsilon(n) \tag{4.8}$$

où

- J est le nombre total de notes jouées (une note est composée d'une fréquence fondamentale et d'harmoniques)
- M_j est le nombre d'harmoniques composant la note numéro j
- $k_{j,m}$ est la fréquence de l'harmonique numéro m de la note numéro j
- $\mathrm{w}(n - i\Delta_n)$ est une fonction de base centrée au temps $i\Delta_n$, $(i = 0, \ldots, I)$ de type gaussienne ou Hamming, etc. Les fenêtres w servent à décomposer l'amplitude non-stationnaire de l'harmonique numéro m de la note numéro i sous la forme de coefficients $\alpha^{\mathrm{s}}_{j,m,i}$ et $\alpha^{\mathrm{c}}_{j,m,i}$, $i = 0, \ldots, I$
- $I + 1$ est le nombre total de fenêtres (fonctions de base) pour décomposer l'amplitude
- $\epsilon(n)$ est un bruit gaussien centré, blanc ou autorégressif.

Dans ce cas, la variable de choix de modèle est $r = \{J, M_1, \ldots, M_J\}$, et sa dimension est elle aussi variable. Ce modèle comporte un grand nombre de paramètres : par exemple, un extrait de 0.5 secondes (échantillonné à 44100Hz), composé de 3 notes ayant chacune 15 harmoniques comporte au total 1145 paramètres à estimer. Cela est cependant possible grâce à un choix judicieux de densités *a priori*. Comme détaillé dans Davy et Godsill (2002a); Godsill et Davy (2002); Davy et Godsill (2002c, 2003), on peut les choisir comme des distributions *a priori* conjuguées, ce qui permet d'intégrer analytiquement les amplitudes et la variance du bruit additif (supposée inconnue). Le problème se ramène donc à l'estimation des 45 fréquences dans l'exemple précédent, en explorant la densité *a posteriori* conjointe du vecteur des fréquences $\mathbf{k}$, du vecteur du nombre d'harmoniques $\mathbf{M}$ et du nombre de notes J, notée $\mathsf{p}(\mathbf{k}, \mathbf{M}, J|\mathbf{x})$. Les distributions *a posteriori* conditionnelles de la variance du bruit additif $\mathsf{p}(\sigma^2|\mathbf{k}, \mathbf{M}, J, \mathbf{x})$ et des amplitudes $\mathsf{p}(\boldsymbol{\alpha}|\sigma^2, \mathbf{k}, \mathbf{M}, J, \mathbf{x})$ sont, elles aussi, disponibles sous forme analytique.

En plus de ce modèle à deux niveaux, nous en avons étudié plusieurs à un seul niveau. L'un porte sur des modèles autorégressifs (Dobigeon *et al.* (2006a)) utilisés pour la segmentation bayésienne conjointe de signaux enregistrés en parallèle, dans lesquels les instants de rupture sont corrélés. Le second reprend le modèle d'estimation spectrale bayésienne présenté dans Andrieu et Doucet (1999), dans lequel nous avons mis en œuvre un algorithme Monte Carlo alternatif aux MCMC à sauts réversibles (Davy *et al.* (2003)). Enfin, dans Davy *et al.* (2000a, 2002a, 2000a); Cottereau *et al.* (2003), nous avons développé des modèles bayésiens et des algorithmes de Monte Carlo sans sauts.

4.1.2 Algorithmes de Monte Carlo à sauts

Les modèles à sauts présentés ci-dessus sont d'une complexité parfois importante, et la mise en œuvre des estimateurs bayésiens correspondants n'est pas directe. Il nous a été nécessaire de développer des algorithmes spécifiques, permettant d'effectivement réaliser l'estimation bayésienne des paramètres, tout en conservant des temps de calcul raisonnables.

Filtrage particulaire adapté aux modèles de Markov à sauts

L'algorithme que nous avons proposé comporte trois phases importantes (outre la phase d'estimation) résumées à l'algorithme 4.5 ci-dessous. La première phase est le calcul de variables auxiliaires, selon l'idée de Pitt et Shephard (1999). Dans notre algorithme, nous utilisons de façon originale l'approximation sans parfum pour estimer la vraisemblance prédictive (à la différence de Pitt et Shephard (1999), qui suggère d'utiliser un mode de la distribution à intégrer). Par ailleurs, l'échantillonnage des particules utilise les idées du filtre particu-

laire sans parfum de van der Merwe *et al.* (2000), tout en tirant parti de quantités calculées auparavant pour les poids auxiliaires.

Algorithme 4.5: Filtre particulaire pour modèle de Markov à sauts

- Initialisation : initialiser chaque particule $\big(\widetilde{\boldsymbol{\theta}}_0^{(i)}, \widetilde{r}_0^{(i)}\big)$ et faire $\widetilde{\omega}_0^{(i)} = 1/N$.
- Pour $n = 1, 2, \ldots$, faire
  1. Rééchantillonnage par variable auxiliaire. Pour chaque particule $i = 1, \ldots, N$, en utilisant l'approximation sans parfum, calculer une estimation $\widehat{\mathsf{p}}(\mathbf{x}_n|\mathbf{x}_{1:n-1}, \widetilde{\boldsymbol{\theta}}_{n-1}^{(i)}, \widetilde{r}_{n-1}^{(i)})$ de la vraisemblance prédictive définie par

$$\mathsf{p}(\mathbf{x}_n|\mathbf{x}_{1:n-1}, \widetilde{\boldsymbol{\theta}}_{n-1}^{(i)}, \widetilde{r}_{n-1}^{(i)}) \;=\; \sum_{r_n \in \mathcal{R}} \pi_{r_n\, \widetilde{r}_{n-1}^{(i)}} \int g_{r_n}(\mathbf{x}_n|\boldsymbol{\theta}_n, \mathbf{x}_{1:n-1})\, f_{\widetilde{r}_{n-1}^{(i)}, r_n}(\boldsymbol{\theta}_n|\widetilde{\boldsymbol{\theta}}_{n-1}^{(i)})\, d\boldsymbol{\theta}_n \quad (4.9)$$

     et calculer les poids auxiliaires

$$\widetilde{\lambda}^{(i)} \;\propto\; \omega_n(i)\widehat{\mathsf{p}}(\mathbf{x}_n|\mathbf{x}_{1:n-1}, \widetilde{\boldsymbol{\theta}}_{n-1}^{(i)}, \widetilde{r}_{n-1}^{(i)}) \quad , \qquad \sum_{i=1}^{N} \widetilde{\lambda}^{(i)} = 1 \quad (4.10)$$

     puis rééchantillonner les particules en dupliquant celles qui ont un poids auxiliaire important, et en supprimant celles qui ont un poids faible.
  2. Mise à jour et prolongation des trajectoires. Pour chaque particule $i = 1, \ldots, N$, faire
     - Echantillonner $\widetilde{r}_n^{(i)}$ selon la loi de probabilité discrète $\mathsf{Q}(\cdot)$ sur $\mathcal{R}$ telle que

$$\mathsf{Q}(r_n) \propto \pi_{r_n\, \widetilde{r}_{n-1}^{(i)}} \int g_{r_n}(\mathbf{x}_n|\boldsymbol{\theta}_n, \mathbf{x}_{1:n-1})\, f_{\widetilde{r}_{n-1}^{(i)}, r_n}(\boldsymbol{\theta}_n|\widetilde{\boldsymbol{\theta}}_{n-1}^{(i)})\, d\boldsymbol{\theta}_n \quad (4.11)$$

       c'est-à-dire proportionnelle à sa contribution à la vraisemblance prédictive, déjà calculée en (4.9).
     - Echantillonner le paramètre selon $\widetilde{\boldsymbol{\theta}}_n^{(i)} \;\sim\; \mathcal{N}\big(\boldsymbol{\theta}_n; \widehat{\boldsymbol{\theta}}_{n|n}^{(i)}, \widehat{\boldsymbol{\Sigma}}_{n|n}^{(i)}\big)$ où la moyenne $\widehat{\boldsymbol{\theta}}_{n|n}^{(i)}$ et la covariance $\widehat{\boldsymbol{\Sigma}}_{n|n}^{(i)}$ sont calculées par le filtre de Kalman sans parfum en utilisant les points d'approximation (*sigma points*) utilisés pour calculer l'estimation $\widehat{\mathsf{p}}(\mathbf{x}_n|\mathbf{x}_{1:n-1}, \widetilde{\boldsymbol{\theta}}_{n-1}^{(i)}, \widetilde{r}_{n-1}^{(i)})$ à l'équation (4.9).
  3. Calcul des poids. Pour chaque particule $i = 1, \ldots, N$, faire

$$\omega_n^{(i)} \;\propto\; \frac{g_{\widetilde{r}_n^{(i)}}(\mathbf{x}_n|\widetilde{\boldsymbol{\theta}}_n^{(i)}, \mathbf{x}_{1:n-1})}{\widehat{\mathsf{p}}(\mathbf{x}_n|\mathbf{x}_{1:n-1}, \widetilde{\boldsymbol{\theta}}_{n-1}^{(i)}, \widetilde{r}_{n-1}^{(i)})} \frac{\pi_{\widetilde{r}_n^{(i)}\, \widetilde{r}_{n-1}^{(i)}}}{\mathsf{Q}(\widetilde{r}_n^{(i)})} \frac{f_{\widetilde{r}_{n-1}^{(i)}, \widetilde{r}_n^{(i)}}(\widetilde{\boldsymbol{\theta}}_n^{(i)}|\widetilde{\boldsymbol{\theta}}_{n-1}^{(i)})}{\mathcal{N}\big(\widetilde{\boldsymbol{\theta}}_n^{(i)}; \widehat{\boldsymbol{\theta}}_{n|n}^{(i)}, \widehat{\boldsymbol{\Sigma}}_{n|n}^{(i)}\big)}. \quad (4.12)$$

     et normaliser les poids de telle sorte que $\sum_{i=1}^{N} \omega_n^{(i)} \;=\; 1$.
  4. Estimation. Par exemple, on peut calculer l'estimation MMAP $\widehat{r}_n$ de r_n (c'est la valeur $r \in \mathcal{R}$ telle que la somme des poids des particules ayant $\widetilde{r}_n^{(i)} = r$ est la plus grande), puis l'estimation MMSE marginale de $\boldsymbol{\theta}_n$ par moyenne pondérée des $\widetilde{\boldsymbol{\theta}}_n^{(i)}$ pour les particules $i = 1, \ldots, N$ telles que $\widetilde{r}_n^{(i)} = \widehat{r}_n$.

Dans cet algorithme, nous faisons appel à l'approximation sans parfum (voir Julier et Uhlmann (2004) pour une description complète de la méthode). Il s'agit d'une technique performante d'approximation déterministe des intégrales, d'inspiration Monte Carlo. L'hypothèse sous-jacente est que la distribution à intégrer est gaussienne (tout au moins, monomodale),

permettant que des points représentatifs de la moyenne et de la covariance soient déterminés (les *sigma points*). Ces points peuvent être utilisés pour approcher des intégrales, ou encore pour construire le filtre de Kalman sans parfum qui est plus précis en général que le filtre de Kalman étendu en cas de modèle non-linéaire.

Algorithme MCMC à doubles sauts réversibles pour modèles harmoniques

Le modèle harmonique dédié à la musique présenté à l'équation (4.8) a une structure hiérarchique à deux niveaux, qui demande un algorithme spécifique. Une version « élémentaire » (mais inefficace) d'algorithme MCMC pour ce modèle comporterait des sauts réversibles de type naissance/mort au niveau des harmoniques d'une note (ajout/retrait de l'harmonique de plus haute fréquence) et au niveau des notes (ajout/retrait d'une note). Cependant, cet algorithme (déjà assez complexe) est inefficace car il reste bloqué dans des « modes locaux », c'est-à-dire que les fréquences et des notes trouvées sont assez éloignées du résultat attendu. Il est donc nécessaire de raffiner les sauts réversibles mis en œuvre. Dans Davy *et al.* (2006c), nous avons proposé la solution suivante :

- Au niveau le plus large, une note est ajoutée ou retirée en bloc selon une loi de proposition construite à partir du signal sonore traité, afin de maximiser la probabilité d'acceptation
- Au niveau de chaque note, une première paire de sauts permet d'ajouter ou retirer d'un coup plusieurs harmoniques les plus hautes. Ce mouvement est une généralisation originale du *birth/death* standard
- Toujours au niveau de chaque note, une seconde paire de sauts permet de multiplier/diviser la fréquence du fondamental par deux, tout en conservant les harmoniques d'avant le mouvement. Cela nécessite de supprimer/ajouter une harmonique sur deux. Pour que cette paire de mouvements reste réversible, l'harmonique de plus haute fréquence peut, en outre, être (aléatoirement) supprimée, ajoutée ou laissée inchangée. Ce saut réversible est lui aussi original, et très utile pour éviter les erreurs d'octave.
- Enfin, au niveau de chaque harmonique, la fréquence est mise à jour par un mélange de lois de proposition locale et globale (c'est une solution standard et efficace). Les amplitudes et la variance du bruit additionnel sont échantillonnés directement selon leurs loi *a posteriori* conditionnelles (gaussienne et inverse gamma respectivement).

En plus de l'estimation des paramètres de chaque note, l'algorithme présenté dans Davy *et al.* (2006c) estime aussi les hyperparamètres, rendant ainsi toute la procédure plus robuste tant du point de vue théorique (modèle plus souple) que pratique (convergence de l'algorithme facilitée). La mise en œuvre nécessite des astuces supplémentaires afin d'assurer un temps de calcul raisonnable. L'une d'entre elle est décrite dans Davy et Idier (2004). Enfin, notons que la procédure finale, efficace et rapide, fait partie des algorithmes MCMC les plus complexes jamais conçus (à ma connaissance).

4.1.3 Autres algorithmes de Monte Carlo

De nombreux modèles bayésiens développés pour des applications du Traitement du Signal s'écrivent comme la somme de composantes ayant toutes la même forme, mais des paramètres différents. Le modèle harmonique ci-dessus en est un exemple, où les composantes sont les harmoniques individuelles, et, à une plus large échelle, les notes. De la même façon, le modèle d'analyse spectrale de Andrieu et Doucet (1999) s'écrit comme une somme de sinusoïdes.

Dans ces cas, en supposant que les amplitudes des composantes sont gaussiennes *a priori*, il est possible de tirer parti de cette structure de modèle pour concevoir l'algorithme MCMC rapide de Davy et Idier (2004). Brièvement, cet algorithme consiste à mettre à jour les paramètres de chaque composante un à la fois conditionnellement aux autres (à la Gibbs), tout en réutilisant au maximum les termes du calcul de la distribution *a posteriori*, notamment les décompositions de Choleski des matrices de covariance. Ces éléments ont été utilisés dans Davy *et al.* (2006c), et ont permis des gain calculatoires très significatifs : dans les version initiales de Davy et Godsill (2002a); Godsill et Davy (2002), l'estimation des notes présentes dans 0.5s de signal nécessitait plusieurs jours, contre quelques dizaines de secondes dans la version accélérée de Davy *et al.* (2006c).

Plus récemment, nous avons abordé une des premières mises en œuvre de l'algorithme d'échantillonage de distributions par échantillonnage préférentiel séquentiel (*Sequential Monte Carlo sampler*) Del Moral *et al.* (2006), pour l'analyse spectrale bayésienne. Brièvement, les algorithmes décrits dans Del Moral *et al.* (2006) sont une alternative aux méthodes MCMC pour l'échantillonnage de distributions, et aux méthodes d'optimisation stochastique de type recuit simulé et algorithmes génétiques. Nous avons mis en œuvre ces algorithmes pour l'analyse spectrale, ce qui a permis de mettre en évidence l'efficacité de l'approche par comparaison aux méthodes MCMC dans des problèmes de Traitement du Signal.

4.2 Méthodes à Noyaux et apprentissage statistique

Cette section présente mes contributions dans le cadre des méthodes à noyaux. Elles concernent la classification de signaux non-stationnaires, l'estimation d'ensembles de mesure minimale (pour la détection de ruptures et d'événements anormaux) et de régression à coefficients positifs, comme détaillé dans les paragraphes suivants.

4.2.1 Classification de signaux non-stationnaires

Dans Gretton *et al.* (2001); Davy *et al.* (2002b), nous avons prolongé les travaux de ma thèse en mettant en place une procédure de classification des signaux non-stationnaires à l'aide de représentations temps-fréquence optimisées et d'algorithmes à vecteurs support. Plus précisément, le travail effectué peut être résumé ainsi :

1. Les signaux non-stationnaires à classer sont transformés sous forme de représentations temps-fréquence de la classe de Cohen, de noyau $\phi(t,f)$, c'est-à-dire, pour le signal $x(t)$,

$$\mathsf{TFR}_x^{\phi}(t,f) \;=\; \iint \mathsf{WV}_x(s-t,\nu-f)\,\phi(s,\nu)\,dsd\nu \tag{4.13}$$

   où $\mathsf{WV}_x \;=\; \int x(t+\tau/2)\,x(t-\tau/2)^{\star}\,\exp(-j2\pi f\tau)d\tau$ est la représentation de Wigner-Ville du signal $x(t)$, voir Flandrin (1999). Comme dans Davy *et al.* (2001b), le noyau est choisi radialement gaussien, dont le contour est défini par des descripteurs de Fourier.

2. Dans le domaine temps-fréquence, un classifieur à vecteurs support est défini grâce à un noyau gaussien, et ses multiplicateurs de Lagrange α_i sont optimisés. On peut remarquer que la transformation qui fournit $\mathsf{TFR}_x^{\phi}(t,f)$ à partir de $x(t)$ peut être vue comme faisant partie du noyau SVM. Ainsi,

$$\mathsf{k}\big(x_1(t),x_2(t)\big) \;=\; \exp -\frac{1}{2\sigma^2}\|\mathsf{TFR}_{x_1}^{\phi}(t,f)-\mathsf{TFR}_{x_2}^{\phi}(t,f)\|_{\mathbb{R}^2}^2 \tag{4.14}$$

3. Les performances de ce classifieur sont directement liées aux choix de $\phi(t,f)$ (les descripteurs de Fourier), et du paramètre de variance σ^2. Tous ces paramètres sont optimisés conjointement par validation croisée. Bien entendu, cette optimisation nécessite un estimateur du vrai risque à partir de l'ensemble d'apprentissage. Cet estimateur est défini ci-dessous.

Afin de définir un estimateur du risque par validation croisée, nous étudions les statistiques de la fonction SVM de classification $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$. Supposons que nous disposons d'un ensemble d'apprentissage (X,Y), et un ensemble de test noté $(\mathsf{X}_\mathsf{t},\mathsf{Y}_\mathsf{t})$. Etant donnée une fonction $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$ apprise sur (X,Y), nous nous intéressons à $\mathsf{p}\big(\mathsf{F}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x}),\mathbf{x}\in\mathsf{X}_\mathsf{t}|\mathsf{y}=+1\big)$ et $\mathsf{p}\big(\mathsf{F}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x}),\mathbf{x}\in\mathsf{X}_\mathsf{t}|\mathsf{y}=-1\big)$ où y est le label de $\mathbf{x}\in\mathsf{X}_\mathsf{t}$. Statistiquement, et comme $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$ est « bien apprise », ces deux densités sont centrées sur une valeur positive et une valeur négative, respectivement. Dans Davy *et al.* (2002b), nous avons fait l'hypothèse que ces deux distributions sont gaussiennes, de moyenne respectives m_{+1} et m_{-1} et de variances σ^2_{+1} et σ^2_{-1}. Sous cette hypothèse, la probabilité d'erreur de $\mathsf{F}_{(\mathsf{X},\mathsf{Y})}$ est donnée par

$$\mathsf{P}_{\mathsf{err}}\big(\mathsf{F}_{(\mathsf{X},\mathsf{Y})}\big) \;\approx\; \int_{-\infty}^{0}\mathcal{N}\big(u;m_{+1},\sigma^2_{+1}\big)du + \int_{0}^{\infty}\mathcal{N}\big(u;m_{-1},\sigma^2_{-1}\big)du \tag{4.15}$$

et permet de réaliser l'optimisation. En pratique, nous estimons $\mathsf{P}_{\mathsf{err}}\big(\mathsf{F}_{(\mathsf{X},\mathsf{Y})}\big)$ par moyenne sur différents ensembles d'apprentissage et de test obtenus en découpant de plusieurs façons (aléatoires) l'ensemble de données qui sont fournies, voir Hastie *et al.* (2001). Cette approche a permis de diviser les taux d'erreur de classification de Davy *et al.* (2001b) par deux. Par ailleurs, l'hypothèse gaussienne faite pour $\mathsf{p}\big(\mathsf{F}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x}),\mathbf{x}\in\mathsf{X}_\mathsf{t}|\mathsf{y}=+1\big)$ et $\mathsf{p}\big(\mathsf{F}_{(\mathsf{X},\mathsf{Y})}(\mathbf{x}),\mathbf{x}\in\mathsf{X}_\mathsf{t}|\mathsf{y}=-1\big)$ est justifiée expérimentalement.

4.2.2 Comparaison d'ensembles de vecteurs et détection de ruptures

De nombreux problèmes de traitement du signal et d'analyse de données requièrent la comparaison d'ensembles de points, sans relation d'ordre. Des exemples important concerent la classification de textures d'images, la vérification de locuteurs ou la détection de ruptures (certaines de ces applications sont présentées au chapitre 5). Dans la suite, nous notons $\mathsf{X}_1 = \{\mathbf{x}_{1,1},\ldots,\mathbf{x}_{1,m_1}\}$ et $\mathsf{X}_2 = \{\mathbf{x}_{2,1},\ldots,\mathbf{x}_{2,m_2}\}$ et il est supposé que ces vecteurs appartiennent tous au même espace, c'est-à-dire $\mathbf{x}_{i,j}\in\mathcal{X}$ quels que soient $i\in\{1,2\}$ et $j\in\{1,\ldots,m_i\}$.

Une des façons les plus simples de les comparer consiste à calculer le *contraste de Fisher* (Duda *et al.* (2001)),

$$\mathsf{K}_{\mathrm{Fisher}}\big(\mathsf{X}_1,\mathsf{X}_2\big) \;=\; \big(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2\big)^\mathsf{T}\big(\widehat{\boldsymbol{\Sigma}}_1+\widehat{\boldsymbol{\Sigma}}_2\big)^{-1}\big(\widehat{\boldsymbol{\mu}}_1-\widehat{\boldsymbol{\mu}}_2\big) \tag{4.16}$$

où $\widehat{\boldsymbol{\mu}}_i$ et $\widehat{\boldsymbol{\Sigma}}_i$ sont les moyennes et matrices de covariance estimées pour l'ensemble X_i, $i=1,2$. Si l'on suppose que les données dans les ensembles X_1 et X_2 sont aléatoires, et distribuées selon des distributions gaussiennes, alors le contraste de Fisher correspond à la séparation entre X_1 et X_2, dans le sens où la probabilité d'erreur[1] diminue lorsque $\mathsf{K}_{\mathrm{Fisher}}\big(\mathsf{X}_1,\mathsf{X}_2\big)$ augmente.

Cette mesure est très utilisée, mais elle connaît plusieurs limitations. Tout d'abord, il est numériquement difficile d'estimer les matrices de covariance lorsque $\mathcal{X}$ est de grande dimension (car il faut beaucoup plus de vecteurs que de dimensions). Ensuite, ce contraste n'est pas défini

[1]La probabilité d'erreur désigne ici la probabilité que, pour un $\mathbf{x}$ donné dans X_i, la densité gaussienne de $\mathsf{X}_{j,j\neq i}$ évaluée en $\mathbf{x}$ est plus grande que la densité de X_i évaluée en $\mathbf{x}$.

sur des espaces $\mathcal{X}$ non numériques. Enfin, l'hypothèse gaussienne faite implicitement peut être très abusive, notamment si les ensembles X_i $(i = 1, 2)$) présentent des région non connexes.

Partant de ce constat, il nous est apparu que des mesures de contraste plus fines devaient être développées. Nous avons abordé la question sous l'angle suivant. Tout d'abord, et contrairement à la plupart des méthodes fournissant une mesure de contraste, nous laissons de côté l'estimation paramétrique des densités ayant généré X_1 et X_2. Ensuite, ayant choisi une approche non-paramétrique[2], nous estimons des *ensembles de mesure minimale* (EMM) plutôt que des densités.

Définition 4.7: Soit $(\mathcal{X}, \mathcal{B}, \mathsf{P})$ un triplet de probabilité, Q une mesure sur $\mathcal{X}$ et une constante $c \in [0, 1]$. Un *ensemble de mesure minimale* de $\mathcal{X}$, noté $\mathcal{R}$, est l'élément de $\mathcal{B}$ tel que

1. $\mathsf{P}(\mathcal{R}) = c$;
2. $\mathsf{Q}(\mathcal{R})$ est minimum. □

L'idée que nous avons poursuivie est de construire des mesures de contraste basées sur les ensembles de mesure minimale estimés. Une méthode algorithmiquement séduisante pour l'estimation d'EMM est la méthode de classification à vecteurs support à une classe, Schölkopf et Smola (2002). Ces estimateurs sont obtenus par minimisation d'un risque régularisé (cf. section 3.2.2). Pour un ensemble de données X (sans labels), le risque régularisé pour l'estimation d'EMM est

$$\mathsf{R}^{\text{reg}}\left[\mathsf{F}\right] \;=\; \sum_{i=1}^{m} \max\left[0, \mathsf{F}(\mathbf{x}_i)\right] + \lambda \widetilde{\Omega}\, \|\mathsf{F}\|_{\mathcal{H}}^2 \tag{4.17}$$

où $\mathsf{F}(\cdot) \in \mathcal{F}$, cf. équation (3.34). Ce problème admet une formulation duale, de forme quadratique sous contraintes linéaires, voir Schölkopf et Smola (2002), et le théorème du représentant (Théorème 3.1, page 35) indique que la solution optimale s'écrit

$$\mathsf{F}_{\mathsf{X}}(\mathbf{x}) \;=\; \sum_{i=1}^{m} \alpha_i \mathsf{k}\big(\mathbf{x}_i, \mathbf{x}\big) + b \tag{4.18}$$

où les α_i, $i = 1, \ldots, m$ et b sont obtenus par optimisation numérique. L'EMM est alors défini par $\mathcal{R}_{\mathsf{X}} = \{\mathbf{x} \in \mathcal{X} \text{ tels que } \mathsf{F}_{\mathsf{X}}(\mathbf{x}) \geq 0\}$. En pratique, on utilise la formulation ν-SVM à une classe, qui autorise des données à être à l'extérieur de la région $\mathcal{R}_{\mathsf{X}}$. En outre, le paramètre ν est directement lié à c, qui définit la mesure de l'EMM. Pour plus de détails sur la relation entre EMM et ν-SVM à une classe, voir Davy *et al.* (2006a).

Afin de construire une mesure de contraste entre EMM, nous utilisons une interprétation géométrique du problème (4.17) représentée à la figure 4.1. Cette interprétation est très puissante : en effet, dans $\mathcal{H}$, la frontière de la région $\mathcal{R}_{\mathsf{X}}$ est incluse dans l'intersection de l'hyperplan $\mathsf{f}_{\mathsf{X}}(\mathbf{x}) + b = 0$ avec l'hypersphère de rayon 1, quelle que soit la forme de $\mathcal{R}_{\mathsf{X}}$ (forme simple ou très sinueuse, avec un ou plusieurs ensembles connexes, etc.). A la figure 4.2, deux estimateurs ν-SVM à une classe sont optimisés pour les ensembles X_1 et X_2, fournissant deux régions $\mathcal{R}_{\mathsf{X}_1}$ et $\mathcal{R}_{\mathsf{X}_2}$, dont la comparaison est très simple : pour construire une mesure de contraste, nous avons proposé dans Desobry et Davy (2003b) et Desobry *et al.* (2005a) de

[2]Le terme « non-paramétrique » fait ici référence à des estimateurs ne faisant pas appel à un modèle paramétrique de densité (gaussienne, mélange de gaussiennes,...).

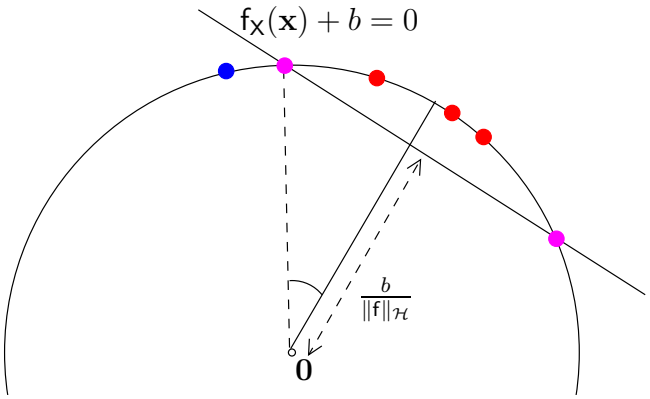


FIG. 4.1 – Représentation géométrique du classifieur à vecteurs support à une classe, dans l'espace $\mathcal{H}$. Pour un noyau normalisé invariant par translation, les fonctions $\mathsf{k}\big(\mathbf{x}, \cdot\big)$ sont toutes de norme unitaire, aussi se trouvent-elles sur une sphère de rayon 1, et de dimension pouvant être infinie. Si l'on définit $\mathsf{f}_{\mathsf{X}}(\cdot) = \sum_{i=1}^{m} \alpha_i \mathsf{k}\big(\mathbf{x}_i, \cdot\big)$, alors $\langle \mathsf{f}_{\mathsf{X}}(\cdot), \mathsf{k}\big(\mathbf{x}, \cdot\big)\rangle_{\mathcal{H}} + b = \mathsf{f}_{\mathsf{X}}(\mathbf{x}) + b = 0$ définit un hyperplan qui est aussi loin que possible de l'origine, tout en séparant les données du demi-espace qui la contient. Dans le cas des ν-SVM à une classe, certaines données sont autorisées à être dans ce demi-espace : ce sont des éléments déclarés « anormaux ». Ici, les points rouges représentent les données déclarées « normales », en violet, ce sont les données « limite », et en bleu ce sont les données « anormales ».

comparer les positions relatives des centres $\mathbf{c}_1$ et $\mathbf{c}_2$, et des points périphériques $\mathbf{p}_1$ et $\mathbf{p}_2$. La nouvelle mesure de contraste est définie par

$$\mathsf{K}_{\mathrm{SVM}}(\mathsf{X}_1, \mathsf{X}_2) \;=\; \frac{\mathsf{d}_{\mathrm{arc}}(\mathbf{c}_1, \mathbf{c}_2)}{\mathsf{d}_{\mathrm{arc}}(\mathbf{c}_1, \mathbf{p}_1) + \mathsf{d}_{\mathrm{arc}}(\mathbf{c}_2, \mathbf{p}_2)} \tag{4.19}$$

où $\mathsf{d}_{\mathrm{arc}}(\cdot, \cdot)$ est la distance sur l'arc de sphère de rayon un, qui correspond ici à l'angle entre les points exprimé en radians. Par ailleurs, nous avons montré que ces distances peuvent être calculées en fonction des paramètres $\alpha_{1,i}$, b_1, $\alpha_{2,i}$ et b_2. La mesure de contraste ainsi définie possède de bonnes propriétés, voir Desobry (2004) et les figures 4.3–4.4.

Application à la détection de ruptures

Cette mesure de contraste peut être utilisée de manière directe pour détecter des ruptures dans une série temporelle $\mathbf{x}_n$, $n = 1, 2, \ldots$. En effet, définissons l'ensemble de vecteurs *passé immédiat* $\mathsf{X}_{1,n} = \{x_{n-m_1}, \ldots, x_{n-1}\}$ et *futur immédiat* $\mathsf{X}_{2,n} = \{x_n, \ldots, x_{n+m_2-1}\}$ relatifs à l'instant n. Il est alors possible de tracer au cours du temps le contraste $\mathsf{K}_{\mathrm{SVM}}(\mathsf{X}_{1,n}, \mathsf{X}_{2,n})$, dont les pics indiquent des rupture. La méthodologie standard de détection de ruptures s'applique alors, voir Desobry et Davy (2003a,b); Desobry (2004); Desobry *et al.* (2005a). Une des caractéristiques essentielles de cette approche est que la série temporelle $\mathbf{x}_n$, $n = 1, 2, \ldots$ à segmenter peut être de grande dimension, ou non numérique. Dans le chapitre 5, nous en présentons quelques applications.

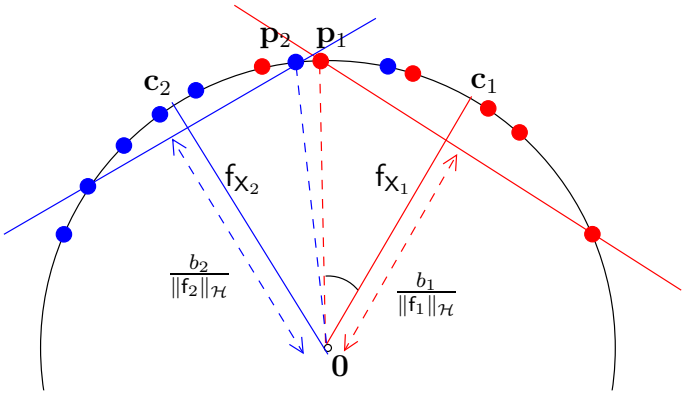


FIG. 4.2 – Représentations géométriques de deux classifieurs à une classe optimisés indépendemment sur deux ensembles de données X_1 (en rouge) et X_2 (en bleu). Chacun de ces classifieurs fournit une fonction f_1 et f_2 dont la prolongation coupe la sphère en $\mathbf{c}_1$ et $centre_2$. Les hyperplans $\mathsf{f}_{\mathsf{X}_1}(\mathbf{x}) + b = 0$ et $\mathsf{f}_{\mathsf{X}_2}(\mathbf{x}) + b = 0$ coupent la sphère en une infinité de points, parmi lesquels on sélectionne $\mathbf{p}_1$ et $\mathbf{p}_2$ dans le plan $(\mathbf{0}, \mathbf{c}_1, \mathbf{c}_2)$.

Calcul itératif des α

La mise en œuvre de la méthode de détection de ruptures requiert le calcul de deux estimateurs SVM d'EMM à chaque instant, ce qui peut être coûteux si l'on recommence à chaque instant. Pour limiter la charge calculatoire, nous avons proposé un algorithme séquentiel qui utilise la solutions SVM à une classe de $\mathsf{X}_{1,n-1}$ comme initialisation pour calculer la solution correspondant à l'ensemble $\mathsf{X}_{1,n}$. Cela est rendu possible car ces deux ensembles ne diffèrent que par le couple $\{\mathbf{x}_{n-m_1-1}, \mathbf{x}_{n-1}\}$. Dans Davy *et al.* (2006b), nous proposons de mettre à jour la solution SVM (les α_i à l'équation (4.18)) en incorporant le nouvel élément $\mathbf{x}_{n-1}$, puis en retirant l'élément ancien $\mathbf{x}_{n-m_1-1}$. Cette méthode à l'avantage de la rapidité et de la stabilité numérique, à la différence de méthodes concurrentes, comme celle de Kivinen *et al.* (2004) où le paramètre b de l'équation (4.18) dérive au cours des itérations.

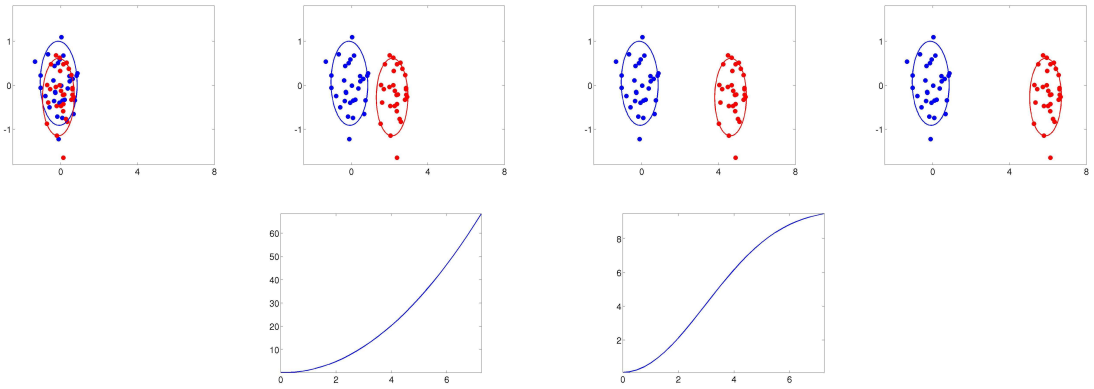
Détection d'événements anormaux en ligne

Un problème voisin de la détection de ruptures est celle de la détection d'événements anormaux en ligne, décrit dans Davy et Godsill (2002b) et Davy *et al.* (2006b). On construit alors un seul ensemble $\mathsf{X}_n = \{\mathbf{x}_{n-m}, \ldots, \mathbf{x}_{n-1}\}$. La solution ν-SVM à une classe à chaque instant $\{\alpha_{1,n}, \ldots, \alpha_{m,n}, b_n\}$ est mise à jour à l'aide de l'algorithme séquentiel évoqué ci-dessus. Un évènement est déclaré déclaré anormal s'il se situe à l'extérieur de la région $\mathcal{R}_{\mathsf{X}_n}$, c'est-à-dire si

$$\sum_{i=1}^{m} \alpha_{i,n} \mathsf{k}\big(\mathbf{x}_{n-m+i-1}, \mathbf{x}_n\big) + b_n < 0 \tag{4.20}$$

où $\mathbf{x}_n$ est l'élément à tester.

Dans Davy *et al.* (2006b), nous fournissons une interprétation géométrique simple dans $\mathcal{H}$ du test (4.20) : il s'agit du rapport du cosinus de l'angle entre $\mathsf{f}_{\mathsf{X}_n}(\cdot) = \sum_{i=1}^{m} \alpha_{i,n} \mathsf{k}\big(\mathbf{x}_{n-m+i-1}, \cdot\big)$



Contraste $\mathsf{K}_{\text{Fisher}}(\mathsf{X}_1, \mathsf{X}_2)$ Contraste $\mathsf{K}_{\text{SVM}}(\mathsf{X}_1, \mathsf{X}_2)$

FIG. 4.3 – Les deux ensembles X_1 (rouge) et X_2 (bleu) sont générés par des distributions gaussiennes, dont les moyennes s'écartent petit à petit. Pour chaque position de la moyenne de X_1, on a tracé le contracte de Fisher $\mathsf{K}_{\text{Fisher}}(\mathsf{X}_1, \mathsf{X}_2)$ et le contraste $\mathsf{K}_{\text{SVM}}(\mathsf{X}_1, \mathsf{X}_2)$. Ils ont un comportement similaire.

et $\mathsf{k}(\mathbf{x}_t, \cdot)$ au cosinus en l'angle entre $\mathsf{f}_{\mathsf{X}_n}(\cdot)$ et $\mathbf{p}_t$, où $\mathbf{p}_t$ est définit comme à la figure 4.2.

Généralisations

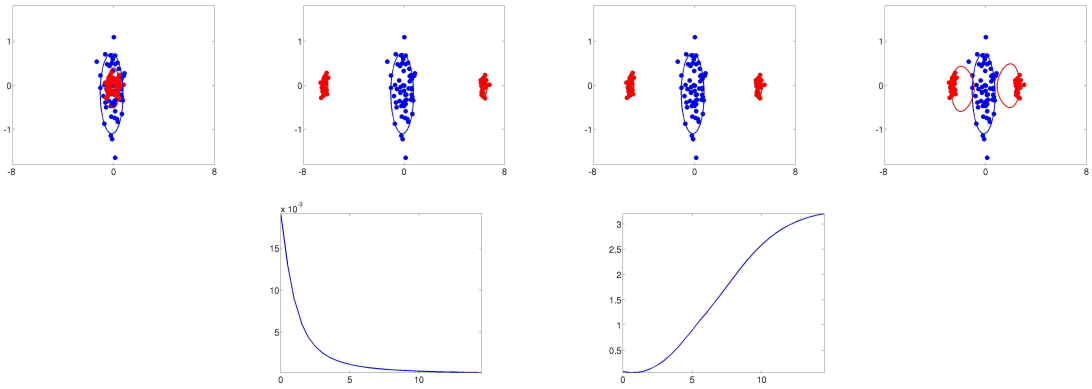
La mesure de contraste $\mathsf{K}_{\text{SVM}}(\mathsf{X}_1, \mathsf{X}_2)$ peut être généralisée de multiples façons. Dans Desobry et Davy (2004), nous proposons toute une classe de tels contrastes, introduits grâce à la notion de *fonctions préservant la métrique* (*Metric preserving functions*, voir Corazza (1999)). Nous invitons le lecteur intéressé à se reporter à Desobry et Davy (2004); Desobry (2004) pour les détails.

Application à la définition de noyaux entre ensembles de points

La famille de mesures de contraste définie ici permet d'aborder de nombreux problèmes de détection et de classification s'écrivant en termes de deux ensembles X_1 et X_2. Toutefois, on ne dispose pas d'un noyau défini-positif $\mathsf{K}(\mathsf{X}_1, \mathsf{X}_2)$ permettant de comparer globalement X_1 et X_2, et qui permettrait l'utilisation de la machinerie des méthodes à noyaux sur des ensembles de vecteurs. Pour définir un tel noyau, il s'agit de considérer X_1 et X_2 comme des objets appartenant à l'espace des ensembles d'éléments de $\mathcal{X}$ sur lequel on cherche à définir $\mathsf{K}(\mathsf{X}_1, \mathsf{X}_2)$. Il ne s'agit plus de définir la matrice noyau $\left[\mathsf{k}(\mathbf{x}_{1,i}, \mathbf{x}2, j)\right]_{i=1,\ldots,m_1, j=1,\ldots,m_2}$ pour $\mathbf{x}_{1,i} \in \mathsf{X}_1$ et $\mathbf{x}_{2,j} \in \mathsf{X}_2$, mais un seul nombre réel $\mathsf{K}(\mathsf{X}_1, \mathsf{X}_2)$.

Un tel noyau permet de traiter des problèmes où les données sont des ensembles de points de tailles potentiellement différentes. Des exemples immédiats sont

- Les coefficients cepstraux extraits sur plusieurs fenêtres pour un signal de parole donné. Avec un tel noyau, on peut repenser la reconnaissance de la parole, ou encore la segmentation en locuteurs
- Les pixels extraits d'une zone d'une image, et ayant une texture donnée. On peut alors effectuer une classification des textures



Contraste $\mathsf{K}_{\text{Fisher}}(\mathsf{X}_1, \mathsf{X}_2)$ Contraste $\mathsf{K}_{\text{SVM}}(\mathsf{X}_1, \mathsf{X}_2)$

FIG. 4.4 – Les deux ensembles X_1 (rouge) et X_2 (bleu) sont générés par des distributions non gaussiennes, dont les moyennes restent les mêmes. Pour chaque position de X_1, on a tracé le contracte de Fisher $\mathsf{K}_{\text{Fisher}}(\mathsf{X}_1, \mathsf{X}_2)$ et le contraste $\mathsf{K}_{\text{SVM}}(\mathsf{X}_1, \mathsf{X}_2)$. Seul le second a le comportement attendu.

Pour que $\mathsf{K}\big(\mathsf{X}_1, \mathsf{X}_2\big)$ soit un noyau, il doit être défini positif. Dans Desobry *et al.* (2005b) nous proposons une stratégie de construction d'un tel noyau, basé sur des ν-SVM à une classe.

4.2.3 Régression à coefficients positifs

Les paragraphes précédents concernent la classification et la détection à l'aide de noyaux. Nous avons également abordé un des aspects principaux de l'apprentissage automatique : la régression. Etant donnés des couples $(\mathsf{X}, \mathsf{Y}) = \{(\mathbf{x}_1, \mathsf{y}_1), \ldots, (\mathbf{x}_m, \mathsf{y}_m)\}$ avec $\mathcal{Y} = \mathbb{R}$, il s'agit de trouver une fonction $\mathsf{F}_{\mathsf{X},\mathsf{Y}}(\cdot)$ telle que $\mathsf{F}_{\mathsf{X},\mathsf{Y}}(\mathbf{x}_i) \approx \mathsf{y}_i$ où la qualité l'approximation est mesurée par une fonction de coût. La moyenne de cette fonction de coût sur l'ensemble (X, Y) permet de définir le risque empirique $\mathsf{R}^{\text{emp}}_{\mathsf{X},\mathsf{Y}}[\mathsf{f}]$ qui mesure l'adéquation de $\mathsf{F}_{\mathsf{X},\mathsf{Y}}(\cdot)$ aux données (X, Y). Dans la formulation SVM, la fonction de coût est de type ϵ-insensible, voir figure 3.3 page 31.

La méthodologie standard consiste à minimiser un risque régularisé, tout comme pour la classification à deux classes, et la solution s'écrit

$$\mathsf{F}_{\mathsf{X},\mathsf{Y}}(\mathbf{x}) \;=\; \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)\mathsf{k}\big(\mathbf{x}_i, \mathbf{x}\big) + b \tag{4.21}$$

car le théorème du représentant s'applique ici encore, voir Schölkopf et Smola (2002). Nous nous sommes intéressés au cas où la solution s'écrit comme une combinaison linéaire de noyaux $\mathsf{k}\big(\mathbf{x}_i, \cdot\big)$, comme à l'équation (4.21), mais avec la contrainte supplémentaire que les coefficients sont positifs. Dans Davy et Wolfe (2005), nous développons un algorithme de résolution de ce problème, avec son application à l'approximation parcimonieuse de représentations temps-fréquence de la classe de Cohen. Imposer des coefficients positifs permet de garder une interprétation énergétique à l'approximation. En outre, nous utilisons le fait que certains noyaux temps-fréquence sont défini positifs, fournissant de façon naturelle le noyau à utiliser pour calculer l'approximation.

Chapitre 5

Contributions applicatives

Ce chapitre détaille mes travaux de recherche liés aux applications. Certaines des méthodes présentées sont de simples mises en œuvre des méthodes génériques présentées au chapitre précédent (certaines ont même initié ces travaux méthodologiques). D'autres ont nécessité des adaptations de méthodes génériques existantes, ainsi que de la mise au point spécifique.

La section 5.1 présente les travaux effectués en relation avec la transcription automatique de la musique. Ces méthodes font grandement appel à l'estimation spectrale, sujet de la section 5.2. La section suivante 5.3 aborde plusieurs contributions dans le domaine du traitement de la parole. Enfin, un autre grand domaine d'application a concerné les problèmes à capteurs multiples, qui sont décrits dans la section 5.4.

5.1 Transcription automatique de la musique

La transcription automatique de la musique consiste à résoudre le problème inverse suivant : étant donné un enregistrement musical (typiquement en stéréophonie), retrouver les paramètres musicaux (notes jouées et durée, instrument utilisé, etc.). Le livre édité par Klapuri et Davy (2006) présente les différents aspects de cette problématique.

Deux contributions à ce domaine ont été élaborées : l'estimation de fréquences fondamentales multiples (en ligne et hors ligne), et la détection de changements des paramètres musicaux.

5.1.1 Estimation de fréquences fondamentales multiples

L'estimation de fréquences fondamentales concerne essentiellement la musique dite « tonale », celle dont le spectre de fréquence comporte des raies régulières. Dans ces travaux, nous avons considéré des modélisations du signal en termes de sinusoïdes. Ainsi, pour un signal parfaitement périodique, de longueur infinie, un modèle naïf de signal musical x s'écrit

$$x(n) \;=\; \sum_{m=1}^{M} \alpha^{\mathrm{s}} \sin(2\pi m k_1 n) + \alpha^{\mathrm{c}} \cos(2\pi m k_1 n) \tag{5.1}$$

où $n = 1, 2, \ldots$ est la variable temporelle discrète. La composante de fréquence k_1 est appelée *le fondamental* et les composantes telles que $m = 2, \ldots, M$ sont *les harmoniques*. Le modèle (5.1) n'est, en fait, pas réaliste. Les signaux musicaux réels individuels ne peuvent pas être entièrement modélisés par des sinusoïdes : par exemple, la respiration d'un joueur de flûte

échappe à cette modélisation, bien qu'elle soit audible dans les enregistrements, voir Goodwin (1996). Comme ces composantes du signal sont hautement variables d'un enregistrement à l'autre, elles sont modélisées statistiquement comme un bruit ϵ, d'où le modèle

$$x(n) \;=\; \sum_{m=1}^{M} \alpha^{\mathrm{s}} \sin(2\pi m k_1 n) + \alpha^{\mathrm{c}} \cos(2\pi m k_1 n) + \epsilon(n) \tag{5.2}$$

Le modèle (5.2) reste trop simple, pour plusieurs raisons : d'abord, un enregistrement musical comporte souvent plusieurs notes jouées simultanément. Ensuite, les amplitudes n'ont aucune raisons d'être constantes au cours du temps. Enfin, les fréquences des harmoniques sont rarement des multiples entiers exacts de la fréquence du fondamental. La figure 5.1 représente un signal de flûte jouant une seule note, où l'on voit clairement le phénomène de non-stationnarité des amplitudes. Dans la suite, nous considérons deux types de modèles, séquentiel et non-séquentiel.

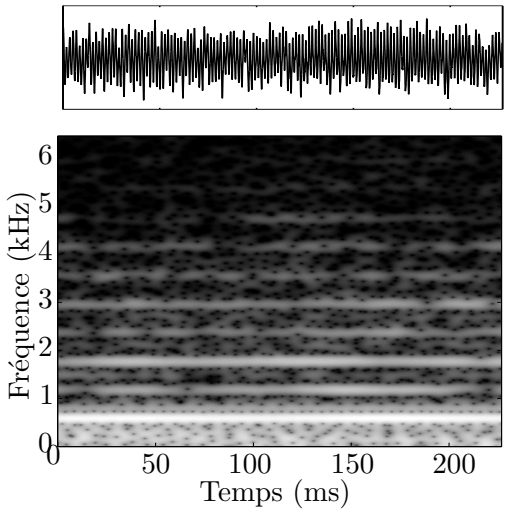


FIG. 5.1 – Signal d'un enregistrement de flûte jouant une seule note, dans le domaine temporel (en haut) et son spectrogramme (en bas).

Modèle non séquentiel

Nous supposons ici que le signal musical a été segmenté de façon à ce que la portion de signal $x(n)$ ne comporte pas de changement de note. Cela peut être effectué par l'algorithme de Desobry *et al.* (2005a), comme indiqué ci-dessous.

Dans Davy *et al.* (2006c), nous avons proposé qu'un signal comportant plusieurs notes soit modélisé par

$$x(n) \;=\; \sum_{j=1}^{J} \sum_{m=1}^{M_j} \alpha^{\mathrm{s}}_{j,m}(n) \sin\big(2\pi k_{j,m}(n) n\big) + \alpha^{\mathrm{c}}_{j,m}(n) \cos\big(2\pi k_{j,m}(n) n\big) + \epsilon(n) \tag{5.3}$$

où J est le nombre total de notes, M_j est le nombre d'harmoniques de la note numéro j ($j = 1, \ldots, J$), $\alpha^{\mathrm{s}}_{j,m}(n)$ et $\alpha^{\mathrm{c}}_{j,m}(n)$ sont les amplitudes (non-stationnaires) de l'harmonique m

de la note j, avec

$$\alpha^{\mathrm{s}}(n) \;=\; \sum_{I=0}^{I} \widetilde{\alpha}_i^{\mathrm{s}} \mathrm{w}[n - i\Delta_n] \;\text{ et }\; \alpha^{\mathrm{c}}(n) \;=\; \sum_{I=0}^{I} \widetilde{\alpha}_i^{\mathrm{c}} \mathrm{w}[n - i\Delta_n] \tag{5.4}$$

où $\widetilde{\alpha}_i^{\mathrm{s}}$, $\widetilde{\alpha}_i^{\mathrm{c}}$ sont les amplitudes associées à chaque fonction de base $\mathrm{w}[n - i\Delta_n]$, de pas Δ_n. La forme de w est choisie de telle sorte que l'amplitude soit lisse au cours du temps (gaussienne, de Hamming, de Hanning, etc.). Les fréquences des harmoniques, données par $k_{j,m}(n)$, peuvent être choisies constantes avec

$$k_{j,m}(n) \;=\; m\, k_{j,1}\, \upsilon_{j,m} \tag{5.5}$$

où $1 + \upsilon_{j,m}$ est le facteur d'inharmonicité. Ce dernier peut être modélisé comme une variable aléatoire gaussienne de moyenne 1, ou par une formule déterministe liée à l'instrument enregistré.

On remarquera que le modèle résultant est lié aux représentations de Gabor (voir par exemple Feichtinger et Strohmer (1998); Gröchenig (2001)), dans lesquelles le signal est projeté sur des sinusoïdes/cosinusoïdes fenêtrées dont les positions en temps et fréquence sont fixées. Ici, la grille temps-fréquence est régulière en temps, et irrégulière en fréquence. Ce type de représentation a été utilisé dans ce contexte dans, par exemple, Davy et Godsill (2002a); Wolfe *et al.* (2004).

Les paramètres du modèle de l'équation (5.4) – s'écrivant aussi sous la forme (4.8) – sont estimés en utilisant l'algorithme MCMC à doubles sauts réversibles évoqué dans le paragraphe 4.1.2, et dans l'article Davy *et al.* (2006c). Les performances de cette méthode d'estimation sont au niveau de l'état de l'art, allant de 100% de bonne estimation dans le cas où $J = 1$ à 80% pour $J = 4$. Les figures 5.2 et 5.3 illustrent les capacité de modélisation de notre modélisation.

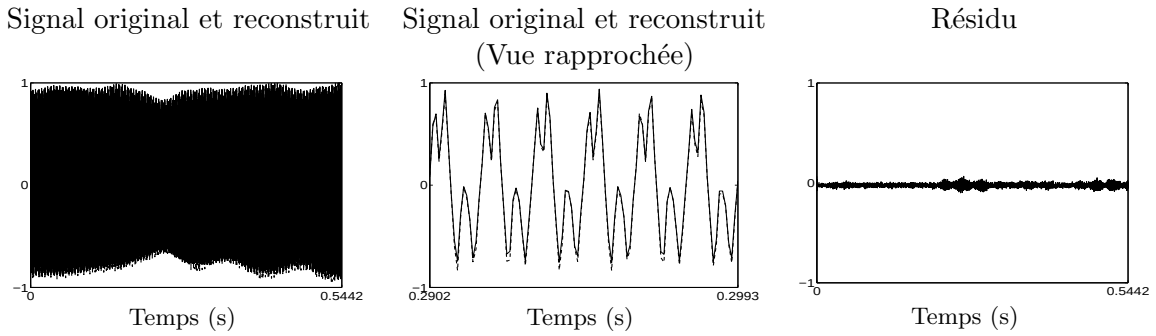


FIG. 5.2 – Exemple d'utilisation du modèle harmonique bayésien. A gauche, un morceau de musique harmonique (violon seul, en trait plein) et le signal reconstruit à partir du modèle harmonique de l'équation (4.8) dont les paramètres ont été estimés avec l'algorithme MCMC (en pointillés). Au centre, la même chose en vue rapprochée. On voit l'excellente précision du modèle pour la reconstruction de la musique harmonique. A droite, le résidu.

Modèle séquentiel

Il est possible de modifier le modèle sinusoïdal (5.2) afin de le mettre sous forme séquentiel. L'approche que nous avons suivie dans Dubois et Davy (2005c); Dubois *et al.* (2005); Dubois

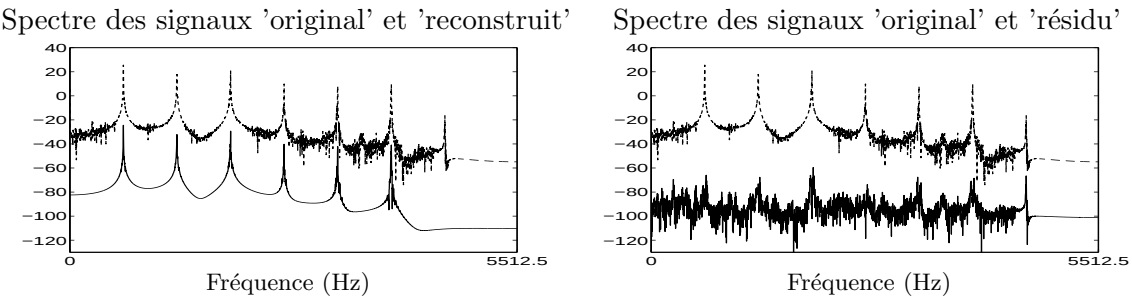


FIG. 5.3 – Résultat d'estimation avec le modèle harmonique bayésien pour le signal de la figure 5.2 présentés dans le domaine fréquentiel (avec un décalage artificiel de -50dB pour une meilleur lisibilité). En pointillés : le spectre du signal original.

et Davy (2005a,b) consiste à définir un modèle de Markov caché, sous la forme suivante

$$
\begin{aligned}
J_n &\sim \mathsf{P}(J_n|J_{n-1}) && (5.6)\\
\mathbf{k}_n &\sim \mathsf{p}(\mathbf{k}_n|\mathbf{k}_{n-1}, J_{n-1:n}) && (5.7)\\
\boldsymbol{\alpha}_n &\sim \mathsf{p}(\boldsymbol{\alpha}_n|\boldsymbol{\alpha}_{n-1}, \mathbf{k}_{n-1:n}, J_{n-1:n}) && (5.8)\\
\mathbf{y}_n &\sim \mathsf{p}(\mathbf{y}_n|\boldsymbol{\alpha}_n, \mathbf{k}_n, J_n) && (5.9)
\end{aligned}
$$

où le vecteur d'observation est une trame du signal original

$$\mathbf{y}_n[i] \;=\; \mathrm{w}[i - n\Delta_n]x(i) \qquad (5.10)$$

qui suit le modèle statistique de sinusoïdes noyées dans un bruit gaussien (5.3). Les vecteurs $\boldsymbol{\alpha}_n$ et $\mathbf{k}_n$ contiennent respectivement les amplitudes et les fréquences des harmoniques des J_n notes pour la trame située à l'instant n. Ces équations définissent un modèle de Markov à sauts. En supposant que la loi d'évolution (5.8) est linéaire et gaussienne, le modèle est linéaire gaussien conditionnellement au nombre de notes et à leurs fréquences. Cela permet de mettre en œuvre un filtre particulaire « rao-blackwellisé », c'est-à-dire un banc de filtres de Kalman en interaction, voir Dubois et Davy (2005b) pour les détails. Les paramètres « non-linéaires » sont estimés par échantillonnage d'importance séquentiel, comme expliqué au chapitre précédent. La figure 5.4, extraite de Dubois *et al.* (2005), montre un résultat d'estimation obtenu par ce filtre. Ce travail est toujours en développement, et une version plus performante de l'algorithme est en cours de mise au point.

5.1.2 Détection de changements des paramètres musicaux

Parmi les tâches à accomplir en vue de transcrire un enregistrement musical, l'une des plus fondamentales concerne les paramètres rythmiques. Les algorithmes d'estimation existants reposent tous sur des détecteurs de changement des paramètres musicaux (changements fréquentiels, brusques hausses/baisses d'amplitude, etc.).

Nous avons appliqué l'algorithme présenté à la section 4.2.2 (page 48), à la détection de changements des paramètres musicaux. Pour cela, nous avons considéré les descripteurs acoustiques suivants

$$\mathbf{x}_n \;=\; \mathsf{SP}_x^{\mathrm{w}}(n-1:n+1, 1:k_{\max}) \qquad (5.11)$$

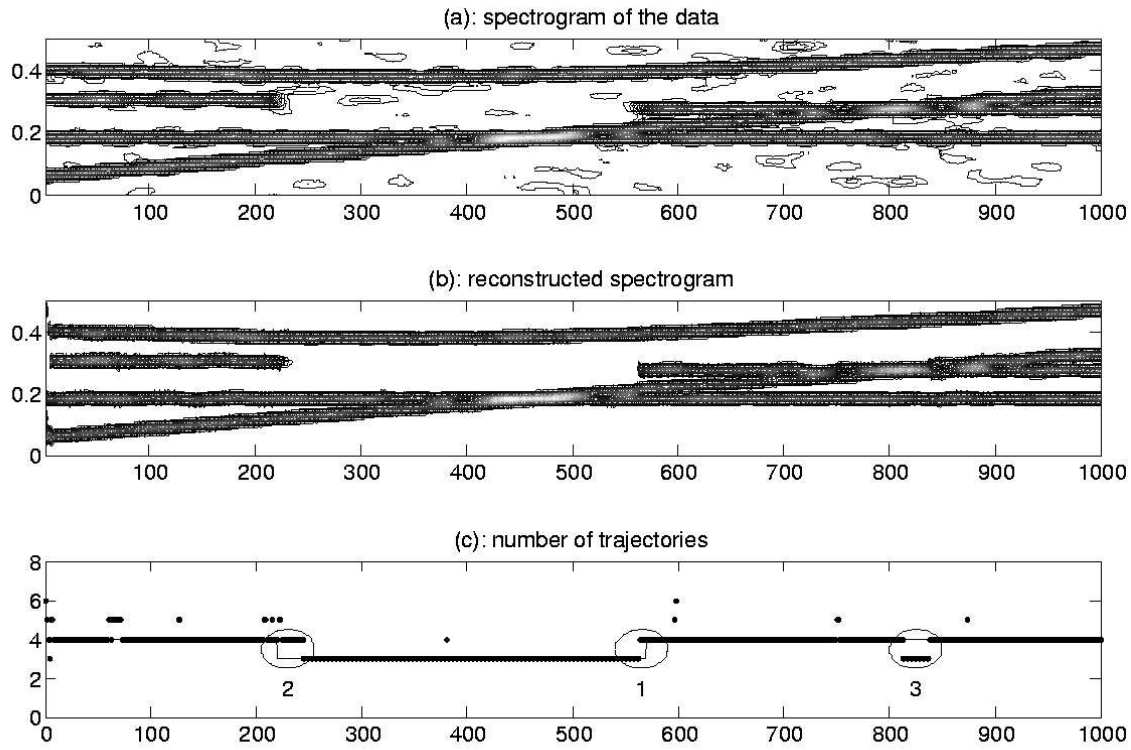


FIG. 5.4 – Résultat d'estimation de trajectoires fréquentielles par filtrage particulaire dans le spectrogramme des données. En haut : spectrogramme original. En bas, spectrogramme reconstruit à l'aide des paramètres estimés. Dans ce cas, le nombre d'harmoniques avait été fixé à 1 (trajectoires individuelles) – extrait de Dubois *et al.* (2005) .

c'est-à-dire trois colonnes (par exemple) du spectrogramme discret du signal à segmenter. Ainsi, l'espace $\mathcal{X}$ sur lequel le noyau utilisé pour la détection des ruptures est défini comme étant typiquement de dimension $3 \times 256 = 768$. Dans Desobry *et al.* (2005a), nous avons segmenté avec succès un enregistrement d'orgue d'église, réputé difficile.

5.2 Estimation spectrale non-stationnaire

Les méthodes présentées ci-dessus font toutes appel à une description du contenu fréquentiel des signaux au cours du temps. Dans tous ces cas, l'élément essentiel est une « bonne » représentation temps-fréquence. Nous avons plusieurs contributions dans ce domaine. Tout d'abord, une méthode bayésienne d'estimation spectrale non-stationnaire à haute résolution (Modèle autorégressif à coefficients variables). Ensuite, une nouvelle méthode de calcul des représentations temps-fréquence à coefficients positifs introduites par Cohen et Posh (1985). Enfin, des représentations temps-fréquence parcimonieuses.

5.2.1 Modèles autorégressif à coefficients variables

Considérons un signal x_n. La modélisation autorégressive à coefficients variables (*Time-Varying AutoRegressive - TVAR)* s'écrit :

$$x_n = \alpha_{1,n} x_{n-1} + \alpha_{2,n} x_{n-2} + \ldots + \alpha_{p_n,n} x_{n-p_n} + \epsilon_n, \tag{5.12}$$

où v_t est l'erreur de modélisation (gaussienne, blanche et centrée) et les coefficients $a_{i,t}$ sont les coefficients AR en nombre K_t indéterminé *a priori*. Pour des raisons de robustesse, on préfère la paramétrisation en termes de pôles instantanés, qui sont les racines du polynôme

$$\chi_n(u) = 1 - \alpha_{1,n} u - \alpha_{2,n} u^2 - \ldots - \alpha_{p_n,n} u^{p_n}. \tag{5.13}$$

Les pôles peuvent être complexes conjugués ou réels. Dans notre approche, décrite dans Andrieu *et al.* (2001a, 2002a,b, 2003), la modélisation porte sur les trajectoires des pôles réels, des fréquences et modules des pôles complexes. En outre, le nombre de paires de pôles complexes conjugués et le nombre de pôles réels sont supposés évoluer selon une chaîne de Markov. On le voit, une telle modélisation est plus physique qu'une modélisation portant sur les trajectoires des coefficients AR $\alpha_{i,n}$ qui sont difficilement interprétables physiquement, et qui ne permettent pas l'augmentation/la diminution de nombre de pôles sans perturber les fréquences. Avec notre modèle, il est possible d'estimer les coefficients AR, ainsi que le nombre de pôles (par MAP marginal). Finalement, les hyperparamètres (variance du bruit d'excitation et variances des bruits d'évolution des modules et arguments des pôles) sont supposés évoluer selon une marche aléatoire log-gaussienne.

La mise en œuvre repose sur le filtre particulaire à sauts markoviens (algorithme 4.5) spécialement développé, qui autorise des sauts markoviens (ajout ou retrait de pôles réel ou de paires de pôles complexes conjugués). On se reportera à Andrieu *et al.* (2003) pour un aperçu des performances de la méthode[1].

5.2.2 Méthode d'estimation spectrale bayésienne

Dans une étude Davy *et al.* (2003) (toujours en cours de développement), nous avons repris le modèle non séquentiel de sinusoïdes+bruit de Andrieu et Doucet (1999) pour l'estimation spectrale stationnaire bayésienne. Il s'agit de la mise en œuvre de l'algorithme de Monte Carlo séquentiel générique de Del Moral *et al.* (2006). Le problème est traité sous deux angles : 1) l'échantillonnage de variables aléatoires suivant la distribution à posteriori et 2) l'estimation MAP vue comme une optimisation de type « recuit simulé » par particules, rappelant les algorithmes génétiques, mais dont les résultats de convergence sont parfaitement maîtrisés.

5.2.3 Représentations temps-fréquence positives

Dans la lettre Davy et Doucet (2003), nous avons proposé de réinterpréter les représentations temps-fréquence positives, à marginales correctes (dites de « Cohen-Posh »). Proposées par Cohen et Posh (1985), ces représentations sont basées sur un résultat théorique de 1980 par Cohen et Zaparovanny. Dans Davy et Doucet (2003), nous avons montré que ce résultat est en fait un cas particulier d'un théorème dû à Sklar en 1959, qui stipule que toute distribution de probabilité de deux variables admet un copula unique. Ainsi, il est possible de définir

[1] La méthode a été programmée en langage C, et le code peut être téléchargé à l'adresse Internet `http://www-lagis.univ-lille1.fr/~davy/code/Davy_TVAR_code.tar.gz`.

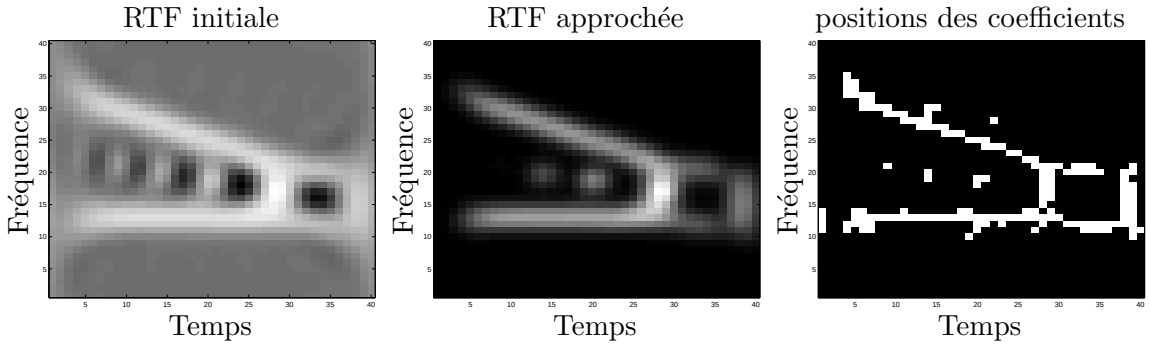


FIG. 5.5 – A gauche : représentation temps-fréquence (RTF) de deux chirps calculée pour un noyau de spectrogramme gaussien. Au centre : RTF approchée par la méthode de régression à coefficients positifs. A droite : position des coefficients non nuls utilisés pour calculer l'approximation. Ces derniers peuvent être interprétés comme une RTF « relocalisée ».

n'importe quelle représentation temps-fréquence positive qui admet des marginales en temps et en fréquence données. La structure de corrélation temps-fréquence est entièrement capturée dans le copula, ce qui nous a permis de proposer une nouvelle méthode de calcul. Cette méthode consiste à calculer le copula du spectrogramme, et l'utiliser pour construire une représentation positive sur la base des « vraies marginales ».

5.2.4 Représentations temps-fréquence parcimonieuses

Des représentations temps-fréquence de la classe de Cohen, calculées pour certains noyaux peuvent être vues comme des fonctions d'un espace de Hilbert à noyau reproduisant sur $\mathbb{R}^2$. Par exemple, le spectrogramme calculé avec une fenêtre gaussienne appartient à l'espace de noyau $\mathsf{k}\big((t_1, f_1), (t_2, f_2)\big)$ gaussien (non circulaire dans le cas général). Ainsi, il nous a semblé naturel de rechercher une approximation parcimonieuse de cette représentation dans la même classe de fonctions, c'est-à-dire le RKHS de noyau $\mathsf{k}\big((t_1, f_1), (t_2, f_2)\big)$. Par ailleurs, par la formulation présentée au paragraphe 4.2.3, nous avons recherché cette approximation sous la somme d'une combinaison linéaire parcimonieuse, à coefficients positifs. Dans Davy et Wolfe (2005), nous avons proposé cette méthode et obtenu les résultats présentés à la figure 5.5.

5.2.5 Détection de défauts d'enceintes acoustiques

Une étude débutée pendant ma thèse, et publiée dans Davy et Doncarli (2002), concerne la détection de défauts d'enceintes acoustiques. Poursuivie dans le cadre d'un contrat entre le CNRS et Neutrik Test Instruments (NTI), basée au Liechtenstein, cette étude utilise des SVM à une classe sur des images temps-fréquence de la classe de Cohen. Le choix de la représentation temps-fréquence des données est cruciale, et le noyau doit être adapté aux données, suivant la méthodologie de Gretton *et al.* (2001).

5.3 Traitement de la parole

Les méthodes d'estimation spectrale évoquées ci-dessus sont bien adaptées au signaux sonores. Parmi ceux-ci, les signaux de parole ont fait l'objet de travaux spécifiques, que nous traitons dans cette section.

L'indexation de documents sonores comporte un certain nombre de sous problèmes. Dans le cas de l'indexation de signaux de parole, l'une des tâches les plus évidentes est celle de la reconnaissance du texte prononcé, sur lequel nous avons travaillé récemment (voir paragraphe 5.3.2 ci-dessous). Une autre tâche importante est celle de l'indexation des locuteurs, qui nécessite une segmentation des signaux audio en plages dans lesquelles un locuteur unique intervient. Les travaux dans cette veine sont décrits dans le paragraphe 5.3.1 ci-dessous. [Notons que l'étude sur la vérification de locuteurs effectuée à la fin de ma thèse et publiée dans Davy (2005a) n'est pas présentée ici.]

5.3.1 Segmentation en locuteurs

Segmenter un signal audio nécessite de déterminer les bornes de segments ayant une certaine cohérence acoustique. Dans le cas de la segmentation en locuteurs, la cohérence correspond à l'identité du locuteur. La segmentation consiste à déterminer des plages temporelles comportant un locuteur unique. Puis, les segments sont regroupés de façon à ce que chaque groupe de segments corresponde à un locuteur.

Dans le cadre de la thèse de Belkacem Fergani, nous avons utilisé la mesure de contraste définie dans Desobry *et al.* (2005a) sur des vecteurs de descripteurs audio, comportant des coefficients cepstraux, leurs dérivées premières et deuxième. Cette mesure de contraste est utilisée à la fois pour la segmentation et le regroupement des segments. Les premiers résultats sont très encourageants. Dans Fergani *et al.* (2006a,c), nous obtenons des résultats comparables – voire supérieurs – aux meilleures techniques, bien que l'approche soit totalement différente.

5.3.2 Nouveau descripteurs pour la reconnaissance

Dans le cadre d'une étude menée en collaboration avec France Telecom R&D, une approche originale pour la reconnaissance de la parole est en cours de développement. En tant que chercheur post-doctorant sous ma responsabilité, Stéphane Rossignol étudie la possibilité de remplacer les descripteurs basés sur des coefficients cepstraux par de nouveaux descripteurs. Ces derniers se fondent sur la recherche de fonctions propres non-linéaires dans un certain espace, et sur l'utilisation de SVM à une classe. Ici encore, les premiers résultats sont au niveau des techniques de référence, bien que l'approche suivie soit, ici encore, très différente.

5.4 Problèmes multicapteurs

Dans les sections précédentes, des travaux relatifs aux signaux audio ont été présentés. Une part plus récente de mes activités concerne des applications très différente, celle de la gestion des systèmes multicapteurs. Plusieurs travaux portent sur le développement et l'utilisation de filtres particulaires adaptés à ce contexte, d'autres sur la détection de ruptures à l'aide de méthodes à noyaux.

5.4.1 Détection de capteurs défaillants

La fusion des informations délivrées par plusieurs capteurs peut être réalisée assez facilement dans le cadre bayésien. En effet, il est souvent très naturel de définir un modèle de

Markov caché du type

$$\begin{aligned} \boldsymbol{\theta}_n &= f_n\big[\boldsymbol{\theta}_{n-1}\big] + \mathbf{v}_n && (5.14)\\ \mathbf{x}_{1,n} &= g_{1,n}\big[\boldsymbol{\theta}_n\big] + \boldsymbol{\epsilon}_{1,n} && (5.15)\\ &\vdots && (5.16)\\ \mathbf{x}_{K,n} &= g_{K,n}\big[\boldsymbol{\theta}_n\big] + \boldsymbol{\epsilon}_{K,n} && (5.17) \end{aligned}$$

où il y a K capteurs. Dans les applications, les capteurs sont des dispositifs physiques relativement fiables, mais sujets au vieillissement, à l'usure et finalement à la défaillance. Par ailleurs, les conditions environnementales de ces capteurs peuvent les perturber (température ou vibrations excessives, ou, pour un capteur optique, présence de fumée, etc.). L'estimation par filtrage dans ce cadre accorde en général une forte confiance au capteur, et incorporer des erreurs de mesure dans l'estimation peut avoir des conséquences dramatiques dans certains cas. Des approches par rejet de mesure ont été développées et consistent à rejeter les mesures qui sont trop en désaccord avec la prédiction, capteur par capteur.

Dans Caron *et al.* (2005) et Caron *et al.* (2006c), nous avons développé une méthodologie différente. Chaque capteur $k = 1, \ldots, K$ se voit assigner une variable discrète $r_{k,n}$ telle que $r_{k,n} = 0$ correspond à la défaillance du capteur, et $r_{k,n} = 1$ correspond à un fonctionnement normal du capteur. Il est possible de définir d'autres valeurs pour $r_{k,n}$, pour d'autres conditions (par exemple, absence/présence de multitrajets pour les capteurs GPS). Alors, les équations d'observation s'écrivent en fonction de $r_{k,n}$.

Afin d'introduire une notion de mémoire de la fiabilité du capteur, une variable vectorielle $\boldsymbol{\alpha}_{k,n}$ indiquant la probabilité de chaque état du capteur (0 ou 1), avec une transition markovienne $p(\boldsymbol{\alpha}_{k,n}|\boldsymbol{\alpha}_{k,n-1})$.

Dans des applications sur données réelles, nous avons montré que ce modèle, et l'algorithme particulaire correspondant, ont des performances d'estimation supérieures aux modèles avec rejet des erreurs. Notamment, notre modèle est capable de « sortir » d'une configuration où le capteur jugé défaillant est en fait non défaillant, tandis que les approches existantes s'enferrent dans l'erreur, voir Caron *et al.* (2005, 2006c).

5.4.2 Détection de mines antipersonnel

La détection des mines antipersonnel se fait aujourd'hui par des opérateurs humains. Développer des appareils autonomes de détection est un enjeu évident, mais difficile. Les prototypes actuels sont équipés de radars à pénétration de sol (GPR), qui scannent le sol et fournissent des données sous la forme d'un signal temporel (amplitude réfléchie en fonction du temps) pour chaque position sur le sol. Les données sont donc de grande dimension, et un algorithme adapté de détection des objets enfouis doit être développé.

Le travail décrit dans Potin *et al.* (2006) combine l'approche de détection de ruptures par SVM développée dans Desobry *et al.* (2005a) avec une méthode de filtrage des images radar du sol (développée par D. Potin, E. Duflos et Ph. Vanheeghe par ailleurs). Les résultats de détection ont été de bonne qualité.

5.5 Bilan

Les applications traitées sont essentiellement liées aux signaux audio. Cependant, les outils nécessaires à ces applications ont été développés dans un soucis de généralité. C'est pourquoi

certains d'entre eux ont pu être appliqués à des domaines n'ayant, *a priori*, rien à voir les uns avec les autres. C'est particulièrement vrai pour le filtre particulaire dédié aux modèles de Markov à sauts (appliqué aux modèles TVAR, à l'estimation de fréquences fondamentales multiples) et l'algorithme de détection de ruptures appliquée à la segmentation musicale, la segmentation en locuteurs et la détection de mines antipersonnel.

Chapitre 6

Synthèse et perspectives

Ce chapitre propose une synthèse articulée autour de la notion de régularisation, qui apparaît *a posteriori* comme un thème central dans mes travaux (section 6.1). C'est en effet dans ce cadre que se situent les approches développées, tant du point de vue de la modélisation mathématique (statistique et fonctionnelle), que du point de vue algorithmique. La suite de ce chapitre est dédiée à quelques perspectives, rassemblées dans la section 6.2.

6.1 Synthèse

La plupart des problèmes présentés dans les chapitres précédents sont en fait des *problèmes inverses* (estimation, classification, détection, etc.), qualifiés de *mal posés*. En tant que tels, ils admettent un grand nombre de solutions, la majorité d'entre elles étant inintéressantes au sens où elles ne répondent pas au problème posé : par exemple, elle ont des performances médiocres face à des données nouvelles. La régularisation permet de faire le tri parmi ces solutions, et d'en sélectionner une – ou plusieurs – qui peuvent être jugées satisfaisantes(s) par l'utilisateur.

6.1.1 Régularisation

La régularisation est l'approche clé dans mes travaux. J'en ai présenté une forme spécifique au chapitre 3. De façon plus générale, elle peut être énoncée comme suit. Tout d'abord, nous disposons de données à analyser. Les problèmes de classification, de détection, d'estimation, de représentation ou de séparation se ramènent à la recherche d'une fonction dans un certain espace, adaptée aux données. Par exemple, pour la classification supervisée, on recherche la séparatrice. Pour l'estimation paramétrique – par exemple, des fréquences de sinusoïdes bruitées, on recherche le couple (modèle, paramètre) qui explique au mieux les données, et c'est une fonction. Dans notre exemple, c'est une fonction du temps s'écrivant comme une somme de sinusoïdes de différentes fréquences. La régularisation se base donc sur les éléments suivants :

- Les **données**, notées X, prises dans leur sens le plus large (c'est-à-dire, qui inclut aussi d'éventuelles étiquettes) ;
- Un **espace de fonctions** où la solution est recherchée, noté $\mathcal{F}$. Cet espace peut être défini par des considérations d'analyse fonctionnelle (par exemple, un RKHS) ou de façon paramétrique (par exemple, l'ensemble de toutes les fonctions temporelles qui

s'écrivent comme une somme de sinusoïdes) ;
- Un **terme régularisant** noté $\Omega(\mathsf{f})$ pour $\mathsf{f} \in \mathcal{F}$, qui pénalise les fonctions trop complexes ou inadéquates dans l'espace de fonctions choisi. Il peut s'agir d'une norme fonctionnelle (par exemple, la norme canonique du RKHS), ou d'un *a priori* sur la dimension du modèle paramétrique. Il peut également s'agir d'un critère de type *minimum description length* ou *Akaike information criterion*.
- Une mesure de la **déviation du modèle par rapport aux données**, notée $\mathbf{M}(\mathsf{X}, \mathsf{f})$. Dans le contexte des méthodes à noyaux, il s'agit du risque empirique, alors que dans le paradigme bayésien, il s'agit de l'opposé du logarithme de la vraisemblance.

Dans les cas traités dans ce manuscrit, la recherche de la solution peut s'exprimer sous la forme

$$\widehat{\mathsf{f}} \;=\; \underset{\mathsf{f}\in\mathcal{F}}{\operatorname{argmin}} \;\; \mathbf{M}(\mathsf{X}, \mathsf{f}) \;+ \lambda\Omega(\mathsf{f}) \tag{6.1}$$

où λ règle le niveau de régularisation. Pour les méthodes à noyaux, la correspondance est évidente. Pour le formalisme bayésien, prendre l'opposé du logarithme de la distribution *a posteriori* conduit à une expression du type (6.1).

Régularisation, SVM et inférence Bayésienne.

La régularisation apparaît de façon explicite dans les méthodes de type « à noyaux », où la fonction solution minimise un risque régularisé. Ici, le terme régularisant n'est autre que la norme de la fonction dans l'espace à noyau. De plus, les travaux de Vapnik établissent un lien direct entre la norme de la fonction et sa complexité au sens de la dimension de Vapnik-Chervonenkis Vapnik (1995). Concernant les méthodes bayésiennes, le régularisateur apparaît également de façon explicite, sous la forme des distributions *a priori*. De nombreux liens ont d'ailleurs été établis formellement entre les méthodes à noyaux et l'inférence bayésienne. Enfin, les approches « fréquentistes » font souvent appel à des vraisemblances pénalisées, qui admettent également une interprétation en terme de régularisation, voir Green (1999).

Notons un dernier exemple de régularisation qui concerne les représentations temps-fréquence : on peut voir la représentation de Wigner-Ville comme « non régularisée », alors que le spectrogramme est régularisé par l'application d'un noyau de lissage.

Nécessaire régularisation.

Il n'est pas possible d'extraire de l'information de données, sans *a priori* sur ce que l'on cherche. Un ensemble de données sans aucune information associée est vide de sens. La connaissance des unités de mesure, des conditions expérimentales, et de ce qui est recherché rend possible leur traitement. Et, obtenir la solution recherchée nécessite de mettre en œuvre une approche régularisée. En effet, il est facile de se convaincre qu'*une infinité de modèles mathématiques expliquent parfaitement un nombre fini de données*. On peut construire un tel ensemble infini en considérant d'abord une fonction qui minimise le terme d'adéquation aux données $M(\mathsf{X}, \mathsf{f})$, puis en la modifiant aux endroits où l'on n'observe pas de données (cela n'a aucune influence sur $M(\mathsf{X}, \mathsf{f})$). Si l'on observe des données partout, c'est que l'espace des données est fini. Dans ce cas, il n'y a aucune fonction à apprendre. Il est donc nécessaire d'en éliminer, en régularisant. Il apparaît donc que l'idée d'une méthode de Traitement du Signal qui « ne fait aucune hypothèse » ou encore, qui « ne nécessite aucun réglage de paramètre » est

vide de sens. Toute méthode *doit* comporter une part d'information extérieure aux données – qui est parfois implicite et donne l'illusion d'absence de modèles ou de paramètres de réglage.

Toutefois, toutes les informations *a priori* ne sont pas également légitimes. Ainsi, celles qui s'inscrivent dans une théorie scientifique (physique, biologique, chimique, etc.) sont certainement plus légitimes que d'autres, plus arbitraires. Et, confronter des modèles construits sur la base de plusieurs types d'*a priori* permet de raffiner la connaissance scientifique.

Pratique de la régularisation.

Face à un problème de Traitement du Signal, la difficulté consiste à choisir le formalisme dans lequel on recherche la solution. Il s'agit donc – implicitement ou explicitement – de choisir un cadre mathématique, c'est-à-dire un espace de fonctions, la forme des termes d'adéquation aux données et de régularisation. On optera par exemple pour une méthode paramétrique bayésienne, ou pour une méthode à noyau, non paramétrique. Dans certains cas, les choix effectués, ainsi que les motivations qui y prévalent, sont difficilement explicitables. Cependant, le problème reste le même : sans sélection *a priori* de la forme de la solution, il est impossible d'aborder un problème de Traitement du Signal.

La pratique montre que le choix de l'espace de fonction, du terme d'adéquation aux données et du terme de régularisation résultent du niveau de connaissances *a priori*, mais aussi de la simplicité de résolution de l'optimisation à l'équation (6.1). Par exemple, les SVM proposent un compromis très attractif entre généralité de la méthode et simplicité de mis en œuvre. Les méthodes bayésienne paramétriques permettent « d'encoder » des informations *a priori* très précises, parfois au prix de calculs numériques complexes. Pour un problème donné, on devra donc toujours déterminer le compromis le plus avantageux entre précision et simplicité.

6.1.2 Point de vue élargi

Déterminer le compromis le plus avantageux entre précision et simplicité. Ne s'agit-il pas la encore de régularisation, où le terme d'adéquation aux données est la « précision », et le terme de régularisation, la « simplicité » ? Ici, l'espace de fonctions $\mathcal{F}$ n'est autre que la panoplie complète des solutions de Traitement du Signal disponibles dans la littérature...

D'une manière plus générale, on peut étendre ce raisonnement aux théories scientifiques : étant donné qu'une infinité de théories sont capables d'expliquer parfaitement les observations collectées par l'humanité depuis ses origines, on choisira la (les) plus « simple(s) ». S'il est relativement facile de vérifier qu'une théorie explique bien les observations considérées, définir sa simplicité est plus délicat. En effet, peut-on dire de la physique quantique qu'elle est simple ? Il me semble qu'il s'agit ici plus d'une considération que je qualifierais d'« esthétique ». Pour moi, un formalisme théorique est esthétique s'il montre une forme de cohérence avec l'esprit humain et admet une formulation – par exemple, mathématique – claire. Lorsqu'une théorie nécessite des « surcouches » de plus en plus complexes, pour rester cohérente avec les observations, il me semble inéluctable qu'un nouveau formalisme plus esthétique vienne à être proposé. Ainsi, la vision copernicienne, simple et belle, du système solaire a-t-elle permis de débarrasser le modèle ptoléméïen de termes correctifs permettant d'expliquer les mouvements rétrogrades apparents de certaines planètes. De même, la théorie de la relativité générale explique mieux que la théorie newtonienne les « anomalies » de l'orbite de Mercure, tout en ayant une formalisation mathématique directe. Ce formalisme ne saurait être jugé simple

par tous, mais l'idée de déformation de l'espace-temps me paraît avoir un aspect esthétique naturel.

En conclusion, et au vu de la notion de régularisation, on comprend mieux que parmi les théories qui expliquent bien les observations, les théories « esthétiques » ont aussi la capacité de *prédire ce qui n'a pas été observé* avec grande précision. Il semble donc qu'il y ait de l'absolu dans cette idée d'esthétique. Faut-il s'en étonner ? L'esprit humain est le fruit du monde qu'il entend étudier, il ne saurait échapper à ses lois.

Aparté. Au cours de l'histoire, et d'une région du monde à l'autre, les théories scientifiques ont pu prendre des formes très différentes. Lorsqu'on les voit comme un compromis entre observations collectées et précision attendue de leurs prédictions (leur capacité à « bien » expliquer le monde, pour une certaine définition de ce « bien » là), ces théories ont toutes leur légitimité. Ce qui ne veut pas dire qu'elles sont toutes acceptables, ou toutes justes. Elles répondent simplement à des compromis différents entre précision et simplicité. Il y a cependant lieu de dénoncer les théories qui se situent hors de ce compromis (ni simples, ni précises, ou trop simples pour expliquer correctement les observations actuelles). Ainsi, la célèbre phrase de Rabelais « science sans conscience n'est que ruine de l'âme » peut-elle être comprise comme un plaidoyer pour la régularisation, où « science » qualifierait la capacité explicative d'une théorie, alors que « conscience » concernerait sa nécessaire esthétique[1] ... (?)

Vérité et modèle. Ces éléments de réflexions conduisent à l'idée que, lorsqu'on analyse des données, on ne découvre pas de *vérité*. Les publications régulières de statistiques économiques, et les débats que cela entraîne dans la société, montrent que l'on peut effectivement « faire dire aux données ce que l'on souhaite ». Ce que l'on découvre, en revanche, résulte du compromis choisi entre lecture des données et connaissances *a priori*. Pour apprendre, il faut connaître – et ce que l'on apprend est « corrompu » par notre connaissance préalable. Aussi, un résultat d'analyse de données, présenté sans la méthode suivie, n'a aucun sens.

6.1.3 Retour aux signaux.

Pour le traitement des signaux, la notion d'esthétique est centrale. En effet, un couple (modèle, algorithme) n'est satisfaisante que s'il fournit des réponses pertinentes au problème posé (en termes d'erreur RMS, de courbes ROC, etc.), mais aussi s'il est élégant des points de vue mathématique et algorithmique. Cet aspect est aussi une garantie de la capacité d'une approche à être généralisée à d'autres contextes, et à bien résister à des situations légèrement différentes de la normale (robustesse). Une solution pragmatique avec des « boîtes », des « flèches » et des « boutons de réglage » est souvent fort efficace dans un contexte donné, mais extrêmement imprécise hors de ce contexte. En revanche, un couple modèle-algorithme précis et élégant sera généralisable. C'est dans cet esprit que j'ai tenté de contribuer aux développements présentés dans ce manuscrit.

6.2 Perspectives

Les suites de ces travaux se situent sur trois échelles de temps. A court terme, tout d'abord, il s'agit de prolonger les travaux présentés dans ce document. Plusieurs domaines sont

[1] L'auteur admet que cette interprétation, si elle est défendable, est certainement partielle...

l'objet d'investigations, et des publications sont en préparation, voir le paragraphe 6.2.1 ci-dessous. A moyen terme (voir paragraphe 6.2.2), la mise en place de l'équipe INRIA-FUTURS « SequeL » orientera une part importante de mon activité, autour du thème de l'apprentissage séquentiel, de l'apprentissage par renforcement et des systèmes multicapteurs. Enfin, à long terme j'envisage que les travaux de recherche restent centrés autour des méthodes statistiques, des aspects algorithmiques et des signaux et systèmes.

6.2.1 Court terme

Les perspectives de travail à court terme sont illustrées par la liste des articles en préparation pour des revues internationales :

- **Construction de noyaux définis positifs entre ensembles de vecteurs**. Un article est en cours de rédaction pour la revue *journal of Machine Learning research* (avec F. Desobry).
- **Représentation temps-fréquences parcimonieuses**. Un article est en préparation pour la revue *Statistica Sinica* (avec P. Wolfe).
- **Estimation spectrale bayésienne par méthodes de Monte Carlo Séquentielles**. Un article est en préparation pour la revue *IEEE Trans. on Signal Processing* (avec A. Doucet et P. Del Moral).
- **Analyse de données par mélanges de Dirichlet de processus gaussiens**. Un article est en préparation (avec E. Jackson et A. Doucet).
- **Le chemin complet de régularisation des SVM à une classe**. un article est en préparation (avec A. Rakotomamonjy).

6.2.2 Moyen terme

L'acceptation par l'INRIA-FUTURS du projet SequeL va orienter la majorité de mes activités futures vers l'apprentissage séquentiel. Il s'agit d'aborder les contextes dans lesquels les données ne sont pas toutes disponibles au même moment pour effectuer un apprentissage, soit du fait de leur séquentialité naturelle, soit du fait de leur trop grande taille. Dans ce contexte, nous nous focaliserons sur l'apprentissage par renforcement, voir Sutton et Barto (1998). L'application étudiée de façon privilégiée sera la gestion des systèmes multicapteurs.

Bibliographie

B. D. O. Anderson et J. B. Moore : *Optimal Filtering*. Prentice Hall, Englewood Cliffs, USA, 1979.

C. Andrieu, M. Davy et A. Doucet : Improved auxiliary particle filtering : Application to time-varying spectral analysis. *In IEEE Statistical Signal Processing Workshop*, p. 309–312, Singapour, août 2001a.

C. Andrieu, M. Davy et A. Doucet : Efficient particle filtering for jump markov systems. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, p. 1625–1628, Orlando, USA, mai 2002a.

C. Andrieu, M. Davy et A. Doucet : A particle filtering technique for jump markov systems. *In European Signal Processing Conference*, Toulouse, France, sept. 2002b.

C. Andrieu, M. Davy et A. Doucet : Efficient Particle Filtering for Jump Markov Systems. Application to Time-Varying Autoregressions. *IEEE Transactions on Signal Processing*, 51 (7):1762–1769, juil. 2003.

C. Andrieu, P. Djuric et A. Doucet : Model selection by MCMC computation. *Signal Processing*, 81(1):19–37, jan. 2001b.

C. Andrieu et A. Doucet : Joint Bayesian detection and estimation of noisy sinusoids via reversible jump MCMC. *IEEE Transactions on Signal Processing*, 47(10):2667–2676, 1999.

N. Aronszajn : Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

A. Berlinet et C. Thomas-Agnan : *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, 2004.

A. Bordes, S. Ertekin, J. Weston et L.Bottou : Fast kernel classifiers with online and active learning. *Journal of Machine Learning Research*, 6:1579–1619, 2005.

F. Caron, M. Davy, A. Doucet, E. Duflos et P. Vanheeghe : Bayesian inference for dynamic models with Dirichlet process mixtures. *In International Conference on Information Fusion*, 2006a. à paraître.

F. Caron, M. Davy, A. Doucet, E. Duflos et P. Vanheeghe : Bayesian inference for linear dynamic models with Dirichlet process mixtures. *IEEE Transactions on Signal Processing*, 2006b.

F. Caron, M. Davy, E. Duflos et P. Vanheeghe : Fusion de capteurs potentiellement defaillants par filtrage particulaire. *In 20ème colloque GRETSI*, Louvain-La-Neuve, Belgium, sept. 2005.

F. Caron, M. Davy, E. Duflos et P. Vanheeghe : Particle filtering for multisensor data fusion with switching observation models. application to land vehicle positioning. *IEEE Transactions on Signal Processing*, 2006c.

L. Cohen et T. Posh : Positive time-frequency distribution functions. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-33(1), February 1985.

P. Corazza : Introduction to Metric-Preserving Functions. *American Mathematics Monthly*, 104(4):309–323, avr. 1999.

H. Cottereau, M. Davy, C. Doncarli et N. Martin : Using ARCAP time-frequency representations for decision. *In European Signal Processing Conference*, Tampere, Finland, sept. 2000.

H. Cottereau, J.-M. Piasco, C. Doncarli et M. Davy : Two approaches for the estimation of time-varying amplitude multichirp signals. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, avr. 2003.

M. Davy : *Noyaux optimisés pour la classification dans le plan temps-fréquence - Proposition d'un algorithme constructif et d'une référence bayésienne basée sur les méthodes MCMC - Application au diagnostic d'enceintes acoustiques.* Thèse de doctorat, Université de Nantes, 2000.

M. Davy : Classification. *In* C. Doncarli et N. Martin, éds : *Décision dans le plan temps-fréquence*, IC2. Hermès, 2003. ISBN 2746207699.

M. Davy : Application de techniques temps-fréquence aux signaux sonores : reconnaissance et diagnostic. *In* F. Hlawatsch et F. Auger, éds : *Temps-fréquence : Concepts et Outils*, IC2. Hermès, 2005a. ISBN 2746210339.

M. Davy : Bayesian separation of harmonic sources. *In Joint Statistical Meeting*, Minneapolis, USA, août 2005b.

M. Davy : An introduction to statistical signal processing and spectrum estimation. *In* Klapuri et Davy (2006).

M. Davy : Multiple F0 frequency estimation based on generative models. *In* Klapuri et Davy (2006).

M. Davy, H. Cottereau et C. Doncarli : Loudspeaker fault detection using time-frequency representations. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5, p. 3329–3332, Salt Lake City, USA, mai 2001a.

M. Davy, P. Del Moral et A. Doucet : Méthodes Monte Carlo séquentielles pour l'analyse spectrale bayésienne. *In 19ème colloque GRETSI*, Paris, France, sept. 2003.

M. Davy, F. Desobry et S. Canu : Estimation of minimum measure sets in reproducing kernel hilbert spaces and applications. *In IEEE International Conference on Audio, Speech and Signal Processing*, Toulouse, France, mai 2006a.

M. Davy, F. Desobry, A. Gretton et C. Doncarli : An online support vector machine for abnormal events detection. *Signal Processing*, 86(8):2009–2025, août 2006b.

M. Davy et C. Doncarli : Optimal kernels of time-frequency representations for signal classification. *In IEEE Symposium on Time-Frequency and Time-Scale Analysis*, p. 581–584, Pittsburgh, USA, 1998.

M. Davy et C. Doncarli : Distances et criteres de contraste dans le plan temps-frequence. *In 19ème colloque GRETSI*, Vannes, France, sept. 1999. In French.

M. Davy et C. Doncarli : A New Nonstationary Test Procedure for Improved Loudspeaker Fault Detection. *Journal of the Audio Engineering Society*, 50(6):458–469, juin 2002.

M. Davy, C. Doncarli et G. F. Boudreaux-Bartels : Improved Optimization of Time-Frequency Based Signal Classifiers. *IEEE Signal Processing Letters*, 8(2):52–57, fév. 2001b.

M. Davy, C. Doncarli et J. Y. Tourneret : Classification of chirp signals using hierarchical Bayesian learning and MCMC methods. *IEEE Transactions on Signal Processing*, 50(2): 377–388, 2002a.

M. Davy, C. Doncarli et J. Tourneret : Supervised classification using MCMC methods. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, p. 33–36, Istanbul, Turkey, juin 2000a.

M. Davy et A. Doucet : Copulas : a new insight into positive time-frequency distributions. *IEEE Signal Processing Letters*, 10(7):215–218, juil. 2003.

M. Davy et S. Godsill : Bayesian harmonic models for musical signal analysis. *In Seventh Valencia International meeting Bayesian statistics 7*, Tenerife, Spain, juin 2002a.

M. Davy et S. Godsill : Detection of abrupt spectral changes using support vector machines. an application to audio signal segmentation. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, p. 1313–1316, Orlando, USA, mai 2002b.

M. Davy et S. Godsill : Bayesian Harmonic Models for Musical Signal Analysis. *In Cambridge Music Processing Colloquium*, Cambridge, UK, mars 2003.

M. Davy, S. Godsill et J. Idier : Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustical Society of America*, 119(4):2498–2517, avr. 2006c.

M. Davy et S. Godsill : Bayesian harmonic models for musical pitch estimation and analysis. Rap. tech. CUED/F-INFENG/TR.431, Department of Engineering, University of Cambridge, Cambridge, UK, avr. 2002c.

M. Davy, A. Gretton, A. Doucet et P. Rayner : Optimised support vector machines for nonstationary signal classification. *IEEE Signal Processing Letters*, 9(12):442–445, déc. 2002b.

M. Davy et J. Idier : Fast MCMC computations for the estimation of sparse processes from noisy observations. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, mai 2004.

M. Davy, B. Leprettre, C. Doncarli et N. Martin : Tracking of spectral lines in an ARCAP time-frequency representation. *In European Signal Processing Conference*, Island of Rhodes, Greece, sept. 1998.

M. Davy, J. Tourneret et C. Doncarli : Méthodes MCMC pour l'apprentissage bayésien. Exposé oral – journée tutoriale du 11 mai, CNES, 2000b.

M. Davy, J. Tourneret et C. Doncarli : Méthodes MCMC pour l'apprentissage bayésien. Exposé oral – réunion pleinière du 5 décembre, GdR-PRC ISIS, 2000c.

M. Davy et P. Wolfe : Une méthode à noyaux pour l'approximation parcimonieuse des représentations temps-fréquence bilinéaires. *In 20ème colloque GRETSI*, Louvain-La-Neuve, Belgium, sept. 2005.

P. Del Moral et A. Doucet : On a class of genealogical and interacting metropolis models. *In* J. Azéma, M. Emery, M. Ledoux et M. Yor, éds : *Séminaire de Probabilités*, vol. XXXVII de *Lecture Notes in Mathematics 1832*, p. 415–446, Berlin, 2003. Springer.

P. Del Moral, A. Doucet et A. Jasra : Sequential Monte Carlo samplers. *Journal of the Royal Statistical Society : Series B*, 2006.

F. Desobry : *Méthodes à noyaux pour la détection de ruptures.* Thèse de doctorat, Université de Nantes, 2004.

F. Desobry et M. Davy : Détection de ruptures en ligne par estimateur SVM de support de densité. *In 19ème colloque GRETSI*, Paris, France, sept. 2003a.

F. Desobry et M. Davy : Support vector-based online detection of abrupt changes. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, Hong Kong, China, avr. 2003b.

F. Desobry et M. Davy : Dissimilarity measures in feature space. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, mai 2004.

F. Desobry, M. Davy et C. Doncarli : An online kernel change detection algorithm. *IEEE Transactions on Signal Processing*, 53(5), mai 2005a.

F. Desobry, M. Davy et W. Fitzgerald : A class of kernels for sets of vectors. *In European Symposium on Artificial Neural Networks*, Bruges, Belgique, mai 2005b.

N. Dobigeon, J. Y. Tourneret et M. Davy : Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach. *IEEE Transactions on Signal Processing*, 2006a.

N. Dobigeon, J. Tourneret et M. Davy : Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach. *In IEEE International Conference on Audio, Speech and Signal Processing*, Toulouse, France, mai 2006b.

C. Doncarli, M. Davy et J. Tourneret : Hierarchical Bayesian classification of chirp signals. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, p. 1565 –1568, Orlando, USA, mai 2002.

A. Doucet, V. Ba-Ngu, C. Andrieu et M. Davy : Particle filtering for multi-target tracking and sensor management. *In International Conference on Information Fusion*, vol. 1, p. 474 –481, 2002.

A. Doucet, N. de Freitas et N. Gordon : *Sequential Monte Carlo Methods in Practice*. Springer, New York, USA, 2001.

C. Dubois et M. Davy : Harmonic tracking using sequential Monte Carlo. *In IEEE Statistical Signal Processing Workshop*, Bordeaux, France, juil. 2005a.

C. Dubois et M. Davy : Joint detection and tracking of time-varying harmonic components : an online Bayesian framework. *IEEE Transactions on Speech and Audio Processing*, 2005b. Submitted.

C. Dubois et M. Davy : Suivi de trajectoires temps-fréquence par filtrage particulaire. *In 20ème colloque GRETSI*, Louvain-La-Neuve, Belgium, sept. 2005c. In French.

C. Dubois, M. Davy et J. Idier : Tracking of time-frequency components using particle filtering. *In IEEE International Conference on Audio, Speech and Signal Processing*, Philadelphia, USA, juil. 2005.

R. O. Duda, P. E. Hart et G. Stork, David : *Pattern classification*. Wiley, New York, USA, second édn, 2001.

H. Feichtinger et T. Strohmer : *Gabor Analysis and Algorithms : Theory and Applications*. Applied and Numerical Harmonic Analysis. Birkhauser, 1998.

B. Fergani, M. Davy et A. Houacine : Application des machines vecteurs support mono-classe l'indexation en locuteurs de documents audio. *In Journées d'étude de la parole*, Dinard, France, juin 2006a.

B. Fergani, M. Davy et A. Houacine : Speaker segmentation using one class support vector machines. *Speech Communication*, juin 2006b. en première lecture.

B. Fergani, M. Davy et A. Houacine : Unsupervised speaker indexing using one-class support vector machines. *In European Signal Processing Conference*, Florence, Italie, 2006c.

P. Flandrin : *Time-Frequency/Time-Scale Analysis*. Academic Press, 1999.

S. Godsill et M. Davy : Bayesian harmonic models for musical pitch estimation and analysis. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, p. 1769–1772, Orlando, USA, mai 2002.

S. Godsill et M. Davy : Bayesian harmonic models for musical pitch analysis. *In 54th Session of the International Statistical Institute (ISI)*, Berlin, Germany, août 2003.

S. Godsill et M. Davy : Bayesian computational models for inharmonicity in musical instruments. *In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, USA, oct. 2005.

M. Goodwin : Residual modeling in music analysis-synthesis. *In IEEE International Conference on Acoustics, Speech, and Signal Processing*, p. 1005–1008, Atlanta, USA, 1996.

N. Gordon, D. Salmond et A. Smith : Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, 140(2):107–113, 1993.

P. Green : Penalized likelihood. *In Encyclopaedia of Statistical Sciences*, vol. 3, p. 578–586, 1999.

A. Gretton, M. Davy, A. Doucet et P. Rayner : Nonstationary signal classification using support vector machines. *In IEEE Statistical Signal Processing Workshop*, p. 305–308, Singapour, août 2001.

K. Gröchenig : *Foundations of time-frequency analysis*. Applied and Numerical Harmonic Analysis. Birkhauser, Boston, MA, 2001. ISBN 0-8176-4022-3.

V. Guigue, A. Rakotomamonjy et S. Canu : Kernel basis pursuit. *In European Conference on Machine Learning*, Porto, Portugal, 2005.

T. Hastie, R. Tibshirani et J. Friedman : *The Elements of Statistical Learning*. Springer, New-York, 2001.

P. Herrera-Boyer, A. Klapuri et M. Davy : Automatic classification of pitched musical instrument sounds. *In* Klapuri et Davy (2006).

A. Johansen, A. Doucet et M. Davy : Maximum likelihood parameter estimation for latent variable models using sequential Monte Carlo. *In IEEE International Conference on Audio, Speech and Signal Processing*, Toulouse, France, mai 2006.

S. J. Julier et J. Uhlmann : Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, mars 2004.

T. Kailath : RKHS approach to detection and estimation problems-part I : Deterministic signals in Gaussian noise. *IEEE Transactions on Information Theory*, 17(5):530–549, sept. 1971.

G. Kitagawa : Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics*, 5(1):1–25, 1996.

J. Kivinen, A. J. Smola et R. C. Williamson : Online Learning with Kernels. *IEEE Transactions on Signal Processing*, 58(8), août 2004.

A. Klapuri et M. Davy, éds. *Signal Processing Methods for Music Trascription*. Springer, New-York, USA, 2006.

H. Laurent et C. Doncarli : Stationarity index for abrupt changes detection in the time-frequency plane. *IEEE Signal Processing Letters*, 5(2):43–45, 1998.

M. Loth, R. Coulom, M. Davy et P. Preux : Least angle temporal difference learning : LATD(λ). *In Journées Francophones Apprentissage, Décision, Contrôle*, Toulouse, France, mai 2006.

N. Martin, C. Doncarli et M. Davy : Méthodes de décision dans un contexte de signaux non stationnaires. *In 3rd SFM-IMEKO-SFA International Conference on Acoustic and Vibratory Surveillance Methods and Diagnostic Techniques*, Senlis, France, oct. 1998.

J. Mercer : Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions Royal Society of London*, A 209:415–446, 1909.

C. S. Ong, X. Mary, S. Canu et A. Smola : Learning with non positive kernels. *In International Conference on Machine Learning*, p. 639–646, 2004.

E. Parzen : Probability density functionals and reproducing kemel Hilbert spaces. p. 155–169. Wiley, New-York (USA), 1963.

M. Pitt et N. Shephard : Filtering via simulation : auxiliary particle filters. *Journal of the American Statistical Association*, 94:590–599, 1999.

D. Potin, P. Vanheeghe, E. Duflos et M. Davy : An abrupt change detection algorithm for buried landmines localization. *IEEE Transactions on Geoscience and Remote Sensing*, 44(2):260–272, fév. 2006.

E. Punskaya, C. Andrieu, A. Doucet et W. Fitzgerald : Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on Signal Processing*, 50(3):747–758, mars 2002.

S. Richardson et P. J. Green : On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society : Series B*, 59(4):731–792, 1997.

C. Robert et G. Casella : *Monte Carlo Statistical Methods*. Springer, New York, USA, 2000.

N. Saito et R. R. Coifman : Local discriminant bases. *In* A. Laine et M. Unser, éds : *Wavelet applications in Signal and Image Processing II*. Proc. SPIE vol 2303, 1994.

B. Schölkopf et A. Smola : *Learning with Kernels*. MIT Press, Cambridge, USA, 2002.

R. Sutton et A. Barto : *Reinforcement learning : an introduction*. MIT Press, 1998.

R. van der Merwe, A. Doucet, N. de Freitas et E. Wan : The unscented particle filter. *In Neural Information Processing Systems*, Denver, USA, 2000.

V. Vapnik : *The Nature of Statistical Learning Theory*. Springer, New-York, 1995.

J. Vermaak, C. Andrieu, A. Doucet et S. Godsill : On-line Bayesian modelling and enhancement of speech signals. Technical Report CUED/F-INFENG/TR.361, Université de Cambridge, jan. 2000.

G. Wahba : *Spline models for observational data*, vol. 59 de *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

P. J. Wolfe, S. J. Godsill et W. Ng : Bayesian variable selection and regularisation for time-frequency surface estimation. *Journal of the Royal Statistical Society : Series B*, 66 (3):575–589, août 2004.

Annexe A

Publications sélectionnées

A.1 Optimised Support Vector Machines for Nonstationary Signal Classification – Davy *et al.* (2002b)

Cet article a été publié dans la revue *IEEE Signal Processing Letters*, Vol. 9, No 12, pp. 442-445, en Décembre 2002.

A.2 Efficient Particle Filtering for Jump Markov Systems. Application to Time-Varying Autoregressions – Andrieu *et al.* (2003)

Cet article a été publié dans la revue *IEEE Trans. on Signal Processing*, Vol. 51, No 7, pp. 1762-1769 en juillet 2003.

A.3 Dissimilarity measures in feature space – Desobry et Davy (2004)

Cet article a été publié à la conférence IEEE ICASSP 2004, qui s'est tenue à Montréal (Canada) en mai 2004.

A.4 An Online Kernel Change Detection Algorithm – Desobry *et al.* (2005a)

Cet article a été publié dans la revue *IEEE Trans. on Signal Processing*, Vol. 53, No. 8 (partie 2), pp. 2961-2974, en août 2005.

A.5 Bayesian Analysis of Polyphonic Western Tonal Music – Davy *et al.* (2006c)

Cet article a été publié dans le *Journal of the Acoustical Society of America*, Volume 119, Numéro 4, pp. 2498-2517 en avril 2006.

A.6 Estimation of minimum measure sets in reproducing kernel Hilbert spaces and applications – Davy *et al.* (2006a)

Cat article sera présenté à la conférence IEEE ICASSP 2006, qui se tiendra à Toulouse en mai 2006.

A.7 Signal Processing Methods for Music Transcription – Klapuri et Davy (2006)