École Doctorale de l'Université de Technologie de Compiègne

Habilitation à Diriger des Recherches

Méthodes à noyau et critères de contraste pour la détection à structure imposée

présentée et soutenue publiquement le 12 décembre 2002

par

Cédric RICHARD

Composition du jury

Président :	I. Nikiforov	Professeur à l'UTT, LM2S, Troyes
Rapporteurs :	F. Castanié	Professeur à l'ENSEEIHT, TeSA, Toulouse
	P. Flandrin	Directeur de Recherche CNRS, laboratoire de physique, ENS Lyon
	B. Picinbono	Professeur émérite à l'Université de Paris-Sud, L2S, Orsay
Examinateurs :	R. Lengellé	Professeur à l'UTT, LM2S, Troyes
	A. Richard	Professeur à l'UHP, CRAN, Nancy

Laboratoire de Modélisation et Sûreté des Systèmes — EA 3173
Université de Technologie de Troyes

Mis en page avec la classe thloria.

Avant propos

Depuis mon intégration en 1995 au sein du Laboratoire de Modélisation et Sûreté des Systèmes de l'Université de Technologie de Troyes, l'ensemble de mes activités de recherche s'est articulé autour du traitement du signal et des théories statistiques de la décision et de l'apprentissage pour la surveillance des systèmes. Plusieurs facettes relatives à la détection et la caractérisation d'événements dans les signaux et systèmes ont été abordées, et ont pu trouver leur prolongement naturel dans l'encadrement de deux stages de DEA et de quatre thèses de doctorat, dont une a été soutenue le 22 octobre 2002. Simultanément, je me suis investi dans les enseignements proposés par le département Génie des Systèmes d'Information et de Décision de l'Université de Technologie de Troyes, principalement sur le thème des télécommunications et réseaux. En vue de l'obtention de l'habilitation à diriger des recherches, ce document retrace mon parcours dans l'enseignement supérieur jusqu'à ce jour, et présente un bilan de mes interactions avec la communauté scientifique.

La première partie de ce manuscrit constitue une notice individuelle dans laquelle je présente mon curriculum vitæ, mon parcours professionnel, mes activités d'enseignement, de recherche et d'encadrement doctoral. Puis j'évoque mes collaborations et contrats industriels, montre mon implication importante dans la communauté du traitement du signal, et décrit mes activités administratives. Une liste de mes publications vient enfin compléter cette présentation.

La deuxième partie du document est consacrée à une description détaillée de mes travaux de recherche, dont le thème central est la détection à structure imposée. Après une introduction dans laquelle j'évoque la nécessité fréquente de recourir à une telle approche, s'enchaînent cinq chapitres au cours desquels j'expose ma vision du problème et les solutions proposées. Une application relative aux signaux électrophysiologiques de sommeil vient finalement illustrer mes propos et ouvrir de nouvelles perspectives, dont une conclusion se fait l'écho.

Table des matières

Première partie

Notice d'activité

Curriculum Vitæ

Cédric RICHARD
né le 24 janvier 1970, à Sarrebourg (Moselle)
nationalité française
célibataire

Domicile
17 rue de Vauluisant
10000 Troyes
tél. : 03.25.73.24.36

Établissement actuel
Université de Technologie de Troyes (UTT)
12 rue Marie Curie, BP 2060, 10010 Troyes cedex
tél. : 03.25.71.58.47 fax. : 03.25.71.56.99
email : cedric.richard@utt.fr

Situation professionnelle

Depuis le 1er septembre 1999, **maître de conférences** à l'Université de Technologie de Troyes
Département Génie des Systèmes d'Information et de Décision (GSID)
Laboratoire de Modélisation et Sûreté des Systèmes (LM2S - EA 3173)

Année universitaire 1998-1999, **Ater à temps complet** à l'Université de Technologie de Troyes
Département Génie des Systèmes d'Information et de Décision (GSID)
Laboratoire de Modélisation et Sûreté des Systèmes (LM2S - EA 3173)

Formation

1998 **Thèse de Doctorat** de l'UTC en Contrôle des Systèmes
« Une méthodologie pour la détection à structure imposée. Applications au plan temps-fréquence », soutenue le 22 décembre 1998.
Mention très honorable avec félicitations du jury,
constitué de T. Denœux (UTC), B. Dubuisson (UTC), P. Duvaut (Rap., ENSEA), P. Flandrin (Rap., ENS LYON), P. Gaillard (Prés., UTT), P. Gonçalvès (INRIA), R. Lengellé (Dir., UTT), I. Nikiforov (UTT).

1994 **Diplôme d'Etudes Approfondies** de l'UTC en Contrôle des Systèmes
Mention Bien, classé 1er sur 42.

1994 **Diplôme d'Ingénieur en Génie Informatique** de l'UTC
Filière Modélisation, Analyse et Commande des Systèmes.

1991 **Deug A** de la Faculté de Saint-Étienne en Sciences pour l'Ingénieur
Mention Bien, classé 1er sur 185 en 1ère année, et 1er sur 50 en 2ème année.

Parcours professionnel

depuis septembre 1998

Enseignant-chercheur à l'UTT, d'abord comme Ater à temps plein (service de maître de conférences) pendant un an, puis comme maître de conférences au département Génie des Systèmes d'Information et de Décision (GSID). Activités de recherches menées au laboratoire LM2S (EA 3173).

décembre 1995 - décembre 1998

Préparation d'une **thèse de doctorat de l'UTC** en Contrôle des Systèmes, au laboratoire LM2S (EA 3173) de l'UTT, et **moniteur de l'enseignement supérieur**. Travaux menés sous la direction du Prof. R. Lengellé, sur le sujet suivant :

« Une méthodologie pour la détection à structure imposée. Applications au plan temps-fréquence. »

Thèse obtenue avec la mention très honorable et les félicitations du jury, constitué de T. Denœux (UTC), B. Dubuisson (UTC), P. Duvaut (Rap., ENSEA), P. Flandrin (Rap., ENS LYON), P. Gaillard (Prés., UTT), P. Gonçalvès (INRIA), R. Lengellé (UTT) et I. Nikiforov (UTT).

septembre 1994 - août 1995

Service National effectué au titre de **scientifique du contingent**. Ingénieur au service Automatique de Giat Industries, Satory, chargé d'étude dans le cadre du projet :

« Automatisation du télépilotage d'un véhicule chenillé lourd. »

L'objet de cette année de recherche a été d'étudier la faisabilité d'un pilote automatique de vitesse pour véhicule chenillé lourd, à partir d'un relevé topographique temps-réel par nappe laser.

mars 1994 - août 1994

Projet de fin d'études d'ingénieur de l'UTC et **stage de DEA** effectués à la DRAS, PSA Peugeot-Citroën, Vélizy, sous la direction de M. Masson (URA CNRS 817, Heudiasyc, UTC), sur le sujet :

« Objectivation de la perception acoustique automobile. »

L'objectif de ce travail a été de définir une première approche méthodologique afin d'établir des corrélations entre le mode de perception auditif d'une population test replacée dans les conditions acoustiques d'un habitacle automobile et un ensemble de paramètres objectifs à déterminer.

Activités d'enseignement

Dès mon arrivée à l'UTT en 1995, je me suis investi dans des enseignements de premier et second cycles aussi divers que l'automatique, les mathématiques et la physique. Ces activités s'inscrivaient alors dans le cadre d'un monitorat. Depuis 1998, mes activités d'enseignement se sont recentrées autour du **domaine des télécommunications et réseaux**. Aussi, je participe aux travaux dirigés et pratiques de traitement du signal et de réseaux destinés aux élèves-ingénieurs de l'UTT, dont un descriptif est présenté ci-après. De plus, j'ai la responsabilité du cours de théorie de l'information, dans le cadre duquel j'ai rédigé **un polycopié**, et co-écrit **un ouvrage pédagogique** dont les références sont les suivantes :

« Théorie et codage de l'information, » cours et exercices corrigés,
C. Richard et P. Cornu, collection Technosup, Editions Ellipses, 224 p., à paraître en 2003.

Parallèlement à cela, j'encadre en moyenne 8 étudiants par an durant leur stage d'un semestre de milieu et de fin d'études d'ingénieur. Je participe également à l'évaluation du stage ouvrier des étudiants de premier cycle.

Théorie et codage de l'information (responsable) - cours, travaux dirigés

Supports de cours : 1 polycopié, 1 ouvrage (voir ci-dessus et notice page suivante).

Parce qu'elle en constitue l'une des principales richesses, l'information est aujourd'hui au cœur de l'entreprise, où elle est par exemple acquise, traitée, stockée ou encore transmise. Dans ce contexte concurrentiel, il s'avère impératif de garantir son intégrité et sa confidentialité, permettre son authentification et assurer de bonnes conditions d'archivage. L'enseignement dispensé n'est pas un cours sur les codes correcteurs, ni sur la cryptographie ou encore la compression. Il est centré sur l'information, sa modélisation et ses traitements les plus courants. Pour cela, le cours proposé s'appuie sur un cadre théorique approprié, celui de la théorie de l'information.

Analyse et traitement du signal - travaux dirigés, travaux pratiques

L'objectif de cet enseignement est de présenter les concepts et outils nécessaires à l'étude des signaux en abordant les thèmes suivants : représentations des signaux déterministes, filtrage, transmission de l'information, signaux et systèmes numériques, signaux aléatoires.

Télécommunications d'entreprise - cours (6 heures), travaux dirigés

Cet enseignement traite de la problématique de communication sous trois angles : réseaux, protocoles, services. Ainsi, on décrit tout d'abord les grandes fonctionnalités des réseaux étendus, dans le cadre desquelles je présente en particulier la fonction transmission durant 6 heures de cours. La partie dédiée aux protocoles est consacrée à la description d'architectures telles que TCP/IP, X25, Frame Relay, RNIS et ATM. Enfin, le thème des services est illustré au travers de différentes applications communicantes (voix, données, image, messagerie).

Ces enseignements représentent une **charge moyenne de 335 heures** (équivalent TP) par an.

Notice d'ouvrage pédagogique

« Théorie et codage de l'information, » cours et exercices corrigés,
C. Richard et P. Cornu, collection Technosup, Editions Ellipses, 224 p., à paraître en 2003.

Objectifs et originalité de l'ouvrage

Parce qu'elle en constitue l'une des principales richesses, l'information est aujourd'hui au cœur de l'entreprise, où elle est par exemple acquise, traitée, stockée ou encore transmise. Dans ce contexte concurrentiel, il s'avère impératif de garantir son intégrité et sa confidentialité, permettre son authentification et assurer de bonnes conditions d'archivage. En regard de ces opérations et des contraintes associées, des traitements spécifiques sont couramment utilisés :

- l'utilisation de *codes détecteurs et correcteurs d'erreurs* en vue de préserver l'intégrité des données, au cours de leur transmission ou stockage par exemple ;
- la mise en œuvre de *méthodes de compression* afin de limiter le temps de transmission ou l'espace de stockage requis ;
- le *chiffrement*, ou *cryptage*, dans le but de garantir la confidentialité de l'information, qu'elle soit stockée ou en cours de transmission, et d'en assurer l'authentification lors de son acquisition par exemple.

Pour pouvoir développer de façon rigoureuse ces différents sujets, et ainsi permettre une confrontation des performances de systèmes réels avec certaines limites théoriques, il est nécessaire d'évoluer dans un cadre théorique approprié : la théorie de l'information. Ainsi, l'ouvrage proposé ne constitue pas un livre sur les codes correcteurs, ni sur la cryptographie ou encore la compression. Il s'agit d'un ouvrage centré sur l'information, sa modélisation et ses traitements les plus courants dans ce monde que l'on qualifie fréquemment de *société de l'information*.

Sujets traités et public concerné

L'ouvrage proposé a pour objectif de familiariser le lecteur avec les aspects des technologies de l'information décrits ci-dessus, présentés de manière rigoureuse et illustrés par des exercices corrigés. Plus précisément, le programme couvre les thèmes suivants :

- théorie de l'information : éléments de probabilité, caractérisation d'une source, d'un langage et d'un codage, modélisation de canaux de transmission ;
- codes détecteurs et correcteurs : éléments d'algèbre discrete, codage et décodage des codes linéaires, codes de Hamming, codes cycliques, exemples de méthodes de codage et compression dédiées ;
- chiffrement : complexité algorithmique et compléments mathématiques, forces et faiblesses des systèmes actuels, contexte réglementaire actuel, cryptosystèmes symétriques et à clé publique, signature électronique.

L'ouvrage proposé s'adresse aux étudiants en technologies de l'information (Génie des Télécommunications et Réseaux, Génie des Systèmes d'Information, etc.) de licence, maîtrise et cycle ingénieur, voire de DESS et DEA. A caractère résolument pédagogique, il concerne également les étudiants d'IUT des mêmes filières désireux d'approfondir leurs connaissances.

Activités de recherche

Thématique et contexte

Le laboratoire de Modélisation et Sûreté des Systèmes (LM2S - EA 3173) de l'UTT a été crée en 1994, soit un an avant mon arrivée. Il regroupe aujourd'hui 19 enseignants-chercheurs permanents, dont 7 sont habilités à diriger des recherches, auxquels s'ajoutent 16 doctorants. Ce laboratoire, sous la responsabilité d'Igor Nikiforov, s'articule autour de 3 axes de recherche :

Axe 1 : Décision et diagnostic en environnement non-stationnaire ;

Axe 2 : Modèles stochastiques pour la fiabilité et la maintenance ;

Axe 3 : Surveillance et contrôle d'environnements complexes et évolutifs par raisonnement cognitif et agents intelligents logiciels.

Depuis mon intégration au sein de l'axe 1, l'ensemble de mes activités de recherche s'articule autour du traitement du signal et de la théorie statistique de la décision pour la surveillance des systèmes. Plusieurs facettes relatives à la détection et à la caractérisation d'événements dans les signaux et systèmes ont été abordées, et ont trouvé un prolongement dans l'encadrement de deux stages de DEA et de quatre thèses de doctorat. La première d'entre elles a été soutenue le 22 octobre 2002. Il est entendu qu'au niveau national, un certain nombre de laboratoires développent des activités de recherche dans ce domaine, par exemple le CRAN, le LAIL, le LSS ou encore l'IRISA. Afin de m'en démarquer, le fil directeur de mes travaux a été jusqu'à présent celui de la détection d'événements lorsque les seules informations disponibles pour l'élaboration du détecteur consistent en un ensemble de réalisations étiquetées des hypothèses en compétition, à savoir les hypothèses « bruit » et « signal+bruit ». Aussi, la méthodologie proposée s'appuie-t-elle sur la théorie statistique de l'apprentissage pour une meilleure compréhension et prise en compte de phénomènes tels que la malédiction de la dimensionnalité. Elle demeure toutefois ancrée dans la théorie statistique de la décision, qui lui offre un cadre théorique approprié. Ces travaux s'inscrivent dans le thème A du GdR-PRC ISIS auquel je participe activement, en particulier dans le cadre de l'analyse et de la décision en environnement non-stationnaire.

MOTS-CLÉ : surveillance des systèmes, traitement du signal, théorie de la décision, théorie de l'apprentissage, analyse temps-fréquence

Bilan statistique des publications

- ouvrage : 1
- chapitres d'ouvrages collectifs : 3
- revues internationales à comité de lecture : 7
- revues nationales à comité de lecture : 3
- conférences à comité de lecture et actes : 23

Encadrement doctoral

Les recherches que je mène en traitement du signal et en théorie statistique de la décision trouvent un écho dans l'encadrement de travaux de thèse et de DEA. Ci-dessous figure une liste de ces activités, sur lesquelles on pourra trouver d'avantage d'informations dans la seconde partie du document.

Doctorat de G. Viardot - soutenu (taux d'encadrement personnel : 50%)

G. Viardot a préparé sa thèse en milieu hospitalier, au sein de la Fondation pour la Recherche en Neuro-Sciences Appliquées à la Psychiatrie (FORENAP) du CHS de Rouffach, encadré par R. Lengellé et moi-même. Une Action Concertée Incitative de 700kF (voir page 11) a permis de financer partiellement ce travail, l'autre partie incombant à FORENAP. L'étudiant a soutenu sa thèse le 22 octobre 2002 à l'UTT, sur le sujet

> « Reconnaissance des formes en présence d'incertitude sur l'expertise. Application à l'étude des phases d'activation transitoires du sommeil chez l'homme. »

Le jury était constitué de

> Guy Carrault, Université de Rennes 1 (rapporteur)
> Catherine Marque, Université de Technologie de Compiègne (rapporteur)
> François Auger, IUT de Saint Nazaire, (examinateur)
> Jean Paul Macher, CHS de Rouffach (examinateur)
> Régis Lengellé, Université de Technologie de Troyes (directeur)
> Cédric Richard, Université de Technologie de Troyes (directeur).

G. Viardot est actuellement ingénieur de recherches chez FORENAP. Durant sa thèse, il a présenté trois communications dans le cadre de conférences à comité de relecture :

G. Viardot, R. Lengellé, C. Richard. Mixture of experts for automated detection of spontaneous phasic arousals in sleep signals. *Proc. IEEE International Conference on Systems, Man and Cybernetics,* IEEE SMC'02, Hamammet, 2002.

G. Viardot, R. Lengellé, C. Richard, A. Coatanhay. Fusion d'avis d'experts et caractérisation de l'expertise. Application à la détection de transitoires dans les signaux physiologiques. *Proc. Colloque GRETSI,* Toulouse, 2001.

G. Viardot, A. Coatanhay, R. Lengellé, C. Richard, L. Staner, A. Muzet, J.-P. Macher. Fusion d'avis divergents d'experts. *Proc. Colloque AcM/MdA,* Grenoble, 2000.

Doctorat de F. Abdallah - en préparation (taux d'encadrement personnel : 50%)

F. Abdallah a débuté la préparation de sa thèse en octobre 2000, encadré par R. Lengellé et moi-même. Le thème de ses recherches est

> « Détection à structure imposée et mesures de contraste. »

Cette thèse est financée par le Conseil Régional de Champagne-Ardennes. A ce jour, F. Abdallah a participé à la rédaction d'une lettre dans une revue internationale de très haut niveau, et présenté quatre communications dans des conférences à comité de sélection.

C. Richard, R. Lengellé, F. Abdallah. Bayes-optimal detectors using relevant second-order criteria. *IEEE Signal Processing Letters,* vol. 9, no. 1, 2002.

F. Abdallah, C. Richard, R. Lengellé. On virtues and vices of second-order measures of quality for binary classification. *Proc. International Conference ANNIE,* Saint Louis, 2002.

F. Abdallah, C. Richard, R. Lengellé. On equivalence between detectors obtained from second-order measures of performance. *Proc. European Signal Processing Conference,* Eusipco'02, Toulouse, 2002.

F. Abdallah, C. Richard, R. Lengellé. A method for designing nonlinear kernel-based discriminant functions from the class of second-order criteria. *Proc. International Conference Asilomar,* Pacific Grove, CA, 2002.

C. Richard, R. Lengellé, F. Abdallah. Optimisation de critères de contraste et des performances en détection. *Proc. Colloque GRETSI,* Toulouse, 2001.

Actuellement, un article est en cours de soumission dans une revue internationale :

F. Abdallah, C. Richard, R. Lengellé. An improved training algorithm for nonlinear kernel discriminants. *IEEE Transactions on Signal Processing,* août 2002 (soumis).

Doctorat de J. Gosme - en préparation (taux d'encadrement personnel : 75%)

J. Gosme a entamé la préparation de sa thèse en octobre 2001, financé par le Conseil Régional de Champagne-Ardennes. Il est encadré par R. Lengellé et moi-même, et bénéficie du concours de P. Gonçalvès (INRIA Rhône-Alpes, Grenoble). Ses recherches portent sur le sujet

« Représentations temps-fréquence et temps-échelle diffusives. »

Durant cette première année de thèse, J. Gosme a présenté une communication dans le cadre d'une conférence à comité de sélection.

J. Gosme, P. Gonçalvès, C. Richard, R. Lengellé. Adaptive diffusion and discriminant analysis for complexity control of time-frequency detectors. *Proc. European Signal Processing Conference,* Eusipco'02, Toulouse, 2002.

Doctorat de I. Costantine - en préparation (taux d'encadrement personnel : 50%)

I. Costantine a débuté ses travaux de thèse en octobre 2002, sur le sujet

« Modélisation des dépendances spatio-temporelles entre l'électrocardiogramme et des cartographies électroencéphalographiques et magnétoencéphalographique. »

Elle est financée par le Fondation pour la Recherche en Neuro-Sciences Appliquées à la Psychiatrie, où elle effectue ses recherches. Elle est encadrée par R. Lengellé et moi-même.

Encadrement de stages de DEA

Aux co-encadrements des travaux de thèse décrits précédemment, viennent s'ajouter les direction et co-direction des stages de DEA suivants.

Stage de DEA de J. Gosme (taux d'encadrement personnel : 50%)

J. Gosme a préparé son stage de DEA sous ma responsabilité et celle de P. Gonçalvès, à l'INRIA Rhône-Alpes, de mars à juillet 2001. Ses travaux s'inscrivaient dans le cadre d'une Opération Jeunes Chercheurs du GdR-PRC ISIS (voir page 11) sur le thème de l'analyse et de la décision par représentations temps-fréquence diffusives. J. Gosme prépare actuellement une thèse de doctorat sous ma responsabilité et celle de R. Lengellé, en collaboration avec P. Gonçalvès.

Stage de DEA de M. Hajj (taux d'encadrement personnel : 100%)

D'avril à septembre 2002, j'ai dirigé le stage de DEA de M. Hajj sur le thème de la regression par Support Vector Machines.

Collaborations et contrats industriels

2002 **Contrat de recherche** avec le CREAS, centre de recherche d'ASCOMETAL, 35kF.

Les travaux que j'ai effectués pour ce centre de recherche concernent la détection de défauts par courants de Foucault sur un train à fils. Ils ont consisté à analyser des signaux acquis sur un banc d'essai, à définir la configuration du capteur pour une prochaine campagne d'essais dans des conditions industrielles (fréquences d'excitation de la bobine, bande d'analyse, gain, etc.), et à mettre en œuvre de premiers traitements (filtrage avec voie de référence, détection). Ce travail de deux semaines a donné lieu à la rédaction d'un rapport (réf. AST-02/06/10-1) intitulé « Étude de faisabilité de la détection de défauts longs par courants de Foucault en mesure absolue ». Une suite devrait prochainement être donnée à ce contrat.

1999 **Action Concertée Incitative** du MENRT, dotée de 700kF sur 2 ans.

Le projet mené, intitulé DAMES pour « Détection automatique de micro-éveils pendant le sommeil », s'inscrit dans le chapitre « Télémédecine et Technologies pour la Santé » des ACI proposées par le MENRT. Les partenaires étaient l'UTT, représentée par R. Lengellé et moi-même, la Fondation pour la Recherche en Neuro-Sciences Appliquées à la Psychiatrie (FORENAP) du CHS de Rouffach, ainsi que la société MEDATEC. La subvention obtenue a permis de financer en partie la thèse de G. Viardot, dont j'ai assuré le co-encadrement.

Depuis 1995 pour ma part, **collaboration** avec le FORENAP du CHS de Rouffach.

Les travaux menés en commun avec cette fondation s'inscrivent dans le cadre d'une convention de collaboration scientifique UTT/FORENAP pour laquelle R. Lengellé et moi-même sommes les représentants locaux. Les thèmes de recherche sont relatifs à la caractérisation et à la détection/classification de signaux électrophysiologiques en environnement non-stationnaire. Depuis 1995, cette collaboration fructueuse a donné lieu à une ACI (voir ci-dessus) et à 2 thèses de doctorat (1 soutenue et 1 en cours).

Rayonnement

L'objet de cette section est de montrer mon implication dans la vie de la communauté du traitement du signal, notamment par le biais de travaux menés de concert avec des chercheurs issus d'autres laboratoires, grâce au concours de structures telles que le GdR-PRC ISIS.

Lecteur/reviewer d'articles
IEEE Transactions on Signal Processing, IEEE Signal Processing Letters, Applied Signal Processing, Traitement du Signal, conférences internationales SCI.

Membre de l'action spécifique « Méthodes à vecteurs support (Support Vector Machines) », département STIC, CNRS, 2002-2003.

Cette action est animée par M. Davy (IRCCyN, Nantes). Elle a été dotée de 67000 euros par le DSTIC, et de 5000 euros par le GdR-PRC ISIS. Elle implique notamment l'organisation de sessions spéciales aux conférences Gretsi'03 et IEEE Icassp'03, ainsi qu'une école d'été.

`http://www.irccyn.ec-nantes.fr/Svm/`

Membre du projet « Méthodes temps-fréquence pour l'analyse des données de l'interféromètre Virgo », IN2P3, 2001-2003.

E. Chassande-Mottin (ILGA, Obs. de la Côte d'Azur) anime ce projet doté de 200kF par l'IN2P3. Les partenaires sont P.-O. Amblard (LIS, Grenoble), F. Auger (GE44, Saint Nazaire), P. Flandrin (ENS Lyon), P. Hello (LAL, Orsay), J.-M. Innocent (LATP-CMI, Marseille), B. Torresani (LATP-CMI, Marseille), J.-Y. Vinet (ILGA, Obs. de la Côte d'Azur) et moi-même.

`http://www.obs-nice.fr/ecm/virgotf/`

Membre de l'action « Nouveaux outils d'analyse et de décision pour les signaux fortement non-stationnaires », projet jeunes chercheurs du GdR-PRC ISIS, 1999-2001.

F. Auger (GE44, Saint Nazaire), M. Davy (IRCCyN, Nantes), P. Gonçalvès (INRIA Rhône-Alpes, Grenoble) et moi-même avons participé à ce projet. Celui-ci a été doté de 50kF par le GdR-PRC ISIS, permettant notamment de financer le stage de DEA de J. Gosme, dont j'ai assuré le co-encadrement à 50% avec P. Gonçalvès.

Présentation de séminaires
LM2S, Troyes (1996, 1997, 1998, 2000) ; GdR-PRC ISIS (1999) ; CRAN, Nancy (1996, 2000) ; LPSI, INSA de Rouen (1998) ; LIS, Grenoble (1998).

Membre de sociétés savantes GdR-PRC ISIS, IEEE, EEA

Bénéficiaire de la prime d'encadrement doctoral et de recherche depuis septembre 2001

Activités administratives

2001 → titulaire de la **Commission des Spécialistes** 61-63$^{\text{ème}}$ section de l'URCA, Reims
2001 → suppléant de la **Commission des Spécialistes** 61$^{\text{ème}}$ section de l'USTL, Lille
2001 → représentant du laboratoire LM2S au **Conseil de l'École Doctorale** de l'UTT
1996-98 membre élu au **Conseil Scientifique** de l'UTT
1996-97 membre élu au **Bureau de Département** GSID de l'UTT

Publications

Ouvrage (1)

C. Richard, P. Cornu. *Théorie et Codage de l'Information.* Paris : Editions Ellipses, collection Technosup, à paraître en 2003.

Chapitres d'ouvrage (3)

C. Richard. Détection par représentations temps-fréquence discrètes. (23 p.) In N. Martin, C. Doncarli, (éds). *Décision dans le Plan Temps-Fréquence.* Paris : Hermès Sciences, Traité IC2, 2002.

R. Lengellé, C. Richard. Apprentissage de règles de décision à structure imposée et contrôle de la complexité. (33 p.) In R. Lengellé, (éd.). *Reconnaissance des Formes et Décision en Signal.* Paris : Hermès Sciences, Traité IC2, 2002.

D. Brie, R. Lengellé, N. Nikiforov, C. Richard. Autres applications. (12 p.) In R. Lengellé, (éd.). *Reconnaissance des Formes et Décision en Signal.* Paris : Hermès Sciences, Traité IC2, 2002.

Revues internationales à comité de lecture (7)

C. Richard. Time-frequency based detection using discrete-time discrete-frequency Wigner distributions. *IEEE Transactions on Signal Processing,* vol. 50, no. 9, p. 2170-2176, 2002.

C. Richard, R. Lengellé, F. Abdallah. Bayes-optimal detectors using relevant second-order criteria. *IEEE Signal Processing Letters,* vol. 9, no. 1, 2002.

C. Richard. Linear redundancy of information carried by the discrete Wigner distribution. *IEEE Transactions on Signal Processing,* vol. 49, no. 11, p. 2536-2544, 2001.

T. Gharbi, D. Barchiesi, O. Bergossi, H. Wioland, C. Richard. Optical near-field analysis by means of time-frequency distributions. Application to the characterization and the separation of the image spectral contents using reassignment method. *Journal of Optical Society of America,* vol. 17, no. 12, p. 2513-2519, 2000.

C. Richard, R. Lengellé. Data-driven design and complexity control of time-frequency detectors. *Signal Processing,* vol. 77, p. 37-48, 1999.

C. Richard, R. Lengellé. Joint time and time-frequency optimal detection of K-complexes in sleep EEG. *Computers and Biomedical Research,* vol. 31, no. 3, p. 209-229, 1998.

C. Richard, R. Lengellé. Joint recursive implementation of time-frequency representations and their modified version by the reassignment method. *Signal Processing,* vol. 60, no. 2, p. 163-179, 1997.

Revues nationales à comité de lecture (3)

A. Coatanhay, C. Richard, L. Staner. Classification en temps-fréquence à partir de données expertisées. Application à la détection des complexes K. *Extraction des Connaissances et*

Apprentissage, Hermès, vol. 1, no. 4, p. 293-304, 2001.

C. Richard, R. Lengellé. Détection automatique de phénomènes transitoires de l'EEG par représentation temps-fréquence. *Innovation et Technologie en Biologie et Médecine,* numéro spécial temps-fréquence, vol. 19, no. 3, p. 167-177, 1998.

C. Richard, R. Lengellé. Détection linéaire optimale. *Décision et Interprétation en Environnement Non-Stationnaire,* rapport du GdR-PRC ISIS, p. 55-56, 1998.

Congrès à comité de lecture et actes (23)

F. Abdallah, C. Richard, R. Lengellé. On virtues and vices of second-order measures of quality for binary classification. *Proc. International Conference ANNIE,* Saint Louis, 2002.

F. Abdallah, C. Richard, R. Lengellé. On equivalence between detectors obtained from second-order measures of performance. *Proc. European Signal Processing Conference,* Eusipco'02, Toulouse, 2002.

F. Abdallah, C. Richard, R. Lengellé. A method for designing nonlinear kernel-based discriminant functions from the class of second-order criteria. *Proc. International Conference Asilomar,* Pacific Grove, CA, 2002.

G. Viardot, R. Lengellé, C. Richard. Mixture of experts for automated detection of spontaneous phasic arousals in sleep signals. *Proc. IEEE International Conference on Systems, Man and Cybernetics,* IEEE SMC'02, Hamammet, 2002.

J. Gosme, P. Gonçalvès, C. Richard, R. Lengellé. Adaptive diffusion and discriminant analysis for complexity control of time-frequency detectors. *Proc. European Signal Processing Conference,* Eusipco'02, Toulouse, 2002.

G. Viardot, R. Lengellé, C. Richard, A. Coatanhay. Fusion d'avis d'experts et caractérisation de l'expertise. Application à la détection de transitoires dans les signaux physiologiques. *Proc. Colloque GRETSI,* Toulouse, 2001.

C. Richard. Discrétisation des détecteurs temps-fréquence : problèmes en découlant et éléments de solution. *Proc. Colloque GRETSI,* Toulouse, 2001.

C. Richard, R. Lengellé, F. Abdallah. Optimisation de critères de contraste et des performances en détection. *Proc. Colloque GRETSI,* Toulouse, 2001.

D. Barchiesi, C. Richard. Time-frequency analysis of near-field optical data for extracting local attributes. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* IEEE ICASSP'01, Salt Lake City, Utah, 2001.

R. Lengellé, C. Richard, S. Millemann. Neural network based membership function estimation. Application to uncertain time-varying systems supervision. *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication,* IEEE ISPACS'00, Honolulu, Hawai, 2000.

A. Coatanhay, C. Richard, R. Lengellé, A. Muzet, J.-P. Macher. Automated detection of K-complexes : characterization of benzodiazepine effects. *Journal of Sleep Research,* vol. 9, no. 1, 2000.

C. Richard, R. Lengellé. On the linear relations connecting the components of the discrete Wigner distribution in the case of real-valued signals. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* IEEE ICASSP'00, p. 85-88, Istambul, 2000.

G. Viardot, A. Coatanhay, R. Lengellé, C. Richard, L. Staner, A. Muzet, J.-P. Macher. Fusion d'avis divergents d'experts. *Proc. Colloque AcM/MdA,* Grenoble, 2000.

A. Coatanhay, C. Richard, R. Lengellé, A. Muzet, J.-P. Macher. Time-frequency detection of K-complexes. Methodology and performance estimation. *Proc. WFSRS 3rd International Congress,* Dresde, 1999.

C. Richard, R. Lengellé. Sur le contrôle de la complexité des détecteurs opérant dans le domaine temps-fréquence par le biais de la fonction de paramétrisation. *Proc. Colloque GRETSI,* p. 905-908, Vannes, 1999.

C. Richard, R. Lengellé. On the dimension of the discrete Wigner-Ville transform range space. Application to time-frequency-based detectors design. *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis,* IEEE TFTS'98, p. 5-8, Pittsburgh, PA, 1998.

C. Richard, R. Lengellé. Two algorithms for designing optimal reduced-bias data-driven time-frequency detectors. *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis,* IEEE TFTS'98, p. 601-604, Pittsburgh, PA, 1998.

C. Richard, R. Lengellé. Structural risk minimization for reduced-bias time-frequency-based detectors design. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* IEEE ICASSP'98, p. 2397-2400, Seattle, WA, 1998.

C. Richard, R. Lengellé. Une nouvelle approche pour la détection linéaire optimale dans le plan temps-fréquence. *Proc. Colloque GRETSI,* p. 659-662, Grenoble, 1997.

C. Richard, R. Lengellé. Joint time and time-frequency optimal detection. *Proc. IEEE-UK Symposium on Applications of Time-Frequency and Time-Scale Methods,* IEEE UK-TFTS'97, p. 29-32, Coventry, 1997.

C. Richard, R. Lengellé. Fast implementation of time-frequency representations modified by the reassignment method. *Proc. International Conference on Signal Processing,* ICSP'96, p. 343-346, Beijing, 1996.

C. Richard, R. Lengellé. Recursive implementation of some time-frequency representations. *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis,* IEEE TFTS'96, p. 313-316, Paris, 1996.

A. Bardot, C. Richard, M. Masson. Analyse de la perception acoustique de véhicules. *Proc. Colloque SIA,* Bruxelles, 1995.

Deuxième partie

Synthèse des travaux scientifiques

Introduction

La théorie statistique de la détection a pour objectif de mener à une prise de décision optimale parmi deux alternatives possibles étant donnée une réalisation d'un processus stochastique. En particulier, dans le cadre de la détection d'un signal noyé dans un bruit, les décisions envisageables consistent à opter pour la validité de l'une ou l'autre des hypothèses suivantes, notées conventionnellement H_0 et H_1 : « l'observation $\boldsymbol{x}$ n'est constituée que de bruit » ou « le signal attendu $\boldsymbol{s}$ est présent dans l'observation $\boldsymbol{x}$ ». A ce problème, on associe généralement une partition $\{\mathcal{X}_0, \mathcal{X}_1\}$ de l'espace des observations $\mathcal{X}$, ainsi qu'une fonction booléenne $d(\boldsymbol{x})$ appelée *test de détection* telle que :

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad \boldsymbol{x} \in \mathcal{X}_1 \qquad (H_1 \text{ supposée vraie}) \\ 0 & \text{si} \quad \boldsymbol{x} \in \mathcal{X}_0 \qquad (H_0 \text{ supposée vraie}). \end{cases}$$

Aussi, la résolution d'un problème de détection est-elle équivalente à la recherche d'une partition de $\mathcal{X}$ qui soit optimale au sens d'un critère préalablement défini. Ce dernier s'inscrit dans une stratégie et peut consister, par exemple, à minimiser le coût moyen d'une décision, ou encore à maximiser la probabilité de détection du signal $\boldsymbol{s}$ pour une probabilité de fausse alarme bornée supérieurement. L'obtention d'un test optimum au sens de ces critères nécessite toutefois une connaissance au moins partielle des lois de vraisemblance conditionnelles de l'observation. Celui-ci repose en effet sur la comparaison du rapport de vraisemblance, éventuellement généralisé si les paramètres des lois sont inconnus, à un seuil. On parle alors de *détection à structure libre*, le détecteur n'étant assujetti à aucune contrainte structurelle mais résultant du choix d'un critère.

Pour une classe assez large de problèmes, l'élaboration d'un modèle probabiliste adéquat n'est cependant pas toujours chose facile. Lorsqu'une expertise des phénomènes observés est envisageable, il peut alors être intéressant de recueillir un ensemble de données pour lesquelles un ou plusieurs experts ont fourni un étiquetage. On note celui-ci

$$\mathcal{A}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\},$$

où chaque étiquette binaire y_k représente la décision idéale à reproduire pour l'observation $\boldsymbol{x}_k$ associée. La constitution d'une telle source d'information, appelée *base d'apprentissage*, nécessite de prendre quelques précautions. Il est ainsi souhaitable qu'elle soit représentative des distributions de probabilité conditionnelle, et qu'elle reflète si possible les probabilités *a priori* des hypothèses en compétition. Néanmoins, l'approche optimale évoquée précédemment demeure inopérante dans un tel contexte, compte tenu de la relative pénurie d'informations à laquelle on est astreint. Une alternative classiquement envisagée consiste à rechercher, au sein d'une famille de détecteurs $\mathcal{D} = \{d(\boldsymbol{x}, \theta) \,:\, \theta \in \Theta\}$ donnée, une solution réalisant la meilleure approximation de y au sens d'une fonctionnelle de risque telle que la probabilité d'erreur. Simple en apparence, cette approche qualifiée de *détection à structure imposée* suppose toutefois que l'on réponde de façon

satisfaisante aux questions qui suivent, après s'être interrogé sur la qualité attendue du résultat par rapport au test du rapport de vraisemblance ou toute autre forme optimum équivalente :

- Comment choisir la classe de détecteurs $\mathcal{D}$?
- Quelles sont les fonctionnelles de risque pertinentes pour le problème traité ? Quelle procédure d'optimisation adopter ?

Le présent mémoire apporte des réponses théoriques et pratiques à ces questions en proposant une méthodologie pour la synthèse de détecteurs à partir d'une base d'exemples. Elle fait appel à la théorie des noyaux reproduisants pour l'élaboration de détecteurs linéaires généralisés dans des espaces transformés de dimension importante, voire infinie, sans qu'aucun calcul n'y soit effectué explicitement. Elle a également recours à l'optimisation du meilleur critère du second ordre pour le problème traité, après s'être assuré que de telles mesures de performance ne constituent en rien un obstacle dans la quête du rapport de vraisemblance. Pour une meilleure prise en compte de phénomènes tels que la malédiction de la dimensionnalité, l'approche proposée s'appuie de plus sur la théorie de l'apprentissage. Ceci lui permet d'offrir des garanties sur les performances en généralisation des détecteurs obtenus, qui sont alors en mesure de rivaliser avec les structures de décision phare du moment : les Support Vector Machines. Enfin, cet éclairage mêlant théories statistiques de la décision et de l'apprentissage donne un point de vue original sur la détection dans un espace transformé particulier : le plan temps-fréquence.

La progression suivante est adoptée pour exposer cette méthodologie :

- Le premier chapitre, à caractère introductif et bibliographique, pose les problèmes rencontrés lors de l'élaboration d'un détecteur à structure imposée en adoptant un formalisme propre à la théorie statistique de l'apprentissage.
- Au vu des contraintes à respecter, le deuxième chapitre propose un choix de détecteurs linéaires généralisés dont la consistance universelle constitue un gage de performances. Puis, on s'assure théoriquement de l'intérêt que présente la classe des critères du second ordre pour la synthèse de structures de détection. Le choix d'un critère particulier n'en demeure pas moins un problème crucial, rarement abordé dans la littérature. La méthode d'optimisation proposée répond à cette attente en exhibant le meilleur critère du second ordre pour le problème traité.
- Des considérations sur les espaces de Hilbert à noyau reproduisant sont proposées au chapitre 3. Celles-ci rendent possible la mise en œuvre des stratégies dégagées au cours du chapitre précédent dans des espaces transformés de dimension très importante sans que les calculs y soient explicitement effectués. Afin de conférer aux détecteurs résultants une robustesse accrue vis-à-vis du phénomène de malédiction de la dimensionnalité, deux procédures inspirées de la théorie statistique de l'apprentissage sont ensuite exposées.
- Grâce à la théorie des noyaux reproduisants et à une construction algébrique originale, le chapitre 4 permet de poser un regard nouveau sur la détection dans un espace transformé particulier : le plan temps-fréquence. Puis on est amené à évoquer les diverses formes que peut prendre la distribution de Wigner discrète, et à comparer trois d'entre elles dans un contexte décisionnel. Ce chapitre s'achève enfin par la présentation d'une méthode visant à contrer les effets néfastes du phénomène de malédiction de la dimensionnalité sur les détecteurs opérant dans le domaine temps-fréquence. Celle-ci repose sur les techniques de diffusion récemment introduites dans le domaine de l'analyse des signaux non-stationnaires.

– Le chapitre 5 est consacré à la mise en œuvre de la méthodologie proposée sur des signaux de sommeil. Après une succincte description des macro et microstructures des enregistrements polysomnographiques humains, on procède à l'élaboration d'outils pour la détection de phénomènes particuliers émaillant l'activité cérébrale nocturne : le complexe K et la phase d'activation transitoire.

Enfin, ce mémoire s'achève sur un bilan des activités de recherches que je mène depuis 1995, et présente de nouvelles perspectives dans les domaines de la détection à structure imposée et de l'analyse temps-fréquence.

Notations

$\mathrm{Re}\{\cdot\}$, $\mathrm{Im}\{\cdot\}$	parties réelle et imaginaire
$\dim\{\cdot\}$	dimension d'un espace vectoriel
$\mathrm{E}\{\cdot\}$	espérance mathématique
$\mathrm{cov}\{\cdot\}$	covariance
$\mathcal{H}$	espace de Hilbert
$\mathcal{X}$	espace des observations
$\mathcal{T}$	espace transformé
n	nombre d'observations $\boldsymbol{x}$ disponibles pour l'apprentissage
l	nombre d'échantillons constituant chaque observation $\boldsymbol{x}$
h	dimension de Vapnik-Chervonenkis
ω_i	classe i, avec $i \in \{0, 1\}$
$\mathcal{A}_n$	base d'apprentissage
$\boldsymbol{x} = (x[1] \dots x[l])^t$	vecteur ou observation
$\boldsymbol{X} = (X[1] \dots X[l])^t$	vecteur aléatoire
$\boldsymbol{m}_i = \mathrm{E}\{\boldsymbol{X}\|\omega_i\}$	espérance conditionnelle de $\boldsymbol{X}$
$\boldsymbol{\Sigma}_i = \mathrm{cov}\{\boldsymbol{X}\|\omega_i\}$	covariance conditionnelle de $\boldsymbol{X}$
$p(\boldsymbol{x}\|\omega_i)$	densité de probabilité conditionnelle
$p(\omega_i\|\boldsymbol{x})$	probabilité *a posteriori* de ω_i sachant $\boldsymbol{x}$
$d(\boldsymbol{x}, \theta)$	détecteur
$\mathcal{D} = \{d(\boldsymbol{x}, \theta) \,:\, \theta \in \Theta\}$	classe de détecteurs
$\lambda(\boldsymbol{x})$, λ_0	statistique et seuil de détection
$\eta_i = \mathrm{E}\{\lambda(\boldsymbol{X})\|\omega_i\}$	espérance conditionnelle de la statistique $\lambda(\boldsymbol{X})$
$\sigma_i^2 = \mathrm{var}\{\lambda(\boldsymbol{X})\|\omega_i\}$	variance conditionnelle de la statistique $\lambda(\boldsymbol{X})$
$J(d)$	fonctionnelle de risque
$J_{emp}(d)$	fonctionnelle de risque empirique
$P_e(d)$	probabilité d'erreur
$P_{emp}(d)$	probabilité d'erreur empirique
$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2)$	noyau reproduisant appliqué aux observations $\boldsymbol{x}_1$ et $\boldsymbol{x}_1$
$\boldsymbol{W}_{x_1 x_2}^{(C)}$	distribution de Wigner discrète classique
$\boldsymbol{W}_{x_1 x_2}^{(R)}$	R-distribution
$\boldsymbol{W}_{x_1 x_2}^{(S)}$	S-distribution

Chapitre 1

Apprentissage de règles de décision à structure imposée

1.1 Introduction et définitions

En théorie statistique de la décision, l'élaboration d'une règle de décision repose sur la connaissance éventuellement partielle d'un modèle probabiliste des observations, et le choix d'un critère de performance. L'application de la règle considérée à toute observation $\boldsymbol{x}$ de $\mathcal{X}$ fournit alors une décision d_i en faveur d'une hypothèse H_i donnée. Cet éclairage conduit très naturellement à associer à ce problème, celui de la recherche d'une partition $\mathcal{X}_0, \ldots, \mathcal{X}_{m-1}$ de $\mathcal{X}$ optimisant le critère de performance préalablement sélectionné. Aussi, la décision d_i est-elle prise pour toute observation $\boldsymbol{x}$ de l'ensemble $\mathcal{X}_i$. Dans le domaine de la reconnaissance des formes, la connaissance d'un modèle probabiliste est remplacée par celle d'un ensemble d'apprentissage $\mathcal{A}_n$. Là encore, l'élaboration d'une règle de décision consiste à rechercher une partition de l'espace des observations $\mathcal{X}$ qui soit optimale au sens du critère de performance choisi. On distingue alors deux approches possibles :

- On définit préalablement la structure de la règle de décision, puis on en optimise les paramètres caractéristiques selon le critère retenu. Cette approche est qualifiée de décision à structure imposée.
- On utilise directement l'ensemble d'apprentissage pour la prise de décision. On parle alors de décision non paramétrique.

Le présent mémoire est principalement consacré à l'élaboration de règles de décision à structure imposée à partir d'une base d'exemples. Comme cela sera précisé par la suite, le problème considéré se traduit par la recherche, au sein d'une famille $\mathcal{D} = \{d(\boldsymbol{x}, \theta) : \theta \in \Theta\}$ où Θ désigne un ensemble de paramètres, d'une fonction réalisant la meilleure approximation de y au sens d'une fonctionnelle de risque de la forme

$$J(d) = \int Q(d(\boldsymbol{x}, \theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy, \tag{1.1}$$

où Q représente le coût associé à chaque couple $(\boldsymbol{x}, y)$. La densité de probabilité $p(\boldsymbol{x}, y)$ étant supposée inconnue, la minimisation de $J(d)$ se traduit dans les faits par celle du risque empirique

$$J_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} Q(d(\boldsymbol{x}_k, \theta), y_k) \tag{1.2}$$

calculable sur les données constituant l'ensemble d'apprentissage $\mathcal{A}_n$. Le risque $J(d)$ peut revêtir de multiples formes selon le type de problème auquel on est confronté. En particulier, lorsqu'il s'agit d'élaborer une structure de décision de probabilité d'erreur minimale, il s'exprime ainsi

$$P_e(d) = \int \mathbb{1}_{d(\boldsymbol{x},\theta)\neq y}\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy, \tag{1.3}$$

où $\mathbb{1}$ désigne la fonction indicatrice. Il s'en suit que le risque empirique associé correspond au nombre d'erreurs d'affectation commises par $d(\boldsymbol{x}, \theta)$ sur l'ensemble d'apprentissage

$$P_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{d(\boldsymbol{x}_k,\theta)\neq y_k}. \tag{1.4}$$

Dans la suite de ce chapitre, nous allons nous intéresser à la validité du principe d'induction consistant à minimiser la probabilité d'erreur empirique P_{emp} pour élaborer une structure de décision à partir d'une base d'exemples, et à son lien avec la minimisation de la probabilité d'erreur P_e. Mais avant, il convient de préciser les contraintes auxquelles doit satisfaire l'ensemble d'apprentissage.

1.1.1 Constitution de l'ensemble d'apprentissage

Soit $\mathcal{A}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ un ensemble d'apprentissage de taille n, où $\boldsymbol{x}_k \in \mathcal{X}$ représente l'observation k supposée de dimension l, et $y_k \in \{0, \ldots, m-1\}$ constitue la décision associée à $\boldsymbol{x}_k$ résultant du processus d'expertise. Ces données sont des réalisations de n couples aléatoires $(\boldsymbol{X}_k, Y_k)$ indépendants, identiquement distribués suivant la densité de probabilité $p(\boldsymbol{x}, y)$ définie sur $\mathbb{R}^l \times \{0, \ldots, m-1\}$, que l'on suppose inconnue. A chaque hypothèse est associée une classe, notée ω. Ainsi, lorsque l'hypothèse H_i est vérifiée pour l'observation $\boldsymbol{x}$, ce que signifie $y = i$, on pose $\boldsymbol{x} \in \omega_i$. La constitution d'un ensemble d'apprentissage nécessite de prendre quelques précautions [Len02] :

- il doit être exhaustif ;
- il doit si possible refléter les probabilités *a priori* des hypothèses en compétition ;
- il doit être représentatif des distributions de probabilité conditionnelle.

La première contrainte n'est pas toujours aisée à vérifier. En effet, si l'on considère par exemple un problème de décision relatif à la surveillance d'un système critique du point de vue de la sécurité, on ne peut disposer en général de données correspondant précisément aux états de dysfonctionnement les plus sérieux. En conséquence, il est important de prévoir une ou plusieurs classes supplémentaires auxquelles seront affectées les observations dont on estimera qu'elles sont sujettes à discussion. On parle de rejet de distance lorsque l'observation n'appartient raisonnablement à aucune classe de $\mathcal{A}_n$, et de rejet d'ambiguïté lorsqu'on estime en revanche que l'observation peut être assignée à plusieurs classes. En ce qui concerne la deuxième contrainte, il n'est pas nécessaire que la constitution de $\mathcal{A}_n$ soit en conformité avec les probabilités *a priori*. Afin de compenser une mauvaise représentativité des classes en compétition, il suffit en effet d'introduire les coûts adéquats dans le critère de performance. Rappelons toutefois que la quantité d'information statistique relative à une classe est d'autant plus grande que l'effectif correspondant est élevé.

Le choix de l'espace de représentation $\mathcal{X}$ est largement dépendant de l'application considérée, et fait en général appel au recueil de l'expertise. Les composantes de $\boldsymbol{X}$ sont appelées variables

ou descripteurs. Elles peuvent être sélectionnées en essayant de prendre en compte une propriété d'invariance attendue au regard de transformations données, ou encore de maximiser une mesure de quantité d'information discriminante. Cette étape d'extraction d'information conditionne largement les performances du système de décision et, lorsqu'elle est bien réalisée, en assure la robustesse. Hélas, il n'existe pas de méthode garantissant l'optimalité du choix de l'espace de représentation. Il est simplement possible de comparer *a posteriori* plusieurs solutions au vu des performances obtenues, de combiner les composantes de $\boldsymbol{X}$ [Fuk90], ou encore d'en sélectionner le meilleur sous-ensemble [Nar77].

Dans le cas des signaux échantillonnés, les échantillons temporels sont supposés constituer une statistique suffisante pour la prise de décision. Le bon sens suggère en conséquence d'élaborer directement la structure de décision sur l'ensemble des échantillons de l'observation, supposés au nombre de l. Ce choix peut cependant s'avérer désastreux si l'on se trouve confronté à la malédiction de la dimensionnalité, phénomène largement décrit dans la littérature [Bel61]. En effet, la densité moyenne dans $\mathcal{X}$ des éléments de $\mathcal{A}_n$ est proportionnelle à n élevé à la puissance $\frac{1}{l}$. Il s'en suit que la taille de l'ensemble d'apprentissage doit croître selon une loi exponentielle avec la dimension de l'espace de représentation si l'on souhaite conserver la même densité d'observations. A titre d'illustration, supposons que pour un problème donné dans $\mathbb{R}$, un ensemble d'apprentissage de cardinal 10 s'avère suffisant. Transposé dans $\mathbb{R}^q$, le même problème nécessite environ 10^q observations pour que la densité demeure comparable. Il convient de reconnaître que ceci devient rapidement irréaliste, même pour de faibles valeurs de q. Cette difficulté et les remèdes que l'on peut y apporter seront étudiés avec d'avantage de détails dans la suite de ce document.

1.1.2 Position du problème

Dans un but didactique, on va à présent se concentrer sur la décision à hypothèse binaire, dite *détection*, sans que cela nuise pour autant au caractère général de la discussion. Soit $(\boldsymbol{X}, Y)$ un couple aléatoire distribué selon la densité de probabilité $p(\boldsymbol{x}, y)$ définie sur $\mathbb{R}^l \times \{0, 1\}$, supposée inaccessible. Le problème considéré consiste en la recherche, au sein d'une classe $\mathcal{D}$ donnée, d'une fonction $d(\boldsymbol{X})$ minimisant la probabilité d'erreur définie par

$$P_e(d) = p(d(\boldsymbol{X}) \neq Y). \tag{1.5}$$

Le détecteur de Bayes d^* avec pour coûts $\{0, 1\}$ optimise ce critère de performance, qui prend alors pour valeur P_e^*, soit

$$d^* = \arg\min_d p(d(\boldsymbol{X}) \neq Y) \tag{1.6}$$

$$P_e^* = p(d^*(\boldsymbol{X}) \neq Y). \tag{1.7}$$

On montre que la règle de décision d^* s'écrit

$$d^*(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad p(\omega_1|\boldsymbol{x}) \geq p(\omega_0|\boldsymbol{x}) \\ 0 & \text{si} \quad p(\omega_1|\boldsymbol{x}) < p(\omega_0|\boldsymbol{x}), \end{cases} \tag{1.8}$$

ou encore, de manière équivalente

$$d^*(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad p(\omega_1|\boldsymbol{x}) \geq 1/2 \\ 0 & \text{si} \quad p(\omega_1|\boldsymbol{x}) < 1/2, \end{cases} \tag{1.9}$$

avec ω_0 et ω_1 désignant les classes associées aux hypothèses en compétition. Disposant dans le cas présent d'une base d'apprentissage $\mathcal{A}_n$, le problème posé consiste à minimiser la probabilité d'erreur conditionnelle

$$P_e(d_n; \mathcal{A}_n) = p(d_n(\boldsymbol{X}, \theta) \neq Y | \mathcal{A}_n), \tag{1.10}$$

qui constitue une variable aléatoire dépendant de $\mathcal{A}_n$. Dans cette expression, $d_n(\boldsymbol{X}, \theta)$ désigne un détecteur de $\mathcal{D}$ déterminé à partir de la base d'exemples disponible, au terme d'un processus appelé *apprentissage*. La probabilité d'erreur caractérisant cette structure est donnée par

$$P_e(d_n) = \mathrm{E}\{P_e(d_n; \mathcal{A}_n)\}. \tag{1.11}$$

Force est de constater que la résolution d'un tel problème est inenvisageable en pratique dans la mesure où le calcul de P_e nécessite la connaissance de la loi de probabilité conjointe $p(\boldsymbol{x}, y)$. Dans les faits, la minimisation de ce critère se traduit par celle de la probabilité d'erreur empirique (1.4), dont l'écriture repose uniquement sur les éléments de $\mathcal{A}_n$. Soit $d_n^*(\boldsymbol{X}, \theta)$ le détecteur optimal de $\mathcal{D}$ au sens de P_{emp}. Comme l'illustre la figure 1.1, l'*erreur de modélisation* définie par l'expression (1.12) permet de caractériser le comportement du récepteur d_n^*.

$$E_{mod}(d_n^*) = P_e(d_n^*) - P_e^*, \tag{1.12}$$

que l'on peut réécrire comme suit

$$E_{mod}(d_n^*) = \underbrace{\left(P_e(d_n^*) - \inf_{d \in \mathcal{D}} P_e(d) \right)}_{E_{est}} + \underbrace{\left(\inf_{d \in \mathcal{D}} P_e(d) - P_e^* \right)}_{E_{app}}. \tag{1.13}$$

La reformulation de E_{mod} met en évidence l'influence de deux sources d'erreur de natures différentes : l'*erreur d'estimation* et l'*erreur d'approximation*. La première, notée E_{est}, reflète conjointement la pertinence du critère de performance sélectionné et l'efficacité du processus d'apprentissage, étant donnés un ensemble de données $\mathcal{A}_n$ et une famille de tests de détection $\mathcal{D}$. La seconde, notée E_{app}, ne dépend que du choix initial de $\mathcal{D}$.

La minimisation de l'erreur de modélisation repose sur la recherche d'un compromis entre ces deux termes antagonistes : l'augmentation du nombre de tests de $\mathcal{D}$ conduit à un accroissement de E_{est} tandis que E_{app} décroît, et inversement. Ce phénomène, qui a été largement étudié dans la littérature depuis le travail fondateur de Vapnik et Chervonenkis sur la théorie statistique de l'apprentissage [Vap71], sera décrit au cours de ce chapitre. Afin de se convaincre de son existence, supposons que $\mathcal{D}$ regroupe l'ensemble des fonctions mesurables $d(\boldsymbol{X})$ de $\mathbb{R}^l$ dans $\{0, 1\}$. Dans ces circonstances, et quelle que soit la densité de probabilité $p(\boldsymbol{x}, y)$, on peut trouver un test de détection dont la probabilité d'erreur empirique est nulle, et dont le résultat est quelconque pour toute observation $\boldsymbol{x}$ n'appartenant pas à $\mathcal{A}_n$:

$$d_n(\boldsymbol{X}) = \begin{cases} y_i \text{ si } \boldsymbol{X} = \boldsymbol{x}_i, \text{ pour tout } i = 1, \ldots, n \\ 0 \text{ ou } 1 \text{ de façon aléatoire, sinon.} \end{cases} \tag{1.14}$$

Ce détecteur conduit ainsi à l'annulation de l'erreur d'approximation compte tenu du choix de $\mathcal{D}$, au détriment de l'erreur d'estimation. On notera que le cas dual consistant à maximiser E_{app} et à annuler E_{est} peut être observé en restreignant $\mathcal{D}$ à un test de détection unique.

La suite de ce chapitre est consacrée aux problèmes rencontrés lors de la conception d'un détecteur à structure imposée. Le recours à des résultats de la théorie statistique de l'apprentissage pour apporter des éléments de réponses au problème de l'optimisation conjointe des erreurs d'estimation et d'approximation constitue l'une des originalités de ce travail.

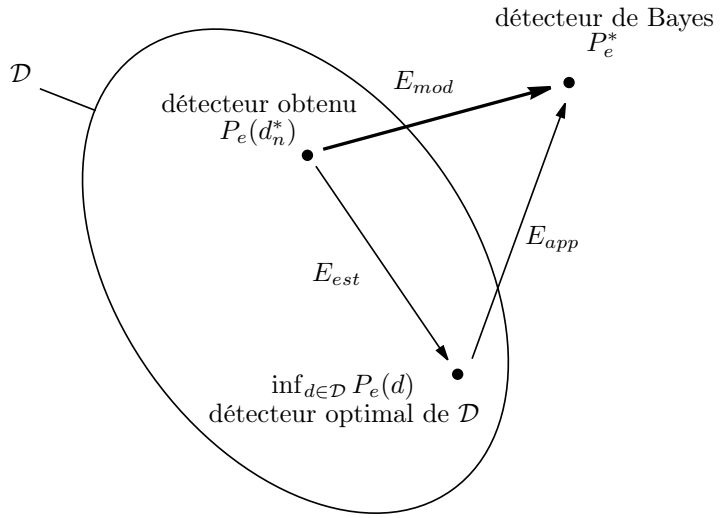

FIG. 1.1 : Recherche d'un détecteur d_n^* dans une classe $\mathcal{D}$ donnée par optimisation de la probabilité d'erreur empirique sur une base d'exemples $\mathcal{A}_n$.

1.1.3 Consistance d'une règle de décision

Il est généralement impossible d'atteindre les performances du détecteur de Bayes avec un détecteur à structure imposée élaboré à partir d'un ensemble d'apprentissage de taille finie n. Cependant, on peut espérer qu'il existe dans la classe $\mathcal{D}$ considérée, une suite $\{d_n^*(\boldsymbol{X},\theta)\}_{n>0}$ de détecteurs optimaux au sens du critère retenu telle que la probabilité d'erreur $P_e(d_n^*)$ puisse être rendue arbitrairement proche de P_e^* lorsque n tend vers l'infini. La notion de consistance d'une règle de décision repose sur cette idée, et définit les types de convergence de l'erreur de modélisation vers zéro.

Définition 1.1.1. *Étant donnée une base d'exemples $\mathcal{A}_n$, une suite $\{d_n^*(\boldsymbol{X},\theta)\}_{n>0}$ de détecteurs optimaux au sens d'un critère donné est dite consistante pour une loi $p(\boldsymbol{x},y)$ si*

$$\lim_{n\to\infty} \mathrm{E}\{P_e(d_n^*;\mathcal{A}_n)\} = P_e^*. \tag{1.15}$$

On dit qu'elle est fortement consistante si

$$\lim_{n\to\infty} P_e(d_n^*;\mathcal{A}_n) = P_e^* \tag{1.16}$$

avec une probabilité égale à 1.

La propriété de consistance assure que la probabilité d'erreur d'un détecteur obtenu par optimisation d'un critère sur une base d'apprentissage $\mathcal{A}_n$ converge vers la borne inférieure fournie par le détecteur de Bayes, lorsque n tend vers l'infini. La suite $\{d_n^*(\boldsymbol{X},\theta)\}_{n>0}$ est dite fortement consistante si cette propriété demeure vraie quelle que soit la réalisation de $\mathcal{A}_n$. On peut enfin distinguer le cas où la propriété de consistance n'est vérifiée que pour une loi $p(\boldsymbol{x},y)$ donnée,

du cas où elle reste vraie indépendamment de celle-ci. La définition suivante prend en compte le caractère parfois universel de la consistance.

Définition 1.1.2. *La suite* $\{d_n^*(\boldsymbol{X}, \theta)\}_{n>0}$ *est dite universellement (fortement) consistante si elle est (fortement) consistante pour toute loi de probabilité* $p(\boldsymbol{x}, y)$.

Cette dernière propriété a été observée pour la première fois en 1977 par Stone [Sto77] dans le cadre de la méthode des *k plus proches voisins*, à la condition que le paramètre k croisse moins vite que la taille n de la base d'apprentissage. Depuis, il a été démontré que d'autres règles de décision y satisfont, en particulier les structures reposant sur une fonction noyau régulière [Dev89] et certains types de détecteurs linéaires généralisés [Dev96] sur lesquels on reviendra.

1.2 Éléments de théorie de l'apprentissage

1.2.1 Consistance du principe d'induction

La présente partie est consacrée à la minimisation de l'erreur d'estimation. Le succès de cette tâche repose sur la *consistance du principe d'induction* visant à substituer à la probabilité d'erreur la probabilité d'erreur empirique. Il convient de noter que cette notion s'inscrit dans un cadre plus large que celui-ci de ces deux critères de performance. En effet, elle est relative aux liens unissant toute fonctionnelle de risque J et le risque empirique J_{emp} associé, dont on rappelle les définitions respectives

$$J(d) = \int Q(d(\boldsymbol{x}, \theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy, \tag{1.17}$$

$$J_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} Q(d(\boldsymbol{x}_k, \theta), y_k) \tag{1.18}$$

où Q représente le coût associé à chaque couple $(\boldsymbol{x}, y)$. La consistance du principe d'induction étudié est définie ainsi :

Définition 1.2.1. *Le principe de minimisation du risque empirique est consistant pour un coût* Q*, une famille de fonctions* $\mathcal{D} = \{d(\boldsymbol{x}, \theta) : \theta \in \Theta\}$ *et une distribution* $p(\boldsymbol{x}, y)$ *si, appliqué à chaque taille* n *d'échantillon, il engendre une suite de fonctions* $\{d_n^*(\boldsymbol{x}, \theta) : \theta \in \Theta\}_{n>0}$ *telle que*

$$J(d_n^*) \underset{n\to\infty}{\xrightarrow{p}} \inf_{d\in\mathcal{D}} J(d) \tag{1.19}$$

$$J_{emp}(d_n^*) \underset{n\to\infty}{\xrightarrow{p}} \inf_{d\in\mathcal{D}} J(d). \tag{1.20}$$

Comme l'illustre la figure 1.2, cette définition signifie que la suite des risques et risques empiriques convergent en probabilité vers la même limite, celle-ci étant la plus faible valeur possible du risque. Bien évidemment, cette notion de consistance ne doit pas être confondue avec celle qui a été présentée au cours de la section précédente, relative aux règles de décision. Fort de cette définition, on peut énoncer un théorème-clé de la théorie statistique de l'apprentissage [Vap95].

Théorème 1.2.1. *Soit* Q *une fonction coût et* $\mathcal{D} = \{d(\boldsymbol{x}, \theta) : \theta \in \Theta\}$ *une famille de fonctions vérifiant la condition suivante*

$$A \leq \int Q(d(\boldsymbol{x}, \theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy \leq B. \tag{1.21}$$

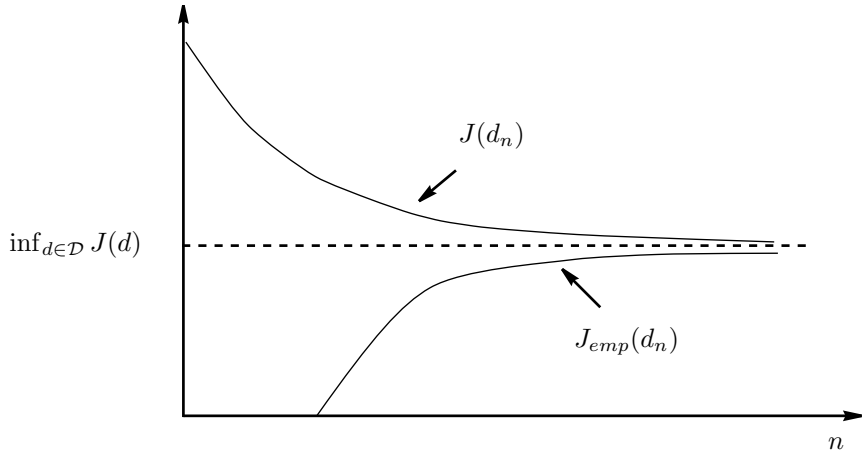

FIG. 1.2 : Consistance du principe d'induction reposant sur la minimisation du risque empirique.

Le principe de minimisation du risque empirique est consistant si et seulement si J_{emp} converge uniformément vers J au sens suivant :

$$\lim_{n\to\infty} p(\sup_{d\in\mathcal{D}}\{J(d) - J_{emp}(d)\} > \varepsilon) = 0, \qquad \forall\varepsilon > 0. \tag{1.22}$$

La borne supérieure figurant dans l'inégalité (1.22) indique que la consistance du principe de minimisation du risque empirique est conditionnée par la fonction de $\mathcal{D}$ dont le comportement est le plus défavorable. Pour cela, la théorie de l'apprentissage considérée est parfois qualifiée *d'analyse au pire cas*. Il est à noter que le théorème 1.2.1 n'est pas applicable en pratique puisque l'évaluation de J nécessite la connaissance de la densité de probabilité $p(\boldsymbol{x}, y)$. Afin de pallier cette difficulté, d'autres conditions nécessaires et suffisantes ont été originellement proposées par Vapnik et Chervonenkis, celles-ci mettant en scène des grandeurs plus aisément calculables.

Par soucis de clarté, on considère dans la suite de cette section que le coût Q prend la forme d'une fonction indicatrice, soit

$$Q(d(\boldsymbol{x},\theta), y) = \mathbb{1}_{d(\boldsymbol{x},\theta)\neq y} \triangleq \begin{cases} 0 & \text{si} \quad y = d(\boldsymbol{x},\theta) \\ 1 & \text{si} \quad y \neq d(\boldsymbol{x},\theta), \end{cases} \tag{1.23}$$

conférant ainsi à J et J_{emp} les qualités de probabilité d'erreur et probabilité d'erreur empirique, respectivement. Une condition suffisante, longtemps restée en vigueur, pour que la relation (1.22) soit satisfaite est que $\mathcal{D}$ soit de cardinal fini [Hof63]. Il s'agit néanmoins d'une exigence très restrictive que ne vérifie pas, par exemple, la famille des structures de détection linéaires. Il a fallu attendre les travaux de Vapnik et Chervonenkis pour que cette propriété soit étendue à certaines classes de cardinalité infinie [Vap71]. Ces auteurs ont en effet démontré une série de résultats fondateurs pour la théorie statistique de l'apprentissage, dont l'un d'eux assure la consistance du principe de minimisation du risque empirique si et seulement si

$$\lim_{n\to\infty} \frac{H_{\mathcal{D}}(n)}{n} = 0, \tag{1.24}$$

où $H_{\mathcal{D}}(n)$ vaut $\mathrm{E}\{\ln N_{\mathcal{D}}(\mathcal{A}_n)\}$. Parce que $N_{\mathcal{D}}(\mathcal{A}_n)$ désigne un nombre d'états définis ci-après, la quantité $H_{\mathcal{D}}(n)$ est nommée à bon escient *VC-entropie* de la classe $\mathcal{D}$. Étant donné θ appartenant

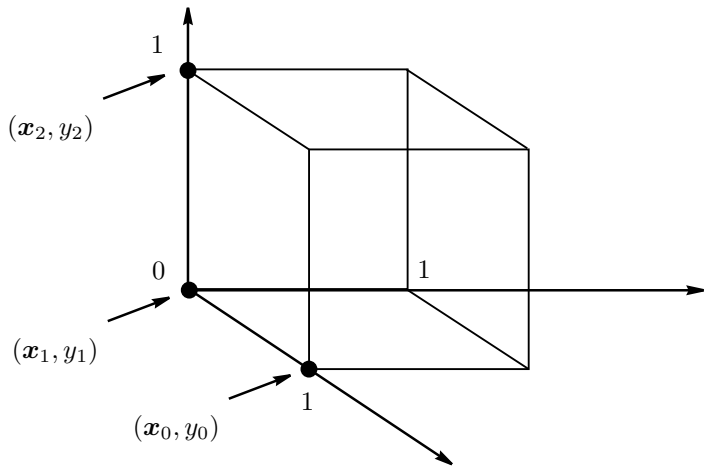

FIG. 1.3 : Représentation des coûts $\{0, 1\}$ associés aux éléments de l'ensemble d'apprentissage $\mathcal{A}_n$ par le biais d'un sous-ensemble de sommets de l'hypercube unité de dimension n, étant donnée une structure de détection $d(\boldsymbol{x}, \theta)$ de $\mathcal{D}$.

à Θ, le calcul du coût (1.23) a pour effet d'associer la valeur 0 ou 1 à chaque paire $(\boldsymbol{x}_k, y_k)$ de l'ensemble d'apprentissage $\mathcal{A}_n$. Comme l'illustre la figure 1.3, l'étiquetage ainsi obtenu peut donc être représenté par un sous-ensemble de n sommets de l'hypercube unité de dimension n. Le paramètre $N_{\mathcal{D}}(\mathcal{A}_n)$ évoqué ci-dessus désigne le nombre de ces configurations distinctes que peut engendrer la classe de détecteurs $\mathcal{D}$, et constitue ainsi un indicateur de la diversité des solutions réalisables. Force est de constater que le calcul de la VC-entropie d'une classe de détecteurs requiert toutefois la connaissance de $p(\boldsymbol{x}, y)$, supposée inaccessible ici. Afin de pallier cette difficulté, a été introduite la *fonction de croissance*

$$G_{\mathcal{D}}(n) = \ln \sup_{\mathcal{A}_n} \{N_{\mathcal{D}}(\mathcal{A}_n)\}, \tag{1.25}$$

qui constitue un majorant de $H_{\mathcal{D}}(n)$ ne nécessitant pas le calcul d'une espérance mathématique. Tout comme la VC-entropie, cette grandeur intervient dans une condition nécessaire et suffisante de consistance du principe de minimisation du risque empirique, qui est :

$$\lim_{n \to \infty} \frac{G_{\mathcal{D}}(n)}{n} = 0. \tag{1.26}$$

L'usage de la fonction de croissance suppose cependant la connaissance de tous les $(\boldsymbol{x}_k, y_k)$ possibles, ce qui est difficilement concevable en pratique. En revanche, on peut démontrer le résultat suivant, attribué indépendamment à Vapnik [Vap71] et Sauer [Sau72], mais connu sous le nom de *lemme de Sauer* :

Théorème 1.2.2. *Toute fonction de croissance vérifie l'égalité*

$$G_{\mathcal{D}}(n) = n \ln 2, \tag{1.27}$$

ou est bornée selon l'inégalité suivante

$$G_{\mathcal{D}}(n) \leq h \left(\ln \frac{n}{h} + 1 \right). \tag{1.28}$$

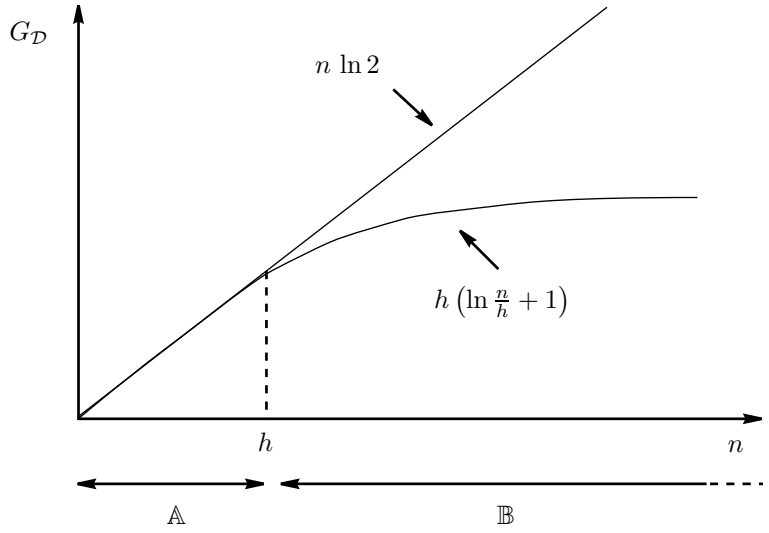

FIG. 1.4 : La fonction de croissance est soit linéaire, soit bornée par une fonction logarithmique. La partie $\mathbb{A}$ de la représentation correspond au cas où $\sup_{\mathcal{A}_n}\{N_{\mathcal{D}}(\mathcal{A}_n)\} = 2^n$, signifiant que les détecteurs de la classe $\mathcal{D}$ peuvent réaliser toutes les dichotomies d'un ensemble de données de taille n. Il n'en est pas ainsi sur la partie $\mathbb{B}$, lorsque $n > h$ avec h fini.

Dans cette expression, h désigne un entier tel que, lorsque $n = h$, on a

$$G_{\mathcal{D}}(h) = h \ln 2, \tag{1.29}$$

$$G_{\mathcal{D}}(h+1) < (h+1) \ln 2. \tag{1.30}$$

En d'autres termes, la fonction de croissance est soit linéaire, soit bornée par une fonction logarithmique comme l'illustre la figure 1.4. La rupture de pente de $G_{\mathcal{D}}(n)$, si elle existe, se produit en une valeur h de la variable n appelée *dimension de Vapnik-Chervonenkis*, ou encore *VC-dimension*. Compte tenu du théorème 1.2.2, h est fini à condition que la fonction de croissance associée soit bornée selon la relation (1.28), et infini si cette dernière est en revanche caractérisée par le comportement linéaire (1.27). La dimension de Vapnik-Chervonenkis est une grandeur fondamentale de la théorie statistique de l'apprentissage caractérisant la diversité des solutions réalisables à partir d'une classe de détecteurs $\mathcal{D}$ donnée. Il est important de remarquer qu'elle n'est en rien liée à la théorie des probabilités. Elle exprime une propriété combinatoire dont les corollaires sont indépendants de la loi $p(\boldsymbol{x}, y)$ gouvernant les observations.

Avant de décrire plus précisément la notion de VC-dimension, il convient de revenir à la problématique de la section en cours, que constitue la consistance du principe de minimisation du risque empirique. D'après les éléments apportés jusqu'à présent, on a

$$\frac{H_{\mathcal{D}}}{n} \leq \frac{G_{\mathcal{D}}}{n} \leq \frac{h\left(\ln \frac{n}{h} + 1\right)}{n}, \tag{1.31}$$

pour $n > h$ avec h fini. Il en résulte directement que le caractère fini de la VC-dimension d'une classe de détecteurs est une condition suffisante pour la consistance du principe d'induction

considéré. Il s'avère que cette condition est également nécessaire [Ehr88], ce qui permet d'énoncer le théorème suivant :

Théorème 1.2.3. *Pour que le principe de minimisation du risque empirique soit consistant indépendamment de la distribution de probabilité gouvernant les observations, il faut et il suffit que la classe de détecteurs considérée soit de VC-dimension finie.*

Au delà de ce résultat majeur qui constitue une réponse qualitative au problème de consistance, les travaux précurseurs de Vapnik et Chervonenkis [Vap71] ont également apporté des enseignements quantitatifs relatifs à la vitesse de convergence de la suite des risque empiriques $J_{emp}(d_n)$ vers la borne inférieure $\inf_{d\in\mathcal{D}} J(d)$. En particulier, dans le contexte de la fonction coût unitaire (1.23), ils ont démontré que

$$p(\sup_{d\in\mathcal{D}}\{P_e(d) - P_{emp}(d)\} > \varepsilon) \leq 4 \sup_{\mathcal{A}_{2n}}\{N_{\mathcal{D}}(\mathcal{A}_{2n})\}\, e^{-n\varepsilon^2/8} \leq 4 \left(\frac{2en}{h}\right)^h e^{-n\varepsilon^2/8}, \qquad (1.32)$$

pour toute famille $\mathcal{D}$ de détecteurs. Il est à noter que la deuxième partie de l'inégalité figurant ci-dessus est un conséquence directe de la relation (1.28). Ce résultat généralise ainsi les théorèmes historiques de Kolmogorov-Smirnov au prix d'une complexité conceptuelle accrue inhérente à la notion de VC-dimension. Depuis sa formulation originelle, l'expression (1.32) a connu de nombreux raffinements intervenant au niveau des constantes, dont le lecteur intéressé pourra trouver un aperçu dans [Vay00].

1.2.2 Dimension de Vapnik-Chervonenkis

La dimension de Vapnik-Chervonenkis caractérise la capacité d'apprentissage des tests de détection d'une classe $\mathcal{D}$ donnée. Comme l'illustre la figure (1.4), elle correspond en effet au nombre maximum de données d'apprentissage que les détecteurs de $\mathcal{D}$ peuvent apprendre sans erreur, quel que soit leur étiquetage. On dit dans ce cas que l'ensemble des observations considéré est *pulvérisé*. On peut à présent énoncer la définition de la dimension de Vapnik-Chervonenkis :

Définition 1.2.2. *La dimension de Vapnik-Chervonenkis d'une classe $\mathcal{D}$ donnée est définie par le plus grand nombre d'éléments $\boldsymbol{x}_k$ de l'espace des réalisations $\mathcal{X}$ dont les détecteurs de $\mathcal{D}$ peuvent réaliser toutes les dichotomies.*

Exemple 1.2.1. A titre d'illustration, on considère la classe $\mathcal{D}$ des détecteurs linéaires opérant dans $\mathbb{R}^l$ définis par $d(\boldsymbol{x},\theta) = \Gamma(\sum_{k=1}^{l} \theta_k\, x(k) + \theta_0)$, les paramètres θ_k étant réels et $\Gamma(\cdot)$ désignant la fonction d'Heaviside. La figure 1.5 permet d'établir que la VC-dimension de $\mathcal{D}$ prend respectivement les valeurs 2 et 3, lorsque l vaut 1 puis 2. Ce résultat peut être aisément étendu à $\mathbb{R}^l$, dans le cadre duquel on montre que $h_{\mathcal{D}} = l + 1$. Bien avant d'être formalisé par Vapnik et Chervonenkis, l'importance d'un tel paramètre a été pressentie par Cover dans le cadre des discriminants linéaires [Cov65]. Celui-ci a en effet observé qu'un discriminant linéaire doté d'un nombre de degrés de liberté comparable à la taille de la base d'apprentissage présente en général de piètres performances, celui-ci ne reflétant alors pas nécessairement la structure des données.

Dans le cas général, il convient toutefois de noter que la VC-dimension d'un modèle diffère du nombre de paramètres qui le caractérise, comme l'atteste la situation extrême suivante.

Exemple 1.2.2. On considère la classe des détecteurs $\{d(x,\theta) = \Gamma(\sin(\theta x)) \,:\, \theta \in \mathbb{R}\}$ opérant dans $\mathbb{R}$, où $\Gamma(\cdot)$ représente la fonction de Heaviside. Il est aisé de démontrer que l'on peut réaliser

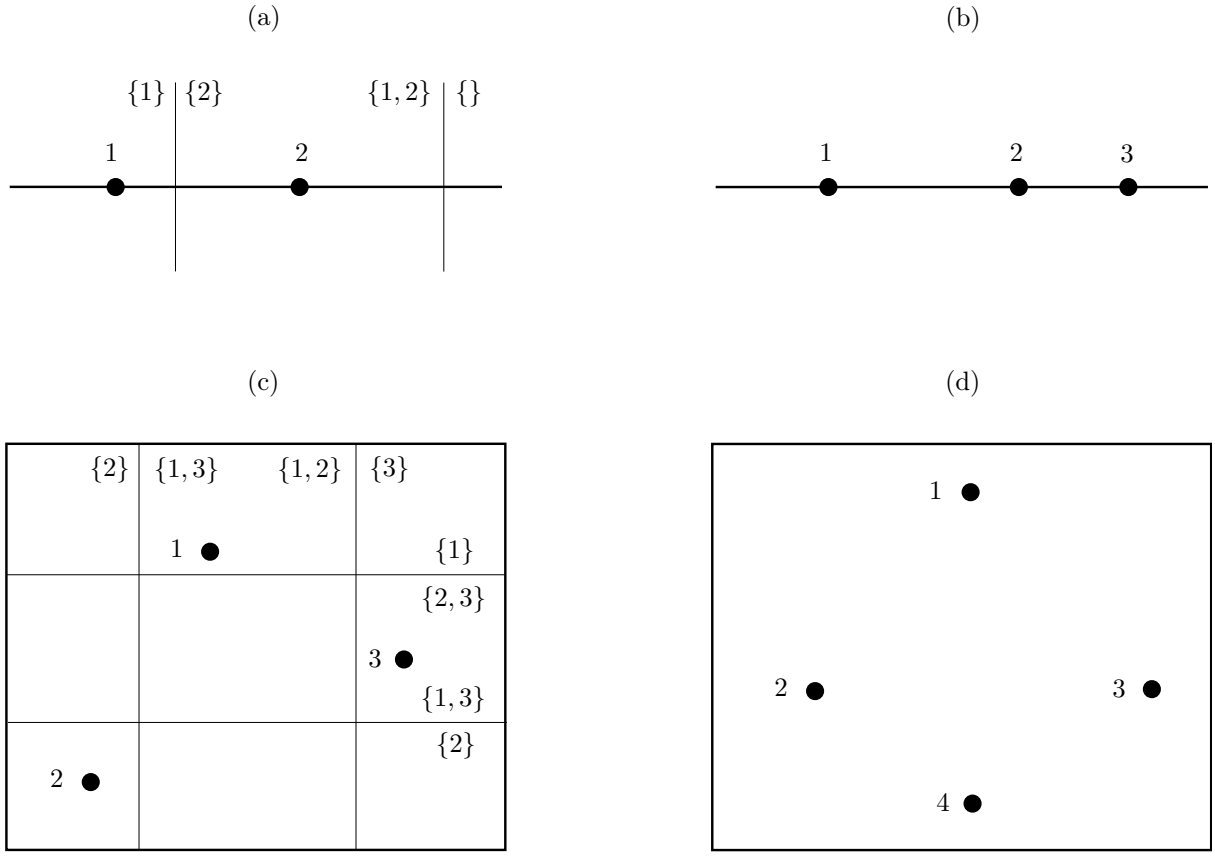

FIG. 1.5 : Détermination de la VC-dimension de la classe $\mathcal{D}$ des discriminants linéaires de $\mathbb{R}^l$. Il est manifeste sur (a) que les discriminants de $\mathbb{R}$ peuvent réaliser toutes les dichotomies de 2 points. En revanche, l'exemple (b) faisant figurer 3 points montre qu'il est impossible d'obtenir les regroupements suivants : $\{1,3\}$ et $\{2\}$. On en conclut donc que la VC-dimension des discriminants linéaires de $\mathbb{R}$ est égale à 2. En considérant les figures (c) et (d), on établit de même que la VC-dimension de la classe des détecteurs linéaires opérant dans $\mathbb{R}^2$ est égale à 3. Il est en effet impossible d'obtenir les regroupements $\{1,4\}$ et $\{2,3\}$ pour l'exemple (d).

une dichotomie idéale de tout ensemble d'apprentissage $\mathcal{A}_n$ grâce à un choix approprié de θ, quel que soit l'étiquetage $\{0,1\}$ des données qui le compose, et quel que soit le cardinal n considéré. Il en résulte que la VC-dimension de cette famille est infinie, malgré le fait qu'un unique paramètre la gouverne.

La section suivante discute de la pertinence du choix d'une famille de détecteurs, compte tenu de sa VC-dimension et du nombre de données d'apprentissage dont on dispose.

1.2.3 Contrôle de la capacité en généralisation

Depuis sa version originelle (1.32), l'inégalité de Vapnik-Chervonenkis a connu de nombreuses évolutions, se traduisant essentiellement par des modifications des termes constants. Toujours dans le contexte de la fonction coût unitaire (1.23), elle prend aujourd'hui la forme suivante

$$p(\sup_{d\in\mathcal{D}}\{P_e(d) - P_{emp}(d)\} > \varepsilon) \leq 4\left(\frac{2en}{h}\right)^h e^{-n\varepsilon^2}. \tag{1.33}$$

A partir de cette expression, il est aisé de déterminer un intervalle de confiance liant la probabilité d'erreur P_e et la probabilité d'erreur empirique P_{emp} pour tout détecteur d'une classe $\mathcal{D}$ donnée. Ainsi, il s'avère que la relation suivante est satisfaite avec une probabilité au moins égale à $1-\varepsilon$

$$P_e(d) \leq P_{emp}(d) + \Phi(n, h, \varepsilon), \tag{1.34}$$

où Φ représente la largeur de l'intervalle de confiance en fonction de n, de la VC-dimension de la famille de détecteurs considérée et du niveau de confiance accordé, selon l'expression

$$\Phi(n, h, \varepsilon) = \sqrt{\frac{h\left(\ln\frac{2n}{h}+1\right) - \ln\frac{\varepsilon}{4}}{n}}. \tag{1.35}$$

Ce résultat indique que la mise en œuvre du principe de minimisation du risque empirique, dont la consistance a été discutée précédemment, est satisfaisante lorsque la quantité $\frac{n}{h}$ est significative. En effet, il s'avère alors que la largeur Φ de l'intervalle de confiance est voisine de zéro, signifiant qu'une valeur minime de la probabilité d'erreur empirique suffit à garantir une probabilité d'erreur faible. Dans le cas contraire, la largeur de l'intervalle de confiance est telle que $P_{emp}(d)$ ne peut donner d'information significative sur $P_e(d)$.

L'unique façon de trouver un compromis satisfaisant entre les deux causes d'erreur que constituent P_{emp} et Φ consiste à contrôler la VC-dimension du modèle, la taille de la base d'exemples étant en général fixée par les conditions expérimentales. Le principe de *minimisation du risque structurel* préconisé par Vapnik pour apporter une solution à ce problème suppose la construction, au sein de la classe $\mathcal{D}$, d'une séquence de sous-ensembles imbriqués $\mathcal{D}_k$

$$\mathcal{D}_1 \subset \ldots \subset \mathcal{D}_k \subset \ldots \subset \mathcal{D}, \tag{1.36}$$

satisfaisant à certaines contraintes techniques énumérées dans [Vap95]. Les VC-dimensions h_k associées, supposées finies, forment alors une suite croissante telle que :

$$h_1 \leq \ldots \leq h_k \leq \ldots \leq h. \tag{1.37}$$

Cette structure étant établie, la phase d'apprentissage est menée en deux étapes comme l'illustre la figure 1.6.

1. Recherche du détecteur d'erreur empirique minimale dans chaque sous-ensemble $\mathcal{D}_k$:

$$d^*_{n,k} = \arg\min_{d\in\mathcal{D}_k} P_{emp}(d). \tag{1.38}$$

2. Sélection du détecteur présentant l'erreur garantie $P_{emp}(d^*_{n,k})+\Phi(n, h_k, \varepsilon)$ la plus favorable :

$$d^*_n = \arg\min_{k\geq 1}\{P_{emp}(d^*_{n,k}) + \Phi(n, h_k, \varepsilon)\}. \tag{1.39}$$

Bien que séduisant, ce principe pose néanmoins quelques problèmes pratiques lors de sa mise en œuvre. En particulier, il nécessite la connaissance de la VC-dimension des ensembles $\mathcal{D}_k$, alors que ce paramètre ne peut être évalué analytiquement que dans les cas les plus simples. De plus, la borne supérieure de P_e que définie l'erreur garantie est généralement surestimée, et le calcul d'un intervalle de confiance plus étroit pose de réelles difficultés. Aussi, une approche généralement préférée consiste à estimer P_e grâce à des techniques de validation croisée ou de ré-échantillonnage par exemple, dont un aperçu est proposé dans la section suivante. On parle alors de *minimisation du risque structurel au sens large.*

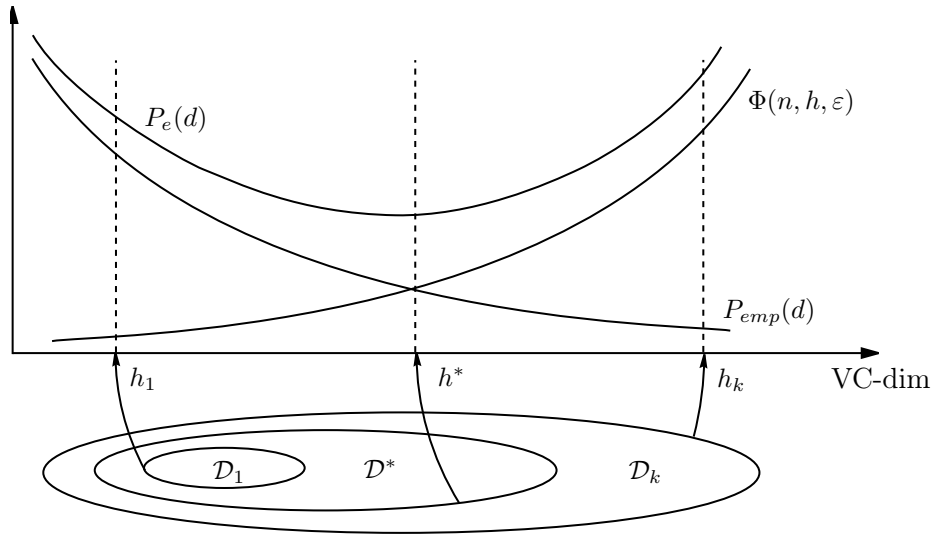

FIG. 1.6 : Principe de minimisation du risque structurel.

1.3 Estimation des performances

La consistance universelle d'une classe de détecteurs constitue un gage de performances. En effet, elle garantit la convergence de la probabilité d'erreur de la structure de décision retenue vers celle du détecteur de Bayes, à mesure que la taille de la base d'apprentissage croît. Aussi satisfaisant que puisse paraître ce résultat, il ne fournit toutefois aucune précision quant à l'écart de performances existant entre ces deux structures. Dans ces circonstances, il apparaît donc nécessaire de pouvoir estimer convenablement la probabilité d'erreur d'un détecteur lorsqu'on ne dispose pour cela que d'un ensemble de données étiquetées. Ce problème est récurrent lorsqu'il s'agit de maîtriser l'influence de la taille de la base d'apprentissage sur les performances d'une structure de détection, en contrôlant la complexité de celle-ci. En effet, l'application *stricto sensu* du principe de minimisation du risque structurel est à proscrire en raison de la largeur excessive de l'intervalle de confiance proposé par la relation (1.34), auquel il est préférable de substituer une estimation de la probabilité d'erreur. On se propose à présent de décrire quelques approches pour y parvenir.

1.3.1 Méthodes de validation croisée

Une première solution consiste à scinder $\mathcal{A}_n$ en deux bases indépendantes, respectivement notées $\mathcal{A}_m = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$ et $\mathcal{A}_{n-m} = \{(\boldsymbol{x}_{m+1}, y_{m+1}), \ldots, (\boldsymbol{x}_n, y_n)\}$. La première est destinée à l'apprentissage de la structure de décision d_m considérée, tandis que la seconde est dédiée à la caractérisation de ses performances. En particulier, une façon naturelle d'estimer la probabilité d'erreur $P_e(d_m)$ repose sur un comptage des individus de $\mathcal{A}_{n-m}$ mal-classés par d_m. Cette approche, dite *holdout* car elle consiste à soustraire des données de l'ensemble d'apprentissage initial $\mathcal{A}_n$ pour effectuer les tests, conduit à la définition de l'estimateur suivant :

$$\widehat{P}_e^{(h)}(d_m) = \frac{1}{n-m} \sum_{k=1}^{n-m} \mathbb{1}_{d_m(\boldsymbol{X}_{m+k}) \neq Y_{m+k}}. \tag{1.40}$$

Il convient de noter que les qualités de l'estimateur $\widehat{P}_e^{(h)}$ dépendent évidemment du nombre de données allouées à l'apprentissage et au test du détecteur. Dans ce registre, on peut montrer que le biais de cet estimateur ne dépend que de la taille m de la base d'apprentissage. En revanche, la variance de celui-ci est dominée par l'influence de la taille $n - m$ de la base de test [Fuk90]. Il en résulte que le choix d'une partition de l'ensemble des données étiquetées disponibles pour l'apprentissage et la validation de la structure de détection découle de la recherche d'un compromis classique de type biais-variance. A cette question non-triviale, vient de plus s'ajouter la contrainte de constituer des bases d'apprentissage et de test identiquement distribuées, ce qui a pour effet de complexifier la tâche à accomplir.

La procédure de *leave-one-out* vise à pallier cette difficulté [Lac68]. Elle consiste à exclure un individu de la base d'apprentissage avant d'élaborer la structure de détection sur les $n - 1$ données restantes. L'individu extrait est dédié au test du détecteur. En répétant n fois cette opération, ce qui permet de tester successivement les n individus disponibles, on peut obtenir une estimation de P_e par simple comptage des erreurs de classement. L'un des avantages de cette approche réside dans le fait qu'elle permet d'écarter la question du partitionnement de $\mathcal{A}_n$. De plus, on note une diminution du biais d'estimation par rapport à la méthode précédente, $n - 1$ données étant consacrées à l'apprentissage. En revanche, on déplore le fait que le *leave-one-out* nécessite l'élaboration successive de n détecteurs, ce qui constitue un sérieux obstacle à sa mise en œuvre.

1.3.2 Méthodes de resubstitution

L'approche *holdout* présentée précédemment nécessite la constitution d'une base de test aussi importante que possible si l'on souhaite limiter la variance d'estimation. Ceci constitue l'un des inconvénients majeurs de la méthode, en particulier lorsque la quantité de données s'avère limitée. Dans ces circonstances, on peut envisager d'utiliser les mêmes individus pour élaborer la structure de décision et la tester. L'estimateur dérivant de cette méthode, dite de *resubstitution*, s'exprime formellement ainsi :

$$\widehat{P}_e^{(r)}(d_n) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{d_n(\boldsymbol{X}_k)\neq Y_k}. \tag{1.41}$$

On montre que cet estimateur présente un biais important, qui a pour conséquence d'offrir une vision optimiste des performances réelles du détecteur. Intuitivement, ceci se justifie par le fait que la structure de détection se comporte nécessairement mieux sur les individus ayant servi à son élaboration que sur des données de test indépendantes.

L'usage de la méthode de ré-échantillonnage *Bootstrap* dans le cadre de l'estimation d'une probabilité d'erreur [Efr79] vise à compenser le biais de l'estimateur (1.41). Cette approche consiste à générer de façon artificielle une base d'apprentissage de taille n en procédant à un tirage aléatoire avec remise de n individus dans la base de données $\mathcal{A}_n$. On note

$$\mathcal{A}_n^{(b)} = \{(\boldsymbol{x}_1^{(b)}, y_1^{(b)}), \ldots, (\boldsymbol{x}_n^{(b)}, y_n^{(b)})\} \tag{1.42}$$

la base d'apprentissage ainsi obtenue et d_n la structure de détection élaborée à partir celle-ci. Parmi les nombreuses solutions déjà proposées et reposant sur le principe du *Bootstrap*, on compte l'estimateur suivant [Dev96] :

$$\widehat{P}_e^{(b)}(d_n) = \widehat{P}_e^{(r)}(d_n) + \Delta(d_n), \tag{1.43}$$

avec

$$\Delta(d_n) = \frac{1}{n}\sum_{k'=1}^{n}\left(1 - \sum_{k=1}^{n} \mathbb{1}_{\boldsymbol{X}_k^{(b)} \neq \boldsymbol{X}_{k'}}\right) \mathbb{1}_{d_n(\boldsymbol{X}_{k'}) \neq Y_{k'}}. \tag{1.44}$$

Généralement, la procédure de ré-échantillonnage est répétée plusieurs fois et $\Delta(d_n)$ est alors remplacée par sa valeur moyenne $\bar{\Delta}(d_n)$. Si le *Bootstrap* se comporte parfois mieux que les autres estimateurs présentés dans cette section, il ne constitue pas pour autant la meilleure des solutions comme certaines études expérimentales ont permis de le constater [Dev96].

Les idées dégagées jusqu'à présent seront mises en œuvre aux cours des chapitres qui suivent, dans le cadre de l'élaboration de détecteurs à partir d'une base d'exemples. Mais avant de clore ce chapitre, il convient d'apporter brièvement quelques éclaircissements sur les liens unissant la détection à structure imposée et la régression au regard de ce qui a été présenté.

1.4 Détection à structure imposée et régression

Les règles (1.8) et (1.9) illustrent le rôle central joué par les probabilités *a posteriori* $p(\omega_0|\boldsymbol{x})$ et $p(\omega_1|\boldsymbol{x})$ en théorie statistique de la décision, faisant de celles-ci un objectif potentiel lorsqu'il s'agit d'élaborer une structure de détection à partir d'une base d'exemples $\mathcal{A}_n$. Pour ce faire, on peut imaginer que la règle de décision considérée repose directement sur une estimation $p_n(\omega_i|\boldsymbol{x};\mathcal{A}_n)$ de $p(\omega_i|\boldsymbol{x})$, notée plus simplement $p_n(\omega_i|\boldsymbol{x})$, et s'écrive

$$d_n(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad p_n(\omega_1|\boldsymbol{x}) \geq 1/2 \\ 0 & \text{si} \quad p_n(\omega_1|\boldsymbol{x}) < 1/2, \end{cases} \tag{1.45}$$

en définissant $p_n(\omega_0|\boldsymbol{x})$ par $1 - p_n(\omega_1|\boldsymbol{x})$. L'objet de la présente section est de comparer cette approche à celle qui a été présentée tout au long de ce chapitre. Pour cela, l'analogie entre l'optimisation d'une telle structure et la résolution d'un problème de régression va être mise en évidence, puis la consistance de la règle de décision ainsi que celle du principe d'induction brièvement étudiées. De ces résultats seront enfin extraits quelques éléments de conclusion.

Connexions avec le problème de régression

Le calcul suivant permet d'établir le lien entre décision et régression

$$\mathrm{E}\{Y|\boldsymbol{X} = \boldsymbol{x}\} = 1\,.\,p(Y = 1|\boldsymbol{x}) + 0\,.\,p(Y = 0|\boldsymbol{x}) = p(\omega_1|\boldsymbol{x}). \tag{1.46}$$

Ainsi, la probabilité *a posteriori* $p(\omega_1|\boldsymbol{x})$ n'est autre que la fonction de régression de la variable Y en $\boldsymbol{x}$, impliquant par là-même qu'elle en est la plus proche au sens de l'erreur quadratique. En conséquence, le détecteur (1.45) peut être élaboré en minimisant la fonctionnelle de risque

$$J(d_n;\mathcal{A}_n) = \int (p_n(\omega_1|\boldsymbol{x}) - y)^2\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy. \tag{1.47}$$

La densité de probabilité $p(\boldsymbol{x}, y)$ étant supposée inconnue, ceci se traduit dans les faits par l'optimisation du risque empirique

$$J_{emp}(d_n;\mathcal{A}_n) = \frac{1}{n}\sum_{k=1}^{n}(p_n(\omega_1|\boldsymbol{x}_k) - y_k)^2, \tag{1.48}$$

calculable à partir des données constituant l'ensemble d'apprentissage $\mathcal{A}_n$. Force est de constater que ce problème est similaire à celui traité au cours des sections précédentes, à l'exception de la fonction coût Q associée à chaque couple $(\boldsymbol{x}, y)$, définie à présent par $(p_n(\omega_1|\boldsymbol{x}_k) - y_k)^2$.

Consistance de la règle de décision

Afin de juger de l'intérêt de la règle (1.45), il convient d'en analyser préalablement la consistance telle qu'elle a été définie en Section 1.1.3. Pour ce faire, on note que le risque J peut se réécrire selon [Dev96]

$$J(d_n;\mathcal{A}_n) = \int (p(\omega_1|\boldsymbol{x}) - y)^2\, p(\boldsymbol{x},y)\, d\boldsymbol{x}\, dy + \int (p_n(\omega_1|\boldsymbol{x}) - p(\omega_1|\boldsymbol{x}))^2\, p(\boldsymbol{x})\, d\boldsymbol{x}. \tag{1.49}$$

On constate que le premier terme de cette expression n'influe pas sur le processus d'optimisation de J puisqu'il est indépendant de $p_n(\omega_1|\boldsymbol{x})$. Quant au second, il vérifie

$$P_e(d_n;\mathcal{A}_n) - P_e^* \leq 2\sqrt{\int (p(\omega_1|\boldsymbol{x}) - p_n(\omega_1|\boldsymbol{x}))^2\, p(\boldsymbol{x})\, d\boldsymbol{x}}, \tag{1.50}$$

que l'on obtient en appliquant l'inégalité de Cauchy-Schwarz à la relation figurant ci-dessous, dont une démonstration est proposée dans [Dev96].

$$P_e(d_n;\mathcal{A}_n) - P_e^* \leq 2\int |p(\omega_1|\boldsymbol{x}) - p_n(\omega_1|\boldsymbol{x})|\; p(\boldsymbol{x})\, d\boldsymbol{x}. \tag{1.51}$$

En conséquence, si l'on est à même de proposer une structure de détection basée sur un estimateur de $p(\omega_1|\boldsymbol{x})$ tel que

$$\int (p_n(\omega_1|\boldsymbol{x}) - p(\omega_1|\boldsymbol{x}))^2\, p(\boldsymbol{x})\, d\boldsymbol{x} \xrightarrow[n\to\infty]{p} 0, \tag{1.52}$$

alors la suite $\{d_n(\boldsymbol{X})\}_{n>0}$ associée est consistante. La propriété de consistance forte est obtenue dès lors que la convergence (1.52) a lieu avec une probabilité égale à 1. Ainsi, les modes de convergence de la probabilité d'erreur $P_e(d_n;\mathcal{A}_n)$ vers la borne inférieure P_e^* fournie par le détecteur de Bayes reposent sur les propriétés de consistance de l'estimateur de $p(\omega_1|\boldsymbol{x})$.

Consistance du principe d'induction, compromis biais-variance

Lorsqu'on dispose d'un ensemble de réalisations $\mathcal{A}_n$ pour seule source d'information *a priori*, le succès de l'approche étudiée repose sur la consistance du principe d'induction visant à substituer à l'erreur quadratique (1.47) le critère empirique (1.48). Il s'agit là d'une problématique récurrente en théorie statistique de l'apprentissage dépassant le cadre de ces deux critères. Aussi, le théorème fondamental 1.2.1 fournissant une condition nécessaire et suffisante de consistance d'un tel principe pour tout couple de risques de la forme (1.1) et (1.2) demeure valide dans le cas présent. En revanche, il n'en est pas de même pour les résultats reposant sur la VC-entropie, la fonction de croissance et autre VC-dimension proposées tout au long de ce chapitre, en raison d'une modification profonde de la nature de la fonction coût Q associée à chaque couple $(\boldsymbol{x},y)$. Si celle-ci prend la forme d'une fonction indicatrice pour des critères tels que la probabilité d'erreur et la probabilité d'erreur empirique, elle désigne une fonction à valeurs dans $\mathbb{R}$ lorsque le problème traité concerne la régression. Ceci nécessite en premier lieu de revoir la définition de la quantité $N_{\mathcal{D}}(\mathcal{A}_n)$, qui représente le nombre de configurations distinctes que peut engendrer une classe de détecteurs $\mathcal{D}$ donnée lorsque Q est une fonction indicatrice. Plusieurs approches ont été proposées afin de généraliser les résultats obtenus dans le cadre des fonctions indicatrices à toute fonction coût à valeurs dans $\mathbb{R}$, dont le lecteur intéressé pourra trouver un aperçu dans [Vap95]. C'est ainsi au prix d'une complexité largement accrue, reposant parfois sur une décomposition préalable de la fonction coût sur un ensemble de fonctions indicatrices, que des conditions de

consistance équivalentes aux expressions (1.24) et (1.26) ont été obtenues. Il en est de même pour l'inégalité (1.32) relative à la vitesse de convergence du risque empirique vers le risque, qui a été déclinée selon plusieurs formes s'adressant à des fonctions coût de natures diverses à valeurs dans $\mathbb{R}$. Il convient de noter que celles-ci expriment toujours la nécessité de trouver un compromis entre la complexité de la structure dont on effectue l'apprentissage, et le nombre de données dont on dispose pour accomplir cette tâche. Ce compromis nécessaire va être à présent mis en évidence dans le cadre de l'estimation de la fonction de régression.

Le problème posé consiste à minimiser l'erreur quadratique $J(d_n;\mathcal{A}_n)$ définie par l'expression (1.47), qui constitue une variable aléatoire dépendant de $\mathcal{A}_n$. Il convient donc d'en calculer l'espérance, ce qui mène au résultat suivant

$$\begin{aligned}\mathrm{E}_{\mathcal{A}_n}\{J(d_n;\mathcal{A}_n)\} = &\int (p(\omega_1|\boldsymbol{x}) - y)^2\, p(\boldsymbol{x},y)\, d\boldsymbol{x}\, dy \\ &+ \int (p(\omega_1|\boldsymbol{x}) - \mathrm{E}_{\mathcal{A}_n}\{p_n(\omega_1|\boldsymbol{x})\})^2\, p(\boldsymbol{x})\, d\boldsymbol{x} + \int \mathrm{var}_{\mathcal{A}_n}\{p_n(\omega_1|\boldsymbol{x})\}\, p(\boldsymbol{x})\, d\boldsymbol{x}. \quad (1.53)\end{aligned}$$

L'erreur quadratique peut ainsi être décomposée selon trois termes de nature différente, qui sont respectivement l'erreur minimale atteignable, un terme de biais et un terme de variance. Étant donné un ensemble d'apprentissage de taille n fixée, on observe généralement que le biais décroît lorsque la complexité du modèle augmente, tandis que la variance croît. La recherche d'une solution nécessite en conséquence de trouver un compromis entre ces deux grandeurs antagonistes, communément appelé *compromis biais-variance*, problème auquel on peut apporter des éléments de réponse pratiques grâce au principe de minimisation du risque structurel.

Cette section visait à proposer une approche alternative pour l'élaboration d'une structure de détection à partir d'une base d'exemples, consistant estimer la fonction de régression plutôt que de minimiser la probabilité d'erreur. Il s'avère que ces deux approches sont étroitement liées, mais que les problèmes posés par un critère tel que l'erreur quadratique sont nombreux et nécessitent de recourir à des outils théoriques complexes. Enfin, le théorème suivant permet de trancher définitivement en faveur d'une minimisation directe de la probabilité d'erreur [Dev96] :

Théorème 1.4.1. *Soit $p_n(\omega_1|\boldsymbol{X})$ un estimateur faiblement convergent de $p(\omega_1|\boldsymbol{X})$, c'est-à-dire vérifiant la propriété*

$$\lim_{n\to\infty} \mathrm{E}\{(p_n(\omega_1|\boldsymbol{X}) - p(\omega_1|\boldsymbol{X}))^2\} = 0. \quad (1.54)$$

Soit d_n une structure de détection reposant sur la comparaison de $p_n(\omega_1|\boldsymbol{x})$ à un seuil. Alors

$$\lim_{n\to\infty} \frac{P_e(d_n) - P_e^*}{\sqrt{\mathrm{E}\{(p_n(\omega_1|\boldsymbol{X}) - p(\omega_1|\boldsymbol{X}))^2\}}} = 0. \quad (1.55)$$

Ce résultat démontre clairement que $P_e(d_n) - P_e^*$ converge plus rapidement vers 0 que l'erreur quadratique. On peut interpréter ce fait en constatant que les performances d'un détecteur dépendent largement de son comportement au voisinage de la frontière entre les classes ω_0 et ω_1. Comme l'illustre la figure 1.7, l'approche consistant à estimer la fonction de régression ne prend pas en compte cette priorité, pouvant même s'avérer plus exigeante que nécessaire, notamment lorsque $p(\omega_i|\boldsymbol{x}) \approx 1$.

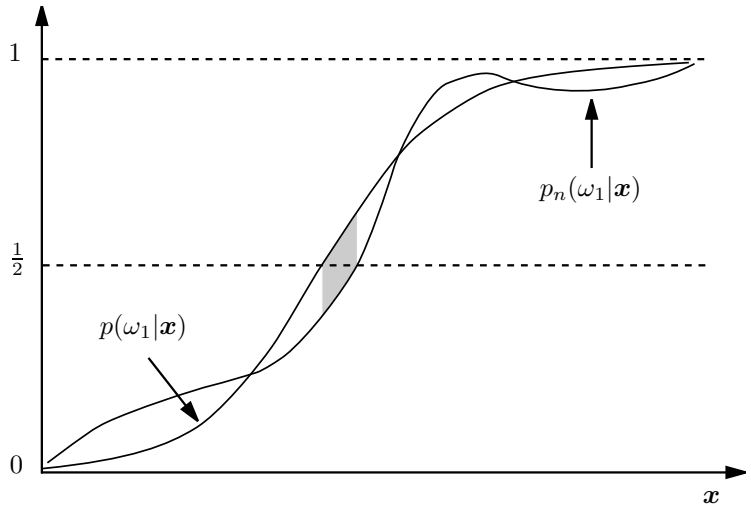

FIG. 1.7 : Exemple schématique montrant qu'il n'est pas nécessaire que l'estimation $p(\omega_1|\boldsymbol{x})$ soit satisfaisante en tout point $\boldsymbol{x}$ pour que le détecteur associé présente de bonnes performances. En effet, celles-ci se dégradent principalement à mesure que l'aire de la zone hachurée croît (d'après [Dev96]).

Chapitre 2

Détecteurs linéaires généralisés et critères de contraste

2.1 Introduction

On distingue au moins deux types d'approches pour la résolution d'un problème de détection. Le premier concerne la détection dite *à structure libre*, pour laquelle l'expression du test résulte de l'application d'un critère de détection et de la connaissance des lois de probabilité de l'observation, conditionnellement aux hypothèses. La deuxième, appelée détection *à structure imposée*, suppose de définir préalablement une classe $\mathcal{D}$ de tests de détection, pour ne retenir que celui qui est optimal au sens d'un critère donné. Ce chapitre est consacré aux problèmes rencontrés lorsqu'on adopte cette dernière approche, et que les éléments de théorie de l'apprentissage proposés précédemment ont permis d'identifier. Aussi, un choix de classes de détecteurs $\mathcal{D}$ autorisant une erreur d'approximation nulle est d'abord présenté dans ce chapitre. Puis, sont sélectionnés des critères de performance pertinents et les algorithmes d'apprentissage qui leur sont associés, au vu de l'objectif à atteindre que constitue la minimisation de l'erreur d'estimation. Enfin, la recherche automatique d'un compromis entre l'erreur d'approximation et l'erreur d'estimation est traitée au chapitre suivant.

2.2 Discriminants linéaires généralisés

Les origines de la détection à structure imposée peuvent être attribuées à R. A. Fisher, qui ébaucha une première méthodologie dès 1936 [Fis36]. Pour ce faire, celui-ci considéra le cas d'un échantillon $\boldsymbol{X}$ distribué selon une loi normale à l dimensions, conditionnellement aux hypothèses en compétition. En notant $\boldsymbol{m}_i$ et $\boldsymbol{\Sigma}_i$ les moments conditionnels d'ordre 1 et 2 de $\boldsymbol{X}$, il montra que la statistique de détection optimale au sens de Bayes est une forme quadratique définie par

$$\lambda(\boldsymbol{x}) = \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_0)^t\,\boldsymbol{\Sigma}_0^{-1}\,(\boldsymbol{x}-\boldsymbol{m}_0) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m}_1)^t\,\boldsymbol{\Sigma}_1^{-1}\,(\boldsymbol{x}-\boldsymbol{m}_1) + \ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}, \tag{2.1}$$

nécessitant l'estimation de $\frac{l(l+3)}{2}$ paramètres. Lorsque $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$, l'expression ci-dessus dégénère en une forme linéaire

$$\lambda(\boldsymbol{x}) = (\boldsymbol{m}_0-\boldsymbol{m}_1)^t\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{x} - \frac{1}{2}(\boldsymbol{m}_0^t\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{m}_0 - \boldsymbol{m}_1^t\,\boldsymbol{\Sigma}^{-1}\,\boldsymbol{m}_1), \tag{2.2}$$

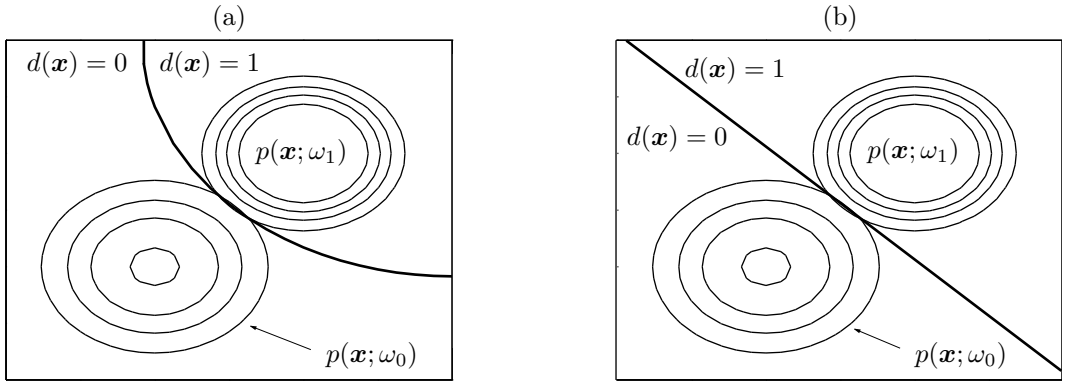

FIG. 2.1 : Détection d'un signal aléatoire gaussien noyé dans un bruit blanc gaussien. Partition de l'espace des observations par (a) un détecteur à structure libre, qui se trouve être de la forme (2.1), et (b) un détecteur linéaire du type (2.2) préconisé par Fisher, en posant $\boldsymbol{\Sigma} = \frac{1}{2}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)$.

ne requérant que la détermination de l paramètres. En notant que l'estimation de $o(l^2)$ paramètres nécessite un nombre conséquent d'observations, parfois irréaliste compte tenu des conditions expérimentales rencontrées, Fisher recommanda d'utiliser également la statistique linéaire (2.2) lorsque $\boldsymbol{\Sigma}_0 \neq \boldsymbol{\Sigma}_1$, en posant

$$\boldsymbol{\Sigma} = \rho\boldsymbol{\Sigma}_0 + (1-\rho)\boldsymbol{\Sigma}_1, \tag{2.3}$$

où ρ est un coefficient à déterminer. Cette approche est illustrée par la figure 2.1. Poursuivant son raisonnement, il préconisa cette même statistique pour traiter le cas non-gaussien, posant ainsi les bases de la décision à structure imposée en accordant une importance particulière aux formes linéaires.

Les détecteurs linéaires présentent l'avantage d'être directement liés à la notion de filtrage et de se prêter à une optimisation aisée. Sous une forme généralisée, on les définit par

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \lambda(\boldsymbol{x}) = \sum_k w_k\,\phi_k(\boldsymbol{x}) - \lambda_0 > 0 \\ 0 & \text{sinon,} \end{cases} \tag{2.4}$$

où ϕ_k est une fonction quelconque de la variable $\boldsymbol{x}$ dans $\mathbb{R}$. Pour certaines classes de fonctions ϕ_k, ces structures de détection possèdent des propriétés de consistance, signifiant qu'elles peuvent produire la décision idéale sous certaines conditions. En particulier, un résultat classique d'analyse adapté au cas présent notifie que, moyennant certaines hypothèses sur $p(\boldsymbol{x}|\omega_0)$ et $p(\boldsymbol{x}|\omega_1)$, le détecteur de Bayes peut s'écrire sous la forme (2.4).

Théorème 2.2.1. *Soit p une loi de vraisemblance et $\{\phi_k\}_{k>0}$ un ensemble orthonormé complet de fonctions bornées appartenant à $\mathcal{L}_2(\mu)$, où μ désigne la mesure de Lebesgue sur $\mathbb{R}^l$. Si p appartient à $\mathcal{L}_2(\mu)$, alors les densités conditionnelles $p(\boldsymbol{x}|\omega_0)$ et $p(\boldsymbol{x}|\omega_1)$ appartiennent à $\mathcal{L}_2(\mu)$, et la statistique optimum $\lambda(\boldsymbol{x}) = p(\omega_1)\,p(\boldsymbol{x}|\omega_1) - p(\omega_0)\,p(\boldsymbol{x}|\omega_0)$ admet le développement suivant :*

$$\lambda(\boldsymbol{x}) = p(\omega_1)\,p(\boldsymbol{x}|\omega_1) - p(\omega_0)\,p(\boldsymbol{x}|\omega_0) = \sum_{k=1}^{+\infty} w_k\,\phi_k(\boldsymbol{x}), \tag{2.5}$$

où les coefficients w_k sont donnés par :

$$w_k = \int \phi_k(\boldsymbol{x})\,\lambda(\boldsymbol{x})\,d\boldsymbol{x}. \tag{2.6}$$

La convergence de ce développement est à considérer en moyenne quadratique :

$$\int \left(\lambda(\boldsymbol{x}) - \sum_{k=1}^{+\infty} w_k \, \phi_k(\boldsymbol{x}) \right)^2 d\boldsymbol{x} = 0. \tag{2.7}$$

Sous certaines conditions portant sur la loi de vraisemblance, le théorème 2.2.1 établit donc la convergence vers 0 de l'erreur d'approximation en considérant un ensemble $\{\phi_k\}_{k>0}$ orthonormé complet de fonctions bornées de $\mathcal{L}_2(\mu)$. Parmi les familles de fonctions possédant une telle propriété, on peut recenser par exemple la base standard des fonctions trigonométriques, les polynômes de Legendre ou encore les bases de Laguerre et de Haar.

Bien que séduisant, ce résultat n'est cependant pas satisfaisant puisqu'il nécessite une certaine connaissance *a priori* de la loi de vraisemblance de l'observation $\boldsymbol{x}$, inconnue par hypothèse. Par les conditions qu'il impose sur cette dernière, il rejette par là-même tout caractère universel de la consistance. Le théorème que l'on va à présent énoncer s'avère beaucoup plus intéressant puisqu'il concerne la convergence de l'erreur de modélisation grâce à un choix approprié de l'ensemble $\{\phi_k\}_{k>0}$ [Dev96]. Contrairement au théorème 2.2.1, ce résultat est issu de la théorie statistique de l'apprentissage et repose en particulier sur l'inégalité de Vapnik-Chervonenkis (1.32).

Théorème 2.2.2. *La suite $\{d_n^*(\boldsymbol{x})\}_{n>0}$ des détecteurs linéaires généralisés (2.4) résultant de la minimisation de l'erreur empirique sur $\mathcal{A}_n$ est universellement fortement consistante si $k \to +\infty$ et $k/n \to 0$ lorsque $n \to +\infty$, et si $\{\phi_k\}_{k>0}$ est une suite de fonctions telle que l'ensemble composé de toutes les combinaisons linéaires des ϕ_k est dense dans $\mathcal{L}_1(\mu)$ sur les boules $\{\boldsymbol{x} : \|\boldsymbol{x}\| \leq B\}$.*

Il convient de noter que les fonctions ϕ_1, ϕ_2, ... définies en particulier par les monômes $x[1]$, ..., $x[l]$, $x[1]\,x[2]$, ... vérifient l'hypothèse de densité mentionnée ci-dessus [Dev96]. Le théorème 2.2.2 donne ainsi une légitimité à certaines classes de détecteurs linéaires généralisés, à condition que la taille de la base d'apprentissage croisse plus vite que le degré des polynômes. En conséquence, dans la suite de ce document et sauf indication contraire, les tests de détection considérés sont du type (2.4). En effet, en plus des garanties de performances que présente la consistance universellement forte de cette classe de détecteurs lorsque l'hypothèse de densité mentionnée dans le théorème 2.2.2 est satisfaite, l'optimisation de telles structures est facilitée par le fait qu'elle s'apparente à la recherche d'un discriminant linéaire dans l'espace des transformées $\phi_k(\boldsymbol{x})$. Ce thème, qui a fait l'objet d'une abondante littérature en reconnaissance des formes, est abordé dans les prochaines sections.

2.3 Taxinomie de critères de performance

L'étude menée précédemment a permis d'identifier une classe de détecteurs $\mathcal{D}$, en l'occurrence celle des structures linéaires généralisées définies par le théorème 2.2.2, autorisant une erreur de modélisation nulle. On rappelle que cette quantité est définie par

$$E_{\text{mod}} = P_e(d_n^*) - P_e^*, \tag{2.8}$$

où d_n^* est le détecteur optimal de $\mathcal{D}$ au sens de la probabilité d'erreur empirique

$$d_n^* = \arg\min_{d \in \mathcal{D}} P_{emp}(d). \tag{2.9}$$

Si ce résultat est intuitivement satisfaisant puisqu'il donne une légitimité à la probabilité d'erreur empirique par rapport à la probabilité d'erreur, on peut regretter qu'il est impossible de

résoudre l'équation (2.9) de façon exacte. En effet, la recherche d'une dichotomie optimale de n points étiquetés de $\mathbb{R}^l$ est un problème démontré NP-complet nécessitant le recours à des méthodes de résolution approchées [Joh78]. Dans ces circonstances, une démarche naturelle pour la détermination d'un optimum local pourrait consister à utiliser une méthode d'optimisation classique. Cependant, la probabilité d'erreur empirique est une fonction constante par morceaux en raison de la nature échantillonnée de $\mathcal{A}_n$, ce qui rend l'opération de minimisation particulièrement délicate. Afin de pallier cette difficulté, de nombreux auteurs ont proposé de modifier quelque peu ce critère, au moyen d'un lissage par exemple, afin de limiter le nombre de minima locaux [Skl79, Wid88]. Dans [Dev96], il est néanmoins fait état du caractère parfois discutable des solutions ainsi obtenues.

2.3.1 Critères de marge

Si l'objet de ce chapitre n'est pas de dresser une liste exhaustive des substituts à la probabilité d'erreur empirique, on ne peut faire l'économie de citer les critères de marge. Ils ont en effet contribué au développement de la reconnaissance des formes avec le *Perceptron* [Ros62], et connaissent depuis peu un nouvel essor avec les *Support Vector Machines* [Cor95]. Il s'agit de fonctions de la distance des échantillons de l'ensemble d'apprentissage à l'hyperplan séparateur recherché. Il est à noter que ce dernier est d'équation $\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x})-\lambda_0=0$ d'après l'expression (2.4).

Le critère du Perceptron est égal à la somme des marges des échantillons mal classés par la règle de décision, c'est à dire [Bis95]

$$J_{per}(d)=\sum_{k\in\mathcal{M}}|\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_k)-\lambda_0|, \tag{2.10}$$

avec $\mathcal{M}=\{k\,:\,d(\boldsymbol{x}_k)\neq y_k\}$ et $\|\boldsymbol{w}\|$ fixé. Ce critère est donc proportionnel à la distance des $\boldsymbol{\phi}(\boldsymbol{x}_k)$ considérés à l'hyperplan séparateur, dont on rappelle qu'elle est donnée par $|\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_k)-\lambda_0|/\|\boldsymbol{w}\|$. On peut montrer que l'algorithme d'apprentissage du Perceptron converge en un nombre fini d'itérations lorsque les classes ω_0 et ω_1 en compétition sont linéairement séparables. Aucune convergence n'est en revanche assurée dans le cas contraire, ce qui a largement contribué à l'abandon de cette approche. Une critique du Perceptron est proposée dans [Min69], accompagnée d'une énumération de problèmes simples ne pouvant être résolus par une telle technique.

Les Support Vector Machines, communément appelées SVM, recherchent la solution garantissant la marge la plus importante possible entre l'hyperplan séparateur et les échantillons de l'ensemble d'apprentissage [Cor95, Vap95, Bur98]. Ce principe est illustré par la figure 2.2. Dans le cas de classes linéairement séparables, le critère maximisé est donc défini par

$$J_{svm}(d)=\min_{1\leq k\leq n}|\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}_k)-\lambda_0|, \tag{2.11}$$

sous la contrainte que tous les échantillons soient convenablement classés par la règle de décision ainsi élaborée, et que $\|\boldsymbol{w}\|=1$. Lorsque les données ne sont pas séparables linéairement, un terme supplémentaire pénalisant les échantillons mal-classés est pris en compte lors de l'optimisation. Bien que d'apparence simpliste, la technique des SVM est directement issue des principes de la théorie statistique de l'apprentissage. Ainsi, un lien entre la marge J_{svm} et les performances en généralisation du discriminant linéaire associé a pu être établi [Vap95], faisant des SVM une mise en œuvre directe du principe de minimisation du risque structurel.

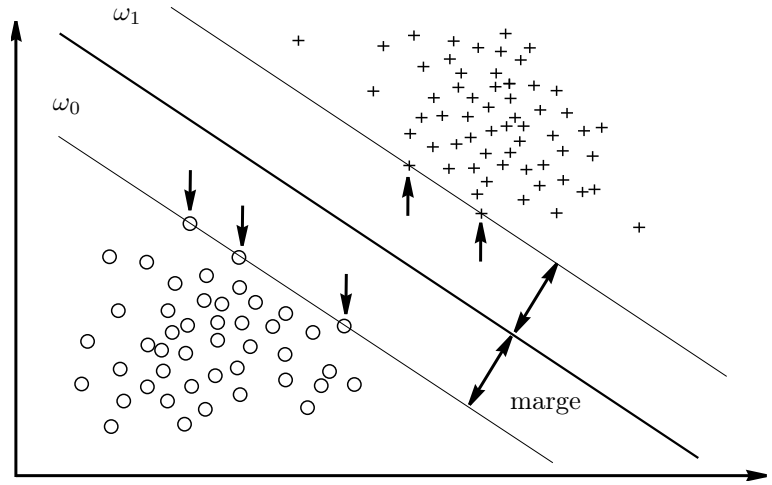

FIG. 2.2 : Principe des SVM dans le cas où les classes ω_0 et ω_1 sont linéairement séparables. Les *Support Vectors*, indiqués par des flèches, désignent les échantillons les plus proches de l'hyperplan séparateur.

2.3.2 Mesures de contraste

Une autre approche consiste à adopter un critère qui quantifie le caractère discriminant d'une statistique de détection $\lambda(\boldsymbol{x})$ pour un problème donné, à la manière d'une mesure de distance entre les lois de probabilité $p(\lambda(\boldsymbol{x})|\omega_0)$ et $p(\lambda(\boldsymbol{x})|\omega_1)$. Certains de ces critères tels que les divergences de Jeffreys et de Kullback-Leibler, l'affinité de Bhattacharyya et autre distance de Kolmogorov, nécessitent la connaissance de $p(\boldsymbol{x}|\omega_0)$ et $p(\boldsymbol{x}|\omega_1)$. En contrepartie, ils fournissent des informations de qualité sur les performances que l'on est en droit d'attendre d'un détecteur, étant donné qu'ils sont étroitement liés à l'erreur de Bayes. D'autres critères, dits *du second ordre*, se contentent en revanche d'une quantité d'information plus modeste, qui se limite aux moments conditionnels du premier et du second ordre de $\lambda(\boldsymbol{x})$, respectivement définis par

$$\eta_i = \mathrm{E}\{\lambda(\boldsymbol{X})|\omega_i\} \tag{2.12}$$
$$\sigma_i^2 = \mathrm{var}\{\lambda(\boldsymbol{X})|\omega_i\}, \tag{2.13}$$

avec $i \in \{0, 1\}$. Ces critères faisant l'objet d'une abondante littérature, on ne mentionnera dans ce travail que les plus connus, laissant au lecteur le soin de consulter par exemple [Gar80, Dud01]. Au sein de cette classe, on compte en particulier le *rapport signal-sur-bruit généralisé*

$$J_{rsbg}(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2) = \frac{(\eta_1 - \eta_0)^2}{\rho\,\sigma_0^2 + (1-\rho)\,\sigma_1^2}, \tag{2.14}$$

dont les propriétés ont été largement étudiées [Gar80, Pic86, Duv87, Ric98(a)]. Il se décline suivant le *critère de Fisher*($\rho = p(\omega_0)$), la *deflexion* ($\rho = 1/2$) et le *rapport signal-sur-bruit* ($\rho = 1$), en notant toutefois que cette nomenclature peut varier selon les références consultées. Comme l'illustre la figure 2.3, l'optimisation de ces critères peut être interprétée comme la maximisation de la dispersion inter-hypothèse Δ_{inter} pratiquée conjointement avec la minimisation de la dispersion intra-hypothèse Δ_{intra}, en posant :

$$\Delta_{\text{inter}} = (\eta_1 - \eta_0)^2 \tag{2.15}$$
$$\Delta_{\text{intra}} = \rho\,\sigma_0^2 + (1-\rho)\,\sigma_1^2. \tag{2.16}$$

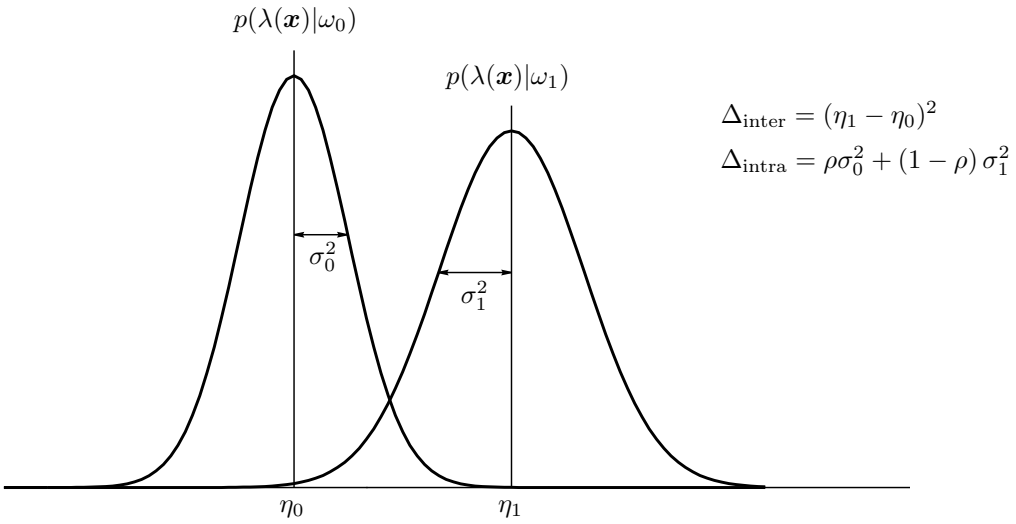

FIG. 2.3 : Définition des dispersions intra-hypothèse et inter-hypothèse. Principe des mesures de contraste ne dépendant que des moments d'ordre 1 et 2 de la statistique de détection.

Ainsi, des arguments simples basés sur des considérations géométriques suffisent à interpréter ces quelques critères en termes de mesure de contraste entre les hypothèses ω_0 et ω_1, que renforcent des propriétés fortes présentées dans la section suivante. Elles sont le fruit de travaux menés avec R. Lengellé dans un premier temps, puis avec le concours d'un doctorant, Fahed Abdallah. Ces résultats ont fait l'objet de plusieurs publications et communications dans des revues et conférences de renom citées tout au long du texte.

2.4 Pertinence des critères du second ordre

Soit $\lambda(\boldsymbol{x}) : \mathbb{R}^l \longrightarrow \mathbb{R}$ une fonction mesurable, et $d(\boldsymbol{x}) : \mathbb{R}^l \longrightarrow \{0,1\}$ une règle de décision reposant sur la statistique $\lambda(\boldsymbol{x})$:

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \lambda(\boldsymbol{x}) > \lambda_0 \\ 0 & \text{sinon.} \end{cases} \tag{2.17}$$

Les théories de Bayes et Neyman-Pearson constituent des approches statistiques classiques pour l'élaboration de détecteurs optimaux de la forme (2.17), au sens du coût moyen minimum dans le premier cas, et de la probabilité de bonne détection maximum pour une probabilité de fausse alarme bornée supérieurement dans le second cas [Van68, Poo94]. Elles conduisent au résultat fondamental qu'une décision optimale est prise en comparant le *rapport de vraisemblance*, défini par $\lambda^*(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_1)/p(\boldsymbol{x}|\omega_0)$, à un seuil λ_0 donné :

$$d^*(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \lambda^*(\boldsymbol{x}) > \lambda_0 \\ 0 & \text{sinon.} \end{cases} \tag{2.18}$$

Il est à noter que la comparaison de toute fonction strictement monotone de $\lambda^*(\boldsymbol{x})$ à un seuil constitue un test équivalent à (2.18) dans le sens où ils ont même courbe COR. Étant donné un problème de détection, on peut en conséquence définir une classe d'équivalence de détecteurs optimaux, dont la particularité est de reposer sur une fonction strictement monotone du rapport de vraisemblance [Pic86].

L'objectif de cette section est d'identifier tous les critères du second ordre $J(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2)$ présentant l'intérêt légitimement souhaitable de garantir une solution optimale au sens des théories statistiques classiques de la décision. Ces critères, dits *pertinents* dans ce document, sont caractérisés par le fait que les statistiques $\lambda(\boldsymbol{x})$ pour lesquelles ils sont optimaux sont des fonctions strictement monotones de $\lambda^*(\boldsymbol{x})$, garantissant ainsi une structure de détection équivalente au test du rapport de vraisemblance (2.18). Cette section est organisée ainsi. Dans un premier temps, on exprime les statistiques de détection obtenues par optimisation des critères du second ordre, en fonction du rapport de vraisemblance. La nécessité que cette fonction soit strictement monotone conduit alors à l'identification de l'ensemble des critères $J(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2)$ recevables. A titre d'illustration, les résultats obtenus sont appliqués à quelques critères connus.

2.4.1 Caractérisation des critères pertinents

Afin de déterminer les statistiques $\lambda(\boldsymbol{x})$ optimum au sens d'un critère du second ordre donné, on procède à un calcul des variations de ce dernier. On obtient ainsi :

$$\delta J = \frac{\partial J}{\partial \eta_0}\,\delta\eta_0 + \frac{\partial J}{\partial \eta_1}\,\delta\eta_1 + \frac{\partial J}{\partial \sigma_0^2}\,\delta\sigma_0^2 + \frac{\partial J}{\partial \sigma_1^2}\,\delta\sigma_1^2. \tag{2.19}$$

En utilisant $\delta\eta_i = \int \delta\lambda(\boldsymbol{x})\, p(\boldsymbol{x}|\omega_i)\, d\boldsymbol{x}$ et $\delta\sigma_i^2 = 2\int (\lambda(\boldsymbol{x}) - \eta_i)\, \delta\lambda(\boldsymbol{x})\, p(\boldsymbol{x}|\omega_i)\, d\boldsymbol{x}$ avec $i \in \{0, 1\}$, il en résulte que :

$$\begin{aligned}\delta J = \int \Bigg[&\frac{\partial J}{\partial \eta_0}\, p(\boldsymbol{x}|\omega_0) + \frac{\partial J}{\partial \eta_1}\, p(\boldsymbol{x}|\omega_1) \\ &+ 2\,(\lambda(\boldsymbol{x}) - \eta_0)\,\frac{\partial J}{\partial \sigma_0^2}\, p(\boldsymbol{x}|\omega_0) + 2\,(\lambda(\boldsymbol{x}) - \eta_1)\,\frac{\partial J}{\partial \sigma_1^2}\, p(\boldsymbol{x}|\omega_1) \Bigg]\, \delta\lambda(\boldsymbol{x})\, d\boldsymbol{x}.\end{aligned} \tag{2.20}$$

Pour que l'on ait $\delta J = 0$ indépendamment de $\delta\lambda$, le terme [.] doit être nul. On obtient ainsi l'expression de la statistique $\lambda(\boldsymbol{x})$ maximisant J, qu'il est possible de réécrire en fonction du rapport de vraisemblance :

$$\lambda(\boldsymbol{x}) = -\frac{1}{2}\,\frac{\frac{\partial J}{\partial \eta_0} + \frac{\partial J}{\partial \eta_1}\lambda^*(\boldsymbol{x})}{\frac{\partial J}{\partial \sigma_0^2} + \frac{\partial J}{\partial \sigma_1^2}\lambda^*(\boldsymbol{x})} + \frac{\eta_0\frac{\partial J}{\partial \sigma_0^2} + \eta_1\frac{\partial J}{\partial \sigma_1^2}\lambda^*(\boldsymbol{x})}{\frac{\partial J}{\partial \sigma_0^2} + \frac{\partial J}{\partial \sigma_1^2}\lambda^*(\boldsymbol{x})}. \tag{2.21}$$

avec $\lambda^*(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_1)/p(\boldsymbol{x}|\omega_0)$. La statistique précédente est équivalente au rapport de vraisemblance si et seulement si il s'agit d'une fonction strictement monotone de $\lambda^*(\boldsymbol{x})$. L'évaluation de la dérivée de $\lambda(\boldsymbol{x})$ par rapport à $\lambda^*(\boldsymbol{x})$ donne immédiatement :

$$\frac{d\lambda}{d\lambda^*}(\boldsymbol{x}) = \frac{(\eta_1 - \eta_0)\,\frac{\partial J}{\partial \sigma_0^2}\,\frac{\partial J}{\partial \sigma_1^2} + \frac{1}{2}\left(\frac{\partial J}{\partial \sigma_1^2}\,\frac{\partial J}{\partial \eta_0} - \frac{\partial J}{\partial \sigma_0^2}\,\frac{\partial J}{\partial \eta_1}\right)}{\left(\frac{\partial J}{\partial \sigma_0^2} + \frac{\partial J}{\partial \sigma_1^2}\,\lambda^*(\boldsymbol{x})\right)^2}. \tag{2.22}$$

On constate donc que $\lambda(\boldsymbol{x})$ défini par l'expression (2.21) est une fonction strictement monotone de $\lambda^*(\boldsymbol{x})$ si et seulement si le numérateur de l'équation (2.22) est différent de 0. Ce résultat conduit directement au théorème suivant [Ric01(a), Ric02(a)].

Théorème 2.4.1. *L'optimisation d'un critère du second ordre $J(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2)$ conduit à une statistique équivalente au rapport de vraisemblance si et seulement si il vérifie la condition*

$$(\eta_1 - \eta_0)\,\frac{\partial J}{\partial \sigma_0^2}\,\frac{\partial J}{\partial \sigma_1^2} + \frac{1}{2}\left(\frac{\partial J}{\partial \sigma_1^2}\,\frac{\partial J}{\partial \eta_0} - \frac{\partial J}{\partial \sigma_0^2}\,\frac{\partial J}{\partial \eta_1}\right) \neq 0. \tag{2.23}$$

Il convient de constater que l'équation aux dérivées partielles (2.23) ne peut être résolue aisément. Elle apparaît toutefois peu contraignante quant au choix d'un critère, et permet d'en vérifier aisément la pertinence. Cette propriété unifie ainsi les nombreux travaux de ce type sur des critères considérés individuellement, généralement le rapport signal-sur-bruit, le critère de Fisher ou encore l'erreur quadratique [Gar80, et réf. incluses]. Sous réserve de montrer qu'il existe des critères ne vérifiant pas la condition (2.23), ce qui est vérifié plus loin, ce résultat nuance également les conclusions présentées dans [Fuk90, p. 141-3], déclarant que tout critère du second ordre conduit à une statistique équivalente au rapport de vraisemblance.

2.4.2 Analyse des critères du type $J((\eta_1 - \eta_0)^2,\, \rho\sigma_1^2 + (1-\rho)\sigma_0^2)$

On considère les critères du second ordre $J(u, v)$ tels que $u = (\eta_1 - \eta_0)^2$ et $v = \rho\sigma_1^2 + (1-\rho)\sigma_0^2$, où ρ désigne un réel de l'intervalle $[0, 1]$. Afin de discuter de la pertinence des critères de cette nature, on fait appel au théorème 2.4.1. Celui-ci nécessite le calcul des dérivées partielles de J par rapport à η_i et σ_i^2 :

$$\frac{\partial J}{\partial \eta_0} = -2\sqrt{u}\,\frac{\partial J}{\partial u} \qquad \frac{\partial J}{\partial \eta_1} = 2\sqrt{u}\,\frac{\partial J}{\partial u} \tag{2.24}$$

$$\frac{\partial J}{\partial \sigma_0^2} = (1-\rho)\,\frac{\partial J}{\partial v} \qquad \frac{\partial J}{\partial \sigma_1^2} = \rho\,\frac{\partial J}{\partial v}. \tag{2.25}$$

En introduisant les expressions (2.24) et (2.25) dans la condition (2.23), on aboutit directement à la conclusion que l'optimisation des critères considérés conduit à un test équivalent à celui du rapport de vraisemblance si et seulement si :

$$\sqrt{u}\,\frac{\partial J}{\partial v}\left[\rho(1-\rho)\frac{\partial J}{\partial v} - \frac{\partial J}{\partial u}\right] \neq 0. \tag{2.26}$$

L'existence de la statistique (2.21) indique que $\partial J/\partial\sigma_0^2$ et $\partial J/\partial\sigma_1^2$ ne peuvent être simultanément égaux à 0. En utilisant les expressions figurant en (2.25), ceci implique que $\partial J/\partial v \neq 0$. Dans ces circonstances, la condition (2.26) devient

$$\rho(1-\rho)\frac{\partial J}{\partial v} \neq \frac{\partial J}{\partial u}. \tag{2.27}$$

Pour tout $\rho \in]0, 1[$, l'équation (2.27) peut être résolue en effectuant les changements de variables $u = u' + v'$ et $v = \rho(1-\rho)(u' - v')$. Il en résulte que les critères de la forme $J(u + v/\rho(1-\rho))$ ne sont pas recevables, dont $J(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2) = \rho(1-\rho)(\eta_1 - \eta_0)^2 + \rho\sigma_1^2 + (1-\rho)\sigma_0^2$ constitue un exemple. On peut par ailleurs vérifier dans ce dernier cas que la statistique (2.21) ne dépend par de $\lambda^*(\boldsymbol{x})$, ce qui est conforme au résultat attendu. Enfin, la résolution de l'équation (2.27) dans le cas où $\rho \in \{0, 1\}$ conduit à $\partial J/\partial u \neq 0$, ce qui signifie que J doit dépendre explicitement de la quantité $(\eta_1 - \eta_0)^2$.

En conservant les notations introduites ci-dessus, le rapport signal-sur-bruit généralisé s'écrit $J_{rsbg}(u, v) = u/v$. A partir de la condition d'admissibilité (2.27) et des solutions exhibées, on vérifie immédiatement qu'il s'agit là d'un critère pertinent [Abd02(a)]. Ce résultat peut être complété en donnant l'expression de toute statistique conjointement optimale au sens de J_{rsbg} et des théories statistiques classiques de la décision, résultat que l'on obtient à partir de la relation (2.21) après avoir évalué $\partial J_{rsbg}/\partial\eta_i$ et $\partial J_{rsbg}/\partial\sigma_i^2$:

$$\lambda(\boldsymbol{x}) = \alpha\,\frac{\lambda^*(\boldsymbol{x}) - 1}{\rho(\lambda^*(\boldsymbol{x}) - 1) + 1} + \beta, \quad \alpha \in \mathbb{R}^*, \quad \beta \in \mathbb{R}. \tag{2.28}$$

où $\lambda^*(\boldsymbol{x})$ désigne le rapport de vraisemblance, soit $\lambda^*(\boldsymbol{x}) = p(\boldsymbol{x}|\omega_1)/p(\boldsymbol{x}|\omega_0)$.

2.4.3 Analyse de l'erreur quadratique moyenne

On complète à présent l'illustration de la condition d'admissibilité (2.23) en considérant l'erreur quadratique moyenne, critère dont on a déjà évoqué l'intérêt au cours du chapitre 1. Étant donnée une observation $\boldsymbol{X}$, on note $\gamma(\boldsymbol{X})$ la valeur désirée pour la statistique $\lambda(\boldsymbol{X})$. L'erreur quadratique moyenne entre cette dernière et la valeur obtenue est définie par

$$\begin{aligned} J_{eq} &= \mathrm{E}\{(\lambda(\boldsymbol{X}) - \gamma(\boldsymbol{X}))^2\} && (2.29)\\ &= \mathrm{E}\{\lambda^2(\boldsymbol{X})\} - 2\mathrm{E}\{\gamma(\boldsymbol{X})\,\lambda(\boldsymbol{X})\} + \mathrm{E}\{\gamma^2(\boldsymbol{X})\}. && (2.30)\end{aligned}$$

S'il existe de multiple façons de choisir pratiquement $\gamma(\boldsymbol{X})$, on pose ici $\gamma(\boldsymbol{X}) = \gamma_0$ si $\boldsymbol{X}$ appartient à ω_0, et $\gamma(\boldsymbol{X}) = \gamma_1$ s'il s'agit d'un élément de ω_1. Afin d'appliquer le théorème d'admissibilité 2.4.1, il convient d'exprimer préalablement J_{eq} en fonction des moments conditionnels η_i et σ_i^2 de $\lambda(\boldsymbol{X})$. Le premier terme $\mathrm{E}\{\lambda^2(\boldsymbol{X})\}$ figurant dans la définition (2.30) peut se réécrire

$$\mathrm{E}\{\lambda^2(\boldsymbol{X})\} = p(\omega_0)\,(\sigma_0^2 + \eta_0^2) + p(\omega_1)\,(\sigma_1^2 + \eta_1^2), \tag{2.31}$$

tandis que le second terme s'exprime ainsi

$$\mathrm{E}\{\gamma(\boldsymbol{X})\,\lambda(\boldsymbol{X})\} = \gamma_0\,p(\omega_0)\,\eta_0 + \gamma_1\,p(\omega_1)\,\eta_1. \tag{2.32}$$

En introduisant les expressions (2.31) et (2.32) dans la définition de J_{eq}, on obtient

$$J_{eq} = p(\omega_0)\,(\sigma_0^2 + \eta_0^2) + p(\omega_1)\,(\sigma_1^2 + \eta_1^2) - 2[\gamma_0\,p(\omega_0)\,\eta_0 + \gamma_1\,p(\omega_1)\,\eta_1] + \mathrm{E}\{\gamma^2(\boldsymbol{X})\}. \tag{2.33}$$

On note que le terme $\mathrm{E}\{\gamma^2(\boldsymbol{X})\}$ est indépendant de $\lambda(\boldsymbol{X})$, et par conséquent de η_i et σ_i^2. L'erreur quadratique moyenne est donc un critère du second ordre dont on peut étudier la pertinence. Après un calcul des dérivées partielles de J_{eq} par rapport à η_i et σ_i^2, la condition de recevabilité (2.23) se traduit par [Abd02(a)]

$$2(\gamma_1 - \gamma_0)\,p(\omega_0)\,p(\omega_1) \neq 0, \tag{2.34}$$

à partir duquel on peut conclure que l'erreur quadratique moyenne est bien un critère pertinent.

Les résultats présentés dans cette section rendent légitime l'usage des critères du second ordre lorsqu'une pénurie relative d'information sur les hypothèses ω_0 et ω_1 conduit à rechercher un détecteur au sein d'une classe $\mathcal{D}$ préalablement fixée. L'efficacité de cette approche est cependant conditionnée par le fait qu'il doit exister dans $\mathcal{D}$ au moins un détecteur équivalent au test du rapport de vraisemblance. Dans le cas contraire, le choix du critère du second ordre a une influence capitale sur les performances de la structure de décision qui en résulte, comme le montre la figure 2.4. La méthode de recherche du critère optimal étudiée dans la section suivante vise à combler cette lacune.

2.5 Sélection du critère optimal

L'étude qui vient d'être menée a permis de démontrer théoriquement l'intérêt que présente la classe des critères du second ordre pour la synthèse de détecteurs linéaires généralisés. Le choix d'un critère particulier n'en demeure pas moins un problème crucial rarement abordé dans la littérature, les solutions arbitrairement retenues étant souvent le rapport signal-sur-bruit ou la déflexion. La méthodologie présentée dans cette section répond à cette attente en permettant

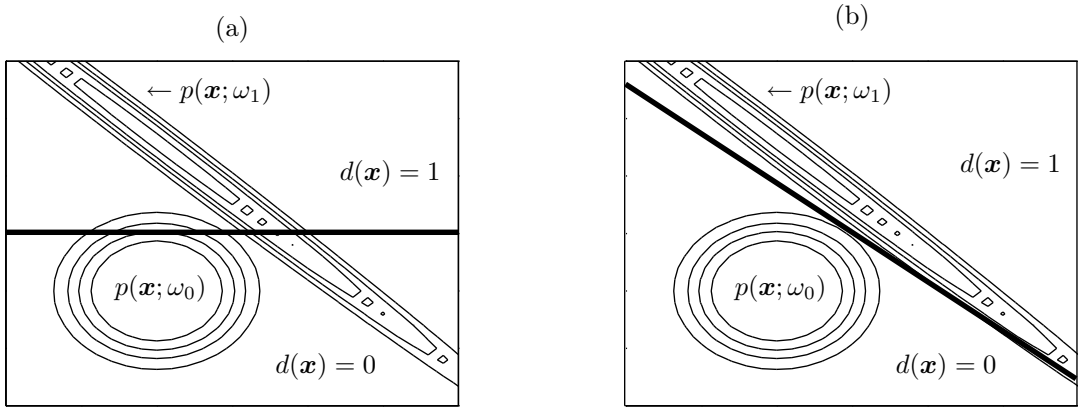

FIG. 2.4 : Détection d'un signal aléatoire gaussien noyé dans un bruit blanc gaussien. Partition de l'espace des réalisations par un détecteur linéaire maximisant (a) le rapport signal-sur-bruit et (b) la déflexion.

de déterminer le meilleur critère du second ordre, c'est-à-dire celui pour lequel le détecteur a une probabilité d'erreur minimale. Cette approche, relativement méconnue, a été initialement proposée pour la recherche de discriminants linéaires en reconnaissance des formes [Pat66]. Elle a par la suite été adaptée au problème de la synthèse de détecteurs à structure imposée, dans le cadre duquel elle a connu de nouveaux développements [Ric98(a)] et apporté un éclairage nouveau sur les critères du second ordre [Abd02(b)]. Les avancées les plus récentes concernent une méthode de synthèse de détecteurs à noyau [Abd02(c), Abd02(d)], qui s'affirme comme une alternative sérieuse aux Support Vector Machines par les performances affichées. Ces divers éléments, décrits en cette fin de chapitre et dans le suivant, sont le résultat de travaux menés personnellement puis avec un doctorant, Fahed Abdallah.

2.5.1 Principe de la méthode

Soit $\lambda(\boldsymbol{x})$ une statistique linéaire généralisée définie par $\lambda(\boldsymbol{x}) = \boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{x}) - \lambda_0$. Le problème de la synthèse d'un détecteur à structure imposée, tel qu'il a été défini précédemment, consiste à rechercher $\boldsymbol{w}$ et λ_0 optimaux au sens d'un critère du second ordre J ne dépendant que des moments conditionnels suivants :

$$\eta_i = \mathrm{E}\{\lambda(\boldsymbol{X})|\omega_i\} = \mathrm{E}\{\boldsymbol{w}\cdot\boldsymbol{\phi}(\boldsymbol{X}) - \lambda_0|\omega_i\} = \boldsymbol{w}\cdot\boldsymbol{m}_i - \lambda_0 \tag{2.35}$$

$$\sigma_i^2 = \mathrm{var}\{\lambda(\boldsymbol{X})|\omega_i\} = \boldsymbol{w}^t\mathrm{E}\left\{(\boldsymbol{X}-\boldsymbol{m}_i)(\boldsymbol{X}-\boldsymbol{m}_i)^t|\omega_i\right\}\boldsymbol{w} = \boldsymbol{w}^t\,\boldsymbol{\Sigma}_i\,\boldsymbol{w}, \tag{2.36}$$

avec $\boldsymbol{m}_i = \mathrm{E}\{\boldsymbol{\phi}(\boldsymbol{X})|\omega_i\}$ et $\boldsymbol{\Sigma}_i = \mathrm{E}\{(\boldsymbol{\phi}(\boldsymbol{X})-\boldsymbol{m}_i)(\boldsymbol{\phi}(\boldsymbol{X})-\boldsymbol{m}_i)^t|\omega_i\}$. Dans ces circonstances, les dérivées de J par rapport à $\boldsymbol{w}$ et λ_0 doivent être nulles, c'est-à-dire

$$\left\{\begin{array}{l} \frac{\partial J}{\partial \boldsymbol{w}} = \frac{\partial J}{\partial \sigma_0^2}\cdot\frac{\partial \sigma_0^2}{\partial \boldsymbol{w}} + \frac{\partial J}{\partial \sigma_1^2}\cdot\frac{\partial \sigma_1^2}{\partial \boldsymbol{w}} + \frac{\partial J}{\partial \eta_0}\cdot\frac{\partial \eta_0}{\partial \boldsymbol{w}} + \frac{\partial J}{\partial \eta_1}\cdot\frac{\partial \eta_1}{\partial \boldsymbol{w}} = 0 \\ \frac{\partial J}{\partial \lambda_0} = \frac{\partial J}{\partial \sigma_0^2}\cdot\frac{\partial \sigma_0^2}{\partial \lambda_0} + \frac{\partial J}{\partial \sigma_1^2}\cdot\frac{\partial \sigma_1^2}{\partial \lambda_0} + \frac{\partial J}{\partial \eta_0}\cdot\frac{\partial \eta_0}{\partial \lambda_0} + \frac{\partial J}{\partial \eta_1}\cdot\frac{\partial \eta_1}{\partial \lambda_0} = 0, \end{array}\right. \tag{2.37}$$

où les dérivées partielles de η_i et σ_i^2 sont données par

$$\frac{\partial \sigma_i^2}{\partial \boldsymbol{w}} = 2\,\boldsymbol{\Sigma}_i\,\boldsymbol{w}, \quad \frac{\partial \eta_i}{\partial \boldsymbol{w}} = \boldsymbol{m}_i, \quad \frac{\partial \sigma_i^2}{\partial \lambda_0} = 0, \quad \frac{\partial \eta_i}{\partial \lambda_0} = -1. \tag{2.38}$$

Ces résultats permettent de réécrire le système (2.37) ainsi :

$$\begin{cases} 2\left[\frac{\partial J}{\partial \sigma_0^2}\boldsymbol{\Sigma}_0 + \frac{\partial J}{\partial \sigma_1^2}\boldsymbol{\Sigma}_1\right]\boldsymbol{w} = -\left[\frac{\partial J}{\partial \eta_0}\boldsymbol{m}_0 + \frac{\partial J}{\partial \eta_1}\boldsymbol{m}_1\right] \\ \frac{\partial J}{\partial \eta_0} + \frac{\partial J}{\partial \eta_1} = 0. \end{cases} \tag{2.39}$$

En introduisant la deuxième équation du système (2.39) dans la première, et en notant que $\boldsymbol{w}$ peut être défini à un coefficient multiplicatif près, on aboutit finalement au système linéaire

$$[\rho\boldsymbol{\Sigma}_0 + (1-\rho)\boldsymbol{\Sigma}_1]\,\boldsymbol{w} = (\boldsymbol{m}_1 - \boldsymbol{m}_0), \tag{2.40}$$

avec

$$\rho = \frac{\frac{\partial J}{\partial \sigma_0^2}}{\frac{\partial J}{\partial \sigma_0^2} + \frac{\partial J}{\partial \sigma_1^2}}. \tag{2.41}$$

Ce résultat est particulièrement intéressant puisqu'il établit que le choix du critère n'intervient dans l'expression de $\boldsymbol{w}$ que par l'intermédiaire d'un paramètre ρ, qu'il est aisé d'optimiser au moyen d'une simple procédure itérative. D'après l'expression (2.41), on note que cette recherche peut être limitée à l'intervalle $[0, 1]$ si l'on restreint le cadre de l'étude aux critères du second ordre dont les sens de variation par rapport à σ_0^2 et σ_1^2 ne sont pas contraires, propriété par ailleurs légitimement souhaitable mais non nécessaire. Quant au terme λ_0, on peut montrer qu'il dépend explicitement de J par le biais de la deuxième équation du système (2.39). Ceci ne constitue cependant pas un obstacle, l'effet de λ_0 étant annulé par la sélection d'un seuil de détection plus approprié. La méthodologie proposée, dite *du critère optimal*, repose sur ces propriétés. Comme l'indique l'algorithme proposé en figure 2.5, elle consiste en l'optimisation conjointe des paramètres ρ et λ_ρ de la structure de détection

$$d_\rho(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \boldsymbol{w}_\rho \cdot \boldsymbol{\phi}(\boldsymbol{x}) - \lambda_\rho > 0 \\ 0 & \text{sinon,} \end{cases} \tag{2.42}$$

de sorte à minimiser, par exemple, la probabilité d'erreur $P_e(d_\rho)$. De cette façon, on détermine un détecteur linéaire optimum au sens du meilleur critère du second ordre, sans qu'il soit nécessaire de l'exhiber. En pratique, cette probabilité d'erreur ne peut être évaluée. On doit alors disposer d'une estimation, que peut fournir par exemple l'erreur empirique sur l'ensemble d'apprentissage $\mathcal{A}_n$. D'autres procédures sont cependant préconisées, parmi lesquelles on recense les méthodes de bootstrap ou encore de validation croisée. Celles-ci ont fait l'objet de la Section 1.3.

1. Initialiser ρ à 0
2. Tant que $\rho \leq 1$, répéter
   - résoudre l'équation (2.40) pour obtenir $\boldsymbol{w}_\rho$
   - déterminer le seuil λ_ρ de sorte à minimiser par exemple $P_e(d_\rho)$
   - mise à jour de ρ : $\rho \leftarrow \rho + \Delta\rho$, avec $\Delta\rho$ préalablement choisi
3. Sélectionner le meilleur détecteur d_ρ obtenu, caractérisé par $(\boldsymbol{w}_\rho, \lambda_\rho)$

FIG. 2.5 : Algorithme de la méthode du critère optimal.

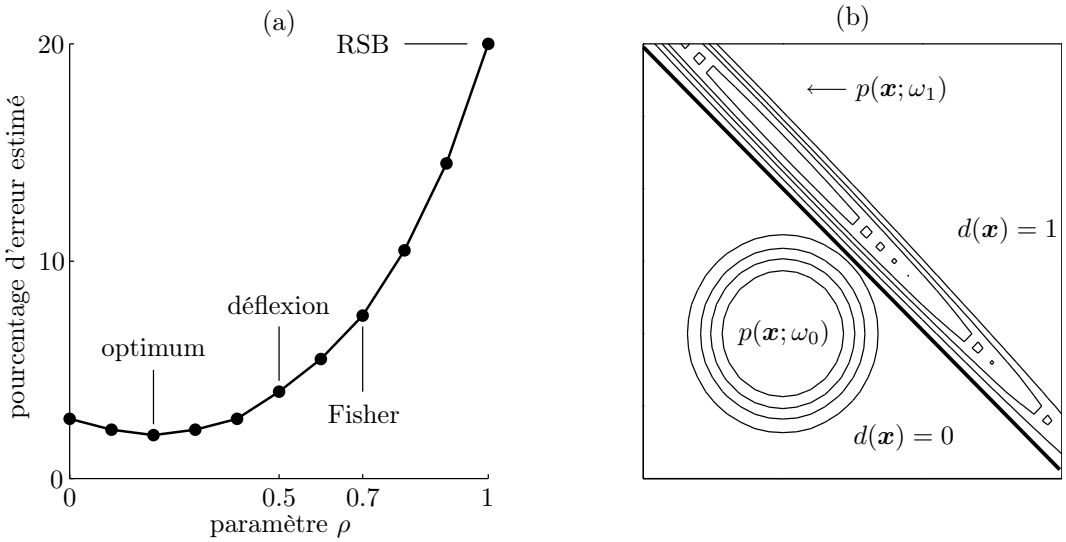

FIG. 2.6 : Détection d'un signal aléatoire gaussien, de probabilité *a priori* $p(\omega_0)$ égale à 0.7, noyé dans un bruit blanc gaussien. (a) La méthode du critère optimal, qui consiste à rechercher le paramètre ρ minimisant la probabilité d'erreur, conduit à un détecteur plus performant que ceux résultant de la maximisation des critères classiques. (b) La partition de l'espace des réalisations effectuée par le détecteur optimal est plus satisfaisante que celles présentées en figure 2.4.

Le détecteur obtenu par la méthode du critère optimal est au moins aussi performant que ceux résultant de la maximisation du rapport signal-sur-bruit, de la déflexion ou encore du critère de Fisher. En effet, il leur correspond à chacun une valeur particulière de ρ, comme l'illustre la figure 2.6. Ainsi, en appliquant la relation (2.41) à la définition du rapport signal-sur-bruit par exemple, puis en supposant $\boldsymbol{\Sigma}_0$ inversible, on retrouve le résultat bien connu suivant :

$$\boldsymbol{w}_{rsb} = \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_0), \tag{2.43}$$

car $\rho_{rsb} = 1$. De façon analogue, on établit pour la déflexion et le critère de Fisher :

$$\boldsymbol{w}_{deflex} = 2\,(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_0) \tag{2.44}$$

$$\boldsymbol{w}_{Fisher} = (p(\omega_0)\,\boldsymbol{\Sigma}_0 + p(\omega_1)\,\boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_0), \tag{2.45}$$

car $\rho_{deflex} = 1/2$ et $\rho_{Fisher} = p(\omega_0)$. Ceci démontre donc clairement que les détecteurs obtenus par maximisation du rapport signal-sur-bruit, de la déflexion ou encore du critère de Fisher présentent *a priori* moins de garanties en termes de performances que celui résultant de la méthode du critère optimal.

2.5.2 Expérimentations

L'objectif de cette section est d'illustrer la méthode du critère optimal en l'appliquant à un problème de détection classique, afin de mettre en lumière différentes situations auxquelles on peut être confronté lors de sa mise en œuvre. Ainsi, on suppose que les l composantes de l'observation sont mutuellement indépendantes et distribuées selon des lois exponentielles, soit

$$p(\boldsymbol{x}|\omega_i) = \prod_{j=1}^{l} \frac{1}{\delta_{ij}} \exp\left(-\frac{1}{\delta_{ij}}\,\boldsymbol{x}[j]\right) \Gamma(\boldsymbol{x}[j]). \tag{2.46}$$

Dans cette expression, $\Gamma(\cdot)$ représente la fonction d'Heaviside, et δ_{ij} désigne le paramètre de la loi exponentielle gouvernant la variable aléatoire $\boldsymbol{X}[j]$, conditionnellement à l'hypothèse ω_i. Il est à noter que l'espérance mathématique et la covariance conditionnelles de $\boldsymbol{X}$ sont respectivement définies par

$$\boldsymbol{m}_i = (\delta_{i1}, \ldots, \delta_{ij}, \ldots, \delta_{il})^t \tag{2.47}$$

$$\boldsymbol{\Sigma}_i = \mathrm{diag}(\delta_{i1}^2, \ldots, \delta_{ij}^2, \ldots, \delta_{il}^2), \tag{2.48}$$

où $\mathrm{diag}(\cdot)$ désigne une matrice diagonale. Sous ces hypothèses, on démontre aisément qu'une statistique équivalente au rapport de vraisemblance est donnée par :

$$\lambda^*(\boldsymbol{x}) = \sum_{j=1}^{l} \left(\frac{1}{\delta_{0j}} - \frac{1}{\delta_{1j}} \right) \boldsymbol{x}[j]. \tag{2.49}$$

Ceci signifie qu'une décision optimum peut être prise en projetant les observations selon une direction définie par

$$\boldsymbol{w}^* = \left(\frac{\delta_{11} - \delta_{01}}{\delta_{11}\,\delta_{01}}, \ldots, \frac{\delta_{1j} - \delta_{0j}}{\delta_{1j}\,\delta_{0j}}, \ldots, \frac{\delta_{1l} - \delta_{0l}}{\delta_{1l}\,\delta_{0l}} \right). \tag{2.50}$$

On recherche à présent une structure de détection par optimisation d'un critère J du second ordre, en raisonnant sur la classe $\mathcal{D}$ des détecteurs linéaires par rapport aux observations. Par la résolution de l'équation (2.40), on aboutit à la direction de projection

$$\boldsymbol{w}_\rho = \left(\frac{\delta_{11} - \delta_{01}}{\delta_{11}^2 - \rho(\delta_{11}^2 - \delta_{01}^2)}, \ldots, \frac{\delta_{1j} - \delta_{0j}}{\delta_{1j}^2 - \rho(\delta_{1j}^2 - \delta_{0j}^2)}, \ldots, \frac{\delta_{1l} - \delta_{0l}}{\delta_{1l}^2 - \rho(\delta_{1l}^2 - \delta_{0l}^2)} \right), \tag{2.51}$$

sous réserve que le dénominateur de chacune des composantes de $\boldsymbol{w}_\rho$ ne s'annule pas. Sinon, $\boldsymbol{w}_\rho$ préalablement normalisé prend la forme d'un vecteur unitaire du repère cartésien adopté. Dans tous les cas, on relève que l'équivalence des solutions obtenues par la méthode du critère optimal et l'approche statistique repose sur l'existence d'une valeur de ρ telle que la colinéarité des vecteurs $\boldsymbol{w}^*$ et $\boldsymbol{w}_\rho$ soit assurée.

Dans un premier temps, on considère le cas d'observations bidimensionnelles ($l = 2$), avec $\delta_{01} \neq \delta_{11}$ et $\delta_{02} \neq \delta_{12}$. La recherche de ρ consiste alors en la résolution d'une équation linéaire à une inconnue. On montre ainsi que les vecteurs $\boldsymbol{w}^*$ et $\boldsymbol{w}_\rho$ sont colinéaires si et seulement si

$$\rho = \frac{\delta_{02}\,\delta_{12}\,\delta_{11}^2 - \delta_{01}\,\delta_{11}\,\delta_{12}^2}{\delta_{02}\,\delta_{12}\,(\delta_{11}^2 - \delta_{01}^2) - \delta_{01}\,\delta_{11}\,(\delta_{12}^2 - \delta_{02}^2)}. \tag{2.52}$$

Par la relation (2.41), ce résultat fournit une condition sur le critère J, que vérifie par exemple le rapport signal-sur-bruit généralisé (2.14) lorsque le paramètre ρ qui le caractérise est défini selon l'expression (2.52). La figure 2.7 illustre l'évolution de $\boldsymbol{w}_\rho$, préalablement normalisé, en fonction de ρ. Celle-ci laisse apparaître deux discontinuités correspondant à l'annulation du dénominateur des composantes de $\boldsymbol{w}_\rho$. On note également la colinéarité de $\boldsymbol{w}_{-\infty}$ et $\boldsymbol{w}_{+\infty}$, qui résulte de leur proportionnalité à $(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_0)$. En effet, on a

$$\boldsymbol{w}_{\pm\infty} \propto \lim_{\rho \to \pm\infty} \frac{1}{\rho} (\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1)^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_0), \tag{2.53}$$

d'après la relation (2.40), à un facteur de normalisation près tel que $\|\boldsymbol{w}_{\pm\infty}\| = 1$.

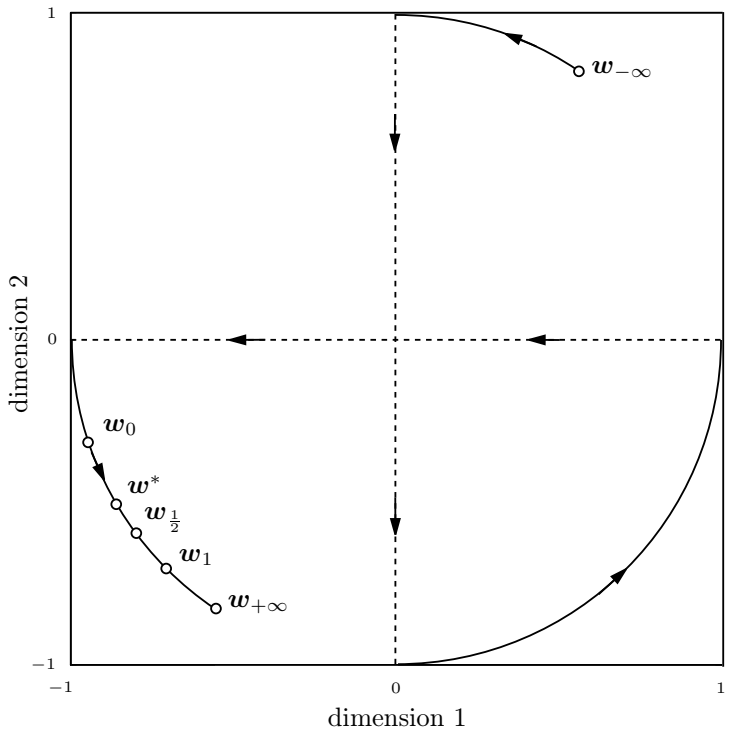

FIG. 2.7 : Evolution de $\boldsymbol{w}_\rho$ en fonction de ρ dans le cas d'observations bidimensionnelles gouvernées par des lois conditionnelles de type exponentiel ($\delta_{01} = 5$, $\delta_{11} = 2$, $\delta_{02} = 3$, $\delta_{12} = 2$).

On suppose à présent que les observations sont de dimension l strictement supérieure à 2, avec $\delta_{0j} \neq \delta_{1j}$ pour tout $j \in \{1, \ldots, l\}$. Comme précédemment, la recherche d'un paramètre ρ assurant la colinéarité de $\boldsymbol{w}^*$ et $\boldsymbol{w}_\rho$ se traduit par la résolution d'un système de $l - 1$ équations linéaires à une inconnue. Sauf situation singulière, il n'existe évidemment pas de solution à ce problème, ce qui signifie que les statistiques obtenues par les deux approches considérées ne sont pas équivalentes. Les figures 2.8, 2.9 et 2.10 illustrent ce scénario dans le cas d'observations tridimensionnelles, en proposant une représentation du vecteur $\boldsymbol{w}_\rho$ sur la sphère unité pour $\rho \in]-\infty, +\infty[$. On peut ainsi observer qu'il n'existe pas pour l'exemple choisi de valeur ρ_0 telle que $\boldsymbol{w}_{\rho_0}$ and $\boldsymbol{w}^*$ soient colinéaires. A titre anecdotique, on relève également la colinéarité de $\boldsymbol{w}_{-\infty}$ et $\boldsymbol{w}_{+\infty}$, que justifie la relation (2.53).

L'expérimentation proposée rappelle simplement qu'au sein d'une classe $\mathcal{D}$ donnée, les règles optimum au sens de Bayes et d'un quelconque critère du second ordre ne coincident pas nécessairement. Afin qu'elles ne forment qu'une, il est nécessaire de relaxer la contrainte pesant sur la structure des détecteurs, sous réserve que le critère du second ordre sélectionné soit pertinent au sens du théorème 2.4.1. Pour tendre naturellement vers cette situation idéale, il convient donc de mettre en œuvre la méthode du critère optimal sur des classes de détecteurs particulièrement riches, tout en contenant les effets néfastes du fléau de la dimensionnalité décrits au chapitre précédent. Ces nouveaux développements font l'objet du troisième chapitre.

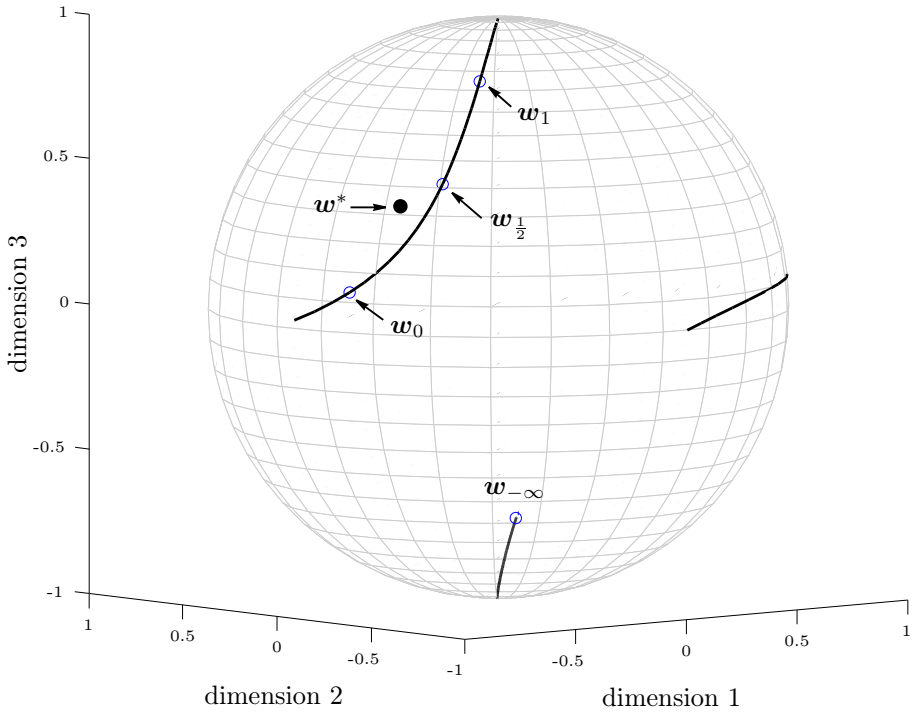

FIG. 2.8 : Evolution de $\boldsymbol{w}_\rho$ en fonction de ρ dans le cas d'observations tridimensionnelles gouvernées par des lois conditionnelles de type exponentiel ($\delta_{01} = 5$, $\delta_{11} = 2$, $\delta_{02} = 3$, $\delta_{12} = 2$, $\delta_{03} = 2$, $\delta_{13} = 3$).

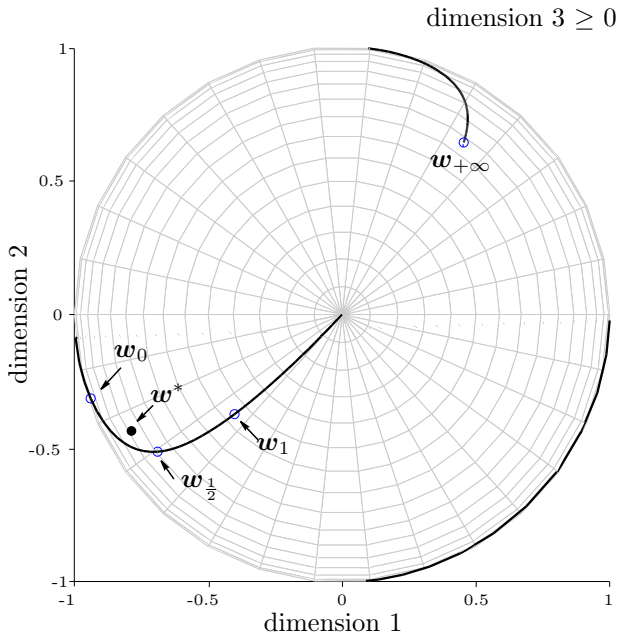

FIG. 2.9 : Idem que la figure 2.8. Vue de dessus de la sphère unité, c'est-à-dire dimension 3 ≥ 0.

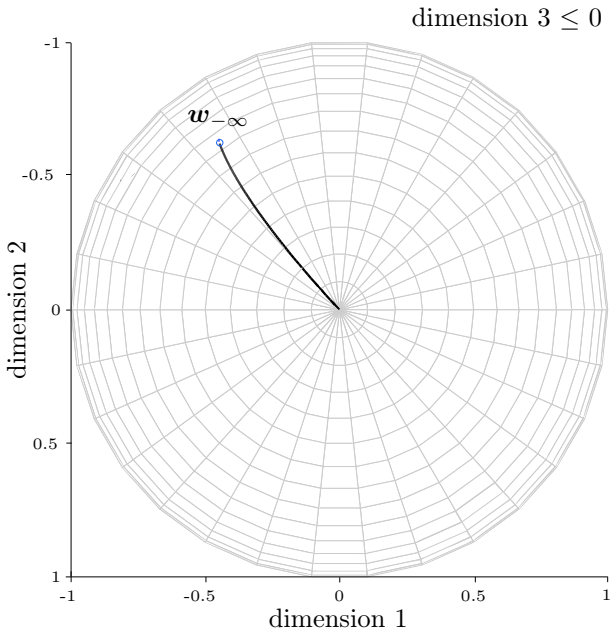

FIG. 2.10 : Idem que la figure 2.8. Vue de dessous de la sphère unité, c'est-à-dire dimension $3 \leq 0$.

Chapitre 3

Détecteurs à noyau reproduisant et contrôle de complexité

3.1 Introduction

L'un des tests de détection les plus élémentaires que l'on puisse imaginer consiste à supposer que la statistique $\lambda(\boldsymbol{x})$ sur laquelle il repose dépend linéairement des l composantes de l'observation, soit :

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \lambda(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} - \lambda_0 > 0 \\ 0 & \text{sinon.} \end{cases} \tag{3.1}$$

Ce type de structure présente l'avantage d'être directement lié à la notion de filtrage, ce qui en facilite l'interprétation. Cependant, il n'offre de solutions optimales que pour une classe réduite de problèmes. En particulier, lorsque les hypothèses en compétition suivent des lois gaussiennes, il est nécessaire que celles-ci aient les mêmes propriétés statistiques d'ordre 2. Afin de répondre de façon satisfaisante à un plus grand nombre de problèmes tout en conservant une structure comparable à (3.1), on peut envisager de plonger préalablement les observations dans un espace de représentation plus adéquat, grâce au concours d'une application $\boldsymbol{\phi}$:

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \lambda(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) - \lambda_0 > 0 \\ 0 & \text{sinon.} \end{cases} \tag{3.2}$$

Pour l'élaboration d'une règle de décision de nature quadratique, ce concept en apparence basique préconise ainsi la mise en œuvre d'une technique de discrimination linéaire sur la représentation $\boldsymbol{\phi}(\boldsymbol{x})$ constituée des composantes de $\boldsymbol{x}$ et de leurs produits croisés. Il est à noter que ce test peut mener, en particulier, à une situation d'optimalité au sens des approches statistiques classiques lorsque les hypothèses en compétition suivent des lois gaussiennes de paramètres distincts. Le nombre de composantes de $\boldsymbol{\phi}(\boldsymbol{x})$, égal à $l(l+3)/2$ et déjà prohibitif au regard de la relative simplicité de la statistique considérée, laisse présager des difficultés pratiques rencontrées lorsque le schéma proposé est adopté pour l'élaboration de règles de décision complexes. Il y a dix ans, des considérations sur les espaces de Hilbert à noyau reproduisant ont toutefois rendu cette pratique possible dans le cadre des Support Vector Machines [Bos92]. En autorisant la synthèse de la règle (3.2) sans jamais évaluer explicitement $\boldsymbol{\phi}(\boldsymbol{x})$, ces structures algébriques ont depuis ouvert la voie des méthodes à noyau [Sho99], que l'on se propose d'explorer dans le cadre de la méthode du critère optimal. Les résultats exposés dans ce chapitre sont issus de travaux menés de concert avec un doctorant, Fahed Abdallah [Abd02(c), Abd02(d)].

Quelle que soit la méthode d'apprentissage adoptée, les performances en généralisation d'un détecteur issu d'une classe $\mathcal{D}$ sont conditionnées par de multiples facteurs tels que sa structure, le nombre de données disponibles et la dimension de l'espace qu'elles engendrent [Guy93, Vap95]. Conformément aux développements du chapitre 1, l'obtention d'une solution performante en termes de probabilité d'erreur nécessite l'adaptation de la VC-dimension de $\mathcal{D}$ à la taille de la base d'apprentissage $\mathcal{A}_n$. Ainsi, les récepteurs dotés d'une capacité d'apprentissage trop importante ont généralement un faible pouvoir de généralisation. Dans le cas contraire, il leur est souvent impossible d'intégrer la totalité de l'information statistique présente dans $\mathcal{A}_n$. Entre ces extrêmes, il existe une configuration optimum vers laquelle il s'agit de tendre en faisant varier la VC-dimension de $\mathcal{D}$, conformément au principe de minimisation du risque structurel. Afin que le lecteur soit en mesure d'accomplir cette tâche, il convient donc de proposer des outils adaptés à la méthode du critère optimal [Ric98(a)], ainsi qu'à sa formulation étendue aux noyaux reproduisants [Abd02(d)]. Mais avant, il est nécessaire de décrire le cadre algébrique dans lequel s'inscrivent les développements à venir.

3.2 Espaces à noyau reproduisant et condition de Mercer

Soit $\mathcal{H}$ un espace fonctionnel hilbertien réel de produit scalaire $< \cdot ; \cdot >_{\mathcal{H}}$, composé de fonctions ϕ continues sur un ensemble $\mathcal{X}$. D'après le théorème de représentation de Riesz [Sai88], il existe une fonction unique de la variable $\boldsymbol{x}_1$, notée $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2)$, telle que

$$\phi(\boldsymbol{x}_2) = < \phi \, ; \kappa(\cdot, \boldsymbol{x}_2) >_{\mathcal{H}}, \qquad \forall \phi \in \mathcal{H}. \tag{3.3}$$

Dans cette expression, $\kappa(\cdot, \boldsymbol{x}_2)$ désigne une fonction définie sur $\mathcal{X}$, obtenue en fixant le second argument de κ à $\boldsymbol{x}_2$. Il en résulte que l'ensemble $\{\kappa(\cdot, \boldsymbol{x}) \, : \, \boldsymbol{x} \in \mathcal{X}\}$ engendre $\mathcal{H}$, et que le produit scalaire $< \cdot ; \cdot >_{\mathcal{H}}$ ne nécessite d'être défini que sur cet ensemble de générateurs. Au vu de cette propriété, κ est appelé *noyau reproduisant* de $\mathcal{H}$. L'équation (3.3) implique de plus que

$$< \kappa(\cdot, \boldsymbol{x}_1) \, ; \kappa(\cdot, \boldsymbol{x}_2) >_{\mathcal{H}} = \kappa(\boldsymbol{x}_2, \boldsymbol{x}_1), \tag{3.4}$$

pour tout $\boldsymbol{x}_1$, $\boldsymbol{x}_2 \in \mathcal{X}$, et met en évidence le caractère symétrique de κ. Ce résultat signifie que $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2)$ fournit le produit scalaire des images dans $\mathcal{H}$ de toute paire d'éléments de l'ensemble $\mathcal{X}$. Afin d'exploiter ce concept sans éventuellement construire explicitement $\mathcal{H}$, il convient à présent de caractériser les noyaux κ ayant un caractère reproduisant.

Selon la théorie de Hilbert-Schmidt [Cou53], toute fonction symétrique $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2)$ de $\mathcal{L}_2(\mathcal{X}^2)$ admet une décomposition de la forme

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_i \gamma_i \, \phi_i(\boldsymbol{x}_1) \, \phi_i(\boldsymbol{x}_2), \tag{3.5}$$

avec $\phi_i \in \mathcal{L}_2(\mathcal{X})$ et $\gamma_i \in \mathbb{R}$. Les éléments ϕ_i et γ_i intervenant dans cette expression correspondent aux fonctions propres et valeurs propres de l'opérateur intégrale défini par le noyau κ, soit :

$$\int \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) \, \phi_i(\boldsymbol{x}_1) \, d\boldsymbol{x}_1 = \gamma_i \, \phi_i(\boldsymbol{x}_2). \tag{3.6}$$

D'après l'expression (3.5), une condition suffisante pour que $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2)$ soit un produit scalaire est que les valeurs propres γ_i soient positives. Il en est ainsi, selon le théorème de Mercer, si et seulement si la condition suivante est satisfaite pour toute fonction f de $\mathcal{L}_2(\mathcal{X})$:

$$\iint \kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) \, f(\boldsymbol{x}_1) \, f(\boldsymbol{x}_2) \, d\boldsymbol{x}_1 \, d\boldsymbol{x}_2 \geq 0. \tag{3.7}$$

Les fonctions κ vérifiant cette relation sont appelées *noyaux de Mercer*. En considérant la relation (3.4), on peut finalement énoncer le théorème suivant [Bos92, Cor95] :

Théorème 3.2.1. *A tout noyau de Mercer κ, on peut associer un espace fonctionnel hilbertien réel $\mathcal{H}$ à noyau reproduisant. Dans ce cadre, on a en particulier $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = <\psi(\boldsymbol{x}_1)\,;\psi(\boldsymbol{x}_2)>_{\mathcal{H}}$, où $\psi(\boldsymbol{x})$ désigne la fonction $\kappa(\cdot, \boldsymbol{x})$ de $\mathcal{H}$, étant donné $\boldsymbol{x}$ un élément de $\mathcal{X}$.*

De façon plus pragmatique, on note d'après l'équation (3.5) que $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{\phi}(\boldsymbol{x}_1) \cdot \boldsymbol{\phi}(\boldsymbol{x}_2)$ avec $\boldsymbol{\phi}(\boldsymbol{x}) = \big(\phi_1(\boldsymbol{x})/\sqrt{\gamma_1},\, \phi_2(\boldsymbol{x})/\sqrt{\gamma_2},\, \ldots\big)$. A titre d'exemple, le noyau polynomial défini par $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = (1 + \boldsymbol{x}_1 \cdot \boldsymbol{x}_2)^q$ vérifie la condition de Mercer [Bos92], ce qui signifie qu'il fournit à moindre coup de calcul le produit scalaire canonique des images de $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ par une application $\boldsymbol{\phi}$. Par la relation (3.6), il est aisé de montrer que les composantes de $\boldsymbol{\phi}(\boldsymbol{x})$ sont les monômes de degré inférieur à q constitués des composantes de $\boldsymbol{x}$. En particulier, pour $\boldsymbol{x} \in \mathbb{R}^2$ et $q = 2$, on a

$$\boldsymbol{\phi}(\boldsymbol{x}) = \Big(1,\, \sqrt{2}\,x[1],\, \sqrt{2}\,x[2],\, \sqrt{2}\,x[1]\,x[2],\, x[1]^2,\, x[2]^2\Big)\,. \tag{3.8}$$

D'autres exemples de noyaux reproduisants sont présentés dans la section suivante, qui est consacrée à leur exploitation pour la synthèse de détecteurs par la méthode du critère optimal.

3.3 Méthode du critère optimal à noyau reproduisant

Une méthode intuitive pour l'élaboration de structures de décision non-linéaires consiste à plonger préalablement les échantillons dans un espace transformé $\mathcal{T}$ au moyen d'une application non-linéaire

$$\begin{array}{rccc} \boldsymbol{\phi}: & \mathbb{R}^l & \longrightarrow & \mathcal{T} \\ & \boldsymbol{x} & \longmapsto & \boldsymbol{\phi}(\boldsymbol{x}), \end{array}$$

puis à mettre en œuvre une technique de discrimination linéaire de son choix sur l'ensemble d'apprentissage $\{(\boldsymbol{\phi}(\boldsymbol{x}_1), y_1), \ldots, (\boldsymbol{\phi}(\boldsymbol{x}_n), y_n)\}$. La construction d'un espace de représentation propice à la recherche d'une solution linéaire motive cette démarche, que justifient les théorèmes 2.2.1 et 2.2.2 relatifs à la consistance des règles de décision de forme linéaire généralisée. Sans précautions préalables, une telle pratique s'accompagne toutefois de problèmes techniques insurmontables dès lors que $\mathcal{T}$ est de dimension élevée, voire infinie. Dans ce contexte, les noyaux reproduisants offrent une élégante parade en autorisant la synthèse d'une structure linéaire dans un espace transformé $\mathcal{T}$, sans jamais y effectuer explicitement les calculs, ni nécessairement connaître la transformation $\boldsymbol{\phi}$ associée [Bos92, Vap95]. La *méthode du critère optimal à noyau reproduisant* exposée à présent [Abd02(c), Abd02(d)] exploite ces propriétés, et généralise les travaux sur le critère de Fisher effectués dans ce contexte [Mik99, Mik01(a), Mik01(b)].

3.3.1 Principe

Soit $J(\eta_0, \eta_1, \sigma_0^2, \sigma_1^2)$ un critère du second ordre. D'après la relation (2.40), la statistique de détection $\lambda(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x})$ opérant dans $\mathcal{T}$ est optimum au sens de J si elle vérifie

$$\Big[\rho\boldsymbol{\Sigma}_0^{\phi} + (1-\rho)\boldsymbol{\Sigma}_1^{\phi}\Big]\,\boldsymbol{w} = (\boldsymbol{m}_1^{\phi} - \boldsymbol{m}_0^{\phi}), \tag{3.9}$$

où $\boldsymbol{m}_i^{\phi}$ et $\boldsymbol{\Sigma}_i^{\phi}$ représentent l'espérance mathématique et la covariance conditionnelles de $\boldsymbol{\phi}(\boldsymbol{X})$. Celles-ci peuvent être estimées sur la base d'apprentissage $\mathcal{A}_n$ par

$$\widehat{\boldsymbol{m}}_i^{\phi} = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \omega_i} \boldsymbol{\phi}(\boldsymbol{x}) \tag{3.10}$$

$$\widehat{\boldsymbol{\Sigma}}_i^{\phi} = \frac{1}{n_i} \sum_{\boldsymbol{x} \in \omega_i} \boldsymbol{\phi}(\boldsymbol{x})\, \boldsymbol{\phi}^t(\boldsymbol{x}) - (\widehat{\boldsymbol{m}}_i^{\phi})\, (\widehat{\boldsymbol{m}}_i^{\phi})^t, \tag{3.11}$$

où n_i désigne le nombre d'échantillons disponibles de la classe ω_i. L'équation (3.9) ne peut être résolue aisément lorsque $\mathcal{T}$ est un espace de dimension importante. On peut toutefois contourner cette difficulté en considérant une fonction $\boldsymbol{\phi}(\cdot)$ associée à un noyau de Mercer κ par la relation

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \boldsymbol{\phi}(\boldsymbol{x}_1) \cdot \boldsymbol{\phi}(\boldsymbol{x}_2). \tag{3.12}$$

La statistique de détection considérée étant définie par le produit scalaire $\boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{X})$, le domaine de recherche de $\boldsymbol{w}$ peut être limité à l'espace linéaire induit par $\boldsymbol{\phi}(\boldsymbol{X})$, toute composante $\boldsymbol{w}^{\perp}$ extraite de l'espace complémentaire étant sans effet sur le résultat. Plus précisément encore, l'information disponible sur l'observation $\boldsymbol{X}$ se bornant aux données $\boldsymbol{x}_k$ de $\mathcal{A}_n$, le domaine de recherche de $\boldsymbol{w}$ peut être restreint à l'espace linéaire induit par les éléments $\boldsymbol{\phi}(\boldsymbol{x}_k)$, soit

$$\boldsymbol{w} = \sum_{k=1}^{n} a[k]\, \boldsymbol{\phi}(\boldsymbol{x}_k) = \boldsymbol{\Phi}\, \boldsymbol{a}. \tag{3.13}$$

Dans cette équation, $\boldsymbol{\Phi}$ représente la matrice $(\boldsymbol{\phi}(\boldsymbol{x}_1) \cdots \boldsymbol{\phi}(\boldsymbol{x}_n))$ et $\boldsymbol{a}$ désigne le vecteur dual de $\boldsymbol{w}$. En multipliant par $\boldsymbol{\Phi}^t$ l'expression (3.9) où l'on a préalablement remplacé $\boldsymbol{m}_i$ et $\boldsymbol{\Sigma}_i$ par les moments estimés (3.10) et (3.11), puis en utilisant la relation (3.13), on obtient

$$\left[\rho\, \boldsymbol{\Phi}^t\, \widehat{\boldsymbol{\Sigma}}_0^{\phi}\, \boldsymbol{\Phi} + (1-\rho)\, \boldsymbol{\Phi}^t\, \widehat{\boldsymbol{\Sigma}}_1^{\phi}\, \boldsymbol{\Phi}\right] \boldsymbol{a} = \boldsymbol{\Phi}^t\, [\widehat{\boldsymbol{m}}_1^{\phi} - \widehat{\boldsymbol{m}}_0^{\phi}]. \tag{3.14}$$

En appliquant la relation de Mercer (3.12), cette équation peut être reformulée ainsi

$$\boldsymbol{\Pi}_{\rho}\, \boldsymbol{a} = \boldsymbol{m}, \tag{3.15}$$

où $\boldsymbol{a}$ est le vecteur dual de $\boldsymbol{w}$ à déterminer. La matrice $\boldsymbol{\Pi}_{\rho}$ est de taille $(n \times n)$ indépendamment du choix de la transformation $\boldsymbol{\phi}(\cdot)$. Elle peut s'écrire sous la forme

$$\boldsymbol{\Pi}_{\rho} = \left[\frac{\rho}{n_0}\, \mathbf{K}_0\, (\mathbf{I} - \mathbf{1}_{n_0})\, \mathbf{K}_0^t + \frac{1-\rho}{n_1}\, \mathbf{K}_1\, (\mathbf{I} - \mathbf{1}_{n_1})\, \mathbf{K}_1^t\right], \tag{3.16}$$

où $\mathbf{K}_i$ est une matrice de taille $(n \times n_i)$ telle que

$$\mathbf{K}_i(p, q) = \kappa(\boldsymbol{x}_p, \boldsymbol{x}_q), \tag{3.17}$$

pour tout $\boldsymbol{x}_p \in (\omega_0 \cup \omega_1)$ et $\boldsymbol{x}_q \in \omega_i$. $\mathbf{I}$ est la matrice identité et $\mathbf{1}_{n_i}$ désigne la matrice dont tous les éléments valent $\frac{1}{n_i}$. Les composantes du vecteur $\boldsymbol{m}$ figurant dans l'expression (3.15) peuvent s'écrire sous la forme

$$m[k] = \frac{1}{n_1} \sum_{\boldsymbol{x} \in \omega_1} \kappa(\boldsymbol{x}, \boldsymbol{x}_k) - \frac{1}{n_0} \sum_{\boldsymbol{x} \in \omega_0} \kappa(\boldsymbol{x}, \boldsymbol{x}_k). \tag{3.18}$$

Des problèmes numériques peuvent survenir lors de la résolution du système linéaire (3.15), dus à un mauvais conditionnement de la matrice $\boldsymbol{\Pi}_{\rho}$. Celui-ci peut être amélioré en ajoutant un terme de régularisation à la matrice incriminée, généralement un multiple $\eta\, \mathbf{I}$ de la matrice identité avec η positif [Fri89, Sho99]. Grâce à la relation (3.13), la règle de décision s'écrit finalement

$$d_{\rho}(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \sum_{k=1}^{n} a_{\rho}[k]\, \kappa(\boldsymbol{x}_k, \boldsymbol{x}) - \lambda_{\rho} > 0 \\ 0 & \text{sinon,} \end{cases} \tag{3.19}$$

où $\boldsymbol{a}_\rho$ est solution de l'équation (3.15). Comme l'illustre l'algorithme présenté en figure 3.1, les paramètres ρ et λ_ρ sont conjointement déterminés de sorte à minimiser, par exemple, la probabilité d'erreur $P_e(d_\rho)$. On aboutit de cette façon à une règle de décision non-linéaire optimum au sens du meilleur critère de contraste, au terme d'une phase d'apprentissage dont la durée est indépendante de la complexité de la statistique. La méthodologie proposée présente également l'intérêt d'une grande flexibilité. Elle offre en effet un large choix de non-linéarités par le biais du noyau reproduisant κ, dont un aperçu va être à présent donné.

3.3.2 Éventail de noyaux reproduisants

L'expression (3.15) à partir de laquelle est élaboré le détecteur montre qu'il n'est pas nécessaire de connaître explicitement l'application $\boldsymbol{\phi}$. Celle-ci est implicitement définie par le choix d'un noyau reproduisant κ, à partir duquel il est possible de couvrir une large classe de non-linéarités. En voici quelques exemples classiques illustrés par la figure 3.2, une liste plus complète pouvant être consultée dans [Vap95, Can02].

Noyaux polynômiaux

Afin d'élaborer une règle de décision basée sur une statistique polynômiale de degré q, on utilise le noyau reproduisant suivant

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = (1 + \boldsymbol{x}_1 \cdot \boldsymbol{x}_2)^q. \tag{3.20}$$

On peut en effet montrer que les composantes de l'application $\boldsymbol{\phi}(\boldsymbol{x})$ associée sont les monômes de degrés inférieurs à q constitués des composantes de $\boldsymbol{x}$. Parce qu'ils sont fonction du produit scalaire des observations, de tels noyaux sont dits *projectifs*.

Noyaux exponentiels radiaux

Les noyaux de type *radial* dépendent de la distance $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$ entre les observations. Ils ont fait l'objet d'une attention particulière dans la littérature en raison du rôle central qu'ils jouent dans les méthodes d'estimation et de classification à base de noyaux ou de potentiels [Dev96]. On compte parmi eux le noyau gaussien, défini par

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2/\beta_0\right), \tag{3.21}$$

---

1. Sélectionner un noyau de Mercer κ et calculer $\boldsymbol{m}$
2. Initialiser ρ à 0
3. Tant que $\rho \leq 1$, répéter
   - calculer $\boldsymbol{\Pi}_\rho$ et résoudre le système $\boldsymbol{\Pi}_\rho\, \boldsymbol{a} = \boldsymbol{m}$ pour obtenir $\boldsymbol{a}_\rho$
   - déterminer le seuil λ_ρ de sorte à minimiser par exemple $P_e(d_\rho)$
   - mise à jour de ρ : $\rho \leftarrow \rho + \Delta\rho$, avec $\Delta\rho$ préalablement choisi
4. Sélectionner le meilleur détecteur d_ρ obtenu, caractérisé par $(\boldsymbol{a}_\rho, \lambda_\rho)$

FIG. 3.1 : Algorithme de la méthode du critère optimal à noyau reproduisant.

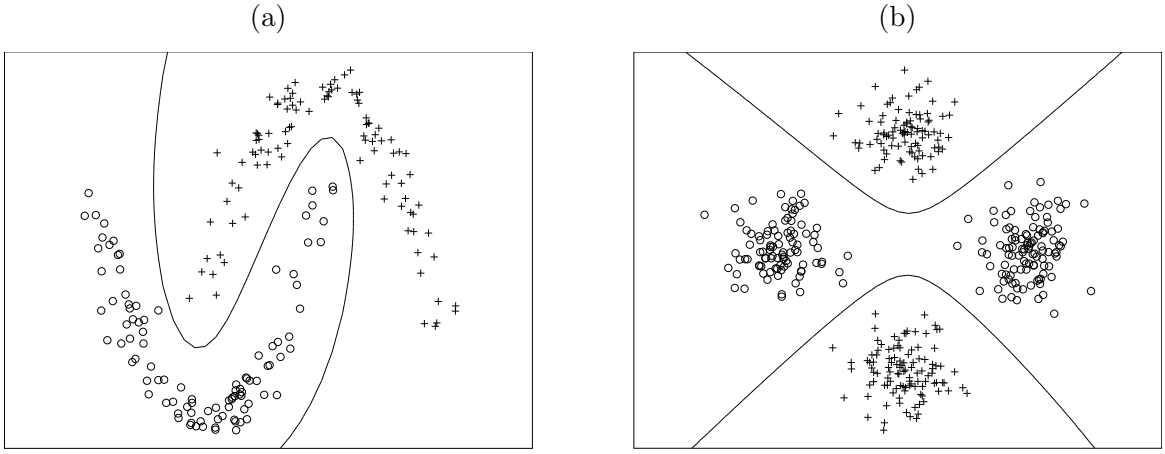

FIG. 3.2 : Exemples de détecteurs à noyau reproduisant : (a) noyau polynômial de degré 3, (b) noyau gaussien de largeur de bande 1. Ces deux problèmes sont issus de [Fuk90], où ils servent à illustrer les limites des critères du second ordre... lorsqu'on adopte une approche linéaire.

où β_0 est appelé *largeur de bande*. Ce noyau est caractérisé par un continuum de valeurs propres, ce qui signifie que les composantes de $\boldsymbol{\phi}$ ne sont pas en nombre fini comme dans l'exemple (3.20). Enfin, le noyau exponentiel (3.22) offre souvent des solutions intéressantes en fournissant une surface de décision linéaire par morceaux dans l'espace des observations.

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp\left(-\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|/\beta_0\right). \tag{3.22}$$

Noyaux sigmoïdaux

On peut élaborer un réseau de neurones à une couche cachée en choisissant le noyau sigmoïdal

$$\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = \tanh(\alpha_0 \left(\boldsymbol{x}_1 \cdot \boldsymbol{x}_2\right) + \beta_0). \tag{3.23}$$

La qualité de noyau reproduisant de κ dépend des paramètres α_0 et β_0 sélectionnés, contrairement aux noyaux polynômiaux et radiaux présentés ci-dessus. S'ils ne sont pas convenablement choisis, il en résulte la perte du cadre rigoureux offert par les espaces de Hilbert à noyau reproduisant. Toutefois, les errements de la pratique font que cette contrainte se trouve parfois relaxée.

3.3.3 Expérimentations

Afin d'illustrer la méthode du critère optimal à noyau reproduisant, celle-ci a été expérimentée sur un ensemble de données synthétiques réparties dans le plan selon des hyperboloïdes. La figure 3.3 présente la frontière de décision associée à un noyau polynômial κ de degré 2. Elle a été obtenue au terme de la recherche du paramètre ρ associé à un détecteur de probabilité d'erreur minimale, comme cela est indiqué par la figure 3.4. Il en résulte une structure optimum au sens du meilleur critère du second ordre, ce que confirment les résultats relatifs à l'emploi du critère de Fisher également reportés sur ces figures. D'autres expérimentations mettant en œuvre le noyau gaussien sur le difficile *problème des deux spirales* sont proposées à la fin de ce chapitre [Lan88]. La souplesse extrême de la méthode et les arguments théoriques qui la soutiennent ne doivent cependant pas faire oublier à l'expérimentateur qu'il n'est que plus exposé au fléau de la dimensionnalité lorsque, frénétiquement, il teste des surfaces de décision de complexité croissante. Les techniques présentées ci-après visent précisément à offrir des garanties de performance en généralisation par un contrôle du nombre de degrés de liberté de la structure considérée.

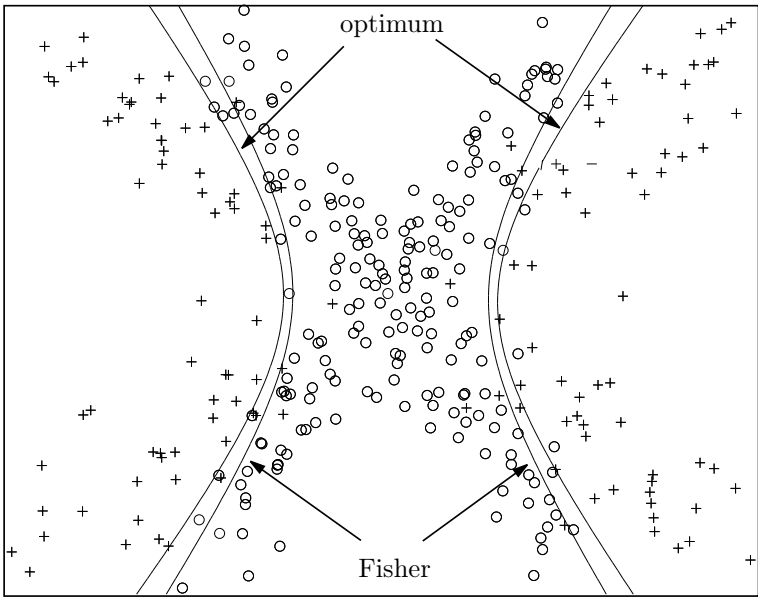

FIG. 3.3 : Frontière de décision obtenue sur des données synthétiques pour un noyau polynômial κ de degré 2. Les observations relatives aux 2 classes en compétition sont indiquées par des « + » et des « ∘ ».

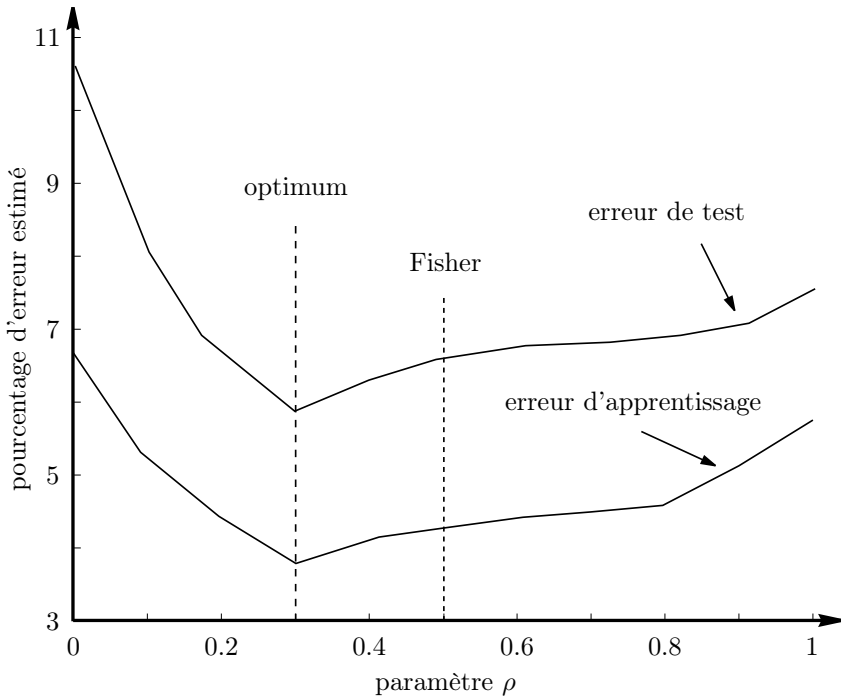

FIG. 3.4 : Recherche du paramètre ρ associé à un détecteur de probabilité d'erreur minimale, pour le problème des distributions hyperboloïdes illustré par la figure 3.3. Les hypothèses en compétition sont supposées équiprobables.

3.4 Contrôle des performances en généralisation

Afin de limiter les effets néfastes de la malédiction de la dimensionnalité, il est nécessaire d'adapter le nombre de paramètres caractéristiques de la statistique de détection à la taille de la base d'apprentissage. L'application du principe de minimisation du risque structurel présenté en Section 1.2.3 suppose que l'on est en mesure de construire une séquence de sous-ensembles imbriqués $\mathcal{D}_i$ au sein de la classe de détecteurs $\mathcal{D}$ considérée, thème qui n'a pas encore été abordé. Afin de combler cette lacune, on va à présent exposer deux stratégies permettant de faire varier la dimension de Vapnik-Chervonenkis des détecteurs linéaires généralisés définis par

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \lambda(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{x}) - \lambda_0 > 0 \\ 0 & \text{sinon,} \end{cases} \tag{3.24}$$

où $\boldsymbol{w}$ et $\boldsymbol{\phi}(\boldsymbol{x})$ sont des éléments de $\mathcal{T}$. Conformément aux résultats énoncés dans l'exemple 1.2.1, on note dans le cas présent que $h_{\mathcal{D}} = \dim(\mathcal{T}) + 1$, où $h_{\mathcal{D}}$ désigne la VC-dimension de $\mathcal{D}$. Les deux techniques proposées [Ric98(a)], qui s'inspirent de procédés de simplification des réseaux de neurones artificiels, consistent en une sélection des composantes de $\boldsymbol{\phi}(\boldsymbol{x})$ influençant significativement le processus d'apprentissage. Chacune d'elles peut être conjointement mise en œuvre avec la méthode du critère optimal (3.9), pour laquelle elle a été originellement prévue [Ric98(a), Ric99(a)], où avec la formulation duale (3.15) de la variante à noyau reproduisant [Abd02(c)]. Ce dernier cas étant retenu dans la présentation qui suit, on rappelle que la règle considérée s'écrit

$$d_\rho(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \sum_{k=1}^{n} a_\rho[k]\, \kappa(\boldsymbol{x}_k, \boldsymbol{x}) - \lambda_\rho > 0 \\ 0 & \text{sinon,} \end{cases} \tag{3.25}$$

où $\boldsymbol{a}_\rho$ désigne une solution du système linéaire $\boldsymbol{\Pi}_\rho\, \boldsymbol{a} = \boldsymbol{m}$.

3.4.1 Méthode variationnelle

Une première approche pour réduire la VC-dimension du détecteur $d_\rho(\boldsymbol{x})$ consiste à annuler certaines composantes du vecteur $\boldsymbol{a}_\rho$, à l'image des techniques d'élagage de réseaux de neurones telles qu'*optimal brain damage* [Lec90], dont s'inspire la méthode proposée. Si on exclut l'idée naïve d'éliminer les plus petites composantes en valeur absolue, on peut envisager d'annuler celles qui entraînent les plus faibles variations de l'erreur quadratique E_ρ suivante

$$E_\rho = \|\boldsymbol{\Pi}_\rho\, \boldsymbol{a} - \boldsymbol{m}\|^2. \tag{3.26}$$

Afin de faciliter les calculs à venir, on choisit d'exprimer E_ρ dans une base de vecteurs propres normalisés de $\boldsymbol{\Pi}_\rho$. On obtient ainsi

$$E_\rho = \sum_k \left[\gamma_\rho[k]\, \tilde{a}[k] - \tilde{m}[k]\right]^2, \tag{3.27}$$

où $\gamma_\rho[k]$ représente la $k^{\text{ème}}$ valeur propre de $\boldsymbol{\Pi}_\rho$. Les vecteurs $\tilde{\boldsymbol{a}}$ et $\tilde{\boldsymbol{m}}$ désignent respectivement $\mathbf{P}_\rho^t\, \boldsymbol{a}$ et $\mathbf{P}_\rho^t\, \boldsymbol{m}$, avec $\mathbf{P}_\rho$ la matrice de passage associée. Une perturbation $\delta\tilde{\boldsymbol{a}}$ modifie la fonction coût E_ρ de la quantité δE_ρ suivante

$$\delta E_\rho = \sum_k \frac{\partial E_\rho}{\partial \tilde{a}[k]}\, \delta\tilde{a}[k] + \frac{1}{2} \sum_k \frac{\partial^2 E_\rho}{\partial \tilde{a}[k]^2}\, \delta\tilde{a}[k]^2 + \frac{1}{2} \sum_{k \neq k'} \frac{\partial^2 E_\rho}{\partial \tilde{a}[k]\, \partial \tilde{a}[k']}\, \delta\tilde{a}[k]\, \delta\tilde{a}[k'] + O\left(\|\delta\tilde{\boldsymbol{a}}\|^2\right). \tag{3.28}$$

On vérifie que le troisième terme du membre de droite de l'équation (3.28) est nul. Il en va de même pour le premier terme lorsque $\tilde{\boldsymbol{a}} = \tilde{\boldsymbol{a}}_\rho \triangleq \mathbf{P}_\rho^t \, \boldsymbol{a}_\rho$, où $\boldsymbol{a}_\rho$ est solution du système $\boldsymbol{\Pi}_\rho \, \boldsymbol{a} = \boldsymbol{m}$. Il en résulte que la variation de l'erreur quadratique en réponse à une perturbation $\delta\tilde{\boldsymbol{a}}_\rho$ s'écrit

$$\delta E_\rho = \sum_k \left(\gamma_\rho[k] \, \delta\tilde{a}_\rho[k]\right)^2 . \tag{3.29}$$

Par conséquent, l'accroissement δE_ρ correspondant à l'annulation de la $k^{\text{ème}}$ composante de $\tilde{\boldsymbol{a}}_\rho$, obtenu en posant $\delta\tilde{a}_\rho[k] = \tilde{a}_\rho[k]$, est égal à

$$\delta E_\rho[k] = \left(\gamma_\rho[k] \, \tilde{a}_\rho[k]\right)^2 . \tag{3.30}$$

La méthode variationnelle permet ainsi de fixer la VC-dimension du détecteur $d_\rho(\boldsymbol{x})$ à une valeur h donnée, en sélectionnant les h composantes de $\tilde{\boldsymbol{a}}_\rho$ les plus significatives au sens de l'erreur quadratique E_ρ grâce au critère (3.30). Ce dernier présente l'avantage d'être indépendant du paramètre ρ comme le montre le développement qui suit, ce qui signifie qu'un même sous-ensemble de composantes $\tilde{a}_\rho[k]$ peut être considéré durant toute la progression de la méthode du critère optimal. Il convient de remarquer que les composantes associées aux valeurs propres nulles peuvent être systématiquement écartées, l'accroissement $\delta E_\rho[k]$ correspondant étant nul.

Soient $\mathbf{A}$ et $\mathbf{B}$ deux matrices symétriques positives. Sous certaines hypothèses peu restrictives pour un problème de détection [Ric98(a)], on montre qu'il existe une matrice non-singulière $\mathbf{P}$ telle que $\mathbf{P}^t \, \mathbf{A} \, \mathbf{P}$ et $\mathbf{P}^t \, \mathbf{B} \, \mathbf{P}$ sont conjointement diagonales [Gol93]. En appliquant ce résultat aux matrices $\boldsymbol{\Phi}^t \, \boldsymbol{\Sigma}_0^\phi \, \boldsymbol{\Phi}$ et $\boldsymbol{\Phi}^t \, \boldsymbol{\Sigma}_1^\phi \, \boldsymbol{\Phi}$ figurant dans l'expression (3.14), on en déduit que la matrice de passage $\mathbf{P}_\rho$ constituée de vecteurs propres de $\boldsymbol{\Pi}_\rho$ ne dépend pas de ρ. Il en résulte donc que le vecteur $\tilde{\boldsymbol{m}} = \mathbf{P}_\rho^t \, \boldsymbol{m}$ est également indépendant de ce paramètre. En constatant enfin que $\gamma_\rho[k] \, \tilde{a}_\rho[k] = \tilde{m}[k]$ lorsque $\gamma_\rho[k]$ est non-nul, on aboutit à

$$\delta E_\rho[k] = \tilde{m}[k]^2, \tag{3.31}$$

ce qui signifie que la variation de E_ρ en réponse à l'annulation de la composante $\tilde{a}_\rho[k]$ ne dépend pas de ρ. En conclusion, les différents sous-espaces sur lesquels le vecteur $\tilde{\boldsymbol{a}}_\rho$ est projeté, de sorte à contrôler la complexité du détecteur $d_\rho(\boldsymbol{x})$ associé, sont indépendants du critère du second ordre considéré. Cette propriété constitue un gage supplémentaire de robustesse pour la solution obtenue.

3.4.2 Méthode de pénalisation

Le second principe proposé repose sur la minimisation de l'erreur quadratique (3.26) sous la contrainte $\|\boldsymbol{a}\|^2 \leq c$, avec $c \in \mathbb{R}_+^*$. S'il n'y a pas d'annulation de composantes de la solution à proprement parler, certaines d'entre elles sont cependant condamnées à tendre vers zéro, suivant le choix du paramètre c. Le principe de cette méthode de pénalisation [Ric98(a), Ric99(a)] s'inspire comme la précédente d'un procédé issu du domaine des réseaux de neurones, appelé *weight decay* [Vap82]. Soit $\mathcal{D}_i$ la classe de détecteurs à noyau reproduisant définie par

$$\mathcal{D}_i = \left\{ d(\boldsymbol{x}) = \Gamma\left(\sum_{k=1}^{n} a[k] \, \kappa(\boldsymbol{x}_k, \boldsymbol{x}) - \lambda_0\right) \, : \, \|\boldsymbol{a}\|^2 \leq c_i, \; c_i \in \mathbb{R}_+^* \right\}, \tag{3.32}$$

où Γ désigne la fonction d'Heaviside. En établissant préalablement une suite de réels c_i telle que $0 < c_1 < c_2 < ...$, on élabore une séquence de sous-ensembles imbriqués $\mathcal{D}_i$ de la forme

$$\mathcal{D}_1 \subset \mathcal{D}_2 \subset \ldots \subset \mathcal{D}. \tag{3.33}$$

Bien que cette dernière soit conforme dans sa structure à la séquence préconisée par le principe de minimisation du risque structurel, elle n'est pas élaborée selon les mêmes règles. Elle repose en effet sur un classement des détecteurs selon le module du vecteur de projection $\boldsymbol{a}$, et non sur la VC-dimension de la classe qui les regroupe. Pourtant, comme cela va être démontré, la finalité visant à faire varier la capacité d'apprentissage des structures de détection pour déterminer celle qui semble la mieux adaptée demeure inchangée.

Afin de simplifier l'exposé qui suit, on se place à présent dans une base de vecteurs propres normalisés de $\boldsymbol{\Pi}_\rho$. Les notations adoptées sont identiques à celles employées pour la description de la méthode variationnelle. On note $\tilde{\boldsymbol{a}}_{\rho,i}$ le vecteur minimisant l'erreur quadratique définie par (3.27) sous la contrainte d'inégalité suivante

$$\|\tilde{\boldsymbol{a}}_{\rho,i}\|^2 \leq c_i, \tag{3.34}$$

de sorte que le détecteur associé appartienne à la classe $\mathcal{D}_i$. D'après le théorème de Kuhn-Tücker, il existe un multiplicateur de Lagrange ξ_i associé au paramètre c_i tel que :

$$\tilde{\boldsymbol{a}}_{\rho,i} = \arg\min\{E(\tilde{\boldsymbol{a}}) + \xi_i\|\tilde{\boldsymbol{a}}\|^2\}. \tag{3.35}$$

En annulant les dérivées partielles de la fonction $E(\tilde{\boldsymbol{a}})+\xi_i\|\tilde{\boldsymbol{a}}\|^2$ par rapport à $\tilde{a}[k]$, on peut établir la relation suivante :

$$\tilde{a}_{\rho,i}[k] = \frac{\gamma_\rho[k]}{(\gamma_\rho[k])^2 + \xi_i}\,\tilde{m}[k] = \frac{(\gamma_\rho[k])^2}{(\gamma_\rho[k])^2 + \xi_i}\,\tilde{a}_\rho[k], \tag{3.36}$$

où $\tilde{\boldsymbol{a}}_\rho$ est solution de l'équation $\boldsymbol{\Pi}_\rho\,\boldsymbol{a} = \boldsymbol{m}$ exprimée dans la base de vecteurs propres considérée. Lorsque $\gamma_\rho[k] \gg \sqrt{\xi_i}$, ce résultat montre ainsi que les composantes de $\tilde{\boldsymbol{a}}_\rho$ ne sont pas affectées par la méthode de pénalisation. Elles tendent en revanche vers 0 quand $\gamma_\rho[k] \ll \sqrt{\xi_i}$. Par conséquent, l'introduction d'une contrainte dans le processus d'apprentissage entraîne une modification de la capacité des détecteurs, que l'on peut alors estimer par l'expression suivante [Moo92] :

$$h_{\mathcal{D}_i} = \left\lfloor \sum_k \frac{(\gamma_\rho[k])^2}{(\gamma_\rho[k])^2 + \xi_i} \right\rfloor, \tag{3.37}$$

où $\lfloor\cdot\rfloor$ désigne l'opérateur *partie entière*. Ce résultat n'est cependant valide que si la matrice $\boldsymbol{\Pi}_\rho$ admet un spectre de valeurs propres étroit.

En agissant sur la dynamique de certaines composantes du vecteur de projection $\boldsymbol{a}$ durant l'apprentissage, les méthodes de sélection qui viennent d'être présentées procèdent de façon analogue pour contrôler la capacité d'apprentissage des détecteurs linéaires généralisés. Bien que très efficace en pratique, la méthode de pénalisation est néanmoins plus délicate à mettre en œuvre que la technique variationnelle [Ric98(a), Ric99(a)]. En effet, elle nécessite l'optimisation conjointe des paramètres ρ et ξ_i lorsqu'on l'associe à la méthode du critère optimal. Pour cette raison, seule l'approche variationnelle est considérée dans la suite de ce document.

3.4.3 Expérimentations

Afin d'illustrer l'efficacité de la méthode variationnelle, on considère le classique et difficile *problème des deux spirales* [Lan88]. Il concerne l'élaboration d'un discriminant entre 2 classes ω_0 et ω_1, représentées chacune par 97 échantillons distribués selon 2 spirales s'enroulant l'une

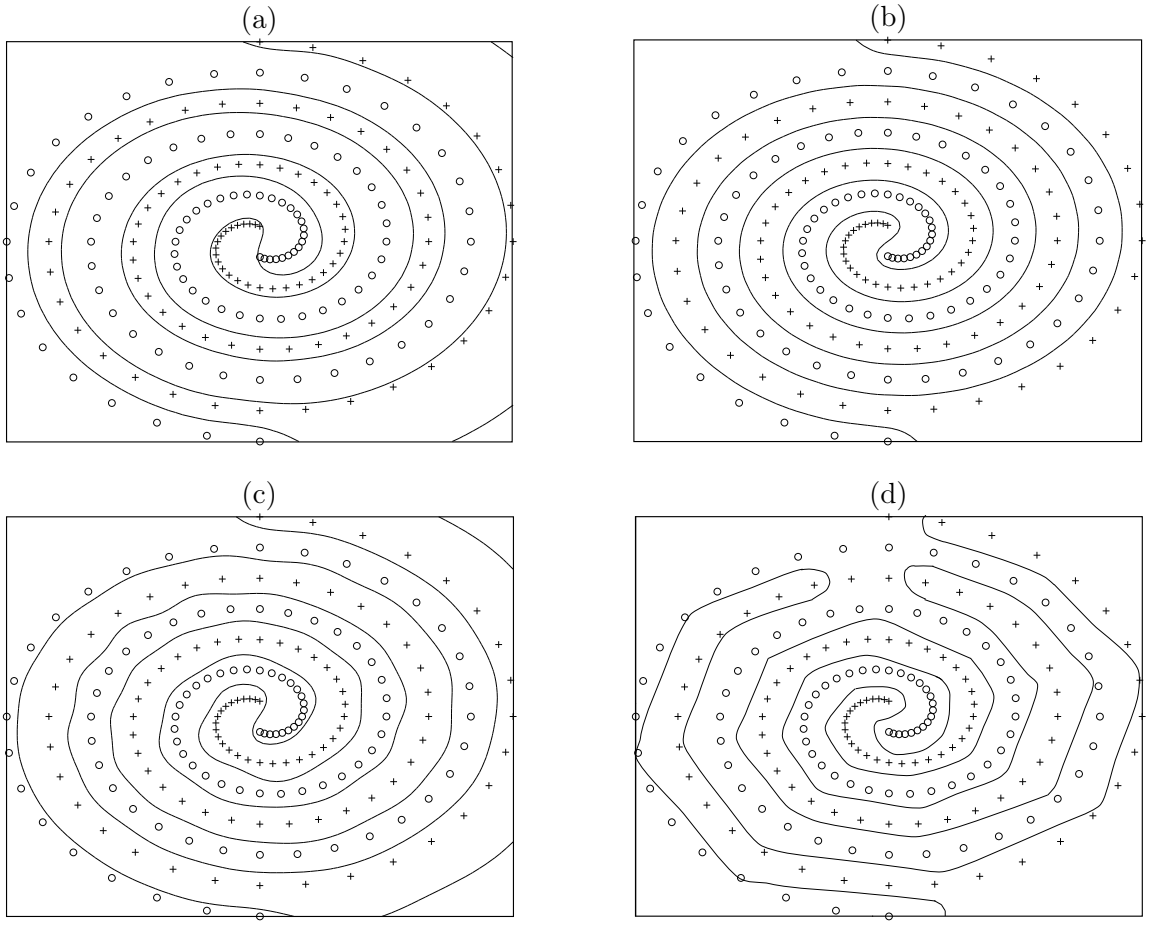

FIG. 3.5 : Comparaison de résultats obtenus avec un noyau gaussien de largeur de bande 1. (a) Sans contrôle de complexité ; (b) et (c) avec contrôle par la méthode variationnelle en retenant respectivement 40, puis 20 composantes significatives au sens de l'erreur quadratique ; (d) avec contrôle de complexité par la méthode OLS, cette solution étant proposée dans [Bil02].

autour de l'autre. La figure 3.5 propose différents résultats obtenus avec un noyau gaussien de largeur de bande 1. La faible marge entre certains échantillons et la frontière de décision présentée en (a) constitue l'une des manifestations caractéristiques du phénomène de sur-apprentissage. Cette configuration, obtenue uniquement avec la méthode du critère optimal, laisse entrevoir une amélioration possible du résultat en ayant recours à une technique de contrôle de complexité. La méthode variationnelle a été appliquée en conséquence. La règle figurant en (b) reposent sur les 40 composantes $\tilde{a}[k]$ les plus significatives au sens de l'erreur quadratique, sur un total de 194. Force est de constater que la régularité de la frontière augure de très bonnes performances en généralisation pour cette solution assez économique. Si la règle présentée en (c) l'est encore d'avantage avec un nombre de composantes significatives porté à 20, on relève néanmoins une dégradation partielle du résultat. A titre de comparaison, (d) présente le discriminant figurant dans [Bil02], obtenu par maximisation du critère de Fisher. Pas moins de 50 composantes significatives ont été sélectionnées par l'algorithme OLS [Che89], chargé du contrôle de complexité, pour obtenir cette solution peu régulière commettant de surcroît des erreurs d'attribution. Aussi, cette comparaison montre-t-elle très nettement la supériorité de l'approche proposée.

3.5 Comparaison avec les Support Vector Machines

La présentation de la méthode du critère optimal à noyau reproduisant serait incomplète si elle ne s'achevait pas sur une comparaison avec la technique phare du moment. Ce sont les Support Vector Machines, couramment appelées SVM, qui constituent des solutions à marge maximale entre l'hyperplan séparateur et les échantillons de l'ensemble d'apprentissage [Bos92, Cor95]. La notion de marge étant particulièrement intuitive dans le cas d'un problème à classes linéairement séparables, cette hypothèse va être retenue dans la première partie de la présentation qui suit, avant d'être abonnée au profit de considérations plus générales. Le lecteur qui serait toutefois intéressé par de plus amples précisions est invité à consulter l'importante littérature sur le sujet, en particulier [Vap95, Sho99].

3.5.1 Algorithme de l'hyperplan optimum

Soit $\mathcal{A}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ un ensemble de données linéairement séparables, où $\boldsymbol{x}_k$ représente une observation et y_k la décision associée résultant du processus d'expertise. Provisoirement, on suppose que cette dernière est un élément de $\{-1, +1\}$, le formalisme classique des Support Vector Machines reposant sur cette convention. On recherche ainsi l'*hyperplan optimum* d'équation $\boldsymbol{w} \cdot \boldsymbol{x} - \lambda_0 = 0$ tel que la distance le séparant des l'échantillons les plus proches soit maximale [Vap82]. Afin de lever toute ambiguïté sur les valeurs de $\boldsymbol{w}$ et λ_0, définies à un facteur près, on impose la contrainte supplémentaire

$$\min_{\boldsymbol{x} \in \mathcal{A}_n} |\boldsymbol{w} \cdot \boldsymbol{x} - \lambda_0| = 1. \tag{3.38}$$

Celle-ci implique directement que

$$\boldsymbol{w} \cdot \boldsymbol{x}_k - \lambda_0 \geq +1, \text{ si } y_k = +1 \tag{3.39}$$

$$\boldsymbol{w} \cdot \boldsymbol{x}_k - \lambda_0 \leq -1, \text{ si } y_k = -1, \tag{3.40}$$

ce que l'on résume par la contrainte générale suivante

$$y_k(\boldsymbol{w} \cdot \boldsymbol{x}_k - \lambda_0) \geq +1. \tag{3.41}$$

On rappelle que la distance δ_k d'une observation $\boldsymbol{x}_k$ à l'hyperplan considéré est donnée par $\delta_k = |\boldsymbol{w} \cdot \boldsymbol{x}_k - \lambda_0| / \|\boldsymbol{w}\|$. En combinant cette expression avec la contrainte (3.41), on aboutit finalement à l'inégalité

$$\delta_k \geq \frac{1}{\|\boldsymbol{w}\|}. \tag{3.42}$$

En conséquence, l'hyperplan recherché minimise $\frac{1}{2}\|\boldsymbol{w}\|^2$ sous la contrainte (3.41). D'après le théorème de Kuhn-Tücker, la solution de ce problème est donnée par le point selle du lagrangien

$$L(\boldsymbol{w}, \lambda_0, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{k=1}^{n} \xi_k \left[y_k(\boldsymbol{w} \cdot \boldsymbol{x}_k - \lambda_0) - 1\right], \tag{3.43}$$

où $\boldsymbol{\xi} = (\xi_1 \cdots \xi_n)$ désigne n multiplicateurs de Lagrange positifs. En son point-selle, L est minimum pour $\boldsymbol{w} = \boldsymbol{w}^*$ et $\lambda_0 = \lambda_0^*$, et maximum pour $\boldsymbol{\xi} = \boldsymbol{\xi}^*$. Les conditions d'annulation de ses dérivées partielles mènent directement aux relations vérifiées par l'hyperplan optimal

$$\boldsymbol{w}^* = \sum_{k=1}^{n} \xi_k^* \, y_k \, \boldsymbol{x}_k \quad \text{et} \quad \sum_{k=1}^{n} \xi_k^* \, y_k = 0. \tag{3.44}$$

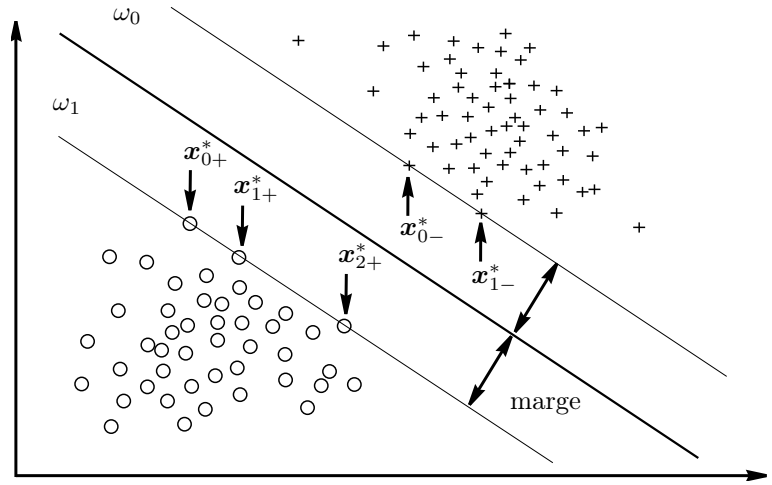

FIG. 3.6 : Principe des SVM dans le cas où les classes ω_0 et ω_1 sont linéairement séparables. Les *Support Vectors*, indiqués par des flèches, désignent les échantillons les plus proches de l'hyperplan séparateur.

Elles permettent également d'exprimer la forme duale du lagrangien

$$W(\boldsymbol{\xi}) = \sum_{k=1}^{n} \xi_k - \frac{1}{2} \sum_{k,k'=1}^{n} \xi_k \, \xi_{k'} \, y_k \, y_{k'} \, (\boldsymbol{x}_k \cdot \boldsymbol{x}_{k'}), \tag{3.45}$$

qu'il faut maximiser sous les contraintes $\sum_{k=1}^{n} \xi_k \, y_k = 0$ et $\xi_k \geq 0$ afin d'obtenir $\boldsymbol{\xi}^*$. Il convient de noter que $\xi_k \neq 0$ si la contrainte (3.41) est effective pour l'échantillon k. Il en résulte alors que $\delta_k = 1$, ce qui signifie qu'il n'existe pas d'observation plus proche de l'hyperplan optimum que $\boldsymbol{x}_k$, comme l'illustre la figure 3.6. De tels échantillons, appelés *support vectors* et notés $\boldsymbol{x}^*_{k-}$ ou $\boldsymbol{x}^*_{k+}$ suivant qu'ils appartiennent à l'une ou l'autre des hypothèses en compétition, permettent en particulier d'évaluer le seuil λ^*_0 par la relation

$$\lambda^*_0 = \frac{1}{2} \left[(\boldsymbol{w}^* \cdot \boldsymbol{x}^*_{i-}) + (\boldsymbol{w}^* \cdot \boldsymbol{x}^*_{j+}) \right], \tag{3.46}$$

étant donnés i et j. Plus généralement, les support vectors constituent l'un des éléments attractifs de l'approche. Ils concentrent en effet l'information discriminante véhiculée par l'ensemble d'apprentissage $\mathcal{A}_n$, l'algorithme pouvant être réitéré sur la simple base de ces observations particulières sans que l'équation de l'hyperplan optimal s'en trouve modifiée. De plus, leur nombre rapporté à la taille de $\mathcal{A}_n$ fournit un majorant du taux d'erreur de la structure de décision constituée [Vap95].

L'équation (3.45) indique que la méthode présentée se prête aisément à l'élaboration d'hyperplans optimums dans les espaces transformés à noyau reproduisant. Conformément aux propriétés de ces structures algébriques, énoncées en début de ce chapitre, il suffit en effet de maximiser la fonction suivante

$$W(\boldsymbol{\xi}) = \sum_{k=1}^{n} \xi_k - \frac{1}{2} \sum_{k,k'=1}^{n} \xi_k \, \xi_{k'} \, y_k \, y_{k'} \, \kappa(\boldsymbol{x}_k, \boldsymbol{x}_{k'}) \tag{3.47}$$

sous les contraintes $\sum_{k=1}^{n} \xi_k \, y_k = 0$ et $\xi_k \geq 0$. Dans cette expression, on rappelle que κ désigne un noyau reproduisant tel que $\kappa(\boldsymbol{x}_k, \boldsymbol{x}_{k'}) = \boldsymbol{\phi}(\boldsymbol{x}_k) \cdot \boldsymbol{\phi}(\boldsymbol{x}_{k'})$. A partir de cette relation et des

équations (3.44) et (3.46), on identifie finalement les paramètres de l'hyperplan optimum

$$\boldsymbol{w}^* \cdot \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{k=1}^{n} \xi_k^* \, y_k \, \kappa(\boldsymbol{x}_k, \boldsymbol{x}), \tag{3.48}$$

$$\lambda_0^* = \frac{1}{2} \sum_{k=1}^{n} \xi_k^* \, y_k \, [\kappa(\boldsymbol{x}_k, \boldsymbol{x}_{i-}^*) + \kappa(\boldsymbol{x}_k, \boldsymbol{x}_{j+}^*)]. \tag{3.49}$$

Bien évidemment, le cas de classes linéairement séparables est rarement rencontré en pratique. Aussi, il s'avère nécessaire d'aménager quelque peu cette approche afin de traiter des problèmes plus généraux, en particulier en introduisant dans la fonction coût $\frac{1}{2}\|\boldsymbol{w}\|^2$ un terme pénalisant les erreurs de classement.

3.5.2 Extension au cas non-séparable

Lorsque les classes ne sont pas linéairement séparables, la solution unanimement adoptée consiste à reformuler les conditions (3.39) et (3.40) sous la forme suivante, avec $\zeta_k \geq 0$ [Cor95] :

$$\boldsymbol{w} \cdot \boldsymbol{x}_k - \lambda_0 \geq +1 - \zeta_k, \text{ si } y_k = +1 \tag{3.50}$$

$$\boldsymbol{w} \cdot \boldsymbol{x}_k - \lambda_0 \leq -1 + \zeta_k, \text{ si } y_k = -1. \tag{3.51}$$

La structure de décision délivre donc une décision erronée pour une observation $\boldsymbol{x}_k$ donnée lorsque le paramètre ζ_k qui lui est associé est supérieur à 1. La fonction $\sum_{k=1}^{n} \zeta_k$ fournit en conséquence une indication sur le nombre d'erreurs commises. Aussi, la fonction coût préconisée par [Cor95] s'exprime ainsi

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + c \sum_{k=1}^{n} \zeta_k, \tag{3.52}$$

qu'il s'agit de minimiser sous les contraintes (3.50) et (3.51), c étant un réel positif préalablement fixé. Comme précédemment, la résolution de ce problème nécessite de recourir au théorème de Kuhn-Tücker afin de déterminer les coefficients ξ_k^* de l'hyperplan optimal $\boldsymbol{w}^* = \sum_{k=1}^{n} \xi_k^* \, y_k \, \boldsymbol{x}_k$. On montre aisément qu'ils sont obtenus en maximisant la forme duale du lagrangien (3.45) sous les contraintes $\sum_{k=1}^{n} \xi_k \, y_k = 0$ et $0 \leq \xi_k \leq c$. Comme précédemment, les support vectors correspondent aux observations les plus proches de l'hyperplan optimal et correctement classées par la structure de décision ainsi constituée.

3.5.3 Expérimentations

Afin de compléter la présentation de la méthode du critère optimal à noyau reproduisant, ce chapitre s'achève par une confrontation avec la technique des Support Vector Machines. Dix problèmes de détection classiquement utilisés pour ce type d'exercice ont ainsi été retenus [Bil02, Mik99]. Téléchargeables à l'adresse `http://www.first.gmd.de/~raetsch/`, ils proviennent des banques de données de l'UCI, de DELVE et de STATLOG. Pour l'ensemble des problèmes considérés, l'essai comparatif a consisté en 30 phases d'apprentissage et de test d'un détecteur à noyau gaussien de largeur de bande unité. A chaque itération du processus, 400 observations ont été aléatoirement prélevées dans la base d'apprentissage afin d'optimiser la direction de projection $\boldsymbol{w}$ et le seuil λ_0 de chacune des structures concurrentes. Puis, 8000 données indépendantes ont été consacrées à l'estimation de leurs performances. Le tableau 3.1 synthétise les résultats obtenus, le recours à la procédure variationnelle dédiée au contrôle de complexité étant

envisagé dans le cas de la méthode du critère optimal. A titre de comparaison, les performances du détecteur à noyau gaussien maximisant le critère de Fisher ont été également indiquées. Ces résultats illustrent la compétitivité de la méthode proposée par rapport à la technique des Support Vector Machines, et démontre à nouveau l'intérêt que présente la recherche du meilleur critère de contraste pour un problème donné. Enfin, cette expérience met en évidence le rôle bénéfique joué par la procédure variationnelle. En effet, le contrôle de complexité pratiqué par celle-ci s'avère particulièrement efficace au vu du nombre n_{cs} réduit de composantes significatives retenues, généralement très inférieur à celui n_{sv} de support vectors. Les performances en détection confirment ce résultat.

	Fisher (%)	critère opt. (%)	critère opt. + var. (%)	SVM (%)	n_{sv}	n_{cs}
banana	10.60	10.59	**10.37**	10.43	132	35
cancer	8.14	6.70	**6.60**	7.10	364	75
diabetes	17.79	17.39	**17.11**	17.68	308	150
german	21.36	20.96	**20.90**	21.06	400	200
heart	4.44	4.41	**4.24**	4.52	388	100
ringnorm	1.53	1.53	**1.50**	1.52	396	20
sonar	32.42	31.61	**31.04**	32.73	364	40
thyroid	0.39	0.25	**0.23**	0.33	156	75
titanic	28.88	28.55	**27.72**	28.88	400	15
waveform	11.14	11.14	11.14	**11.07**	400	400

TAB. 3.1 : Comparaison des taux d'erreur obtenus avec la technique des Support Vector Machines et la méthode du critère optimal à noyau reproduisant, avec ou sans contrôle de complexité dans ce dernier cas. Les deux dernières colonnes indiquent le nombre de support vectors et de composantes significatives retenues par la procédure variationnelle, notés n_{sv} et n_{cs} respectivement.

3.6 Conclusion

Une puissante technique pour la synthèse de détecteurs à structure imposée a été proposée au cours de ce chapitre. Celle-ci repose sur l'optimisation du meilleur critère du second ordre pour le problème traité, et exploite les noyaux de Mercer afin de conférer un caractère non-linéaire à la règle de décision élaborée. Pour une meilleure prise en compte de phénomènes tels que la malédiction de la dimensionnalité, la méthodologie proposée s'appuie également sur la théorie statistique de l'apprentissage. Ceci lui permet d'offrir des garanties sur les performances en généralisation des détecteurs obtenus, qui sont alors en mesure de rivaliser avec les Support Vector Machines. Cet éclairage mêlant théories statistiques de la décision et de l'apprentissage permet à présent d'envisager l'étude de domaines transformés particuliers pour la résolution de problèmes de décision. Aussi le chapitre 4 s'intéresse-t-il à un espace de représentation classiquement considéré en analyse des signaux non-stationnaires : le plan temps-fréquence.

Chapitre 4

Détection dans un espace transformé : le plan temps-fréquence

4.1 Introduction

Parce qu'elles constituent un mode de caractérisation spectrale locale nécessaire à l'analyse des signaux non-stationnaires, les représentations temps-fréquence jouent un rôle fondamental en traitement du signal. Parmi les nombreuses solutions proposées, la distribution de Wigner définie ci-dessous est souvent privilégiée en raison des multiples propriétés qu'elle vérifie, parmi lesquelles on recense celles relatives à ses distributions marginales ou encore à son support [Fla98].

$$W_{x_1x_2}(t,f) = \int_{\mathbb{R}} x_1(t+\tau/2)\, x_2^*(t-\tau/2)\, e^{-j2\pi f\tau}\, d\tau. \tag{4.1}$$

Cette distribution fournit également un espace de représentation privilégié pour la résolution de problèmes de détection, principalement en raison de sa covariance par rapport aux translations dans le plan temps-fréquence et de son unitarité [Fla88, Say95].

4.1.1 Distributions de Wigner discrètes

Afin d'implémenter la distribution (4.1) sur un calculateur, il est nécessaire d'en donner une définition pour laquelle les variables temporelle t et fréquentielle f prennent des valeurs discrètes. De nombreux principes de discrétisation ont été étudiés à ce jour afin de résoudre ce problème non trivial. Généralement, la distribution de Wigner discrète est considérée sous la forme suivante

$$W^{(C)}_{x_1x_2}[t,f] = \sum_{\tau=0}^{l-1} x_1[t+\tau]\, x_2^*[t-\tau]\, e^{-j\frac{2\pi}{l}f\tau}, \tag{4.2}$$

où t et f sont à valeurs dans l'ensemble $\{0,\ldots,l-1\}$, le paramètre l représentant le nombre d'échantillons des signaux analysés. Cette définition, dite *classique* dans le cadre de ce document, présente cependant certains inconvénients, l'un des principaux étant de ne fournir qu'une représentation fréquentielle des signaux sur la bande normalisée $[-\frac{1}{4},\frac{1}{4}]$. Par rapport à la distribution de Wigner à variables continues, la perte de certaines propriétés fondamentales telles que l'unitarité est également à déplorer. Récemment, d'importants efforts ont été faits afin de développer une distribution de Wigner discrète autorisant une analyse complète de la bande spectrale $[-\frac{1}{2},\frac{1}{2}]$. Une liste relativement exhaustive des solutions envisagées est consultable dans [Cos01].

Dans [Rch98] par exemple, Richman *et col.* ont eu recours à la théorie des groupes. La distribution de Wigner discrète résultant de leur étude, ici appelée *R-distribution*, vérifie des propriétés analogues à celles satisfaites par son homologue à variables continues. En utilisant la notation $(a)_l$ pour désigner *a modulo l*, celle-ci est définie par

$$W_{x_1x_2}^{(R)}[t,f] = \frac{1}{l}\sum_{\tau=0}^{l-1}\sum_{\nu=0}^{l-1}\sum_{t'=0}^{l-1}\rho_l[\nu,\tau]\,x_1[(t'+\tau)_l]\,x_2^*[t']\,e^{-j\frac{2\pi}{l}(t\nu+f\tau-t'\nu)}. \tag{4.3}$$

Il est à noter que deux expressions définissent la fonction $\rho_l[\nu,\tau]$ figurant ci-dessus, que le lecteur intéressé pourra consulter dans [Rch98], la sélection de l'une ou l'autre dépendant de la parité de l. Dans [One99(a), One99(b)], les auteurs ont privilégié une approche axiomatique. Leur démarche mène à la conclusion que, selon leur définition, la distribution de Wigner discrète n'existe que pour les signaux dont le nombre d'échantillons l est impair. Dans ce cas, ils aboutissent également à l'expression (4.3). Parmi les nombreux travaux sur le sujet, on peut encore et finalement citer ceux de Stanković sur la *S-distribution* [Sta94, Sta01]. Cette dernière repose sur la transformée de Fourier à court terme, comme le précise la définition suivante

$$W_{x_1x_2}^{(S)}[t,f] = \frac{1}{l}\sum_{\nu=-b_f}^{b_f} F_{x_1}[t,f+\nu]\,F_{x_2}^*[t,f-\nu], \tag{4.4}$$

avec $b_f = \min\{f,\, l-f-1\}$ de sorte que les conditions $(f+\nu \leq l-1)$ et $(f-\nu \geq 0)$ soient vérifiées, étant donné f. On rappelle que la transformée de Fourier à court terme peut s'exprimer ainsi

$$F_x[t,f] = \sum_{\tau} x[\tau]\,g^*[\tau-t]\,e^{-j\frac{2\pi}{l}f\tau}, \tag{4.5}$$

où g désigne une fenêtre d'analyse. Aussi la S-distribution fournit-elle une représentation spectrale de la bande $[-\frac{1}{2},\frac{1}{2}]$ tout comme la R-distribution, ce qui procure à ces deux solutions un avantage substantiel sur l'approche traditionnelle (4.2).

4.1.2 Détection par représentation temps-fréquence

Les méthodes temps-fréquence, en particulier la distribution de Wigner, ont souvent été associées à des structures décisionnelles en vertu du point de vue intéressant qu'elles offrent sur les signaux non-stationnaires. Ce type d'approche couvre des domaines d'application aussi variés que l'acoustique [Dav00], l'astrophysique [Cha98], le biomédical [Ric98(a)] ou encore les radars [Lem95]. Si l'on écarte les solutions reposant sur l'extraction d'attributs des représentations tels que la position d'un pic ou l'encombrement d'une composante, la dernière décennie a été consacrée au développement d'une théorie pour la détection optimum par représentation temps-fréquence. En particulier, Flandrin a caractérisé des scenarii pour lesquels des structures de détection opérant dans le plan temps-fréquence revêtent un caractère optimal [Fla88]. Plus récemment, Sayeed et Jones ont présenté des détecteurs temps-fréquence optimaux pour la détection de signaux aléatoires gaussiens en présence de bruit gaussien, avec pour paramètres de nuisance l'instant d'arrivée et la fréquence Doppler de l'événement à détecter [Say95]. Enfin, Matz et Hlawatsch ont proposé une simplification de ces structures dans [Mat98]. La mise en œuvre de tels tests n'a été abordée que très récemment dans la littérature [Ric01(b), Ric02(b)], celle-ci se limitant auparavant à des considérations sur la définition classique (4.2) de la distribution de Wigner discrète [Mar97]. Ce sujet n'est toutefois pas dépourvu d'intérêt, les procédures de discrétisation des distributions temps-fréquence étant multiples et, lorsqu'elles s'accompagnent de la perte de propriétés, potentiellement lourdes de conséquences sur les performances en détection.

4.1.3 Position du problème

Le présent chapitre est consacré à l'étude de détecteurs opérant dans l'espace transformé que constitue le plan temps-fréquence. Ne pouvant prendre en considération l'ensemble des solutions proposées pour la distribution de Wigner discrète, la discussion qui suit a été volontairement limitée aux trois définitions qui viennent d'être présentées, celles-ci offrant une vision représentative des situations qui peuvent être rencontrées. Aussi, on aborde dans un premier temps le sujet du point de vue de la détection à structure libre, par le biais du problème de détection académique considéré dans [Fla88]. On étudie alors l'existence de solutions temps-fréquence reposant sur les distributions de Wigner discrètes considérées. Puis, on se consacre plus longuement au thème de la détection à structure imposée par représentation temps-fréquence. A cette occasion, on discute du choix de la définition qui pourrait garantir les meilleures performances en détection, gardant à l'esprit les effets néfastes induits par le phénomène de malédiction de la dimensionnalité. Ce chapitre s'achève par la présentation d'une méthode pour le contrôle de la complexité des détecteurs opérant dans le plan temps-fréquence. Elle est le fruit du travail d'un doctorant, J. Gosme [Gos02], et d'une collaboration avec P. Gonçalvès (INRIA Rhône-Alpes, Grenoble). Le contenu de ce chapitre, original dans sa totalité, a fait l'objet de plusieurs publications et communications dans des revues et conférences de renom, en particulier [Ric01(b), Ric02(b)].

4.2 Détection à structure libre

L'objectif de cette section est de montrer que la définition attribuée à la distribution de Wigner discrète a un impact important sur les performances de la structure de détection à laquelle elle est associée. Le problème sélectionné pour cela est défini ainsi

$$\begin{cases} \omega_0 : \boldsymbol{x} = \boldsymbol{b} \\ \omega_1 : \boldsymbol{x} = \boldsymbol{b} + \boldsymbol{s}, \end{cases} \tag{4.6}$$

où $\boldsymbol{x}$ désigne une observation discrète de longueur l et $\boldsymbol{s}$ le signal à détecter. Ce dernier est supposé gaussien, d'espérance mathématique $\boldsymbol{m}$ et de covariance $\boldsymbol{\Sigma}$. Le bruit $\boldsymbol{b}$ dans lequel il est noyé est quant à lui blanc, gaussien, centré et de variance σ_0^2. Le problème (4.6) possède une solution dans le plan temps-fréquence, dont on trouvera une description complète dans [Fla88]. Cette dernière a été étudiée en supposant le temps et la fréquence continues. On se propose à présent de traiter la question dans le cas discret, en considérant une base orthonormée de vecteurs propres de $\boldsymbol{\Sigma}$. Aussi on note

$$\tilde{x}[k] = \sum_{t=0}^{l-1} x[t]\,\psi_k^*[t] \qquad \tilde{m}[k] = \sum_{t=0}^{l-1} m[t]\,\psi_k^*[t], \tag{4.7}$$

où $\boldsymbol{\psi}_k$ désigne le $k^{\text{ème}}$ élément de la base considérée, de valeur propre associée γ_k. Dans ces conditions, le problème (4.6) admet une solution classique s'exprimant ainsi

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \tilde{\lambda}_R(\boldsymbol{x}) + \tilde{\lambda}_D(\boldsymbol{x}) \geq \lambda_0 \\ 0 & \text{sinon,} \end{cases} \tag{4.8}$$

où λ_0 désigne un seuil donné, et

$$\tilde{\lambda}_R(\boldsymbol{x}) = \frac{1}{\sigma_0^2} \sum_{k=0}^{l-1} \frac{\gamma_k}{\gamma_k + \sigma_0^2} \left|\tilde{x}[k]\right|^2, \tag{4.9}$$

$$\tilde{\lambda}_D(\boldsymbol{x}) = 2 \sum_{k=0}^{l-1} \frac{1}{\gamma_k + \sigma_0^2} \operatorname{Re}\left\{\tilde{x}[k]\,\tilde{m}^*[k]\right\}. \tag{4.10}$$

On peut constater que ce résultat est équivalent à celui obtenu dans le cas continu [Fla88], hormis le fait que l'expression (4.7) remplace une décomposition de Karhunen-Loève de $\boldsymbol{x}$ et $\boldsymbol{m}$ sur une base de fonctions propres de $\boldsymbol{\Sigma}$. Ce parallèle peut être complété en recherchant une formulation temps-fréquence de la règle de décision (4.8). En adoptant une démarche analogue à celle présentée dans [Fla88], on est ainsi amené à constater que

$$\lambda_R(\boldsymbol{x}) = \frac{1}{l\sigma_0^2} \sum_{t,f=0}^{l-1} W_x^{(\cdot)}[t,f] \left(\sum_{k=0}^{l-1} \frac{\gamma_k}{\gamma_k + \sigma_0^2} W_{\psi_k}^{(\cdot)}[t,f] \right) \tag{4.11}$$

$$\lambda_D(\boldsymbol{x}) = \frac{2}{l} \sum_{t,f=0}^{l-1} \mathrm{Re}\left\{ W_{xm}^{(\cdot)}[t,f] \right\} \left(\sum_{k=0}^{l-1} \frac{1}{\gamma_k + \sigma_0^2} W_{\psi_k}^{(\cdot)}[t,f] \right) \tag{4.12}$$

faisant intervenir une distribution de Wigner discrète, sont des formes équivalentes des expressions (4.9) et (4.10) respectivement, à condition que la loi de conservation du produit scalaire suivante soit satisfaite :

$$\sum_{t,f=0}^{l-1} W_{x_1 z_1}^{(\cdot)}[t,f] \left(W_{x_2 z_2}^{(\cdot)}[t,f] \right)^* = l \left(\sum_{t=0}^{l-1} x_1[t]\, x_2^*[t] \right) \left(\sum_{t=0}^{l-1} z_1[t]\, z_2^*[t] \right)^* . \tag{4.13}$$

Si une relation similaire existe dans le cas continu, la validité de cette propriété dépend en revanche, dans le cas discret, de la définition attribuée à la distribution de Wigner. Aussi, on démontre aisément que seule la R-distribution y satisfait parmi les trois solutions étudiées. Elle est par conséquent la seule en mesure de garantir l'optimalité de la statistique de décision combinant les expressions (4.11) et (4.12). A cela s'ajoute la propriété de covariance par rapport aux translations dans le plan temps-fréquence qu'elle vérifie, ce qui en fait un mode de représentation privilégié pour la prise en compte de l'instant d'arrivée et de la fréquence Doppler de l'événement à détecter, en termes de paramètres de nuisance [Say95].

En conclusion, l'optimalité d'une règle de décision établie sur la base de la distribution de Wigner continue n'est pas garantie lorsqu'elle est transposée sans précautions préalables au cas des signaux discrets, ce qu'illustre la figure 4.1. Il s'avère donc indispensable d'étudier préalablement les propriétés de la distribution de Wigner discrète adoptée, la R-distribution pouvant constituer un choix judicieux de ce point de vue.

4.3 Détection à structure imposée

L'élaboration d'un détecteur optimal requiert la connaissance des propriétés statistiques de l'échantillon, conditionnellement à chacune des hypothèses en compétition. Celles-ci étant généralement inaccessibles lorsqu'on sort d'un cadre théorique tel que celui qui vient d'être dressé, on est couramment amené à leur substituer un autre type d'information *a priori*. Aussi un ensemble d'apprentissage $\mathcal{A}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ peut-il être utilisé à des fins d'élaboration d'une règle de décision. Pour ce faire, une démarche envisageable mais néanmoins sous-optimale consiste à 1) sélectionner une classe de détecteurs $\mathcal{D}$, puis 2) rechercher dans $\mathcal{D}$ la structure de détection qui minimise un critère de performance donné. Ce point de vue propre à la détection à structure imposée est à présent adopté dans le contexte temps-fréquence. Dans un soucis de cohérence avec les développements des chapitres précédents, cette partie est consacrée à la classe $\mathcal{D}$ des détecteurs linéaires opérant sur une distribution de Wigner discrète, c'est-à-dire

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \sum_{t,f=0}^{l-1} A[t,f]\, W_x^{(\cdot)}[t,f] \geq \lambda_0 \\ 0 & \text{sinon.} \end{cases} \tag{4.14}$$

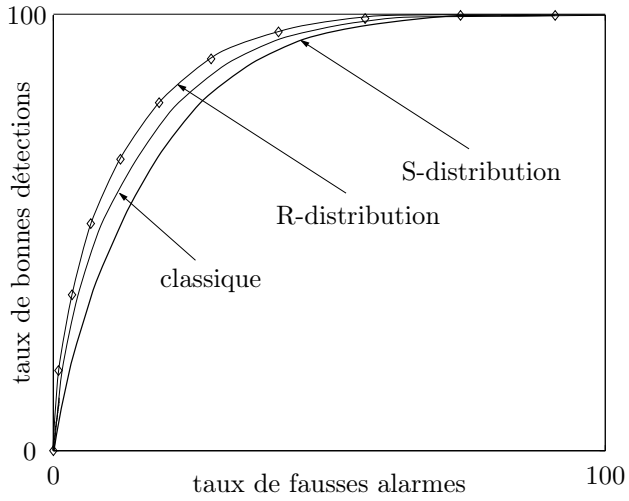

FIG. 4.1 : Détection d'un signal aléatoire gaussien noyé dans un bruit blanc, gaussien et centré. Sont ici comparées les courbes COR des détecteurs du type (4.11)-(4.12) opérant sur les trois distributions de Wigner discrètes considérées. On note que les performances de la structure associée à la R-distribution sont identiques à celles du détecteur de Bayes, ces dernières étant matérialisées par les $\Diamond$. La loi de conservation du produit scalaire à laquelle elle satisfait justifie cette optimalité.

Dans cette expression, les $A[t,f]$ et λ_0 désignent les paramètres devant être déterminés à partir de l'information *a priori* disponible. Il est à noter que ce type de structure de détection a été largement étudié dans la littérature. Il confère en effet au détecteur quadratique classique une interprétabilité accrue, et facilite le traitement des paramètres de nuisance que peuvent constituer l'instant d'arrivée et la fréquence Doppler du signal à détecter. On peut cependant déplorer les nombreux calculs requis par la représentation des observations dans le plan temps-fréquence, étape préalable à la prise de décision. Aussi l'objectif de cette partie est-il tout d'abord de mettre à profit la théorie des noyaux reproduisants afin de synthétiser, puis mettre en œuvre la structure (4.14), sans nécessairement calculer de représentations temps-fréquence. Enfin, on discute du choix d'une distribution de Wigner discrète qui pourrait garantir les meilleures performances en détection.

4.3.1 Détection dans le plan temps-fréquence et noyaux reproduisants

Un espace signal $\mathcal{X} \subset \mathbb{C}^l$ est dit linéaire sur $\mathbb{C}$ s'il vérifie la propriété suivante : quels que soient $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ des éléments de $\mathcal{X}$ et (α,β) un couple de scalaires complexes, l'élément défini par $(\alpha\,\boldsymbol{x}_1 + \beta\,\boldsymbol{x}_2)$ appartient également à $\mathcal{X}$. Soit $\{\boldsymbol{s}_q\}$ un sous-ensemble non vide de $\mathcal{X}$. On dit qu'il constitue une base de $\mathcal{X}$ si les $\boldsymbol{s}_q$ sont linéairement indépendants et s'ils engendrent $\mathcal{X}$. La dimension de cet espace est alors donnée par le cardinal de la base $\{\boldsymbol{s}_q\}$. De plus, cette dernière est dite orthonormée si $\boldsymbol{s}_q \cdot \boldsymbol{s}_{q'} = \delta_{qq'}$, où $\boldsymbol{s}_q \cdot \boldsymbol{s}_{q'}$ désigne le produit scalaire des deux éléments considérés, et $\delta_{qq'}$ le symbole de Kronecker. Tout élément $\boldsymbol{x}$ de $\mathcal{X}$ peut alors être représenté par $\boldsymbol{x} = \sum_q \alpha_q\,\boldsymbol{s}_q$, avec $\alpha_q = \boldsymbol{x} \cdot \boldsymbol{s}_q$. Dans le cadre de ce chapitre, on suppose que $\mathcal{X}$ est l'espace linéaire sur $\mathbb{C}$ des signaux complexes de longueur l.

Soit $\boldsymbol{\phi}^{(\cdot)}$ l'application associant la distribution de Wigner discrète $\boldsymbol{W}_x^{(\cdot)}$ à tout signal complexe $\boldsymbol{x}$. On désigne par $\mathcal{T}^{(\cdot)}$ l'espace image de $\boldsymbol{\phi}^{(\cdot)}$, soit

$$\mathcal{T}^{(\cdot)} \triangleq \left\{ \boldsymbol{W}_x^{(\cdot)} : \boldsymbol{W}_x^{(\cdot)} = \boldsymbol{\phi}^{(\cdot)}(\boldsymbol{x}),\, \boldsymbol{x} \in \mathbb{C}^l \right\}. \tag{4.15}$$

On constate que $\mathcal{T}^{(\cdot)}$ n'est pas un espace linéaire puisque toute combinaison linéaire de distributions de Wigner discrètes n'est pas nécessairement une distribution de Wigner discrète valide. Aussi, on lui associe l'espace linéaire regroupant toutes les combinaisons linéaires de distributions de Wigner discrètes sur $\mathbb{R}$, que l'on note $\tilde{\mathcal{T}}^{(\cdot)}$. Ce dernier est appelé *espace linéaire induit* [Hla92]. Enfin, on confère à $\tilde{\mathcal{T}}^{(\cdot)}$ le statut d'espace euclidien en le munissant du produit scalaire suivant

$$\boldsymbol{W}_1^{(\cdot)} \cdot \boldsymbol{W}_2^{(\cdot)} = \sum_{t=0}^{l-1} \sum_{f=0}^{l-1} W_1^{(\cdot)}[t,f]\, W_2^{(\cdot)}[t,f]. \tag{4.16}$$

Ainsi voit-on se profiler une structure algébrique similaire à celles rencontrées lors du chapitre 3. Afin de parachever le parallèle avec ces précédents développements, on choisit de reformuler préalablement la règle de décision (4.14) ainsi

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \boldsymbol{A} \cdot \boldsymbol{W}_x^{(\cdot)} \geq \lambda_0 \\ 0 & \text{sinon.} \end{cases} \tag{4.17}$$

On note que la recherche de $\boldsymbol{A}$ peut être limitée à $\tilde{\mathcal{T}}^{(\cdot)}$, toute composante $\boldsymbol{A}^{\perp}$ extraite de l'espace complémentaire $\tilde{\mathcal{T}}^{\perp(\cdot)}$ dans $\mathbb{R}^{l^2}$ étant sans conséquence sur la statistique de détection, compte tenu de la forme de celle-ci. Plus précisément encore, l'information disponible sur l'observation $\boldsymbol{X}$ se bornant aux données $\boldsymbol{x}_k$ de $\mathcal{A}_n$, le domaine de définition de $\boldsymbol{A}$ peut être restreint à l'espace linéaire induit par les éléments $\boldsymbol{W}_{x_k}^{(\cdot)}$, soit

$$\boldsymbol{A} = \sum_{k=1}^{n} a_k\, \boldsymbol{W}_{x_k}^{(\cdot)}, \tag{4.18}$$

où les paramètres a_k sont à déterminer à partir de l'information *a priori* dont on dispose. Il s'agit là d'un point de vue original sur la détection par représentation temps-fréquence, menant directement à la formulation à noyau suivante

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \sum_{k=1}^{n} a_k\, \kappa^{(\cdot)}(\boldsymbol{x}_k, \boldsymbol{x}) \geq \lambda_0 \\ 0 & \text{sinon.} \end{cases} \tag{4.19}$$

Dans cette expression, le noyau $\kappa^{(\cdot)}(\boldsymbol{x}_k, \boldsymbol{x})$ représente le produit scalaire des éléments $\boldsymbol{x}_k$ et $\boldsymbol{x}$ dans l'espace transformé $\tilde{\mathcal{T}}^{(\cdot)}$. Force est de constater que son évaluation ne requière pas nécessairement le calcul explicite de représentations temps-fréquence. Dans le cas de la R-distribution par exemple, $\kappa^{(R)}$ revêt une forme particulièrement simple puisque $\kappa^{(R)}(\boldsymbol{x}_k, \boldsymbol{x}) = l\,(\boldsymbol{x}_k \cdot \boldsymbol{x})^2$, d'après la loi de conservation (4.13). Ne présentant pas de difficulté, le calcul des noyaux associés aux distributions classique (4.2) et de Stanković (4.4) est quant à lui laissé au soin du lecteur.

L'une des conséquences directes de la formulation (4.19) du détecteur (4.17) est l'existence avérée de noyaux reproduisants offrant des propriétés de covariance vis à vis de certaines transformations, ici les translations en temps et en fréquence de l'observation. Aussi me semble-t-il opportun, dans le cadre de travaux ultérieurs, de caractériser de tels noyaux sans nécessairement connaître l'espace transformé qui leur est associé. En effet, les structures ainsi constituée permettraient de traiter avec efficience certains problèmes de détection sur lesquels pèsent des paramètres de nuisance, par exemple un décalage Doppler ou un changement d'échelle. A ma connaissance, ce thème n'a encore fait l'objet d'aucun développement dans le domaine des Support Vector Machines et autres méthodes à noyau.

4.3.2 Distribution classique vs. R-distribution

Afin de discuter du choix d'une distribution de Wigner discrète qui pourrait garantir aux structures de détection (4.17) et (4.19) les meilleures performances, il convient de comparer les classes de détecteurs que l'on peut générer dans chacun des cas [Ric02(b)]. Pour cela, il est nécessaire de caractériser l'espace transformé $\tilde{\mathcal{T}}^{(\cdot)}$ associé à chacune des distributions considérées, qui correspond également au domaine de recherche de la référence $\boldsymbol{A}$ figurant dans l'expression (4.17).

On note Δ_{t_0} l'impulsion unité définie par $\Delta_{t_0}[t] = 1$ si $t = t_0$, 0 sinon. Par construction, la distribution classique (4.2) peut être réécrite selon

$$\boldsymbol{W}_x^{(C)} = \sum_{(t,\tau)\in\mathcal{J}} x[t+\tau]\, x^*[t-\tau]\, \boldsymbol{W}^{(C)}_{\Delta_{(t+\tau)}\Delta_{(t-\tau)}}, \tag{4.20}$$

où $\mathcal{J}$ désigne l'ensemble des paires (t,τ) telles que $(t+\tau)$ et $(t-\tau)$ appartiennent conjointement à $\{0,\ldots,l-1\}$. On remarque que les éléments générateurs mis en évidence ci-dessus n'appartiennent pas à l'espace euclidien $\tilde{\mathcal{T}}^{(C)}$, et que l'expression (4.20) ne constitue pas une combinaison linéaire sur le corps $\mathbb{R}$. Aussi ce résultat ne donne-t-il pas entière satisfaction puisqu'il n'offre pas directement d'ensemble d'éléments que l'on pourrait qualifier *a posteriori* de base de l'espace euclidien considéré. Toutefois, un calcul peu engageant mais simple permet de montrer que

$$4\,\boldsymbol{W}^{(C)}_{x_1x_2} = (1-j)\left[\boldsymbol{W}^{(C)}_{x_1+x_2} - \boldsymbol{W}^{(C)}_{x_1+jx_2}\right] + (1+j)\left[\boldsymbol{W}^{(C)}_{x_2+x_1} - \boldsymbol{W}^{(C)}_{x_2+jx_1}\right]. \tag{4.21}$$

En combinant cette relation avec l'expression (4.20), puis en considérant la partie réelle du résultat, on aboutit à

$$\boldsymbol{W}_x^{(C)} = \frac{1}{2}\sum_{(t,\tau)\in\mathcal{J}} \left(\mathrm{Re}\{x[t+\tau]\, x^*[t-\tau]\} + \mathrm{Im}\{x[t+\tau]\, x^*[t-\tau]\}\right) \left(\boldsymbol{W}^{(C)}_{\Delta_{(t+\tau)}+\Delta_{(t-\tau)}} - \boldsymbol{W}^{(C)}_{\Delta_{(t+\tau)}+j\Delta_{(t-\tau)}}\right). \tag{4.22}$$

Il en résulte une famille génératrice de l'espace euclidien $\tilde{\mathcal{T}}^{(C)}$, donnée par

$$\mathcal{B}^{(C)} = \left\{\boldsymbol{W}^{(C)}_{\Delta_{(t+\tau)}+\Delta_{(t-\tau)}} - \boldsymbol{W}^{(C)}_{\Delta_{(t+\tau)}+j\Delta_{(t-\tau)}} \,:\, (t,\tau)\in\mathcal{J}\right\}. \tag{4.23}$$

On constate qu'il s'agit d'une famille libre en montrant que les éléments qui la composent sont orthogonaux deux à deux, ce qui implique que $\mathcal{B}^{(C)}$ constitue une base orthogonale de $\tilde{\mathcal{T}}^{(C)}$. L'occasion est à présent offerte de déterminer la dimension de cet espace en évaluant le cardinal de la base obtenue. En combinant $0 \le t+\tau \le l-1$ et $0 \le t-\tau \le l-1$, on obtient que $-t \le \tau \le t$ si $0 \le t \le \lfloor\frac{l-1}{2}\rfloor$, et $-(l-t-1) \le \tau \le (l-t-1)$ sinon, où $\lfloor\cdot\rfloor$ représente la fonction *partie entière*. Ces inégalités fournissent directement le résultat

$$\dim(\tilde{\mathcal{T}}^{(C)}) = \sum_{t=0}^{\lfloor\frac{l-1}{2}\rfloor}[2t+1] + \sum_{t=\lfloor\frac{l-1}{2}\rfloor+1}^{l-1}[2(l-t-1)+1], \tag{4.24}$$

soit encore que la dimension de $\tilde{\mathcal{T}}^{(C)}$ est égale à $\lfloor\frac{l^2+1}{2}\rfloor$. Parce que cette valeur est strictement inférieure au nombre l^2 de composantes constituant les représentations, on en déduit que l'information qu'elles véhiculent est linéairement redondante. Ceci signifie qu'il est possible d'exhiber

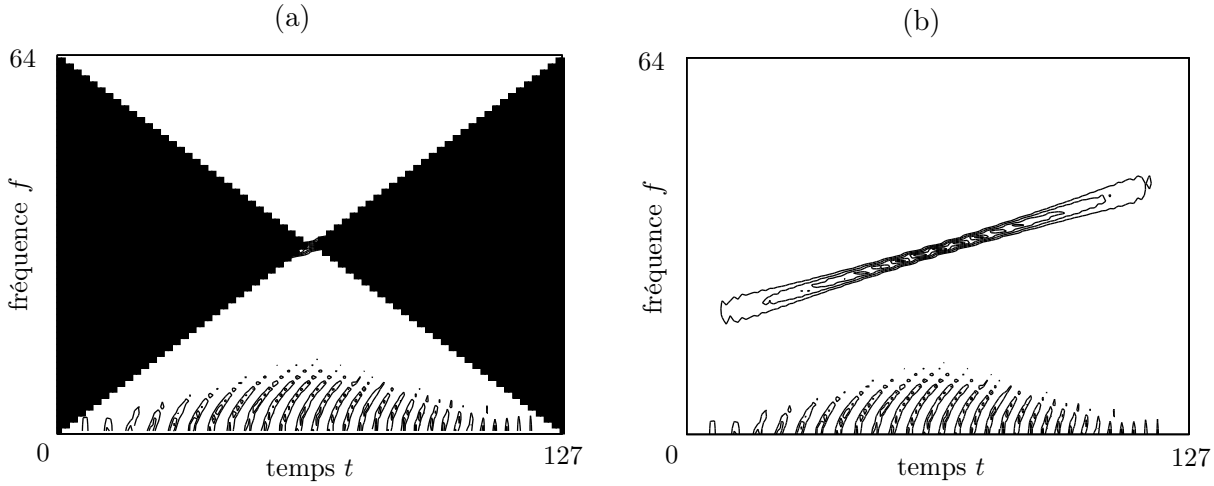

FIG. 4.2 : Illustration de la redondance informationnelle de la distribution $\boldsymbol{W}_x^{(C)}$. L'information non-masquée dans (a) permet de reconstruire en totalité la représentation (b) via une application linéaire. Le cas ici considéré est celui d'un signal réel, traité dans [Ric01(b)].

des familles regroupant seulement $\lfloor \frac{l^2+1}{2} \rfloor$ composantes du plan temps-fréquence et concentrant pourtant la totalité de l'information, les autres composantes pouvant être déduites des premières par une application linéaire. Une étude approfondie de cette propriété, qu'illustre la figure 4.2, et de ses conséquences en détection est proposée dans [Ric01(b)].

On s'intéresse maintenant à la R-distribution. Dans le but d'exhiber une famille génératrice de l'espace induit qui lui est associé, on procède comme précédemment en considérant le développement suivant [Ric02(b)]

$$\boldsymbol{W}_x^{(R)} = \sum_{t,\tau=0}^{l-1} x[(t+\tau)_l]\, x^*[t]\, \boldsymbol{W}^{(R)}_{\Delta_{(t+\tau)_l}\Delta_t}. \tag{4.25}$$

En utilisant la relation (4.21), qui s'applique également au cas présent, on aboutit à une famille génératrice de $\tilde{\mathcal{T}}^{(R)}$

$$\mathcal{B}^{(R)} = \left\{ \boldsymbol{W}^{(R)}_{\Delta_{(t+\tau)_l}+\Delta_t} - \boldsymbol{W}^{(R)}_{\Delta_{(t+\tau)_l}+j\Delta_t} \,:\, 0 \le t,\tau \le l-1 \right\}. \tag{4.26}$$

On est amené à constater qu'il s'agit d'une base orthogonale de cet espace euclidien en montrant que les éléments de $\mathcal{B}^{(R)}$ sont orthogonaux deux à deux. Il s'en suit que

$$\dim(\tilde{\mathcal{T}}^{(R)}) = \mathrm{card}(\mathcal{B}^{(R)}) = l^2, \tag{4.27}$$

ce qui signifie que les l^2 composantes de $\boldsymbol{W}_x^{(R)}$ sont linéairement indépendantes. Ceci entraîne directement que la famille $\mathcal{D}^{(R)}$ des détecteurs linéaires opérant sur la R-distribution est confondue avec $\mathcal{Q}$, la classe des détecteurs quadratiques définis par

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \sum_{t,\tau=0}^{l-1} x[t]\, Q[t,\tau]\, x^*[\tau] \ge \lambda_0 \\ 0 & \text{sinon,} \end{cases} \tag{4.28}$$

où $\boldsymbol{Q}$ est une matrice hermitienne assurant le caractère réel de la statistique de détection. Par conséquent, la classe $\mathcal{D}^{(C)}$ est incluse dans $\mathcal{D}^{(R)}$ car $\mathcal{T}^{(C)}$ est un sous-espace de $\mathcal{T}^{(R)}$. En d'autres termes, $\mathcal{D}^{(R)}$ propose le plus large éventail de solutions, et est théoriquement toujours à même de fournir un détecteur au moins aussi performant que toute structure de détection issue de $\mathcal{D}^{(C)}$.

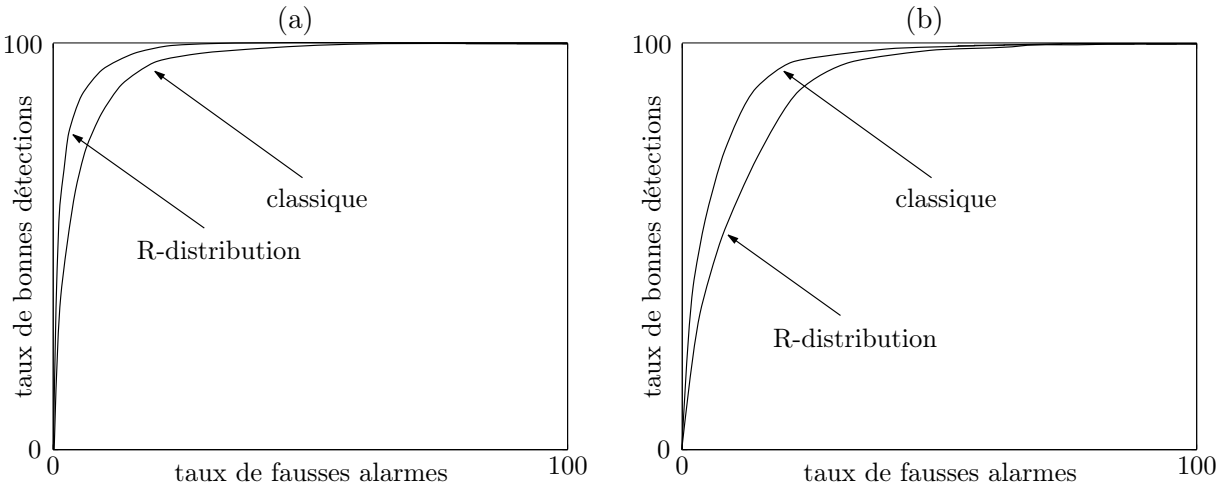

FIG. 4.3 : Performances de détecteurs linéaires opérant sur la distribution classique et la R-distribution, synthétisés avec l'algorithme du critère optimal à partir d'ensembles d'apprentissage constitués de (a) 12000 individus et (b) 200 individus.

4.3.3 Influence de la malédiction de la dimensionnalité

Comme on a pu l'évoquer à de multiples reprises, diverses stratégies peuvent être adoptées pour la résolution d'un problème de détection, selon la nature de l'information *a priori* à laquelle on accède. On suppose ici disposer d'une base d'apprentissage pour pouvoir ajuster les paramètres caractéristiques d'une structure de détection. Lorsqu'on adopte une telle démarche, les performances du détecteur obtenu sont conditionnées par l'adéquation existant entre la complexité de celui-ci et la taille de la base d'apprentissage. Ainsi, les récepteurs dotés d'un nombre de degrés de liberté trop important auront un faible pouvoir de généralisation. Dans le cas contraire, ces derniers seront incapables d'intégrer la totalité de l'information discriminante présente dans l'ensemble d'apprentissage. Entre ces extrêmes, il existe une complexité optimale pour laquelle la probabilité d'erreur du détecteur est minimale. Ce comportement de la probabilité d'erreur, exposé au cours du chapitre 1 et qualifié de malédiction de la dimensionnalité, a été formellement étudié par Vapnik et Chervonenkis [Vap71]. Pour ce faire, ces auteurs ont été amenés à définir la complexité d'une structure de détection au moyen d'une quantité appelée VC-dimension, ici notée h, qu'il convient d'adapter à la taille de l'ensemble d'apprentissage de sorte à contenir les effets néfastes de la malédiction de la dimensionnalité.

Généralement, l'estimation de la VC-dimension d'une famille de détecteurs constitue une tâche difficile. On rappelle toutefois que $h = l + 1$ dans le cas des structures de détection linéaires, où l représente la dimension de l'espace engendré par les observations, ou par leur représentation temps-fréquence si les détecteurs considérés opèrent dans ce domaine. A partir des équations (4.24) et (4.27), on obtient ainsi directement que $h^{(R)} = l^2 + 1$ dans le cas de détecteurs linéaires opérant sur la R-distribution, tandis que $h^{(C)} = \lfloor (l^2 + 1)/2 \rfloor + 1$ lorsque cette même structure est associée à la distribution classique. On constate donc que $h^{(R)} > h^{(C)}$, ce qui signifie que les détecteurs de la classe $\mathcal{D}^{(R)}$ sont d'avantage sujets au phénomène de malédiction de la dimensionnalité que ceux de $\mathcal{D}^{(C)}$. Les éléments de théorie proposés jusqu'à présent ont toujours plaidé en faveur de la R-distribution. En pratique, force est de constater que la comparaison des classes $\mathcal{D}^{(R)}$ et $\mathcal{D}^{(C)}$ peut cependant être à l'avantage de la dernière, et cela malgré la perte d'information statistique. Ce phénomène, d'autant plus manifeste que l'ensemble

d'apprentissage est de faible cardinalité, est à présent mis en évidence au moyen de simulations. Le problème considéré est celui de la détection d'un signal $s[t\,;\phi_0)$ constitué de 16 échantillons, où ϕ_0 représente une phase aléatoire uniformément distribuée sur l'intervalle $[-\pi,\pi[$. Celui-ci est noyé dans un bruit blanc additif $w(t)$ distribué selon $(1-\epsilon)\mathcal{N}(0,\sigma^2)+\epsilon\mathcal{N}(0,K^2\sigma^2)$ avec $\epsilon = 0.5$ et $K = 5$, où $\mathcal{N}(0,\sigma^2)$ désigne la loi normale de moyenne nulle et de variance σ^2. Le rapport signal-sur-bruit est fixé à -6 dB. Dans un premier temps, des détecteurs temps-fréquence linéaires opérant sur les deux distributions de Wigner discrètes considérées ont été élaborés grâce à l'algorithme du critère optimal, à partir d'un ensemble d'apprentissage constitué de 12000 individus. La comparaison des performances des solutions obtenues, illustrée par la figure 4.3.(a), est conforme aux éléments de théorie présentés précédemment : $\mathcal{D}^{(R)}$ est toujours en mesure de fournir une solution au moins aussi performante que $\mathcal{D}^{(C)}$, à condition que les effets de la malédiction de la dimensionnalité demeurent négligeables. Dans un second temps, l'expérience a été renouvelée avec une base d'apprentissage constituée de 200 individus. La figure 4.3.(b) montre que le détecteur linéaire associé à la R-distribution présente de moins bonnes performances que celui opérant sur le distribution classique, bien que $\mathcal{D}^{(C)}$ soit un sous-ensemble de $\mathcal{D}^{(R)}$. Comme cela a été longuement décrit précédemment, la malédiction de la dimensionnalité justifie ce surprenant résultat.

4.3.4 Une alternative intéressante : la S-distribution

Nul n'est besoin d'insister à nouveau sur la nécessité de contrôler la complexité d'une structure de détection lorsque son élaboration repose sur un ensemble d'apprentissage de taille finie. Il existe de multiples façons de s'acquitter de cette tâche, parmi lesquelles on recense les procédures *weight decay* et *optimal brain damage* décrites au cours du chapitre précédent. On rappelle que la première est une méthode de régularisation tandis que la seconde consiste indirectement en une projection des données dans un espace de dimension réduite. Ce dernier principe a été mis en œuvre à de multiples reprises dans le cadre de la détection par représentations temps-fréquence, sous la forme d'un lissage préalable des distributions. Ainsi fait-on par exemple appel dans [Ric99(b)] à un spectrogramme dont on a optimisé la fenêtre d'analyse plutôt qu'à la distribution de Wigner, ou encore détourne-t-on dans [Dav01] la méthode du noyau optimal de Jones et Baraniuk [Jon95] de son usage originel. Il convient de noter que le mode opératoire de ce type de techniques, qui permet d'améliorer sensiblement les performances en généralisation d'une structure de décision opérant dans le plan temps-fréquence, a pour la première fois été justifié par des considérations issues de la théorie statistique de l'apprentissage dans [Ric98(a)]. Avec ces grands principes en toile de fond, l'objectif de cette section est de montrer que la S-distribution peut être adoptée à des fins de contrôle de complexité, ce qui lui procure un avantage substantiel sur la distribution classique et la R-distribution.

En traitant le cas des détecteurs linéaires opérant sur les distributions classique et de Richman *et col.*, on a montré que $h^{(C)} = \lfloor\frac{l^2+1}{2}\rfloor + 1$ et $h^{(R)} = l^2 + 1$, où l désigne la longueur des signaux complexes étudiés. En pratique, il est donc nécessaire de recourir à un post-traitement des représentations temps-fréquence afin d'adapter la complexité des détecteurs opérant sur celles-ci à la taille de la base d'apprentissage. Dans le cas de la S-distribution, il apparaît clairement que l'on peut tirer profit de la fenêtre de lissage $g[t]$ figurant dans la définition (4.5) afin de contrôler la dimension de l'espace induit $\tilde{\mathcal{T}}^{(S)}$. Si le calcul de cette dimension n'est pas traité dans ce document, on peut toutefois remarquer que

1) $\dim(\tilde{\mathcal{T}}^{(S)}) = l$ si $l_g = 1$

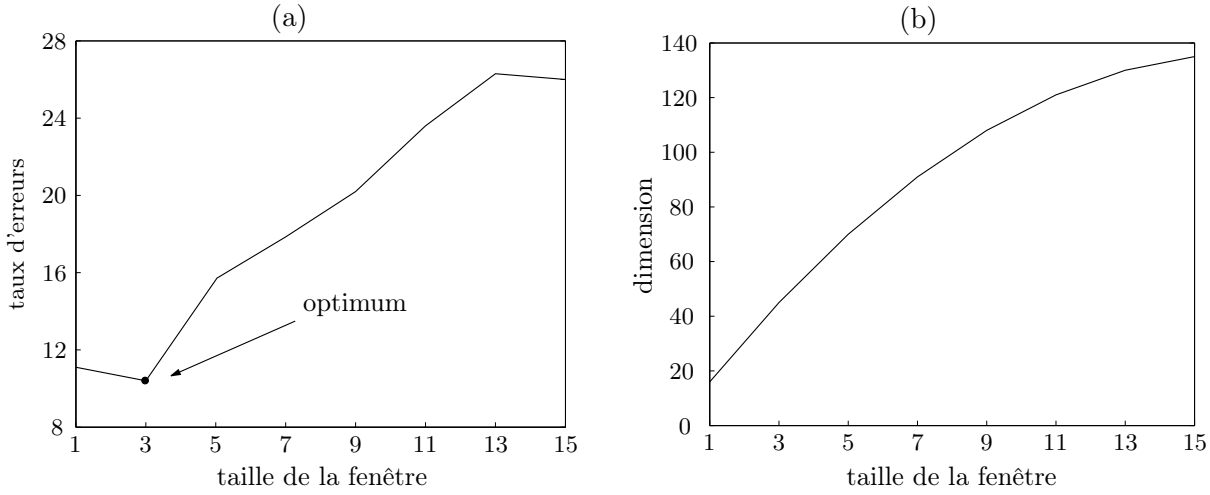

FIG. 4.4 : Contrôle de la complexité d'un détecteur linéaire opérant sur la S-distribution. Ces graphiques représentent l'évolution (a) d'une estimation de l'erreur de généralisation et (b) de la dimension de l'espace induit $\tilde{\mathcal{T}}^{(C)}$, en fonction de la taille l_g de la fenêtre d'analyse $g[t]$. L'ensemble d'apprentissage utilisé est identique à celui considéré dans la figure 4.3.

2) $\dim(\tilde{\mathcal{T}}^{(S)}) \leq l^2$,

où l_g désigne la taille de la fenêtre d'analyse $g[t]$. En effet, $W_x^{(S)}[t,f]$ est proportionnel à $\|x(t)\|^2$ lorsque $l_g = 1$, ce qui mène à 1). La taille l^2 des représentations justifie 2), et implique directement que la classe $\mathcal{D}^{(S)}$ des détecteurs linéaires opérant sur la S-distribution est incluse dans $\mathcal{D}^{(R)}$. Afin d'illustrer l'aptitude de cette distribution à contrôler la capacité de la structure de détection qui lui est associée, on considère le problème décrit en fin de section précédente. On rappelle qu'il concerne la détection d'un signal $s[t\,;\phi_0)$, composé de 16 échantillons et de phase initiale aléatoire ϕ_0, noyé dans un bruit blanc suivant une loi de type $(1-\epsilon)\mathcal{N}(0,\sigma^2) + \epsilon\mathcal{N}(0,K^2\sigma^2)$, à partir d'une base d'apprentissage constituée de 200 individus. Les figures 4.4.(a) et 4.4.(b) représentent les variations de performances du détecteur obtenu par la méthode du critère optimal, ainsi que la dimension de l'espace induit $\tilde{\mathcal{T}}^{(S)}$, en fonction de la taille l_g de la fenêtre rectangulaire $g[t]$. On observe qu'un taux d'erreur minimal est atteint pour $l_g = 3$, ce qui correspond à $\dim(\tilde{\mathcal{T}}^{(S)}) = 43$. Comme indiqué par la figure 4.5, la structure obtenue est plus performante que celles associées aux distributions classique et de Richman *et col.* Dans ces deux cas, on note en particulier que $\dim(\tilde{\mathcal{T}}^{(C)}) = 128$ et $\dim(\tilde{\mathcal{T}}^{(R)}) = 256$, ce qui justifie la plus grande vulnérabilité des structures concernées vis à vis de la malédiction de la dimensionnalité.

Dans [Sta94, Sta01], les auteurs montrent que les termes interférentiels inhérents au caractère quadratique de la S-distribution, et nuisant à sa lisibilité, peuvent être atténués en ajustant le paramètre b_f figurant dans la définition (4.4). Avant de clore le sujet, on montre à présent que b_f n'influe aucunement sur les performances d'un détecteur linéaire opérant sur la S-distribution. Étant donnée une fenêtre d'analyse $g[t]$ de largeur fixée, on note que $F_x[t,f+\nu] = F_{x_{+\nu}}[t,f]$ avec $x_{+\nu}[t] = x[t]\,\exp(-2j\pi\nu t/l)$, où $F_x[t,f]$ représente la transformée de Fourier à court terme. Par construction, la S-distribution peut donc s'exprimer ainsi

$$W_{x_1x_2}^{(S)}[t,f] = \sum_{\nu=-b_f}^{b_f} S_{x_{1+\nu}x_{2-\nu}}[t,f], \tag{4.29}$$

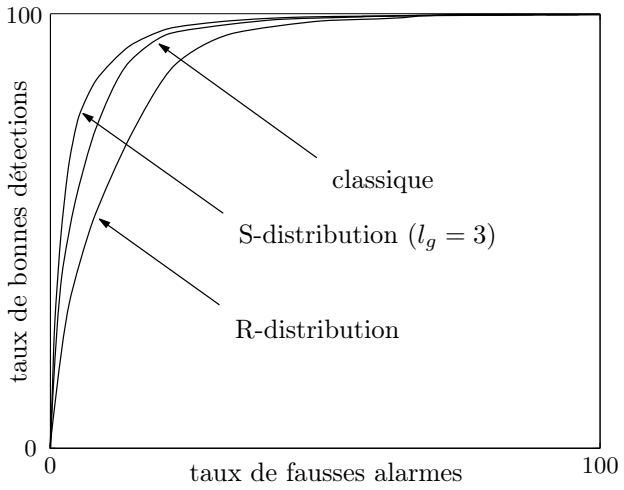

FIG. 4.5 : Performances des détecteurs opérant sur les distributions classique, de Richman *et col.* et de Stanković. Ceux-ci ont été synthétisés à l'aide de la méthode du critère optimal, à partir d'une base d'apprentissage composée de 200 individus.

où $S_{x_{1+\nu}x_{2-\nu}}[t,f]$ désigne une composante du spectrogramme $\frac{1}{l}\,F_{x_{1+\nu}}[t,f]\,F^*_{x_{2-\nu}}[t,f]$. Dans le cas de la distribution de Wigner classique, il a été précédemment évoqué l'existence de composantes du plan temps-fréquence où se concentre l'intégralité de l'information disponible. Ce même point de vue peut évidemment être adopté pour la S-distribution et le spectrogramme, dont les liens sont soulignés par l'expression (4.29). On déduit en particulier de cette relation que les deux distributions considérées concentrent l'information qu'elles véhiculent en de mêmes lieux temps-fréquence, étant donnés b_f et une fenêtre d'analyse $g[t]$. Il en résulte que les classes de détecteurs linéaires opérant sur le spectrogramme et la S-distribution coincident, ce qui démontre l'absence d'incidence du paramètre b_f sur la capacité d'apprentissage de telles structures. En conséquence, b_f peut être totalement dédié à l'amélioration de la lisibilité des représentations, sans que cela influe sur le processus de décision.

On a montré que l'on peut tirer profit de la fenêtre d'analyse $g[t]$ pour contrôler la dimension de l'espace induit par la S-distribution, et améliorer ainsi les performances d'une structure de détection qui lui serait associée. On peut cependant regretter qu'un tel procédé réserve le même sort à l'information capitale pour la résolution d'un problème de détection, et à celle qui ne l'est pas. La prochaine section est consacrée à la présentation d'une technique de filtrage des représentations temps-fréquence à même de faire une telle distinction, pour une meilleure préservation de l'information discriminante.

4.4 Contrôle de complexité par diffusion des représentations

Durant la dernière décennie, la communauté du traitement d'image a connu l'émergence de nouveaux outils reposant sur la théorie des équations aux dérivées partielles. Ceux-ci s'avèrent particulièrement performants pour l'amélioration, la restauration ou encore l'analyse multi-échelle d'images [Wei96]. De plus, certains vérifient des propriétés intéressantes telles que l'existence et l'unicité de solution [Cat92], et peuvent être mis en œuvre au moyen de schémas numériques très efficaces [Alv94]. Enfin, il est à noter que ces procédés font l'objet d'un certain nombre de travaux unificateurs [Der96].

Les opérations de diffusion, parmi lesquelles on compte l'équation de la chaleur, font partie intégrante des techniques qui viennent d'être évoquées [Wei97]. Elles expriment un processus physique visant à équilibrer des différences de concentration, sans qu'il y ait création ou perte de matière. Sous une forme non-homogène, elles offrent la possibilité d'intégrer de la connaissance *a priori* par le biais d'une fonction de conductance dépendant des coordonnées spatiales. Cette dernière a pour rôle de guider l'évolution du processus de diffusion au cours du temps, permettant un traitement différencié des diverses zones d'information présentes dans une image. Les techniques de diffusion ont été récemment introduites dans le domaine de l'analyse des signaux non-stationnaires et peu de travaux leur sont encore consacrés [Gon98, Gos02]. Dans [Gon98], il est question d'améliorer la lisibilité de représentations temps-fréquence bruitées et encombrées de termes interférentiels. La méthode proposée, s'inspirant de la diffusion de Perona et Malik [Per90, Per94], est confrontée aux approches les plus performantes [Aug95, Jon95] et son efficacité clairement démontrée. Les avancées les plus récentes concernent l'application de techniques de diffusion pour le contrôle de la complexité de détecteurs opérant dans le plan temps-fréquence. Les résultats obtenus [Gos02], décrits en cette fin de chapitre, sont le fruit de travaux menés avec un doctorant, Julien Gosme, et Paulo Gonçalvès de l'INRIA Rhône-Alpes.

4.4.1 Opérateurs de diffusion en analyse temps-fréquence

Il a été établi que la fonction de Green de l'équation de la chaleur, dont on rappelle qu'elle est définie par

$$\frac{\partial H(u,v\,;\,\tau)}{\partial \tau} = c\,\Delta H(u,v\,;\,\tau), \tag{4.30}$$

où $h(u,v)$ désigne l'état initial du système et c un paramètre réel, est donnée par le noyau gaussien $k(u,v\,;\,\tau) = (2\pi\tau)^{-1}\,\exp(-(u^2+v^2)/2\tau)$. Ceci signifie que la solution de l'équation (4.30) peut être obtenue, à chaque instant τ, par convolution bi-dimensionnelle de $k(u,v\,;\,\tau)$ avec $h(u,v)$. Dans [Koe84], Koenderink a introduit cette conception du filtrage dans le domaine du traitement d'image, et apporté ainsi un point de vue nouveau sur les classiques techniques de restauration en les assimilant à une diffusion linéaire homogène. Par la suite, le schéma (4.30) a connu de nombreux raffinements, par exemple en y introduisant une fonction de conductance dépendant des coordonnées spatiales u et v, et de l'instant de diffusion τ. Le processus de Perona et Malik repose sur ce principe [Per90, Per94], et joue un rôle central au sein de la famille des opérateurs de diffusion non-homogènes. Il est défini par

$$\frac{\partial H(u,v;\tau)}{\partial \tau} = \mathrm{div}\{c(|\nabla H|)\,\nabla H\}. \tag{4.31}$$

Dans cette expression, $c(\cdot)$ désigne une fonction positive décroissante de $|\nabla H|$ dont le rôle est d'inhiber la diffusion dans les régions où le contraste est marqué, par exemple

$$c(|\nabla H|) = \frac{1}{\sqrt{1+|\nabla H|^2/\delta^2}} \tag{4.32}$$

$$c(|\nabla H|) = \exp(|\nabla H|^2/\delta^2), \tag{4.33}$$

où δ est un paramètre de normalisation. Aussi cette technique adaptative permet-elle, par exemple, d'éradiquer le bruit des images tout en y préservant les contours.

Inspirés par cette démarche, Gonçalvès et Payot [Gon98] ont récemment établi une connection entre l'analyse temps-fréquence et le formalisme des équations aux dérivées partielles. Ils ont ainsi montré que la diffusion de la distribution de Wigner (4.1) par l'équation de la chaleur (4.30) conduit, à l'instant $\tau = (4\pi)^{-1}$, à un spectrogramme $S_x(t,f)$ dont la fenêtre de lissage est une fonction gaussienne d'écart-type $\sigma = (2\sqrt{\pi})^{-1}$. Poursuivant le raisonnement de [Per90], ces mêmes auteurs ont suggéré d'adopter le schéma (4.31) comme technique adaptative de lissage non-homogène des représentations temps-fréquence. Appliqué à la distribution de Wigner, le procédé proposé s'exprime ainsi

$$\left\{ \begin{array}{l} W_x(t,f;\tau=0) = W_x(t,f) \\ \frac{\partial W_x(t,f;\tau)}{\partial \tau} = \mathrm{div}\{c_x(t,f)\,\nabla W_x(t,f;\tau)\}, \end{array} \right. \tag{4.34}$$

où $c_x(t,f)$ désigne une fonction de conductance à déterminer en fonction de l'objectif visé. A titre de remarque, celle-ci peut éventuellement dépendre de l'instant τ, le processus de diffusion étant alors qualifié de non-linéaire. Dans [Gon98], le schéma (4.34) a été mis en œuvre pour l'amélioration de la lisibilité des représentations temps-fréquence encombrées de termes interférentiels. Afin d'isoler ces composantes non-désirées dans le plan temps-fréquence, la fonction de conductance suivante a été proposée

$$c_x(t,f) = \left(1 + \left(\frac{S_x(t,f)}{\delta}\right)^{\alpha}\right)^{-1}, \tag{4.35}$$

avec $\alpha \geq 0$ et $\delta > 0$. Cette dernière repose ainsi sur le spectrogramme, dont l'une des caractéristiques est le contenu interférentiel réduit. Aussi c_x inhibe-t-elle le processus de lissage dans les régions du plan temps-fréquence correspondant au support du spectrogramme, où se trouvent localisées les composantes du signal. Inversement, elle favorise la diffusion en des lieux (t,f) où le spectrogramme est caractérisé par de faibles valeurs, susceptibles d'être le siège des composantes interférentielles de la distribution de Wigner traitée. Parce que la diffusion adaptative est une technique itérative, il convient de lui associer un critère d'arrêt. La solution préconisée par Gonçalvès et Payot est une mesure d'entropie [Gon98]. Il est à noter que cette référence propose également un exemple de réalisation numérique de la procédure de diffusion (4.34), dont le lecteur est invité à s'inspirer pour la mise en œuvre de ce qui suit.

4.4.2 Application à la détection par représentation temps-fréquence

Précédemment, il a été évoqué qu'une amélioration des performances d'un détecteur opérant dans le plan temps-fréquence peut être observée lorsque les représentations sont préalablement filtrées, malgré la perte d'information statistique qui en résulte. Évidemment, une dégradation minimum de l'information discriminante au cours de l'opération conditionne le succès de celle-ci. On se propose à présent d'élaborer un processus de diffusion respectant cette contrainte, pour un contrôle approprié de la dimension de l'espace transformé. Les variantes offertes par les opérateurs de diffusion étant multiples [Der96, Wei97], le cadre de cette étude est restreint au schéma (4.34) de Perona et Malik.

Il existe de nombreuses techniques pour extraire l'information discriminante d'un ensemble d'apprentissage [Fuk90]. Compte tenu de sa simplicité de mise en œuvre, l'analyse factorielle discriminante est couramment retenue pour s'acquitter de cette tâche [Sap90]. Dans un problème à c classes ω_i, cette approche consiste à rechercher une transformation linéaire maximisant le

critère $J = \mathrm{trace}(\mathbf{\Sigma}_{\mathrm{intra}}^{-1}\,\mathbf{\Sigma}_{\mathrm{inter}})$, où $\mathbf{\Sigma}_{\mathrm{inter}}$ et $\mathbf{\Sigma}_{\mathrm{intra}}$ désignent les matrices de covariance inter-classe et intra-classe définies par

$$\begin{aligned}\mathbf{\Sigma}_{\mathrm{intra}} &= \sum_{i=1}^{c} p(\omega_i)\,\mathrm{E}\{(\boldsymbol{X}-\boldsymbol{m}_i)\,(\boldsymbol{X}-\boldsymbol{m}_i)^t|\omega_i\}\\ \mathbf{\Sigma}_{\mathrm{inter}} &= \sum_{i=1}^{c} p(\omega_i)\,(\boldsymbol{m}_i-\boldsymbol{m})\,(\boldsymbol{m}_i-\boldsymbol{m})^t,\end{aligned}$$

avec $\boldsymbol{m} = \mathrm{E}\{\boldsymbol{X}\}$ et $\boldsymbol{m}_i = \mathrm{E}\{\boldsymbol{X}|\omega_i\}$. Le lecteur intéressé par de plus amples précisions sur la maximisation de J est invité à consulter [Fuk90]. On retiendra que dans le cas de 2 classes, ce critère est équivalent à celui de Fisher (2.14), signifiant que la transformation recherchée consiste en une projection des données sur l'axe défini par le vecteur $\boldsymbol{w}_{\mathrm{Fisher}} = \mathbf{\Sigma}_{\mathrm{intra}}^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_0)$.

Afin de préserver l'information discriminante durant un processus de diffusion, implicitement défini sous une forme discrète compte tenu de la nature des données traitées, on suggère de recourir à une représentation temps-fréquence de $\boldsymbol{w}_{\mathrm{Fisher}}$, à l'image du spectrogramme utilisé précédemment pour inhiber la procédure en lieu et place des composantes du signal. On recommande en particulier une représentation de Wigner de celui-ci pour les propriétés intéressantes de localisation auxquelles elle satisfait. Il est à noter que $|W_{w_{\mathrm{Fisher}}}[t,f]|$ est d'autant plus grand que l'information en (t,f) est discriminante et ne doit subir d'altération. Aussi on suggère d'adopter la fonction de conductance suivante

$$c[t,f] = \left(1+\left(\frac{|W^{(\cdot)}_{w_{\mathrm{Fisher}}}[t,f]|}{\delta}\right)^{\alpha}\right)^{-1}, \tag{4.36}$$

avec $\alpha \geq 0$ et $\delta > 0$. Afin d'illustrer cette approche, dont la figure 4.6 reprend la chronologie, on considère le problème de la détection d'un signal $s[t\,;\phi_0)$ constitué de 16 échantillons, où ϕ_0 représente une phase aléatoire uniformément distribuée sur l'intervalle $[-\pi,\pi[$. Celui-ci est noyé dans un bruit blanc gaussien, le rapport signal-sur-bruit étant fixé à -1 dB. Pour apporter une solution à ce problème, on dispose d'une base d'apprentissage constituée de 100 individus équitablement répartis entre les classes ω_0 et ω_1. La distribution de Wigner classique a été arbitrairement choisie pour représenter ceux-ci dans le plan temps-fréquence, et élaborer la fonction de conductance (4.36). La figure 4.7.(a) montre la décroissance de la dimension de l'espace transformé au cours du processus de diffusion, tandis que la figure 4.7.(b) reproduit dans le même temps l'évolution du taux d'erreur d'un détecteur linéaire y opérant. Il convient de noter que cette dernière laisse clairement apparaître les phénomènes de sur-apprentissage, puis de sous-apprentissage. On observe qu'un taux d'erreur minimal de 24% est atteint au terme de 350 itérations du processus de diffusion, la dimension de l'espace transformé étant alors 29. L'approche présentée permet ainsi d'améliorer de façon notoire les performances du système considéré, le taux d'erreur de celui-ci étant initialement de 44%. Enfin, on relève le caractère régularisant du processus de diffusion, celui-ci entraînant une réduction du taux d'erreur de 12% durant les 100 premières itérations sans modifier la dimension de l'espace transformé. La figure 4.8 reprend ces résultats et les compare aux performances du filtre adapté temps-fréquence.

Par ces premiers résultats encourageants, les travaux amont qui ont été entrepris appellent à une recherche plus approfondie de nouvelles méthodes d'analyse des signaux non-stationnaires au regard des possibilités offertes par les équations aux dérivées partielles. Aussi envisage-t-on d'étudier les classes de distributions obtenues par application d'opérateurs de diffusion sur des

1. Effectuer une analyse factorielle discriminante de la base d'apprentissage $\mathcal{A}_n$.
2. Évaluer la fonction de conductance (4.36).
3. Itérer la procédure de diffusion pour chaque élément de $\mathcal{A}_n$.
4. Élaborer un détecteur et retourner en 3.
5. En fin de diffusion, sélectionner le détecteur le plus performant.

FIG. 4.6 : Algorithme pour le contrôle de complexité par diffusion des représentations.

éléments des groupes de Cohen et Affine, en s'interrogeant sur les propriétés et les bénéfices qu'on peut en attendre. Pour ce faire, on prévoie d'aborder l'étude de différents modèles de diffusion existants (isotrope, anisotrope, homogène, non-homogène) afin de compléter les travaux déjà menés sur la diffusion de Perona et Malik dans le cadre de l'analyse temps-fréquence. La caractérisation de fonctions de conductance adaptées à des problématiques particulières telles que l'amélioration de la lisibilité, le débruitage ou encore la classification constitue également un axe de recherche possible. Enfin, l'étude des propriétés des représentations obtenues, comme par exemple celles de covariance en translation et dilatation, apparaît comme une direction de travail intéressante. Il est à noter que le thème des représentations diffusives a fait l'objet d'un projet « jeunes chercheurs » du GdR-PRC ISIS en 1999-2001, auquel j'ai activement participé.

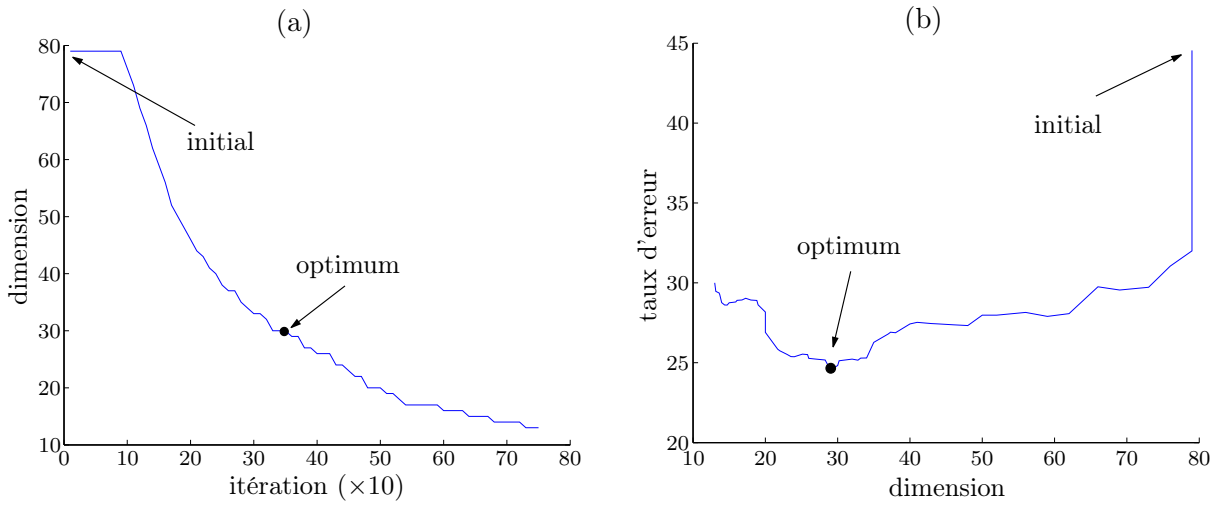

FIG. 4.7 : Contrôle de la capacité d'apprentissage d'un détecteur linéaire opérant sur la distribution de Wigner classique, par diffusion des représentations. (a) Evolution de la dimension de l'espace transformé et (b) des performances de la structure de détection au cours du processus de diffusion.

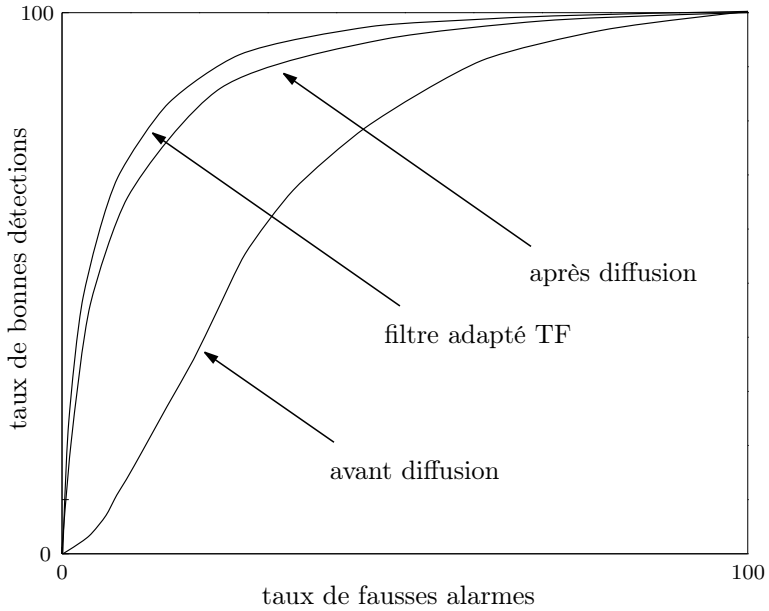

FIG. 4.8 : Courbes COR de la structure de détection avant et après diffusion. Comparaison avec les performances du détecteur de Bayes.

Chapitre 5

Applications aux signaux de sommeil

5.1 Introduction

L'étude du sommeil a débuté dans les années 30 avec l'établissement d'une première classification en stades de l'activité cérébrale humaine à partir de l'électroencéphalogramme [Loo37]. A l'heure actuelle, les efforts se concentrent d'avantage sur l'analyse des phénomènes transitoires, l'un des principaux objectifs étant la compréhension des mécanismes générateurs du sommeil et de l'activité cérébrale nocturne. Dans ce domaine de recherche, l'utilisation de moyens automatiques de traitement de l'information demeure un facteur de progrès important en raison du volume considérable de données à analyser [Gai73, Ray86, Sch93]. En effet, l'expertise des signaux de sommeil, extrêmement coûteuse et astreignante lorsqu'elle doit être pratiquée visuellement, ne peut être raisonnablement envisagée à grande échelle, à moins d'être automatisée.

Reposant sur l'approche proposée tout au long de ce mémoire, le présent chapitre a comme premier objectif d'élaborer un outil pour la détection d'un signal transitoire particulier de l'électroencéphalogramme de sommeil : le *complexe K*. Le choix de ce problème se justifie par une assez grande difficulté à le résoudre en raison d'une ressemblance marquée de l'événement considéré avec d'autres phénomènes observés en sommeil profond [Ric98(a)], en particulier les *bouffées d'ondes delta*. La seconde application abordée vise à ouvrir de nouvelles perspectives de recherche quant à la méthodologie proposée, en montrant que la constitution d'un ensemble d'apprentissage peut parfois être sujette à discussion. Le problème traité dans ce but concerne la détection d'un autre phénomène issu de l'activité cérébrale nocturne, la *phase d'activation transitoire* [Via02(a)]. Par sa complexité et les nombreux paramètres physiologiques qu'elle intègre, la définition de cet événement est sujette à l'interprétation de chaque cotateur [Asd92]. Il en résulte d'inévitables désaccords lors de la constitution de l'ensemble d'apprentissage. Aussi décrit-on ici une approche pour la fusion d'avis en vue de l'obtention d'un étiquetage unique des données, permettant de surcroît une caractérisation du comportement de chaque cotateur par rapport aux autres membres du groupe. Mais avant, il convient de procéder à une description du contexte dans lequel s'inscrivent les applications traitées, en particulier en présentant succinctement la macrostructure du sommeil et les principaux événements que l'on peut y rencontrer.

Les résultats présentés dans cet ultime chapitre sont le fruit de travaux d'abord menés avec R. Lengellé [Ric98(a)], puis avec un doctorant, G. Viardot [Via02(a)]. Ils ont fait l'objet de plusieurs publications dans des revues, par exemple [Ric98(b), Ric99(a)], et de communications dans des conférences nationales et internationales telles que [Via00, Via01, Via02(b)]. Ils sont l'aboutissement d'une longue collaboration avec la Fondation pour la Recherche en Neuro-sciences

Appliquées à la Psychiatrie, qui s'est vue dotée en 1999 d'une subvention de 700kF dans le cadre de l'Action Concertée Incitative « Télémédecine et technologies pour la santé ». Le sujet de ce projet était la détection des phases d'activation transitoire durant le sommeil.

5.2 Macro et microstructures du sommeil

Depuis l'avènement des techniques modernes d'enregistrement du sommeil, l'étude de ce dernier repose sur une analyse préalable de sa macrostructure. L'unité temporelle de cette opération est la *page*, dont la durée de vingt ou trente secondes constitue un héritage des historiques enregistrements sur papier. Chacune de ces pages se voit ainsi attribuée un stade de sommeil parmi cinq possibles, selon une règle standardisée dans [Rec68]. Pour accomplir cette tâche, trois types d'enregistrements sont nécessaires.

L'électroencéphalogramme (EEG). Il reflète l'activité du système nerveux central par l'intermédiaire des variations de potentiel électrique engendrées au niveau du scalp, mesurées en des points dont la position est standardisée [Bro95].

L'électromyogramme (EMG). Il caractérise le tonus musculaire général du patient par l'intermédiaire des différences de potentiel générées par l'activité musculaire au niveau du menton.

L'électrooculogramme (EOG). Il se rapporte aux mouvements horizontaux et verticaux des globes oculaires saisis par des électrodes collées autour des yeux.

Ces informations sont généralement complétées par l'acquisition simultanée de données cardiovasculaires, respiratoires et musculaires périphériques.

L'électroencéphalogramme joue un rôle central dans la cotation des signaux de sommeil en raison de la diversité des événements l'émaillant tout au long de la nuit. On y recense en particulier les bouffées d'ondes delta, thêta, alpha ou bêta, que l'on identifie par leur support fréquentiel [Cim97]. Certaines interviennent comme suit dans la classification du sommeil en cinq stades.

Stade 1. Fragmentation des ondes alpha et survenue des ondes thêta.

Stade 2. Apparition d'éléments transitoires tels que les fuseaux de sommeil et les complexes K.

Stade 3. Présence sur 20% à 50% du tracé d'ondes lentes de grande amplitude, dites delta.

Stade 4. Présence de bouffées d'ondes delta sur plus de 50% du tracé.

Sommeil paradoxal. Atonie musculaire et mouvements oculaires rapides.

La figure 5.1 illustre cette nomenclature. Conventionnellement, le *sommeil léger* regroupe les stades 1 et 2, tandis que le *sommeil profond* est constitué des stades 3 et 4. Si cette classification fournit des informations essentielles pour l'identification de certaines anomalies du sommeil, elle conduit néanmoins à un diagnostic incomplet. Elle ne tient en effet pas explicitement compte, par exemple, des fréquences d'apparition des phénomènes transitoires. Pourtant, ceux-ci constituent des marqueurs importants du fonctionnement nocturne des systèmes de régulation de l'homme, de la qualité du sommeil et de pathologies non nécessairement liées à l'appareil nerveux central.

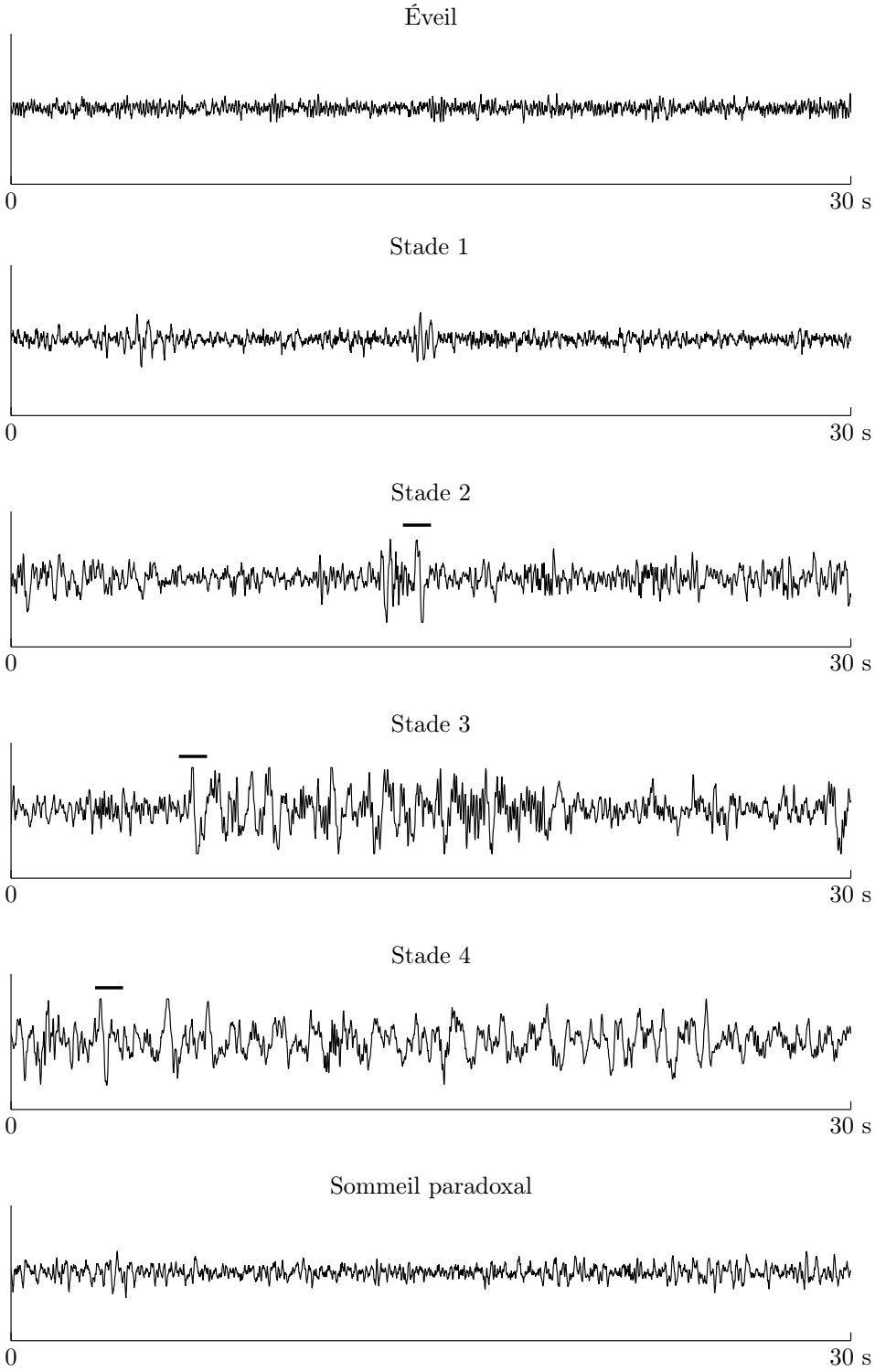

FIG. 5.1 : Illustration des cinq stades de sommeil au moyen de séquences de 30 secondes d'électroencéphalogramme. Les segments horizontaux figurant sur les tracés 2, 3 et 4 indiquent la présence d'un événement transitoire particulier : le complexe K.

Pour affiner la compréhension du sommeil, il est indispensable de compléter sa segmentation en stades par une étude de sa microstructure. Celle-ci est constituée d'événements qualifiés de *phasiques* ou *transitoires* dans la littérature, voire de *grapho-éléments*. Ci-dessous figure une description succincte des plus fréquemment rencontrés dans l'électroencéphalogramme de sommeil.

Les pointes vertex

Comme l'illustre la figure 5.2, les pointes vertex sont des impulsions négatives survenant en stade 1, durant la phase d'endormissement, et dont l'amplitude augmente avec l'approfondissement du sommeil. Elles sont considérées comme précurseurs des complexes K, et ne concernent que la zone vertex. Ces phénomènes phasiques peuvent apparaître en réponse à des stimuli externes, ou se manifester spontanément.

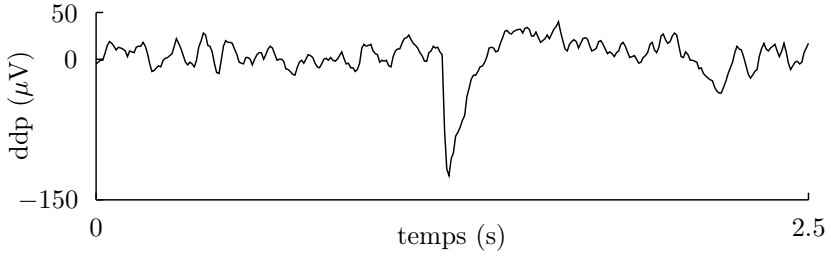

FIG. 5.2 : Exemple de pointe vertex.

Les fuseaux de sommeil

Ces événements constituent l'un des critères de définition du stade 2. Ils apparaissent sur l'électroencéphalogramme sous la forme de signaux transitoires quasi-sinusoïdaux, de fréquence comprise entre 12 et 14 Hz, et de durée variant entre 0.5 et 1 seconde chez l'adulte. Un exemple est présenté en figure 5.3. On les associe volontiers à un mécanisme neurologique protégeant l'organisme des perturbations externes du sommeil, bien que cette fonction ne fasse pas toujours l'unanimité.

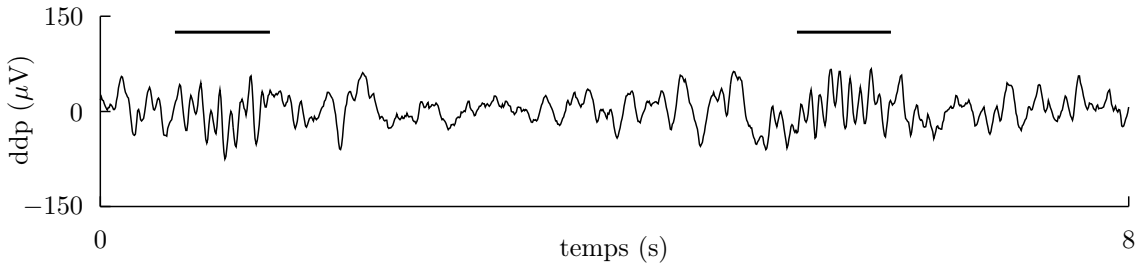

FIG. 5.3 : Exemples de fuseaux de sommeil.

Les complexes K

En apparaissant dès le stade 2, le complexe K constitue l'un des principaux marqueurs du début du sommeil avec les fuseaux. Il est généralement d'aspect biphasique, comprenant une onde positive rapide suivie d'une onde négative de grande amplitude, comme on peut le constater sur la figure 5.4. Il dure environ 1 seconde, présente un support fréquentiel compris entre 1 et 4 Hz, et se distingue aisément de l'électroencéphalogramme de fond en stade 2 par son amplitude, comme l'indique la figure 5.1. Cependant, en raison d'une similitude marquée avec d'autres phénomènes non-stationnaires observés en sommeil profond, tels que les bouffées d'ondes delta présentées en figures 5.1 et 5.4, il reste très difficile à isoler en stades 3 et 4. Ceci justifie l'intérêt de méthodes statistiques pour tenter d'apporter une réponse satisfaisante au problème de sa détection.

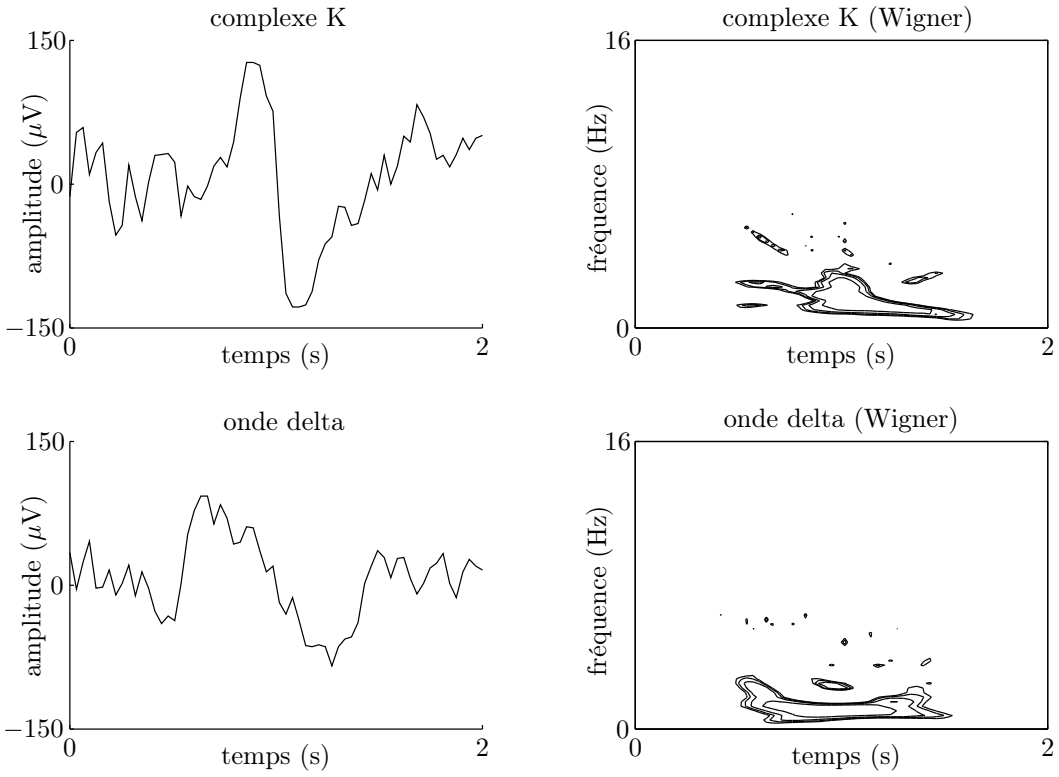

FIG. 5.4 : Complexe K et onde delta représentés dans les domaines temporel et temps-fréquence.

Les phases d'activation transitoire

De tels événements impliquent l'activation de l'ensemble de l'organisme. Comme l'illustre la figure 5.5, les phases d'activation transitoire se traduisent essentiellement par une augmentation de la fréquence des ondes de l'électroencéphalogramme qui s'apparentent alors à des rythmes d'éveil, le renforcement du tonus musculaire, l'apparition de mouvements corporels, ou encore l'accélération du rythme cardiaque. Elles sont souvent décrites comme un passage brutal d'un stade de sommeil profond vers un stade plus léger, ou du sommeil paradoxal vers l'éveil. Si elles surviennent avec une fréquence accrue en sommeil paradoxal, elles peuvent toutefois être observées au cours de tous les stades.

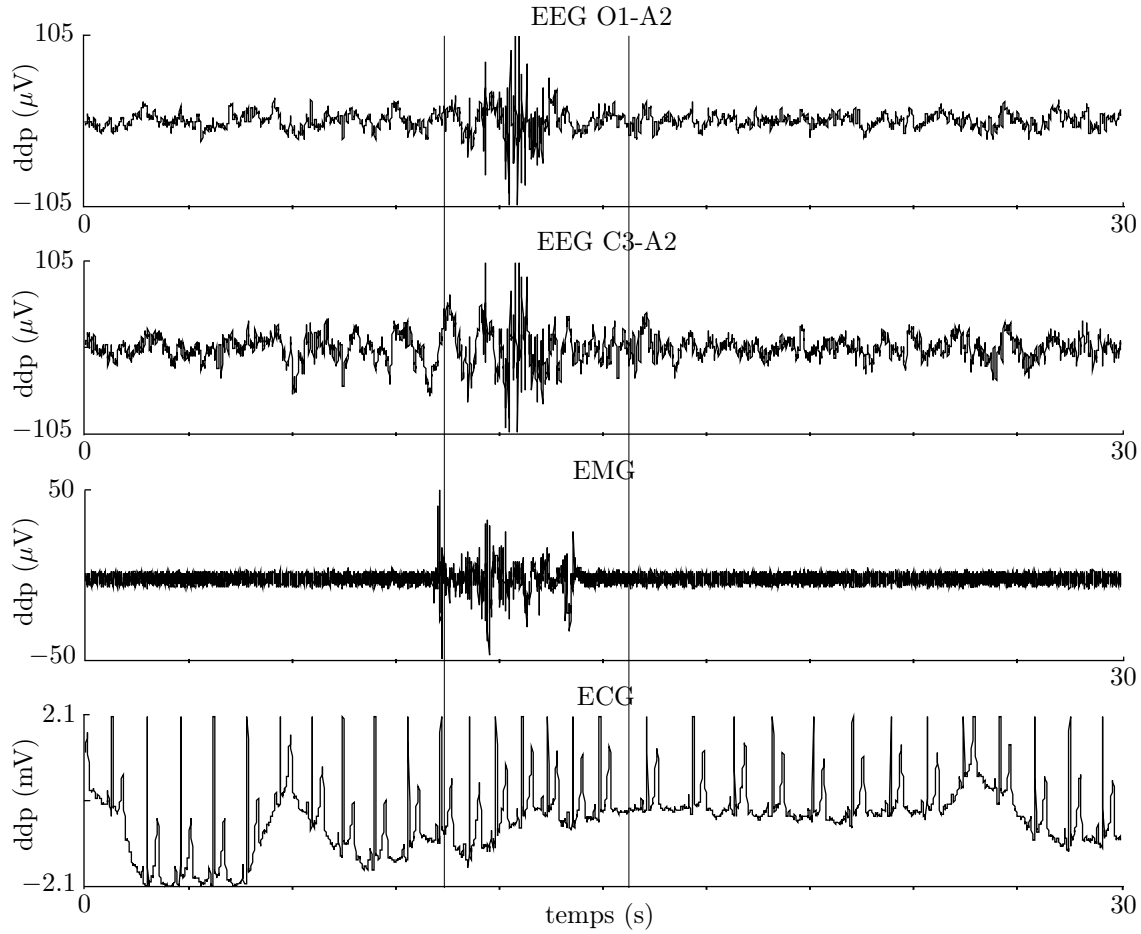

FIG. 5.5 : Exemple académique de phase d'activation transitoire visible sur deux électroencéphalogrammes mesurés en des points différents, un électromyogramme et un électrocardiogramme.

5.3 Détection automatique de complexes K

Si le rôle physiologique du complexe K demeure aujourd'hui aussi abstrait que sa dénomination, une majorité d'électrophysiologistes s'accorde à considérer ce grapho-élément comme une réponse à des stimuli d'origine externe ou interne. Aussi le complexe K jouerait-il pour certains le rôle de mécanisme protecteur du sommeil, le dormeur reprenant contact avec son environnement pour une courte durée durant laquelle ses capacités de traitement de l'information se trouvent améliorées, sans que cela nuise pour autant à la qualité de son repos [Hal93, Wau95]. Pour d'autres, il constitue au contraire un phénomène en étroite relation avec les mécanismes d'activation, voire d'éveil [Bas92]. Afin de lever le voile sur le sens de ce phénomène phasique et d'acquérir une meilleure compréhension de son rôle éventuel dans certains troubles du sommeil, de nombreux systèmes de détection automatique ont été proposés dans la littérature. Certains reposent sur des techniques de filtrage [She77, Dar91]. D'autres consistent en une extraction préalable de paramètres du signal électroencéphalographique [Bre70], ou d'une représentation temps-fréquence de celui-ci [Cim97], la prise de décision pouvant incomber à un réseau de neurones par exemple [Ban92, Des94]. L'objectif de cette section est d'élaborer un détecteur de complexes K grâce à la méthode du critère optimal à noyau reproduisant, et d'étendre

par là-même de précédents travaux sur la détection de ces événements dans le plan temps-fréquence [Ric98(a), Ric98(b)] à des espaces transformés plus vastes.

L'acquisition des électroencéphalogrammes utilisés dans le cadre de cette étude a été effectuée sur la dérivation dite Cz, à une fréquence d'échantillonnage de 128 Hz. Un ensemble de complexes K a d'abord été constitué par un groupe de cinq experts. Ces électrophysiologistes confirmés ont travaillé individuellement sur trois enregistrements de huit heures environ, effectuant visuellement la cotation des signaux proposés. Du fait d'une certaine disparité dans les résultats qu'ils ont obtenus, on n'a retenu pour cette étude que les événements sélectionnés par *au moins deux* d'entre eux afin de constituer un ensemble de référence pour le complexe K. Puis une base de phénomènes présentant des similitudes marquées avec ce dernier a été créée, en majorité avec des bouffées d'ondes delta telles que celle présentée en figure 5.4. Tous ces événements transitoires ont alors été synchronisés avant d'être segmentés sur des durées de 2 secondes. Finalement, les données ont été décimées d'un facteur 4, une fréquence d'échantillonnage de 32 Hz étant compatible avec les supports fréquentiels du complexe K, des ondes rapides qui peuvent se superposer à lui, ainsi que des bouffées d'ondes delta. De cette façon, une base de données composée de 1215 complexes K et de 1196 événements de nature différente a été constituée en vue de l'élaboration d'une règle de décision de la forme

$$d(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \sum_{k=1}^{n} a_k \, \kappa(\boldsymbol{x}_k, \boldsymbol{x}) - \lambda_0 > 0 \\ 0 & \text{sinon,} \end{cases} \tag{5.1}$$

où chaque $\boldsymbol{x}_k$ désigne un élément de l'ensemble d'apprentissage et κ un noyau. Il est à noter que la restriction de cette base aux 609 complexes K sélectionnés par *au moins trois* experts parmi les cinq consultés a déjà été utilisée dans une précédente étude [Ric98(a), Coa01]. Le tableau 5.1 rappelle les excellentes performances qui avaient alors été obtenues grâce à une structure quadratique élaborée au moyen de la méthode du critère optimal. On peut remarquer que celle-ci compte parmi les solutions les plus performantes décrites dans la littérature, même s'il convient de noter que toutes n'ont pas été testées dans les mêmes conditions. Dans le cadre de ce document, on a volontairement opté pour un problème plus difficile que celui figurant dans [Ric98(a)], en retenant également des complexes K n'ayant été reconnus comme tel que par deux experts. En effet, il s'agit ici d'illustrer au mieux la diversité des performances obtenues selon le choix du noyau κ, et de l'usage ou non d'une procédure de contrôle de complexité.

Au cours de 30 phases successives d'apprentissage et de test, 400 complexes K et 400 ondes delta ont été prélevés aléatoirement dans la base d'apprentissage afin d'optimiser la structure (5.1) à l'aide de la méthode du critère optimal à noyau reproduisant, les données restantes étant dédiées à l'estimation des performances. Dans ce contexte, différents noyaux reproduisants ont été

Travaux	Approche	{bonnes détections ; fausses alarmes}
Da Rosa [Dar91]	modélisation	{89% ; 49%}
Destiné [Des94]	réseau de neurones	{64% ; 4%}
Bankman [Ban92]	réseau de neurones	{90% ; 8%}
Cimetière [Cim97]	temps-fréquence et kPPV	{85% ; 15%}
Richard [Ric98(a)]	structure imposée	{90% ; 4%}

TAB. 5.1 : Performances de quelques détecteurs de complexes K existants.

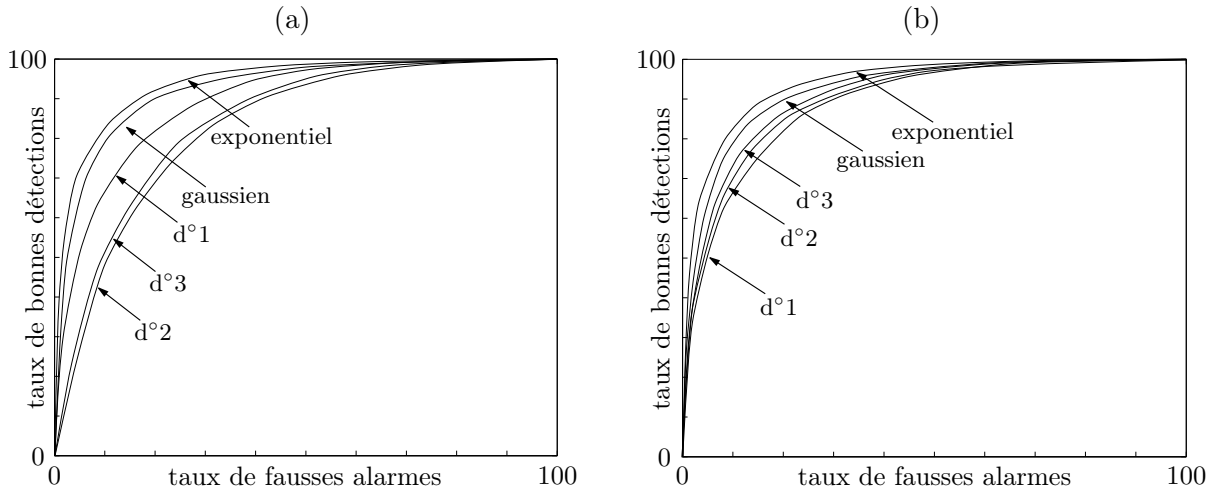

FIG. 5.6 : Performances des détecteurs de complexes K obtenus par la méthode du critère optimal à noyau reproduisant, (a) sans et (b) avec contrôle de complexité.

noyau	poly. (d°1)	poly. (d°2)	poly. (d°3)	gaussien	exponentiel
taux d'erreur (sans cont.)	18.80	23.88	22.88	14.14	13.28
taux d'erreur (avec cont.)	18.66	17.37	16.82	13.98	13.17
nb de composantes retenues	55	202	229	7	5

TAB. 5.2 : Taux d'erreur atteints suivant le choix du noyau.

sélectionnés, en particulier les noyaux polynômiaux de degré 1, 2 et 3, ainsi que les noyaux exponentiel et gaussien définis en section 3.3.2. Le tableau 5.2 synthétise les résultats obtenus suivant l'usage ou non de la procédure variationnelle dédiée au contrôle de complexité. La figure 5.6 présente quant à elle les courbes COR associées. Force est de constater que les détecteurs considérés subissent les manifestations néfastes de la malédiction de la dimensionnalité, que l'on parvient à contrer grâce à la procédure prévue à cet effet. On note par ailleurs le faible nombre de composantes sélectionnées par cette dernière lorsqu'elle est appliquée aux détecteurs à noyaux radiaux. Il n'est donc pas étonnant que ces structures affichent finalement les meilleures performances en généralisation.

5.4 Vers une détection des phases d'activation transitoire

Les phases d'activation transitoire ont pour la première fois été étudiées chez l'homme dans [Shi71]. Il s'agit d'activations corticales, spontanées ou provoquées, présentant des similitudes marquées avec les réactions d'éveil observées lorsque des stimuli extérieurs viennent troubler le sommeil. Les règles de cotation permettant de les isoler dans les signaux polysomnographiques sont particulièrement complexes, comme l'attestent les critères fixés par l'American Sleep Disorder Association il y a dix ans. En voici quelques éléments, choisis parmi les deux pages de définitions que comporte la référence [Asd92] :

– une phase d'activation transitoire est associée à une modification du contenu fréquentiel de l'EEG durant au moins trois secondes. Ce changement doit se produire sur la bande fréquentielle $[4, 13[$ Hz, ou plus rarement au delà de 16 Hz ;

- un renforcement du tonus musculaire de base accompagne nécessairement une phase d'activation transitoire. Il ne s'agit cependant pas d'une condition suffisante ;
- une phase d'activation transitoire ne peut être cotée que si un stade de sommeil stable est installé depuis au moins 10 secondes. Deux événements de cette nature ne peuvent donc se produire à moins de 10 secondes d'intervalle ;
- la cotation des phases d'activation transitoire est indépendante du stade de sommeil attribué à la page ;
- une phase d'activation transitoire ne peut être cotée en sommeil paradoxal que si il y a renforcement du tonus musculaire ;
- un changement de stade s'accompagne d'une modification du spectre de l'EEG qui ne doit pas être confondue avec une phase d'activation transitoire. Cette dernière peut cependant induire un changement de stade.

De nouvelles règles sont venues depuis s'ajouter tacitement à cette définition, par exemple une possible accélération du rythme cardiaque telle qu'on peut l'observer sur la figure 5.5 [Sta00]. D'autres règles ont en revanche été assouplies, certaines relatives à la chronologie et à la durée des événements composant les phases d'activation transitoire en particulier [Nic95]. De nombreuses études montrent enfin la grande variabilité des indicateurs considérés suivant l'âge, le sexe, le capital génétique, les pathologies et l'environnement du dormeur [Nic95, Eat99, Str99]. Aussi résulte-t-il de cette multiplicité de recommandations ni nécessaires, ni suffisantes, un travail de cotation laissant une part importante à l'interprétation de chacun, et entraînant de nombreuses divergences entre experts [Via02(a)]. Il convient de remarquer que de telles difficultés sont récurrentes dans de nombreux domaines d'application n'ayant pas trait à l'environnement médical, par exemple celui de la caractérisation du confort automobile [Len00]. En présence de telles incertitudes, force est de reconnaître que l'élaboration d'une règle de décision performante s'avère délicate et invite à la prudence lors de la constitution de l'ensemble d'apprentissage. Si l'on peut parfois aboutir à des résultats satisfaisants en sélectionnant les événements retenus par une majorité d'experts, comme l'atteste le problème de la détection du complexe K décrit précédemment, ce principe de vote ne donne pas toujours entière satisfaction [Lam97]. Co-encadré par R. Lengellé et moi-même, G. Viardot s'est précisément intéressé à ce problème de reconnaissance des formes en présence d'incertitude sur l'expertise dans le cadre d'une thèse de doctorat, qu'il a soutenue le 22 octobre 2002 à l'Université de Technologie de Troyes. Son travail a consisté en l'élaboration d'une méthode pour la fusion d'avis afin d'obtenir un étiquetage unique des données d'apprentissage, permettant de surcroît une caractérisation du comportement de chaque cotateur par rapport aux autres membres du groupe. Cette section en expose les principaux résultats et présente de nouvelles perspectives de recherche dans cette voie.

5.4.1 Incertitudes sur l'expertise et fusion d'avis divergents

Selon [Pal94], on compte trois sources d'incertitude responsables du désaccord entre experts : *l'incertitude probabiliste*, *l'incertitude résolutionnelle* et *l'incertitude floue*. La première est liée au caractère aléatoire des observations. Elle est souvent associée au bruit de mesure ou à des fluctuations aléatoires du système observé. Aussi convient-il de fournir un nombre important d'événements à chaque expert afin de limiter les effets néfastes de celle-ci. La complexité de la tâche du cotateur est également conditionnée par l'accès à l'information pertinente qui lui est offert. Si un nombre trop réduit de paramètres disponibles constitue un obstacle certain

à une expertise correcte, une profusion d'indicateurs peut également s'avérer problématique, en particulier lorsque ceux-ci sont contradictoires vis-à-vis des événements à caractériser. De ces deux extrêmes peut ainsi naître une incertitude résolutionnelle importante. L'incertitude floue est quant à elle liée à la façon dont l'expertise est recueillie, puis codée. Un codage riche autorise une description fine de chaque événement mais complique la synthèse des expertises. Un choix restreint à deux alternatives induit en revanche une perte d'information importante et, par manque de nuances, favorise les désaccords entre cotateurs.

L'un des objectifs de la fusion d'expertises est l'obtention d'un avis unique à partir des opinions émises par plusieurs experts. Diverses méthodes peuvent répondre à cette attente, le choix en faveur de l'une d'elles dépendant de la technique adoptée pour modéliser le savoir des experts. Par exemple, celui-ci peut être représenté par un réseau bayésien [Zig00], un graphe sémantique [Ste95], un groupe de prédicats [Lev00] ou encore, de façon plus directe, instancié par un ensemble de données étiquetées. Conformément aux hypothèses de travail en vigueur dans ce manuscrit, la présente section considère exclusivement cette dernière forme d'expression d'avis dans le cadre d'un problème de détection. On note $\{\delta_1, \ldots, \delta_e\}$ l'ensemble des décisions fournies par e experts quant au degré d'appartenance d'une observation à l'une des deux classes en compétition, avec $\delta_i \in [0, 1]$. La sélection d'une règle de fusion $f(\cdot)$ repose sur le type de comportement que l'on souhaite reproduire, après avoir pris connaissance de celui des experts. Ainsi, $f(\cdot)$ est dit *à comportement constant indépendant du contexte* s'il ne dépend que des avis δ_i et vérifie l'une des trois inégalités suivantes, pour tout e-uplet $\{\delta_1, \ldots, \delta_e\}$ [Blo96] :

1. *Opérateur conjonctif* ou *t-norme* :

$$f(\delta_1, \ldots, \delta_e) \leq \min(\delta_1, \ldots, \delta_e) \tag{5.2}$$

   exemple : $f(\delta_1, \delta_2) = \delta_1 \cdot \delta_2$

2. *Opérateur disjonctif* ou *t-conorme* :

$$f(\delta_1, \ldots, \delta_e) \geq \max(\delta_1, \ldots, \delta_e) \tag{5.3}$$

   exemple : $f(\delta_1, \delta_2) = \delta_1 + \delta_2 - \delta_1 \cdot \delta_2$

3. *Opérateur de compromis* :

$$\min(\delta_1, \ldots, \delta_e) < f(\delta_1, \ldots, \delta_e) < \max(\delta_1, \ldots, \delta_e) \tag{5.4}$$

   exemple : $f(\delta_1, \delta_2) = w\delta_1 + (1 - w)\delta_2$ avec $w \in [0, 1]$.

On peut remarquer que les opérateurs (5.2) et (5.3) ont un caractère résolument pessimiste et optimiste, respectivement, ce qui ne donne que plus de sens à la notion de compromis que véhicule l'opérateur (5.4). Le choix en faveur de l'un de ces comportements n'est donc pas neutre. Il doit prendre en compte la qualité des sources d'information que constituent les experts, ainsi que le niveau de discordance de leurs avis. Lorsque certains experts ont des jugements sûrs, il convient d'adopter un opérateur reproduisant ceux-ci. Les t-conormes s'avèrent alors indiquées. Lorsque les sources d'information sont en revanche peu fiables, il est prudent de tempérer les jugements en choisissant une t-norme. Plus nuancé, un opérateur de compromis peut constituer une alternative intéressante quand la fiabilité des sources n'est pas connue *a priori*. Enfin, cette présentation serait incomplète si l'on n'évoquait pas les opérateurs de fusion indépendants du

contexte dont le comportement conjonctif, disjonctif ou consensuel dépend des valeurs prises par $\{\delta_1, \ldots, \delta_e\}$. Ils sont dits *à comportement variable*. Lorsque $f(\cdot)$ ne dépend pas uniquement des avis formulés par les experts, mais repose également sur des informations additionnelles telles que l'observation, l'opérateur considéré est dit *dépendant du contexte*. Il est à noter que cette propriété peut permettre, par exemple, de pondérer les jugements de chaque expert selon les régions de l'espace des observations. Évidemment, ces informations ne sont généralement pas disponibles *a priori*. Elles sont toutefois véhiculées par les données expertisées et peuvent, dans une certaine mesure, être extraites de celles-ci.

5.4.2 Une méthode pour la fusion d'expertises

Lorsqu'une expertise est convenablement menée, il est admis que les décisions proposées par le cotateur sont stochastiquement liées aux observations. Par analogie, une règle de fusion doit veiller à entretenir ou rétablir une telle relation, malgré les divergences d'opinions. Dans le cadre de sa thèse de doctorat, G. Viardot a proposé une méthodologie reposant sur ce principe pour l'élaboration d'opérateurs de fusion. Non sans rappeler l'apprentissage de détecteurs à structure imposée, elle consiste à rechercher au sein d'une famille $\mathcal{F} = \{f(\delta_1(\boldsymbol{x}), \ldots, \delta_e(\boldsymbol{x}), \theta) \,:\, \theta \in \Theta\}$ donnée, une solution $f^*(\cdot)$ maximisant une mesure de dépendance entre deux variables aléatoires. Aussi cette approche suscite-t-elle les deux interrogations suivantes

- Quelle règle de fusion choisir ?
- Quelle mesure de dépendance stochastique adopter ?

Après une étude de l'existant sur les principales familles d'opérateurs de fusion indépendants ou dépendants du contexte, à comportement constant ou variable, on a été amené à retenir les structures linéaires définies par

$$f(\delta_1(\boldsymbol{x}), \ldots, \delta_e(\boldsymbol{x}), \theta) = \sum_{i=1}^{e} \theta_i \, \delta_i(\boldsymbol{x}), \tag{5.5}$$

sous la contrainte de normalisation $\sum_{i=1}^{e} \theta_i = 1$. L'exploration d'une telle classe correspond en effet à la recherche d'une solution consensuelle [Blo96], propriété légitiment souhaitable lorsqu'on ne dispose pas d'informations *a priori* quant à la qualité des expertises. De plus, les coefficients de pondération θ_i sont par construction liés à la fiabilité de chaque expert, conférant au résultat une interprétabilité accrue échappant à des solutions plus complexes [Len00]. Ceux-ci sont généralement estimés à partir d'un jeu de données ré-étiquetées lors d'une réunion de consensus. La pertinence de chaque expert peut alors, par exemple, être quantifiée indépendamment de celle des autres en calculant un taux d'accord entre sa cotation et le classement de consensus, ou encore faire l'objet d'une procédure d'apprentissage supervisé [Now91, Vog99, Moe00]. Par son caractère non-supervisé, l'approche proposée ne nécessite pas de recourir à un avis de référence, ce qui la rend particulièrement attractive. En effet, la règle (5.5) est directement élaborée par maximisation d'une mesure de dépendance entre le résultat de la fusion des avis et les observations, en l'occurrence *l'information mutuelle* dans le cadre de ce travail [Cov91]. Issue de la théorie de l'information, celle-ci a été retenue parce qu'elle constitue une mesure de la quantité d'information partagée par deux variables aléatoires ne préjugeant pas du type de relation les unissant. Ceci lui procure un avantage certain sur de nombreux critères, par exemple les mesures de corrélation. Ainsi, la règle de fusion recherchée est solution du problème suivant [Via00]

$$f^* = \arg\max_{f \in \mathcal{F}} I(f(\delta_1(\boldsymbol{X}), \ldots, \delta_e(\boldsymbol{X})), \boldsymbol{X}), \tag{5.6}$$

où $I(\cdot\,,\cdot)$ désigne l'information mutuelle, dont on rappelle qu'elle est définie par

$$I(X,Y) = \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)}. \tag{5.7}$$

Dans le cadre de l'application qui suit, il a été fait appel à la méthode de Parzen pour l'estimation de l'information mutuelle et à un algorithme génétique pour la résolution du problème (5.6), compte tenu du caractère non-convexe du critère considéré par rapport aux paramètres θ_i.

Dans le but de tester la pertinence de la méthode de fusion proposée, douze enregistrements de nuit complète ont été confiés de façon indépendante à quatre électrophysiologistes afin qu'ils identifient les phases d'activation transitoire. Pour ce faire, ceux-ci disposaient d'un électroencéphalogramme central (C3-A2) et occipital (O1-A2), d'un électromyogramme mesuré au menton ainsi que d'un double électrooculogramme constitué des dérivations horizontale et verticale. Compte tenu des nombreux désaccords dans ce type d'exercice [Via02(a)], il a été convenu que les cotateurs pourraient exprimer un avis gradué quant à chaque événement afin de limiter l'incertitude floue évoquée précédemment. Aussi leur a-t-on offert la possibilité d'user de l'une des trois modalités mutuellement exclusives « sûre », « possible » ou « éventuelle » pour chaque phase d'activation transitoire identifiée, comme l'illustre la figure 5.8. Le codage sur trois bits de ces avis, indiqué dans le tableau 5.3, a nécessité de modifier la définition de l'opérateur de fusion (5.5) selon

$$f(\delta_{11}(\boldsymbol{x}),\ldots,\delta_{ec}(\boldsymbol{x}),\theta) = \sum_{i=1}^{e}\sum_{j=1}^{c}\theta_{ij}\,\delta_{ij}(\boldsymbol{x}) \tag{5.8}$$

sous la contrainte de normalisation $\sum_{i=1}^{e}\sum_{j=1}^{c}\theta_{ij} = 1$, où e désigne le nombre d'experts ($e = 4$) et c le nombre de modalités de codage ($c = 3$).

Les paramètres physiologiques susceptibles de caractériser les phases d'activation transitoire sont nombreux, comme l'attestent la définition initiale [Shi71] et les règles de cotation figurant dans [Asd92]. Dans une première expérimentation [Via01], les données d'apprentissage $\boldsymbol{x}_k$ ont été uniquement constituées à partir de la puissance de l'électroencéphalogramme (C3-A2) dans la bande $[4,13[$ Hz, et de celle de l'électromyogramme dans la bande $[10,60[$ Hz, comme le préconisent les études [Nic95] et [Dec99]. La figure 5.8 représente le résultat de la fusion des avis d'experts au vu de ces données objectives. On observe qu'il est conforme aux observations que l'on peut faire sur les signaux physiologiques, et qu'il présente les caractéristiques d'un étiquetage consensuel comme attendu. La pondération θ_{ij} associée à chaque expert et à chacune des modalités de cotation est présentée en figure 5.7. Celle-ci montre que les qualificatifs « sûre », « possible » et « éventuel » n'ont pas été utilisés de la même façon par les cotateurs, l'expert 1 ayant par exemple eu recours au terme « possible » dans des cas plus hasardeux que les experts 3 et 4. On relève également que l'expert 2 semble avoir usé dans une moindre mesure des paramètres physiologiques classiques considérés dans cette étude, à l'inverse des autres

	δ_{i1}	δ_{i2}	δ_{i3}
certain	1	0	0
possible	0	1	0
éventuel	0	0	1

TAB. 5.3 : Codage de l'avis de l'expert i

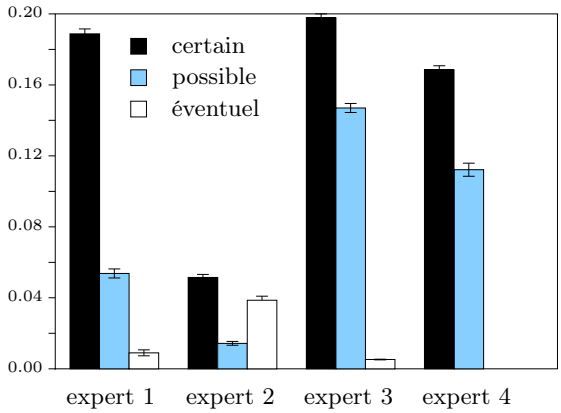

FIG. 5.7 : Pondération des modalités associées à chaque expert après fusion de leurs jugements.

membres du groupe. Une discussion *a posteriori* a confirmé cette analyse, permettant par là-même d'identifier une source de nombreux désaccords. Réciproquement, la méthode proposée permet de porter un regard critique sur des paramètres objectifs extraits des observations, au vu des expertises. Afin d'illustrer cette propriété, on s'est intéressé aux bandes de fréquence préconisées par [Shi71, Asd92] quant aux manifestations des phases d'activation transitoire dans l'électroencéphalogramme, à savoir $[4, 13[$ Hz et au delà de 16 Hz plus rarement. On a ainsi pu montrer que les cotateurs associent aux événements considérés un contenu fréquentiel plus élevé qu'attendu. Ils les identifient en effet dans la bande $[17, 35[$ Hz, résultat cohérent avec une réalité physiologique puisqu'il correspond précisément au rythme électroencéphalographique d'éveil. Il convient de noter que ce support fréquentiel a été obtenu au terme d'une optimisation conjointe de l'information mutuelle et des caractéristiques d'un banc de filtres [Via02(b)] opérant sur les dérivations (C3-A2) et (O1-A2) de l'électroencéphalogramme. Ainsi peut-t-on simultanément élaborer la règle de fusion et optimiser l'espace de représentation à partir d'un critère unique.

5.5 Conclusion et perspectives

Le développement de moyens automatiques de traitement des signaux polysomnographiques constitue un facteur de progrès important dans la compréhension du sommeil et des nombreuses pathologies dont il est le siège. La tâche s'avère toutefois délicate, la complexité de sa macro-structure et la rareté de certains grapho-éléments mettant souvent à mal les algorithmes les plus éprouvés. Malgré cette situation pour le moins hostile, la méthode du critère optimal a permis par le passé d'apporter des solutions satisfaisantes au difficile problème de la détection du complexe K. Grâce à des considérations sur les noyaux reproduisants, on a récemment pu développer chez celle-ci de nouvelles ressources pour l'élaboration de structures décisionnelles encore plus performantes. Des difficultés subsistaient toutefois quant à la prise en compte d'incertitudes sur l'expertise lors de la constitution de l'ensemble d'apprentissage. Aussi s'est-on intéressé au développement d'une méthode non-supervisée de fusion d'avis, autorisant une caractérisation de la pertinence de chaque expert et de l'espace des observations. Celle-ci a été mise à l'épreuve dans le cadre de la cotation des phases d'activation transitoire, exercice réputé pour les conflits d'experts qu'il engendre. Il s'agit là d'un premier pas vers un schéma général auquel on aspire, qui autoriserait la mise en œuvre de la méthode du critère optimal à noyau reproduisant sur des données faisant l'objet d'étiquetages multiples.

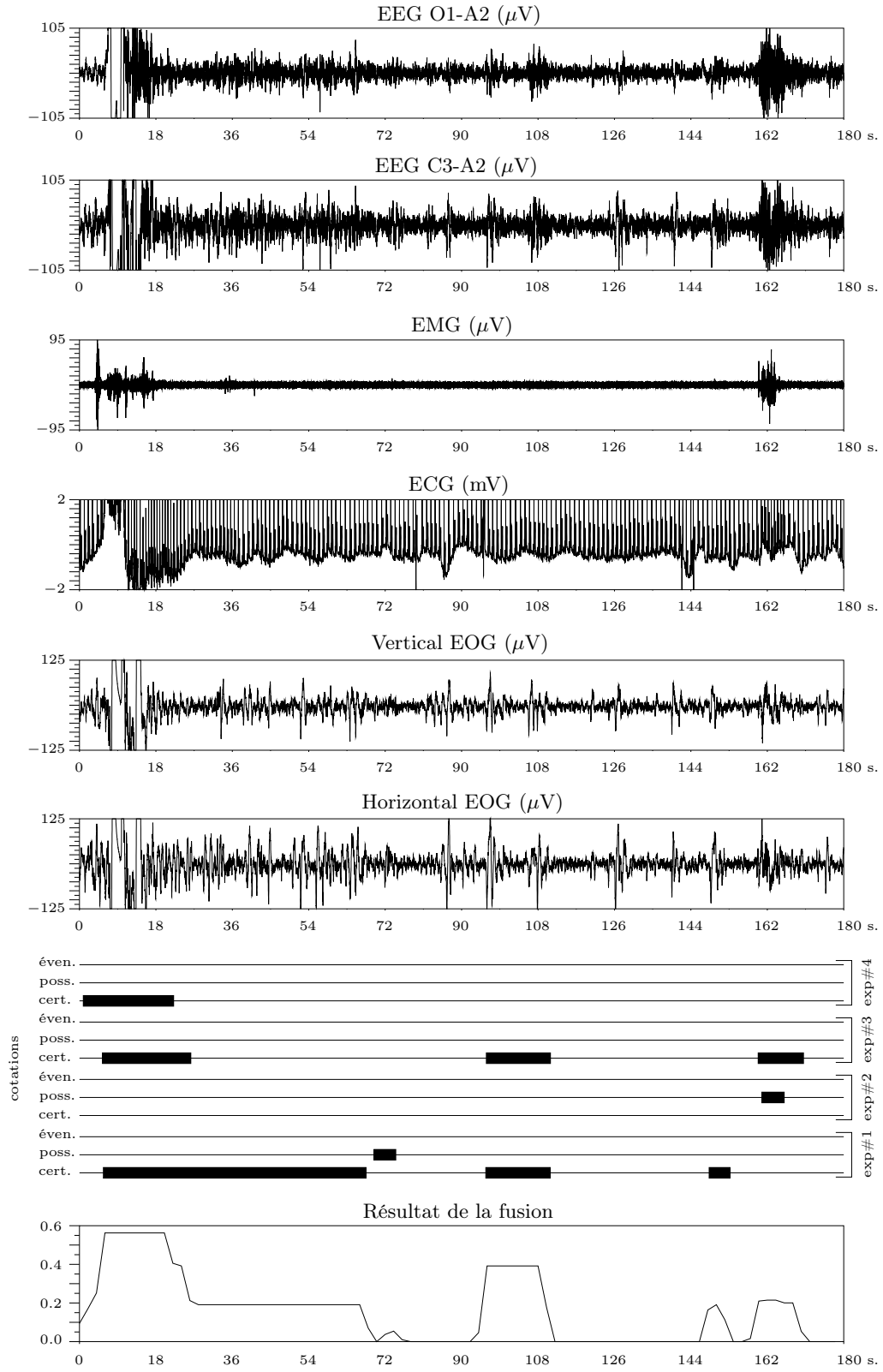

FIG. 5.8 : Caractérisation des phases d'activation transitoire : signaux polysomnographiques, cotations des experts et résultat de la fusion.

Conclusion

Depuis mon intégration en 1995 au sein du Laboratoire de Modélisation et Sûreté des Systèmes de l'Université de Technologie de Troyes, le thème directeur de mes travaux a été celui de la synthèse de structures de détection lorsque les seules informations disponibles consistent en un ensemble de réalisations étiquetées de chacune des hypothèses en compétition, à savoir les hypothèses « bruit » et « signal + bruit ». Le potentiel des approches temps-fréquence pour l'analyse et la décision en environnement non-stationnaire a également été étudié. Si ces recherches ont un caractère essentiellement méthodologique, leur support applicatif n'a pas été négligé pour autant, comme mes travaux sur les signaux de sommeil l'attestent par exemple. Cette conclusion reprend un certain nombre de sujets abordés au cours de ces années, dont ce mémoire a donné un aperçu, en les replaçant dans une perspective thématique et chronologique.

Apprentissage de règles de décision à structure imposée

La résolution optimale d'un problème de détection nécessite une connaissance au moins partielle des lois de probabilité conditionnelles régissant l'observation. Dans un grand nombre d'applications, on ne jouit cependant pas du confort absolu que constitue ce savoir, les phénomènes étant complexes et difficilement interprétables. Une démarche envisageable, néanmoins sous-optimale, consiste à définir préalablement la nature de la structure de détection, puis à en optimiser les paramètres caractéristiques selon un critère donné, à partir de l'information *a priori* dont on dispose. Dans ce contexte, l'élaboration d'un détecteur pose donc les questions suivantes :

- Quelle structure de détection choisir ?
- Quel critère de performance optimiser ?

L'un des objectifs de mes travaux a été de proposer des éléments de réponse à chacune de ces interrogations lorsqu'on ne dispose, pour seule source d'information *a priori*, que d'un ensemble de signaux étiquetés.

En réponse à la première question, il a été montré que deux arguments au moins plaident en faveur de certaines catégories de détecteurs linéaires généralisés, par exemple celle des détecteurs polynômiaux. Le premier argument est fondé sur les garanties de performance que présente la consistance universellement forte de ces classes. Le second, plus pragmatique, repose sur la constatation que la recherche d'un détecteur de ce type est équivalente à l'optimisation d'un discriminant linéaire dans un espace transformé, thème faisant l'objet d'une abondante littérature dans le domaine de la reconnaissance des formes. Les critères du second ordre, dit aussi *de contraste*, apportent une réponse intéressante à la seconde question. En effet, on montre que l'optimisation de nombre d'entre eux peut mener à un détecteur équivalent au test du rapport de vraisemblance. Lorsque cette convergence ne peut cependant avoir lieu, par exemple lorsque

la forme retenue pour la structure s'avère inappropriée, on constate que le choix du critère a une influence prépondérante sur le niveau de performance atteint par cette dernière. La stratégie présentée a permis de pallier cette difficulté en menant au meilleur critère de contraste pour le problème traité, c'est-à-dire celui pour lequel le détecteur résultant a une probabilité d'erreur minimale. Aussi cette approche est-elle appelée *méthode du critère optimal*. L'ensemble de ces travaux a été valorisé par des chapitres d'ouvrages collectifs [Bri02, Len02], des publications dans des revues internationales [Ric98(b), Ric99(a)] et des communications dans le cadre de conférences avec actes [Ric97(a), Ric97(b)]. Afin d'en cerner les qualités et d'en repousser les limites, préciser formellement les liens forts existant entre la méthodologie qui vient d'être évoquée et l'approche paramétrique classique était une nécessité. Ce thème a fait l'objet des premiers travaux de F. Abdallah dans le cadre d'une thèse de doctorat co-encadrée à 50% avec R. Lengellé, qui a débuté au mois de novembre 2000 grâce à un financement du Conseil Régional. Les résultats obtenus ont été valorisés par une publication dans une revue internationale d'excellent niveau [Ric02(a)], ainsi que par des communications dans des conférences avec actes [Ric01(a), Abd02(b), Abd02(a)]. Puis F. Abdallah a apporté de nouveaux développements à la méthode du critère optimal en faisant appel à la théorie des espaces de Hilbert à noyau reproduisant. Elle offre à présent la possibilité d'élaborer des détecteurs linéaires généralisés dans des espaces transformés de dimension très importante, sans qu'aucun calcul y soit effectué explicitement. Les tests pratiqués montre que les détecteurs résultants sont en mesure de concurrencer les structures décisionnelles phare du moment : les Support Vector Machines. Ce travail a été en partie valorisé par une communication dans une conférence américaine avec actes [Abd02(c)]. Un article de synthèse sur cette nouvelle approche, dite *du critère optimal à noyau reproduisant*, a été soumis pour publication dans une revue internationale [Abd02(d)]. Une extension de celle-ci au cas multi-hypothèses est d'ores et déjà à l'étude.

Obtenir une probabilité d'erreur satisfaisante avec un détecteur élaboré au moyen d'une base d'apprentissage nécessite de trouver une adéquation entre la complexité de celui-ci, formellement définie par la dimension de Vapnik-Chervonenkis ou VC-dimension, et le nombre de signaux étiquetés dont on dispose. En effet, les récepteurs dotés d'une complexité trop importante ont généralement un faible pouvoir de généralisation. Dans le cas contraire, ces derniers peuvent être incapables d'intégrer la totalité de l'information discriminante présente dans l'ensemble d'apprentissage. En conséquence, j'ai été amené à proposer trois méthodes pour l'optimisation de la VC-dimension des détecteurs linéaires généralisés, dont deux ont été exposées dans le cadre de ce mémoire. Leur mode opératoire consiste en un contrôle approprié de la dimension de l'espace de représentation des observations, conformément au principe de minimisation du risque structurel proposé par Vapnik dans le cadre de la théorie statistique de l'apprentissage. Il en résulte une robustesse accrue des structures de détection traitées vis à vis du phénomène de malédiction de la dimensionnalité [Ric98(d), Ric98(f)]. Récemment, F. Abdallah a adapté ces procédures à la méthode du critère optimal à noyau reproduisant. Un test effectué sur un problème classique, celui des deux spirales imbriquées [Lan88], a permis de constater la très nette supériorité de notre approche par rapport à une technique récemment publiée [Bil02]. Ces résultats ont été présentés dans ce mémoire, et constituent l'un des objets de l'article de synthèse [Abd02(d)] évoqué précédemment, en cours de soumission.

L'ensemble de ces travaux s'inscrit dans la continuité de la révolution que connaît le domaine de la reconnaissance des formes depuis le milieu des années 90, avec l'avènement des noyaux reproduisants pour la résolution de problèmes de détection/classification et régression. Ceux-ci permettent en effet de développer un caractère non-linéaire dans certains traitements linéaires,

sans en affecter la facilité de mise en œuvre. Il semble que le domaine du traitement du signal puisse devenir un terrain d'application de cette nouvelle technique, ce à quoi je vais essayer de contribuer. Dans ce contexte, il est à noter ma participation à l'action spécifique « Méthodes à vecteurs de support » animée par M. Davy (IRCCyN, Nantes). Elle a été dotée de 67000 euros par le DSTIC et de 5000 euros par le GdR-PRC ISIS, pour la période 2002-2003. Elle implique notamment l'organisation de sessions spéciales aux conférences Gretsi'03 et IEEE Icassp'03, ainsi qu'une école d'été.

Analyse et décision en environnement non-stationnaire

En vertu du point de vue intéressant qu'elles offrent sur les signaux non-stationnaires, les méthodes temps-fréquence ont souvent été associées à des structures décisionnelles. Aussi ai-je fréquemment privilégié ce mode de représentation lors de la mise en œuvre de la méthode du critère optimal, et des techniques de contrôle de complexité qui lui sont associées [Ric98(a)]. Dans le cadre de problèmes relatifs aux signaux de sommeil en particulier, les performances des détecteurs ainsi obtenus se sont avérées très satisfaisantes en comparaison de solutions proposées dans la littérature [Ric98(c), Coa01], bien que l'ensemble d'apprentissage fut parfois de taille modeste. Ceci illustre la validité des choix qui ont été faits pour aboutir à la méthodologie proposée, que des propriétés théoriques fortes viennent évidemment renforcer. Il est à noter que nos investigations sur les structures à noyau permettent à présent d'élaborer des détecteurs opérant dans ce type d'espace transformé, sans qu'aucune représentation temps-fréquence ne soit calculée. L'une des conséquences de cette remarque est l'existence avérée de noyaux reproduisants offrant des propriétés de covariance vis à vis de certaines transformations, ici les translations en temps et en fréquence de l'observation. Aussi me semble-t-il opportun de caractériser ceux-ci sans connaître explicitement l'espace transformé qui leur est associé, afin de traiter avec efficience certains problèmes de détection sur lesquels pèsent des paramètres de nuisance. A ma connaissance, ce thème n'a encore fait l'objet d'aucun développement.

Le lien rigoureux établi entre les structures à noyau et les détecteurs opérant dans le plan temps-fréquence a nécessité la définition préalable d'un cadre algébrique approprié [Ric01(b)]. Il s'agit en l'occurrence d'un espace euclidien engendré par les représentations temps-fréquence discrètes des éléments d'un espace signal donné. Cette notion d'espace linéaire de représentations temps-fréquence ne va pas sans évoquer les travaux de Hlawatsch sur la représentation temps-fréquence d'un espace vectoriel [Hla98], et de ses développements en décision, estimation ou encore synthèse de signaux. A moyen terme, j'envisage également d'aborder de telles problématiques au jour de la structure algébrique que j'ai établie. Jusqu'à présent, ce cadre a permis d'étudier la redondance informationnelle de certaines représentations discrètes [Ric01(b)]. A l'aide d'arguments issus de la théorie de l'apprentissage, j'ai également pu justifier le fait qu'un contrôle efficace de la complexité des détecteurs opérant dans le plan temps-fréquence puisse être obtenu par un lissage approprié des représentations [Ric02(b), Ric02(c)]. Ce concept a été mis en œuvre dans le cadre d'une action « jeunes chercheurs » du GdR-PRC ISIS placée sous la responsabilité de F. Auger (GE44, Saint-Nazaire), à laquelle j'ai participé durant la période 1999-2001. Intitulé « Nouveaux outils d'analyse et de décision pour les signaux fortement non-stationnaires », ce projet a été consacré aux représentations temps-fréquence et temps-échelle diffusives. On rappelle que celles-ci résultent de l'action d'un processus de diffusion sur des représentations conventionnelles, que supervise une fonction de conductance convenablement choisie. J. Gosme a effectué sont stage de DEA dans le cadre de cette action sur le thème de la détection par représentations diffusives, co-dirigé à 50% par P. Gonçalvès (INRIA, Grenoble) et moi-même. Durant

cette période, il a étudié un schéma de diffusion permettant de révéler les zones d'information discriminante dans une représentations temps-fréquence ou temps-échelle, offrant ainsi la possibilité de contrôler efficacement la complexité d'un détecteur opérant sur celles-ci [Gos02]. Depuis le mois d'octobre 2001, J. Gosme prépare une thèse de doctorat sous ma responsabilité (75%) et celle de R. Lengellé (25%), en collaboration étroite avec P. Gonçalvès. Dans la continuité de son sujet de DEA, il étudie les principaux modèles de diffusion existants (isotrope, anisotrope, homogène, non-homogène) qu'il applique aux distributions des groupes de Cohen et Affine. Dans ce contexte, il essaye de caractériser les fonctions de conductance adaptées à des problématiques particulières telles que l'amélioration de la lisibilité, le débruitage ou encore la classification, et s'interroge sur les propriétés des classes de représentations obtenues.

Parallèlement à ces activités, j'ai eu l'opportunité de me joindre à un groupe de chercheurs renommés dans le cadre du projet « Méthodes temps-fréquence pour l'analyse des données de l'interféromètre Virgo », couvrant la période 2001-2003 et placé sous la responsabilité d'É. Chassande-Mottin (Obs. de la Côte d'Azur). Enfin, je collabore occasionnellement avec le Laboratoire de Nanotechnologie et d'Instrumentation Optique de l'Université de Technologie de Troyes, dans le cadre duquel je contribue à l'analyse temps-fréquence de signaux de champs proche optique. Ces échanges pluridisciplinaires ont donné lieu à des publications dans une revue [Gha00] et une conférence internationales de haut niveau [Bar01].

Application à la détection dans les signaux de sommeil

Les résultats en décision statistique et en analyse temps-fréquence que j'ai présentés ont pour la plupart été validés sur des problèmes relatifs à la détection de phénomènes transitoires dans les signaux polysomnographiques [Ric98(c), Coa01]. A la suite de cela, certains ont été adoptés pour l'étude des effets de substances chimiques sur le sommeil [Coa00]. Ce support applicatif s'inscrit dans une coopération scientifique menée depuis plusieurs années avec la Fondation pour la Recherche en Neurosciences Appliquées à la Psychiatrie, sur le thème de la caractérisation et de la détection/classification d'événements dans les signaux électrophysiologiques. En octobre 1999, cette collaboration s'est vue dotée d'une subvention de 700kF dans le cadre de l'Action Concertée Incitative « Télémédecine et technologies pour la santé », pour le projet « Détection automatique de micro-éveils pendant le sommeil ». Cette aide a permis de financer en partie la thèse de G. Viardot, soutenue le 22 octobre 2002, dont j'ai assuré le co-encadrement à 50% avec R. Lengellé. Le doctorant a principalement consacré ses recherches à la fusion d'avis d'experts et à la caractérisation de l'expertise. Plus précisément, l'élaboration d'une règle de décision à partir d'une base d'apprentissage suppose en général que l'étiquetage des données utilisées est exempt d'erreurs. Cette situation idéale n'est cependant pas toujours réaliste, en particulier lorsque la définition du phénomène étudié est sujette à l'interprétation du cotateur [Len00]. Les travaux de G. Viardot ont consisté à proposer une méthode pour la fusion d'expertises divergentes, qui autorise une analyse *a posteriori* de la pertinence de chacun des experts sollicités et du choix de l'espace de représentation des observations. Elle repose sur la maximisation de l'information mutuelle normalisée entre les observations et une fonctionnelle des avis émis par les experts [Via00, Via01]. Cette approche a été appliquée avec succès à la caractérisation de l'expertise des phases d'activation transitoires, ou micro-éveils, dans les signaux de sommeil [Via02(b)]. Il s'agit là d'un premier pas vers un schéma général auquel j'aspire, qui consisterait en la mise en œuvre de la méthode du critère optimal à noyau reproduisant sur des données faisant l'objet d'étiquetages multiples.

Bibliographie

[Abd02(a)] F. Abdallah, C. Richard, R. Lengellé. On equivalence between detectors obtained from second-order measures of performance. *Proc. European Signal Processing Conference,* Eusipco'02, Toulouse, 2002.

[Abd02(b)] F. Abdallah, C. Richard, R. Lengellé. On virtues and vices of second-order measures of quality for binary classification. *Proc. International Conference Annie,* Saint Louis, 2002.

[Abd02(c)] F. Abdallah, C. Richard, R. Lengellé. A method for designing nonlinear kernel-based discriminant functions from the class of second-order criteria. *Proc. International Conference Asilomar,* Pacific Grove, CA, 2002.

[Abd02(d)] F. Abdallah, C. Richard, R. Lengellé. An improved training algorithm for nonlinear kernel discriminants. *IEEE Transactions on Signal Processing,* en cours de soumission, 2002.

[Alv93] L. Alvarez, F. Guichard, P. L. Lions, J. M. Morel. Axioms and fundamental equations of image processing. *Archieve for Rational Mechanics,* vol. 123, no. 3, p. 199-257, 1993.

[Alv94] L. Alvarez, L. Mazorra. Signal and image restauration using shock filters and anisotropic diffusion. *SIAM,* vol. 31, no. 2, p. 590-605, 1994.

[Asd92] American Sleep Disorders Association (ASDA). EEG arousals : scoring rules and examples. A preliminary report from sleep disorders atlas task force of the ASDA. *Sleep,* vol. 15, no. 2, p. 173-184, 1992.

[Aug95] F. Auger, P. Flandrin. Improving the readability of time-frequency and time-scale representations by reassignment methods. *IEEE Transactions on Signal Processing,* vol. 43, p. 1068-1089, 1995.

[Ban92] I. N. Bankman, V. G. Sigillito, R. A. Wise, P. L. Smith. Feature-based detection of K-complex wave in the human electroencephalogram using neural networks. *IEEE Transactions on Biomedical Engineering,* vol. 39, no. 12, p. 1605-1610, 1992.

[Bar01] D. Barchiesi, C. Richard. Time-frequency analysis of near-field optical data for extracting local attributes. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* IEEE ICASSP'01, Salt Lake City, Utah, 2001.

[Bas92] C. Bastien, K. Campbell. The evoked K-complex : all-or-none phenomenon ? *Sleep,* vol. 15, no. 3, p. 236-245, 1992.

[Bel61] R. Bellman. *Adaptive Control Processes : a Guided Tour.* Princeton : Princeton University Press, 1961.

[Bil02] S. A. Billings, K. L. Lee. Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm. *Neural Networks,* vol. 15, p. 263-270, 2002.

[Bis95] C. M. Bishop. *Neural networks for pattern recognition.* Oxford : Oxford University Press, 1995.

[Blo96] I. Bloch. Information combination operators for data fusion : a comparative review with classification. *IEEE Transactions on Systems, Man and Cybernetics,* vol. 26, no. 1, p. 52-67, 1996.

[Bos92] B. E. Boser, I. M. Guyon, V. N. Vapnik. A training algorithm for optimal margin classifiers. *Proc. 5th Annual Workshop on Computational Learning Theory,* p. 144-152, 1992.

[Bre70] G. Bremer, J. R. Smith, I. Karakan. Automatic detection of the K-complex in sleep electroencephalograms. *IEEE Transactions on Biomedical Engineering,* vol. 17, p. 314-323, 1970.

[Bri02] D. Brie, R. Lengellé, N. Nikiforov, C. Richard. Autres applications. (12 p.) In R. Lengellé, (éd.). *Reconnaissance des Formes et Décision en Signal.* Paris : Hermès Sciences, Traité IC2, 2002.

[Bro95] J. Bronzino. *Biomedical Engineering Handbook.* IEEE Press, 1995.

[Bur98] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery,* vol. 2, no. 2, p. 121-167, 1998.

[Can02] S. Canu. Modèles connexionistes et SVM pour la décision. In R. Lengellé, (éd.). *Décision et Reconnaissance des Formes en Signal.* Paris : Hermès Sciences, Traité IC2, 2002.

[Cat92] F. Catte, T. Coll, P. L. Lions, J. M. Morel. Image selective smoothing and edge detection by nonlinear diffusion *SIAM,* vol. 29, no. 1, p. 182-193, 1992.

[Cha98] E. Chassande-Mottin. *Méthodes de Réallocation dans le Plan Temps-Fréquence pour l'Analyse et le Traitement de Signaux Non-Stationnaires.* Thèse de doctorat, Ecole Normale Supérieure de Lyon, 1998.

[Che89] S. Chen, S. A. Billings, W. Luo. Orthogonal least squares methods and their application to nonlinear system identification. *International Journal of Control,* vol. 50, no. 5, p. 1873-1886, 1989.

[Cim97] A. Cimetiere. *Méthodes Temporelles et Temps-Fréquence pour la Reconnaissance Automatique des Complexes K de l'EEG de Sommeil.* Thèse de doctorat, Université de Technologie de Compiègne, 1997.

[Coa00] A. Coatanhay, C. Richard, R. Lengellé, A. Muzet, J.-P. Macher. Automated detection of K-complexes : characterization of benzodiazepine effects. *Journal of Sleep Research,* vol. 9, no. 1, 2000.

[Coa01] A. Coatanhay, C. Richard, L. Staner. Classification en temps-fréquence à partir de données expertisées. Application à la détection des complexes K. *Extraction des Connaissances et Apprentissage,* Hermès, vol. 1, no. 4, p. 293-304, 2001.

[Cor95] C. Cortes, V. Vapnik. Support-vector networks. *Machine Learning,* vol. 20, p. 1-25, 1995.

[Cos01] A. H. Costa, G. F. Boudreaux-Bartel. An overview of aliasing errors in discrete-time formulations of time-frequency representations. *IEEE Transactions on Signal Processing,* vol. 47, p. 1463-1474, 2001.

[Cou53] R. Courant, D. Hilbert. *Methods of Mathematical Physics.* New York : Interscience, 1953.

[Cov65] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities for finite length signals. *IEEE Transactions on Electron. Comp.,* vol. 10, p. 326-334, 1965.

[Cov91] T. M. Cover, J. A. Thomas. *Elements of Information Theory.* New York : Wiley, 1991.

[Dar91] A. C. Da Rosa, B. Kemp, T. Paiva. A model-based detector of vertex sharp waves and K-complexes in sleep electroencephalogramm. *Electroencephalography and Clinical Neurophysiology,* vol. 78, p. 71-79, 1991.

[Dav00] M. Davy. *Noyaux Optimisés pour la Classification dans le Plan Temps-Fréquence.* Thèse de doctorat, Université de Nantes, 2000.

[Dav01] M. Davy, C. Doncarli, G. F. Boudreaux-Bartels. Improved optimization of time-frequency based signal classifiers. *IEEE Signal Processing Letters,* vol. 8, p. 52-57, 2001.

[Dec99] F. De Carli, L. Nobili, P. Gelcich, F. Ferrillo. A method for the automatic detection of arousals during sleep. *Sleep,* vol. 22, no. 5, p. 561-572, 1999.

[Der96] R. Deriche, O. Faugeras. Les EDP en traitement des images et vision par ordinateur. *Traitement du Signal,* vol. 13, no. 6, 1996.

[Des94] J. Destiné, B. Beckers, M. Fombellida, R. Poirrier, D. Dive, G. Franck. Utilisation des réseaux de neurones artificiels pour la reconnaissance de grapho-éléments phasiques dans le cadre de l'analyse du sommeil. *Proc. Symposium International 25 ans d'Analyse Automatique de Sommeil,* Genève, 1994.

[Dev89] L. Devroye, A. Krzyzak. An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference,* vol. 23, p. 71-82, 1989.

[Dev96] L. Devroye, L. Györfi, G. Lugosi. *A Probabilistic Theory of Pattern Recognition.* New York : Springer-Verlag, 1996.

[Dub90] B. Dubuisson. *Diagnostic et Reconnaissance des Formes.* Paris : Hermès, 1990.

[Dud01] R. O. Duda, P. E. Hart, D. G. Stork. *Pattern Classification.* New York : Wiley and Sons, 2001.

[Duv87] P. Duvaut. *Contraste et Détection. Application à la Quantification et aux Filtres de Volterra Optimaux pour la Détection.* Thèse de Doctorat, Université Paris 11, Centre d'Orsay, 1987.

[Eat99] E. Eaton, K. Hume, P. Stone, P. Stone, A. Woodcock. Respiratory paradox as an indicator of arousal from non-REM sleep. *Sleep,* vol. 22, no. 8, 1999.

[Efr79] B. Efron. Bootstrap methods : another look at jacknife. *Annals of Statistics,* vol. 7, p. 1-26, 1979.

[Ehr88] A. Ehrenfeucht, D. Haussler, M. Kearns, L. G. Valiant. A general lower bound on the number of examples needed for learning. *Proc. 1st Workshop on Computational Learning Theory,* Palo Alto : Morgan Kaufmann Publishers, p. 42-45, 1988.

[Fis36] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Ann. Eugenics,* vol. 7, no. 2, p. 111-132, 1936.

[Fla88] P. Flandrin. A time-frequency formulation of optimum detection. *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 36, p. 1377-1384, 1988.

[Fla98] P. Flandrin. *Temps-Fréquence.* Paris : Hermès, seconde édition, 1998.

[Fri89] J. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association,* vol. 84, no. 405, p. 165-175, 1989.

[Fuk90] K. Fukunaga. *An Introduction to Statistical Pattern Recognition.* London : Academic Press, 2ème édition, 1990.

[Gar80] W. A. Gardner. A unifying view of second-order measures of quality for signal classification. *IEEE Transactions on Communications,* vol. 28, no. 6, p. 807-816, 1980.

[Gai73] J. M. Gaillard, R. Tissot. Principles of automatic analysis of sleep records with hybrid systems. *Computers and Biomedical Research,* vol. 6, p. 1-13, 1973.

[Gha00] T. Gharbi, D. Barchiesi, O. Bergossi, H. Wioland, C. Richard. Optical near-field analysis by means of time-frequency distributions. Application to the characterization and the separation of the image spectral contents using reassignment method. *Journal of Optical Society of America,* vol. 17, no. 12, p. 2513-2519, 2000.

[Gol93] G. H. Golub and C. F. Van Loan.. *Matrix Computations.* London : The Johns Hopkins University Press, 1993.

[Gon98] P. Gonçalvès, E. Payot. Adaptive diffusion equation for time-frequency representations. *Proc. IEEE Digital Signal Processing Workshop,* 1999.

[Gos02] J. Gosme, P. Gonçalvès, C. Richard, R. Lengellé. Adaptive diffusion and discriminant analysis for complexity control of time-frequency detectors. *Proc. European Signal Processing Conference,* Eusipco'02, Toulouse, 2002.

[Guy93] I. Guyon, B. Boser, V. Vapnik. Automatic capacity tuning of very large VC-dimension classifiers. In S. J. Hanson, J. D. Cowan, C. L. Giles, (éds). *Proc. Advances in Neural Information Processing Systems.* San Mateo : Morgan Kaufmann, vol. 5, p. 147-155, 1993.

[Hal93] P. Halasz, W. Krattenthaler. Arousals without awakening. Dynamic aspect of sleep. *Physiol. and Behav.,* vol. 54, p. 795-802, 1993.

[Her94] J. Hérault, C. Jutten. *Réseaux Neuronaux et Traitement du Signal.* Paris : Hermès, 1994.

[Hla92] F. Hlawatsch, W. Krattenthaler. Bilinear signal synthesis. *IEEE Transactions on Acoustics, Speech and Signal Processing,* vol. 40, p. 352-363, 1992.

[Hla98] F. Hlawatsch. *Time-Frequency Analysis and Synthesis of Linear Signal Space.* Norwell : Kluwer Academic Press, 1998.

[Hof63] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association,* vol. 58, p. 13-30, 1963.

[Jon95] D. L. Jones, R. G. Baraniuk. An adaptive optimal kernel time-frequency representation *IEEE Transactions on Signal Processing,* vol. 43, no. 10, p. 2361-2371, 1995.

[Joh78] D. Johnson, F. Preparata. The densest hemisphere problem. *Theoretical Computer Science,* vol. 6, no. 1, p. 93-107, 1978.

[Koe84] J. J. Koenderink. The structure of images. *Biological Cybernetics,* vol. 50, p. 363-370, 1984.

[Lac68] P. Lachenbruch, M. Mickey. Estimation of error rates in discriminant analysis. *Technometrics,* vol. 10, p. 1-11, 1968.

[Lam97] L. Lam, C. Y. Suen. Application of majority voting to pattern recognition : an analysis of its behavior and performance. *IEEE Transactions on Systems, Man and Cybernetics,* vol. 27, no. 5, p. 553-568, 1997.

[Lan88] K. J. Lang, M. J. Witbrock. Learning to tell two spirals apart. *Proc. Connectionnist Summer Schools.* San Mateo : Morgan Kaufmann, 1988.

[Lec90] Y. LeCun, J. S. Denker, S. A. Solla. Optimal brain damage. *Proc. Advances in Neural Information Processing Systems.* vol. 2, p. 598-605, 1990.

[Lem95] O. Lemoine. *Détection des Signaux Non-Stationnaires par Représentation Temps-Fréquence.* Thèse de doctorat, Université de Nice-Sophia Antipolis, 1995.

[Len00] R. Lengellé, C. Richard, S. Millemann. Neural network based membership function estimation. Application to uncertain time-varying systems supervision. *Proc. IEEE International Symposium on Intelligent Signal Processing and Communication,* IEEE ISPACS'00, Honolulu, Hawai, 2000.

[Len02] R. Lengellé, C. Richard. Apprentissage de règles de décision à structure imposée et contrôle de la complexité. (33 p.) In R. Lengellé, (éd.). *Reconnaissance des Formes et Décision en Signal.* Paris : Hermès Sciences, Traité IC2, 2002.

[Lev00] H. Levesque, G. Lakemeyer. *The Logic of Knowledge Bases.* Cambridge : MIT Press, 2000.

[Loo37] A. L. Loomis, E. N. Harvey, G. A. Hobart. Cerebral states during sleep as studied by human brain potentials. *Journal of experimental psychology,* vol. 21, p. 127-144, 1937.

[Mar97] N. MARINOVITCH. The singular value decomposition of the Wigner distribution and its applications. In W. MECKLENBRÄUKER, F. HLAWATSCH, (éds). *The Wigner distribution : theory and applications in signal processing.* Amsterdam : Elsevier, 1997.

[Mat98] G. MATZ, F. HLAWATSCH. Time-frequency methods for signal detection with application to the detection of knock in car engines. *Proc. IEEE-SP Workshop on Statistical Signal and Array Processing,* p. 196-199, Portland, 1998.

[Mik99] S. MIKA, G. RÄTSCH, J. WESTON, B. SCHÖLKOPF, K. R. MÜLLER. Fisher discrimiant analysis with kernels. In Y. H. HU, J. LARSEN, E. WILSON, S. DOUGLAS, (éds). *Proc. Advances in Neural Information Processing Systems.* San Mateo : Morgan Kaufmann, p. 41-48, 1999.

[Mik01(a)] S. MIKA, G. RÄTSCH, K. R. MÜLLER. A mathematical programming approach to the kernel Fisher algorithm. In T. K. LEEN, T. G. DIETTERICH, V. TRESP, (éds). *Proc. Advances in Neural Information Processing Systems.* Cambridge : MIT Press, p. 591-597, 2001.

[Mik01(b)] S. MIKA, A. J. SMOLA, B. SCHÖLKOPF. An improved training algorithm for kernel Fisher discriminants. In T. JAAKKOLA, T. RICHARDSON, (éds). *Proc. AISTATS.* San Mateo : Morgan Kaufmann, p. 98-104, 2001.

[Min69] M. L. MINSKI, S. A. PAPERT. *Perceptrons.* Cambridge : MIT Press, 1969. Réédition étendue, 1990.

[Moe00] P. MOERLAND. *Mixtures Models for Unsupervised and Supervised Learning.* Thèse de Doctorat, EPFL, Lausanne, 2000.

[Moo92] J. E. MOODY. Generalization, weight decay and architecture selection for nonlinear learning systems. In J. E. MOODY, S. J. HANSON, R. P. LIPPMANN, (éds). *Proc. Advances in Neural Information Processing Systems.* San Mateo : Morgan Kaufmann, 1992.

[Mul01] K. R. MÜLLER, S. MIKA, G. RÄTSCH, K. TSUDA, B. SCHÖLKOPF. An introduction to kernel-based learning algorithms. *IEEE Transactions on Neural networks,* vol. 12, no. 2, p. 181-201, 2001.

[Nar77] P. NARENDRA, K. FUKUNAGA. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers,* vol. 26, p. 917-922, 1977.

[Nic95] A. NICOLAS. *Étude des Caractéristiques Électrophysiologiques des Phases d'Activation Transitoire au cours du Sommeil Humain.* Thèse de Doctorat, Université Louis Pasteur, Strasbourg, 1995.

[Now91] S. J. NOWLAN, G. E. HINTON. Evaluation of adaptive mixtures of competing experts. In R. P. LIPPMANN, J. E. MOODY, D. S. TOURETZKY, (éds). *Proc. Advances in Neural Information Processing Systems.* San Mateo : Morgan Kaufmann, 1991.

[One99(a)] J. C. O'NEILL, P. FLANDRIN, W. J. WILLIAMS. On the existence of discrete Wigner distributions. *IEEE Signal Processing Letters,* vol. 6, p. 304-306, 1999.

[One99(b)] J. C. O'NEILL, W. J. WILLIAMS. Shift covariant time-frequency distributions of discrete signals. *IEEE Transactions on Signal Processing,* vol. 47, p. 789-799, 1999.

[Pal94] N. Pal, J. Bezdek. Measuring fuzzy uncertainty. *IEEE Transactions on Fuzzy Systems,* vol. 2, no. 2, p. 107-118, 1994.

[Pat66] D. W. Patterson, R. L. Mattson. A method of finding linear discriminant functions for a class of performance criteria. *IEEE Transactions on Information Theory,* vol. 12, no. 3, p. 380-387, 1966.

[Per90] P. Perona, J. Malik. A method of finding linear discriminant functions for a class of performance criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 12, no. 7, p. 629-639, 1990.

[Per94] P. Perona, T. Shiota, J. Malik. Anisotropic diffusion. In B. M. ter Haar Romeny, (éd.). *Geometry-Driven Diffusion in Computer Vision.* New York : Kluwer Academic, 1994.

[Pey86] F. Peyrin, P. Prost. A unified definition for the discrete-time discrete-frequency, and discrete time-frequency Wigner distributions. *IEEE Transactions on Signal Processing,* vol. 34, p. 858-867, 1986.

[Pic86] B. Picinbono, P. Duvaut. Detection and contrast. In C. Baker, (éd.). *Stochastic Processes in Underwater Acoustics.* New York : Springer-Verlag, 1986.

[Poo94] H. V. Poor. *An Introduction to Signal Detection and Estimation.* New York : Springer-Verlag, 1994.

[Ray86] S. Ray, W. D. Lee, C. D. Morgan, W. Airth-Kindree. Computer sleep stage scoring. An expert system approach. *International Journal of Biomedical Computing,* vol. 19, 1986.

[Rec68] A. Rechtschaffen, A. Kales. *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages in Human Subjects.* Washington DC : US Government Printing Office, Public Health Series, 1968.

[Ric97(a)] C. Richard, R. Lengellé. Une nouvelle approche pour la détection linéaire optimale dans le plan temps-fréquence. *Proc. Colloque GRETSI,* p. 659-662, Grenoble, 1997.

[Ric97(b)] C. Richard, R. Lengellé. Joint time and time-frequency optimal detection. *Proc. IEEE-UK Symposium on Applications of Time-Frequency and Time-Scale Methods,* IEEE UK-TFTS'97, p. 29-32, Coventry, 1997.

[Ric98(a)] C. Richard. *Une Méthodologie pour la Détection à Structure Imposée. Applications au Plan Temps-Fréquence.* Thèse de Doctorat, Université de Technologie de Compiègne, 1998.

[Ric98(b)] C. Richard, R. Lengellé. Joint time and time-frequency optimal detection of K-complexes in sleep EEG. *Computers and Biomedical Research,* vol. 31, no. 3, p. 209-229, 1998.

[Ric98(c)] C. Richard, R. Lengellé. Détection automatique de phénomènes transitoires de l'EEG par représentation temps-fréquence. *Innovation et Technologie en Biologie et Médecine,* numéro spécial temps-fréquence, vol. 19, no. 3, p. 167-177, 1998.

[Ric98(d)] C. RICHARD, R. LENGELLÉ. Two algorithms for designing optimal reduced-bias data-driven time-frequency detectors. *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis,* IEEE TFTS'98, p. 601-604, Pittsburgh, PA, 1998.

[Ric98(e)] C. RICHARD, R. LENGELLÉ. On the dimension of the discrete Wigner-Ville transform range space. Application to time-frequency-based detectors design. *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis,* IEEE TFTS'98, p. 5-8, Pittsburgh, PA, 1998.

[Ric98(f)] C. RICHARD, R. LENGELLÉ. Structural risk minimization for reduced-bias time-frequency-based detectors design. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* IEEE ICASSP'98, p. 2397-2400, Seattle, WA, 1998.

[Ric99(a)] C. RICHARD, R. LENGELLÉ. Data-driven design and complexity control of time-frequency detectors. *Signal Processing,* vol. 77, p. 37-48, 1999.

[Ric99(b)] C. RICHARD, R. LENGELLÉ. Sur le contrôle de la complexité des détecteurs opérant dans le domaine temps-fréquence par le biais de la fonction de paramétrisation. *Proc. Colloque GRETSI,* p. 905-908, Vannes, 1999.

[Ric00(a)] C. RICHARD, R. LENGELLÉ. On the linear relations connecting the components of the discrete Wigner distribution in the case of real-valued signals. *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing,* IEEE ICASSP'00, p. 85-88, Istambul, 2000.

[Ric01(a)] C. RICHARD, R. LENGELLÉ, F. ABDALLAH. Optimisation de critères de contraste et des performances en détection. *Proc. Colloque GRETSI*, Toulouse, 2001.

[Ric01(b)] C. RICHARD. Linear redundancy of information carried by the discrete Wigner distribution. *IEEE Transactions on Signal Processing,* vol. 49, p. 2536-2544, 2001.

[Ric02(a)] C. RICHARD, R. LENGELLÉ, F. ABDALLAH. Bayes-optimal detectors using relevant second-order criteria. *IEEE Signal Processing Letters,* vol. 9, no. 1, 2002.

[Ric02(b)] C. RICHARD. Time-frequency based detection using discrete-time discrete-frequency Wigner distributions. *IEEE Transactions on Signal Processing,* vol. 50, no. 9, p. 2170-2176, 2002.

[Ric02(c)] C. RICHARD. Détection par représentations temps-fréquence discrètes. (23 p.) In N. MARTIN, C. DONCARLI, (éds). *Décision dans le Plan Temps-Fréquence.* Paris : Hermès Sciences, Traité IC2, 2002.

[Ric02(d)] C. RICHARD. Discrétisation des détecteurs temps-fréquence : problèmes en découlant et éléments de solution. *Proc. Colloque GRETSI,* Toulouse, 2001.

[Rch98] M. S. RICHMAN, T. W. PARKS, R. G. SHENOY. Discrete-time, discrete-frequency, time-frequency analysis. *IEEE Transactions on Signal Processing,* vol. 46, p. 1517-1527, 1998.

[Ris89] J. RISSANEN. *Stochastic Complexity in Statistical Inquiry.* Singapore : World Scientific, 1989.

[Ros62] F. Rosenblat. *Principles of Neurodynamics : Perceptron and Theory of Brain Mechanisms.* Washington DC : Spartan Books, 1962.

[Sai88] S. Saitoh. *Theory of Reproducing Kernel and its Applications.* Harlow : Longman Scientific & Technical, 1988.

[Sap90] G. Saporta. *Probabilité, analyse des données et statistique.* Paris : Éditions Technip, 1990.

[Sau72] N. Sauer. On the density of family of sets. *Journal of Combinatorial Theory Series A,* vol. 13, p. 145-147, 1972.

[Sha89] M. Shaw, B. Gaines. Comparing conceptual structures : consensus, conflict, correspondance and contrast. *Knowledge Acquisition,* vol. 1, p. 341-363, 1989.

[Say95] A. M. Sayeed, D. L. Jones. Optimal detection using bilinear time-frequency and time-scale representations. *IEEE Transactions on Signal Processing,* vol. 43, p. 2872-2883, 1995.

[Sch93] N. Schaltenbrand, R. Lengellé, J.-P. Macher. Neural network model : application to automatic analysis of human sleep. *Computers and Biomedical Research,* vol. 26, p. 157-171, 1993.

[Shi71] J. Schieber, A. Muzet, J. Ferriere. Les phases d'activation transitoire spontanées au cours du sommeil normal chez l'homme. *Archives des Sciences Physiologiques,* vol. 25, no. 4, p. 443-465, 1971.

[Sho99] B. Schölkopf, C. Burges, A. Smola. *Advances in Kernel Methods-Support Vector Learning.* Cambridge : MIT Press, 1999.

[She77] O. Sherif, B. Pagurek, S. Mahmouhd, R. Broughton. Detection of K-complex in the sleep EEG. *Proc. Int. Electrical and Electronic Conf. and Exp.,* vol. 84, p. 204-205, 1977.

[Skl79] J. Sklansky, G. Wassel. *Pattern Classifiers and Trainable Machines.* New York : Springer-Verlag, 1979.

[Sta00] L. Staner, G. Viardot, A. Coatanhay, A. Muzet, R. Luthringer, J.-C. Roegel, J.-P. Macher. Microarousal and heart rate variability during human sleep. *Proc. Conference of the European Sleep Research Society,* Istambul, 2000.

[Sta94] L. Stanković. A method for time-frequency signal analysis. *IEEE Transactions on Signal Processing,* vol. 42, p. 225-229, 1994.

[Sta01] L. Stanković, I. Djurović. A note on *"An overview of aliasing errors in discrete-time formulations of time-frequency representations". IEEE Transactions on Signal Processing,* vol. 49, p. 257-259, 2001.

[Ste95] M. Stefik. *Knowledge Systems.* San Francisco : Morgan Kaufmann, 1995.

[Sto77] C. Stone. Consistent non parametric regression. *Annals of Statistics,* vol. 8, p. 1348-1360, 1977.

[Str99] J. Stradling, D. Pitson, C. Barbour, R. Davies. Variation in the arousal pattern after obstructive events in obstructive apnea. *American Journal of Respiratory and Critical Care Medecine,* vol. 159, p. 130-139, 1999.

[Van68] H. L. Van Trees. *Detection, Estimation, and Modulation Theory*, vol. 1. New York : John Wiley & Sons, 1968.

[Vap71] V. Vapnik, A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications,* vol. 16, p. 264-280, 1971.

[Vap74] V. Vapnik, A. Chervonenkis. *Theory of Pattern Recognition.* Moscou : Nauka, 1974.

[Vap82] V. Vapnik. *Estimation of Dependencies based on Empirical Data.* New York : Springer-Verlag, 1982.

[Vap89] V. Vapnik, A. Chervonenkis. The necessary and sufficient conditions for consistency of the method of empirical risk minimization (in Russian). *Yearbook of the Academy of Sciences of the USSR on Recognition, Classification and Forecasting,* vol. 2, p. 217-249, 1989.

[Vap95] V. Vapnik. *The Nature of Statistical Learning Theory.* New York : Springer-Verlag, 1995.

[Vay00] N. Vayatis. *Inégalités de Vapnik-Chervonenkis et Mesures de Complexité.* Thèse de Doctorat, Ecole Polytechnique, 2000.

[Via00] G. Viardot, A. Coatanhay, R. Lengellé, C. Richard, L. Staner, A. Muzet, J.-P. Macher. Fusion d'avis divergents d'experts. *Proc. Colloque AcM/MdA,* Grenoble, 2000.

[Via01] G. Viardot, R. Lengellé, C. Richard, A. Coatanhay. Fusion d'avis d'experts et caractérisation de l'expertise. Application à la détection de transitoires dans les signaux physiologiques. *Proc. Colloque GRETSI,* Toulouse, 2001.

[Via02(a)] G. Viardot. *Reconnaissance des Formes en présence d'Incertitudes sur l'Expertise. Application à la Détection des Phase d'Activation Transitoires du Sommeil chez l'Homme.* Thèse de Doctorat, Université de Technologie de Troyes, 2002.

[Via02(b)] G. Viardot, R. Lengellé, C. Richard. Mixture of experts for automated detection of spontaneous phasic arousals in sleep signals. *Proc. IEEE International Conference on Systems, Man and Cybernetics,* IEEE SMC'02, Hamammet, 2002.

[Vog99] C. C. Vogt, G. W. Cotrel. Fusion via a linear combination of scores. *Information Retrieval,* vol. 3, no. 1, p. 151-173, 1999.

[Was72] G. Wassel, J. Sklansky. Training a one-dimensional classifier to minimize the probability of error. *IEEE Transactions on Systems, Man and Cybernetics,* vol. 2, p. 533-541, 1972.

[Wau95] A. Wauquier, L. Aloe, A. Declerck. K-complexes : are they signs of arousal or sleep protective ? *Journal of Sleep Research,* vol. 4, p. 138-143, 1995.

[Wei96] J. Weickert. *Anisotropic diffusion in image processing.* Ph.D. thesis, University of Kaiserslautern, 1996.

[Wei97] J. Weickert. Review of nonlinear diffusion filtering. In B. ter Haar Romeny, L. Florack, J. Koenderink, M. Viergever, (éds). *Scale-Space Theory in Computer Vision, Lecture Notes in Computer Science,* vol. 1252, p. 3-28. Berlin : Springer, 1997.

[Wid88] B. Widrow, M. Hoff. Adaptive switching circuits. In J. A. Anderson, E. Rosenfeld, (éds). *Neurocomputing : Fundations of Research.* Cambridge : MIT Press, 1988.

[Zig00] D. Zighed, R. Rakotomolala. *Graphes d'Induction.* Paris : Hermès, 2000.

Première annexe

C. Richard. Time-frequency based detection using discrete-time discrete-frequency Wigner distributions. *IEEE Transactions on Signal Processing,* vol. 50, no. 9, p. 2170-2176, 2002.

Deuxième annexe

C. Richard, R. Lengellé, F. Abdallah. Bayes-optimal detectors using relevant second-order criteria. *IEEE Signal Processing Letters,* vol. 9, no. 1, 2002.

Troisième annexe

C. Richard. Linear redundancy of information carried by the discrete Wigner distribution. *IEEE Transactions on Signal Processing,* vol. 49, no. 11, p. 2536-2544, 2001.

Quatrième annexe

C. Richard, R. Lengellé. Data-driven design and complexity control of time-frequency detectors. *Signal Processing,* vol. 77, p. 37-48, 1999.

Résumé

Le présent mémoire propose une méthodologie pour la synthèse de détecteurs à structure imposée à partir d'une base d'exemples. Elle fait appel à la théorie des noyaux reproduisants pour l'élaboration de détecteurs linéaires généralisés dans des espaces transformés de dimension importante, voire infinie, sans qu'aucun calcul n'y soit effectué explicitement. Elle a également recours à l'optimisation du meilleur critère du second ordre pour le problème traité, après s'être assuré que de telles mesures de performance ne constituent en rien un obstacle dans la quête du rapport de vraisemblance. Pour une meilleure prise en compte de phénomènes tels que la malédiction de la dimensionnalité, l'approche proposée s'appuie de plus sur la théorie de l'apprentissage. Ceci lui permet d'offrir des garanties sur les performances en généralisation des détecteurs obtenus, qui sont alors en mesure de rivaliser avec les structures de décision phare du moment : les Support Vector Machines. Enfin, cet éclairage mêlant théories statistiques de la décision et de l'apprentissage donne un point de vue original sur la détection dans des espaces transformés particuliers tels que le plan temps-fréquence. La méthodologie proposée est finalement illustrée dans le cadre de la détection d'événements dans des signaux de sommeil.

Mots-clés: théories de la décision et de l'apprentissage, critères de contraste, noyaux reproduisants, analyse temps-fréquence, signaux de sommeil.