Étude d'algorithmes d'apprentissage artificiel pour la prédiction de la syncope chez l'homme

Mathieu Feuilloy

université angers

Année : 2009
N° d'ordre : 971

Étude d'algorithmes d'apprentissage artificiel pour la prédiction de la syncope chez l'homme

Thèse de doctorat

Spécialité : Informatique

École Doctorale STIM

Présentée et soutenue publiquement

Le 8 juillet 2009

À Angers

Par **Mathieu FEUILLOY**

Devant le jury ci-dessous :

Rapporteurs :	Guy Carrault,	Professeur à l'Université de Rennes 1
	Antoine Cornuéjols,	Professeur à AgroParisTech
Examinateurs :	Béatrice Duval,	Maître de Conférences à l'Université d'Angers
	Philippe Leray,	Professeur à l'Université de Nantes
Directeur de thèse :	Pascal Nicolas,	Professeur à l'Université d'Angers
Co-encadrant :	Daniel Schang,	Enseignant-chercheur au Groupe ESEO

Laboratoire d'Étude et de Recherche en Informatique d'Angers
2, Bd Lavoisier, 49045 Angers Cedex 01

Remerciements

Je dédicace cette thèse à

Table des matières

Notations mathématiques **1**

Acronymes **2**

Introduction Générale **3**

Partie I État de l'art

1 Apprentissage artificiel et discrimination **11**
1.1 Introduction . . . 11
1.1.1 Apprentissage artificiel . . . 11
1.1.2 Reconnaissance de formes . . . 13
1.1.3 Introduction à la discrimination . . . 14
1.2 Approches probabilistes pour la classification . . . 16
1.2.1 Introduction . . . 16
1.2.2 Vision bayésienne . . . 16
1.2.3 Estimation des densités de probabilité . . . 18
1.2.4 Classifieurs basés sur la théorie de Bayes . . . 19
1.2.5 Conclusions . . . 22
1.3 Classification Linéaire . . . 22
1.3.1 Introduction . . . 22
1.3.2 Problème à deux classes . . . 23
1.3.3 Extension de la discrimination à plus de deux classes . . . 35
1.3.4 Conclusions . . . 37
1.4 Classification non linéaire . . . 38
1.4.1 Introduction . . . 38
1.4.2 Réseaux de neurones artificiels . . . 38
1.4.3 *Support vector machines* non linéaires . . . 53
1.4.4 Conclusions . . . 56
1.5 Résumé et discussions . . . 56

2 Réduction de la dimensionnalité **59**
2.1 Introduction . . . 59
2.2 Prétraitement . . . 60
2.2.1 Introduction . . . 60

2.2.2 Données aberrantes (*outliers*) . . . 61
2.2.3 Normalisation des données . . . 62
2.2.4 Données manquantes . . . 63
2.2.5 Conclusions . . . 64
2.3 Extraction de caractéristiques . . . 65
2.3.1 Introduction . . . 65
2.3.2 Approches linéaires . . . 65
2.3.3 Approches non linéaires pour la réduction de la dimensionnalité . . . 71
2.3.4 Conclusions . . . 76
2.4 Sélection de variables . . . 77
2.4.1 Introduction . . . 77
2.4.2 Critères d'évaluation . . . 79
2.4.3 Génération de sous-ensembles : procédures de recherche . . . 85
2.4.4 Conclusions . . . 96
2.5 Résumé et discussions . . . 97

3 Évaluation et comparaison de modèles 99
3.1 Introduction . . . 99
3.2 Mesures de la qualité d'un modèle . . . 100
3.3 Évaluation de la performance . . . 101
3.3.1 Facteurs influençant la généralisation . . . 101
3.3.2 Méthodes d'estimation . . . 105
3.3.3 Intervalle de confiance . . . 106
3.4 Mesures de performance d'un test diagnostique . . . 108
3.4.1 Indices de performance . . . 108
3.4.2 Courbes de ROC . . . 112
3.5 Comparaisons de modèles par analyse des courbes de ROC . . . 113
3.6 Conclusions . . . 114

Partie II Contributions

4 Problématique étudiée : la prédiction de la syncope 117
4.1 Introduction . . . 117
4.2 Investigations et démarches diagnostiques . . . 119
4.2.1 Test d'inclinaison : *Head-Upright Tilt-Test* . . . 120
4.2.2 Signaux de mesures : électrocardiogramme et signal d'impédancemétrie thoracique . . . 121
4.3 État de l'art sur la prédiction de la syncope . . . 124
4.4 Conclusions . . . 125

5 Études expérimentales pour la prédiction de la syncope 127
5.1 Introduction . . . 127
5.2 Cadres expérimentaux . . . 127
5.3 Recherche d'indices prédictifs du résultat du *tilt-test* durant la période de repos . . . 131
5.3.1 Analyse exhaustive des sous-ensembles pertinents de variables initiales . . . 131
5.3.2 Extraction de caractéristiques pertinentes par combinaison des variables initiales . . . 136

5.3.3 Discussions et conclusions . . . 143
5.4 Recherche d'indices prédictifs du résultat du *tilt-test* durant les deux périodes de l'examen : couchée et basculée . . . 146
5.4.1 Introduction . . . 146
5.4.2 Méthodes . . . 148
5.4.3 Expérimentations et résultats . . . 150
5.4.4 Discussions et conclusions . . . 155
5.5 Évaluation de la pertinence du signal d'impédancemétrie thoracique dans la prédiction de la syncope . . . 158
5.5.1 Introduction . . . 158
5.5.2 Prétraitement et extraction des complexes *dZ* . . . 159
5.5.3 Sélection des complexes par minimisation de l'erreur quadratique moyenne totale . . . 161
5.5.4 Sélection des complexes par optimisation manuelle du rapport signal sur bruit et évaluation de nouvelles caractéristiques prédictives . . . 162
5.5.5 Amélioration du processus de sélection des complexes par optimisation automatique du rapport signal sur bruit . . . 165
5.5.6 Discussions et conclusions . . . 177
5.6 Discussions . . . 179

6 Nouvelle approche pour l'extraction d'informations et l'interprétation des méthodes de projection non linéaire 183
6.1 Introduction . . . 183
6.2 Fondements . . . 184
6.3 Extraction de la contribution des variables dans le cas d'une analyse en composantes principales . . . 185
6.3.1 Description de la méthode . . . 185
6.3.2 Validation expérimentale . . . 187
6.4 Extraction de la contribution des variables dans le cas d'une réduction de dimension non linéaire . . . 187
6.4.1 Introduction . . . 187
6.4.2 Evaluation et vérification de la projection par estimation des « pseudo-valeurs propres » . . . 188
6.4.3 Estimation des « vecteurs propres locaux » . . . 189
6.4.4 Extension de la procédure d'estimation de la qualité de la représentation aux différentes méthodes de projections . . . 192
6.4.5 Validation expérimentale . . . 192
6.4.6 Renforcement de la pertinence de la lecture des estimations des contributions des variables dans les composantes non linéaires . . . 192
6.5 Application expérimentale à la prédiction de la syncope . . . 195
6.5.1 Introduction . . . 195
6.5.2 Rappel du contexte . . . 196
6.5.3 Extraction de la contribution des composantes curvilignes liées à la prédiction de la syncope . . . 197
6.6 Discussions et conclusions . . . 198

Conclusion Générale **201**

Annexes **207**

A Compléments mathématiques **209**
A.1 Algorithme de rétropropagation 209
A.2 Estimation des pentes et des aires sur le signal d'impédancemétrie thoracique et sa dérivée 212
A.2.1 Détermination des pentes ($Slope_{norm}$) 212
A.2.2 Détermination des aires ($Area_{norm}$) 212
A.3 Estimation de la probabilité d'erreur de classification 213

B Exemples illustratifs de l'apprentissage du OU-exclusif **217**
B.1 Résolution du problème *XOR* par des réseaux de neurones 217
B.2 Résolution du problème *XOR* par un réseau RBF 220
B.3 Résolution du problème *XOR* par les SVM 221

C Illustration de la pertinence des indices mesurant la performance d'un modèle **225**
C.1 Influence du déséquilibre entre les classes : prévalence de la maladie 225
C.2 Construction de la courbe de ROC 227

D Méthodes de sélection de variables combinant sélection séquentielle et algorithmes génétiques **229**
D.1 Rappel de l'approche pour la sélection de variables 229
D.2 Cadres expérimental et méthodologique 231
D.3 Résultats expérimentaux 232
D.3.1 Analyses des résultats 240
D.4 Discussions et conclusions 243

Liste des figures **245**

Liste des tables **251**

Liste des algorithmes **255**

Index **257**

Bibliographie commentée **261**

Références bibliographiques **263**

Résumé / Abstract **278**

Notations mathématiques

Dans ce manuscrit, nous adoptons la notation suivante, de manière à harmoniser l'écriture avec la plupart des références citées dans le domaine de l'apprentissage machine.

Les caractères sans gras représentent un scalaire, comme n. Les caractères minuscules en gras sont des vecteurs **colonnes**, comme $\mathbf{b}$ et les caractères majuscules en gras sont des matrices, comme $\mathbf{C}$. Par conséquent, la notation $[b_1, \ldots, b_n]$ indique un vecteur ligne de n éléments. Enfin, le vecteur transposé est noté $\mathbf{b} = (b_1, \ldots, b_n)^T$. D'autre part, en considérant une matrice $\mathbf{C}$, nous noterons c_{ij}, l'élément de la i-ème ligne et de la j-ème colonne de $\mathbf{C}$.

Les ensembles et sous-ensembles de données se noteront $\mathcal{X}$, et les ensembles et sous-ensembles de variables se noteront $\langle SS \rangle$. La combinaison de q éléments parmi p se notera C_p^q.

La liste ci-dessous récapitule les principales notations utilisées dans ce manuscrit :

n	nombre d'observations
p	nombre de variables
q	nombre de variables/caractéristiques sélectionnées/générées
$p(\mathbf{x})$	densité de probabilité
$P(\mathbf{x})$	probabilité
$\mathcal{X}$	ensemble des données
$\mathcal{X}_A$	sous-ensemble de données d'apprentissage
$\mathcal{X}_V$	sous-ensemble de données de validation
$\mathcal{X}_T$	sous-ensemble de données de test
$\mathcal{SV}$	ensemble des vecteurs de support
$\langle FS \rangle$	ensemble des variables
$\langle SS \rangle$	sous-ensemble de variables

Acronymes

ACC	Analyse en Composantes Curvilignes
ACP	Analyse en Composantes Principales
AFD	Analyse Factorielle Discriminante
AG	Algorithmes Génétiques
ANOVA	*ANalysis Of VAriance*
AUC	*Area Under the ROC Curves*
B&B	*Branch and Bound*
BFGS	*Broyden-Fletcher-Goldfard-Shanno*
bpm	battements par minute
CP	Composante Principale
EQM	Erreur Quadratique Moyenne
EQMT	Erreur Quadratique Moyenne Totale
FDL	Fonction Discriminante Linéaire
FDQ	Fonction Discriminante Quadratique
FDR	*Fisher Discriminant Ratio*
IC	Intervalle de Confiance
k-ppv	k-plus proches voisins
KDD	*Knowledge Discovery Data*
LLE	*Locally Linear Embedding*
LRS	*plus-L minus-R Selection*
GROG	Groupes Régionaux d'Observation de la Grippe
GTM	*Generative Topographic Mapping*
HUTT	*Head-Upright Tilt-Test*
MDS	*Multidimensional Scaling*
NLM	*Nonlinear Mapping*
OBD	*Optimal Brain Damage*
OBS	*Optimal Brain Surgeon*
PMC	Perceptron Multicouches
DSP	Densité spectrale de puissance
RBF	*Radial Basis Function*
RdF	Reconnaissance de Formes
RGSS	*Random Generation plus Sequential Selection*
ROC	*Receiver Operating Characteristic*
RNA	Réseaux de Neurones Artificiels
RSB	Rapport Signal sur Bruit
RV	Rapport de Vraisemblance (positif : RV^+ et négatif : RV^-)
SBiS	Sélection bidirectionnelle séquentielle – *Sequential Bidirectional Selection*
SBS	Sélection descendante séquentielle – *Sequential Backward Selection*
SFBS	Sélection descendante flottante séquentielle – *Sequential Floating Backward Selection*
SFFS	Sélection ascendante flottante séquentielle – *Sequential Floating Forward Selection*
SFS	Sélection ascendante séquentielle – *Sequential Forward Selection*
SOM	*Self-Organizing Map*
SVM	*Support Vector Machine*
TEVG	Temps d'Éjection Ventriculaire Gauche
XOR	OU-exclusif – *eXclusive OR*

Introduction Générale

Les avancées technologiques ont facilité l'acquisition et le recueil de nombreuses données, notamment dans le domaine médical lors d'examens de patients. Ces données peuvent alors être utilisées comme support de décision médicale, conduisant aux développements d'outils capables de les analyser et de les traiter ; dans la littérature, nous trouvons régulièrement la notion « d'aide à la décision ». Depuis de nombreuses années maintenant, l'aide au diagnostic médical s'est développée et a gagné en popularité ; ces systèmes sont même considérés comme étant essentiels dans beaucoup de disciplines médicales [Miller, 1994; Coiera, 2003]. En pratique, il existe déjà de nombreuses applications qui permettent d'assister les cliniciens dans leurs démarches diagnostiques [Huguier and Flahault, 2003]. Par ailleurs, les systèmes reposant sur les techniques issues de l'apprentissage artificiel sont de plus en plus élaborés [Coiera, 2003][1]. Ils permettent par exemple, d'assister les médecins dans la surveillance d'un monitoring cardiaque, en générant des alertes suite à la détection d'un trouble du rythme cardiaque [Pan and Tompkins, 1985; Karsai and Sztipanovits, 1999; Christov, 2004; Dubois, 2004]. L'aboutissement de ce type de processus, passe nécessairement par une série d'étapes, qui réunies, forment ce que nous pouvons appeler un processus de reconnaissance de formes (RdF). Originaire de l'ingénierie, de par ses applications liées à la reconnaissance d'images et de la parole, la RdF s'est développée au travers de l'informatique depuis ces vingt dernières années et les techniques qui en résultent, sont aujourd'hui employées dans des domaines très variés. La démarche classique d'un système de RdF consiste à opérer suivant le schéma de la figure 1, introduit par [Belaïd and Belaïd, 1992].

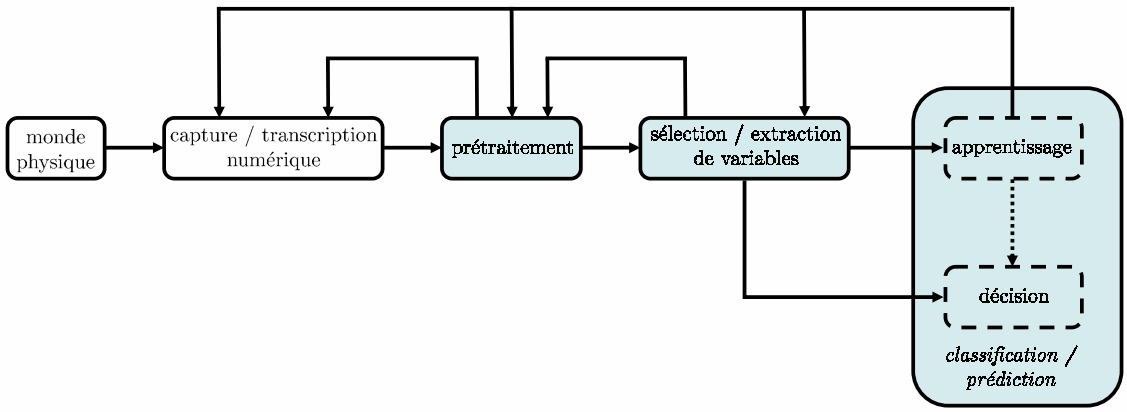


FIG. 1 – Schéma général d'un système de reconnaissance de formes [Belaïd and Belaïd, 1992].

Ce type de processus utilise des méthodes issues de l'apprentissage artificiel, appelé parfois « apprentissage machine » en référence à l'appellation anglaise *machine learning*. [Cornuéjols and Miclet, 2002] définissent l'apprentissage artificiel comme un ensemble de méthodes permettant d'extraire des connaissances à partir d'observations disponibles et de les utiliser afin de chercher de nouvelles informations, ou de décrire différemment les observations. Ainsi, appliqué à la RdF, l'apprentissage artificiel utilise un ensemble d'observations reflétant le comportement d'un processus ou d'un phénomène, afin d'en extraire des règles ou des modèles, de manière à permettre la prise de décision « automatique », c'est-à-dire, sans l'aide d'un opérateur. Parmi les différents types d'apprentissage, l'apprentissage supervisé s'adresse certainement à la majorité des applications, où chaque observation est présentée sous forme d'un couple (**observation**, **étiquette**).

[1] E. Coiera propose de nombreux exemples à l'adresse suivante http ://www.coiera.com/ailist/list.html .

Dans une tâche de discrimination, les étiquettes sont nommées les **classes** et l'objectif de l'apprentissage est de construire un modèle ou un ensemble de règles permettant d'affecter pour une nouvelle observation, la classe parmi toutes les classes préalablement définies.

Comme le montre la figure 1, un processus de RdF nécessite l'élaboration de plusieurs tâches, dont :

- le **prétraitement**, qui consiste à préparer les données avant leur analyse ;
- l'**extraction** et la **sélection de variables**, qui consistent à « fouiller » dans l'ensemble de données, afin d'en sélectionner et d'en extraire l'information pertinente ;
- l'**apprentissage**, qui consiste à éclairer la **décision** à l'aide de connaissances *a priori*, en générant des règles.

Ces tâches sont incontournables, et ont donc été largement étudiées dans nos travaux de thèse. Notons également que la tâche d'**extraction/sélection de variables** est commune aux processus de fouille de données qui, en partageant des méthodes de la RdF, correspond à l'autre grand axe de l'apprentissage artificiel.

Problématique

Nos travaux se sont portés sur la prédiction de la syncope chez l'homme, et plus précisément sur des patients ayant des épisodes de syncopes fréquents et inexpliqués. La syncope est un terme médical désignant l'évanouissement. Elle se caractérise par une perte subite et brève de connaissance et du tonus postural, suivie par un retour spontané à un état de conscience normal [Kapoor, 2000]. Une apparition isolée d'une syncope ne constitue pas nécessairement un problème, mais peut le devenir lorsque les épisodes d'évanouissement sont répétés. En effet, la syncope cause rarement de graves traumatismes, cependant les préoccupations et les angoisses liées aux risques traumatiques et à la crainte de récidive peuvent influer sur la qualité de vie des patients. Aussi, les syncopes inexpliquées sont les plus handicapantes, car l'absence de diagnostic empêche de prescrire un traitement approprié pour prévenir l'apparition des symptômes.

Malgré les nombreux examens effectués, il arrive que l'origine de la syncope ne soit pas clairement identifiée. Dans ce cas, et, lorsque les épisodes sont répétés, le patient peut être amené à réaliser le test de la table d'inclinaison [Benditt *et al.*, 1996]. Cet examen, appelé *tilt-test*, est une méthode reconnue pour recréer les conditions dans lesquelles le patient ressent les symptômes. Effectué à jeun, ce test débute par une période de repos (environ 10 minutes) où le patient doit rester allongé sur la table d'examen en position horizontale ; cette période stabilise les mesures à recueillir. Suite à la phase de stabilisation et sous l'action d'un moteur électrique, la table s'incline à un angle compris entre 60° et 80° pendant une durée pouvant atteindre 45 minutes. Ainsi, le passage brutal de la position allongée à la position debout peut provoquer l'apparition des symptômes de la syncope, et dans ce cas, le test est considéré comme positif.

En provoquant le mécanisme des malaises, ce test permet d'apporter des éléments de réponse afin d'adapter le traitement du patient. Cependant, le principal problème de ce test est sa durée. En effet, du personnel médical est monopolisé pendant près d'une heure si aucun symptôme n'apparaît. Dès lors, pour des raisons de coût et de bien-être des patients, il paraît important de pouvoir réduire la durée de l'examen. C'est dans cet objectif que s'inscrivent les études réalisées dans ces recherches, qui tentent de prédire l'apparition des signes de la syncope avant que l'examen n'arrive à son terme, et éviter ainsi aux patients de ressentir les symptômes.

Motivations et contexte

La problématique qui nous a été proposée porte sur la prédiction du résultat de l'examen du *tilt-test* pour des patients sujets à l'apparition récurrente de syncopes inexpliquées. Dans le cadre de ce travail, l'objectif des médecins est d'obtenir avant la fin de l'examen du *tilt-test*, un verdict quant à son résultat : présence ou non de symptômes liés à l'apparition de syncopes. La réduction de la durée de l'examen est, comme évoquée précédemment, envisagée pour des raisons de coût et de bien-être des patients. Aussi, les syncopes inexpliquées laissent de nombreuses questions sans réponses sur les causes, les facteurs ou les éléments qui interviennent dans l'apparition des symptômes.

Les réalisations effectuées dans le cadre de cette thèse sont fondées sur des études menées au service de cardiologie du CHU d'Angers. La mise en œuvre de nos travaux s'est heurtée à plusieurs difficultés, dues en partie aux particularités du milieu médical. En effet, en menant plusieurs études, les médecins ont pu recueillir un nombre important de variables sur un ensemble de patients qui est considéré par le milieu médical comme étant relativement conséquent[2]. Cependant, en apprentissage artificiel, le nombre d'observations (dans notre cas les patients) est essentiel et doit être le plus important possible, car il influence considérablement la précision des performances de discrimination. Dès lors, l'un des enjeux de cette thèse est de construire des modèles de discrimination efficaces à partir d'un nombre de données limitées.

C'est aussi dans ce contexte particulier, où le nombre de données est limité, que la tâche d'extraction et de sélection de variables agit considérablement sur les performances finales du modèle de discrimination. Aussi, [Dreyfus *et al.*, 2002] recommandent d'élaborer des modèles les plus parcimonieux possibles, entraînant forcément une réduction de la dimensionnalité du problème. Cette observation nous a poussé à étudier particulièrement cette tâche d'extraction et de sélection de variables.

Contributions

Les travaux effectués ont donné lieu à plusieurs contributions. Celles-ci peuvent être rattachées aux études expérimentales, liées à la prédiction du résultat du *tilt-test* ou aux propositions théoriques et algorithmiques, qui découlèrent des observations expérimentales. En effet, certaines difficultés rencontrées ont nécessité le développement de nouvelles techniques.

Dans un premier temps, nous avons comparé l'efficacité de nombreuses méthodes disponibles dans la littérature, pour effectuer les tâches de sélection/extraction de variables et de discrimination. Le processus d'expérimentation utilisé a permis d'évaluer l'efficacité, et l'influence du choix des méthodes pour chaque tâche, de manière indépendante et complémentaire. Ces études réalisées lors des phases de repos et d'inclinaison du *tilt-test* ont permis d'observer plusieurs variables pertinentes, notamment celles fondées sur un signal physiologique (signal d'impédancemétrie thoracique[3]). Ce dernier, bien que peu utilisé dans la prise en charge courante des patients, s'est révélé porteur d'informations utiles quant à la prédiction de la syncope. C'est ainsi que nous avons développé un traitement particulier pour ce signal, assurant une extraction efficace de l'information permettant de déterminer précocement le résultat du *tilt-test*. Les performances de

[2]Notons que dans le milieu médical, les « échantillons » de patients participant à des études cliniques dépassent rarement la centaine, et dans de tels cas, ils sont considérés comme très satisfaisants.

[3]Le signal d'impédancemétrie thoracique est basé sur la mesure de l'impédance électrique du thorax, afin d'analyser, durant chaque battement, l'hémodynamique cardiaque. À l'image de l'électrocardiogramme, ce signal permet d'évaluer le volume d'éjection systolique durant chaque cycle cardiaque et donc de déterminer le débit cardiaque.

discrimination obtenues par les modèles élaborés tout au long de ces travaux ont pu être très favorablement comparés aux performances d'autres études. Aussi, dans un souci d'apporter de la clarté dans la description et l'interprétation de nos modèles de discrimination, nous avons usé de méthodes couramment employées en fouille de données. Nos bonnes performances ont pu être obtenues par le recours à une analyse minutieuse des ensembles de données. Ces bons résultats ont naturellement donné lieu à des contributions, dont les deux principales sont liées à la tâche de sélection/extraction de variables.

La première contribution est apparue lors de la comparaison de nombreuses méthodes de sélection de variables dans les études expérimentales, qui a révélé la difficulté de ces méthodes à réaliser une sélection efficace et rapide. En d'autres termes, il est apparu globalement que ces méthodes possèdent un compromis rapidité/efficacité loin d'être satisfaisant. Parmi ces méthodes, les approches heuristiques pour la sélection de type séquentiel ont retenu notre attention, car il est apparu que ce compromis pouvait être fortement amélioré. Dès lors, afin d'enrichir les méthodes de sélection séquentielle, nous les avons combinées à une méthode non déterministe, en l'occurrence les algorithmes génétiques. Les observations faites, une fois ces nouvelles méthodes comparées à celles de la littérature, ont permis de constater leur grande capacité à sélectionner des sous-ensembles de variables optimisant la qualité de la discrimination, surpassant les performances des meilleures méthodes de sélection. En outre, cette combinaison de méthodes accroît leurs performances et réduit considérablement le nombre de variables sélectionnées, tout en minimisant le coût calculatoire. Ceci améliore, par conséquent, le compromis rapidité/efficacité. Aussi, cette approche de sélection de variables, fondée sur une combinaison de techniques heuristiques et non déterministes, a montré sa reproductibilité lors d'expérimentations sur d'autres ensembles de données.

La seconde contribution repose sur une démarche plus théorique, elle est apparue au cours de nos recherches sur l'utilisation des processus de projection non linéaire. En effet, nous avons pu constater la difficulté pour interpréter ces processus, et plus particulièrement les composantes résultantes de ces projections. Notons que l'utilisation d'une de ces méthodes (l'analyse en composantes curvilignes, ACC) a permis d'obtenir un modèle performant, capable de prédire efficacement l'apparition des symptômes de la syncope durant la phase de repos du *tilt-test*. Or, en l'état actuel des choses, il n'est pas possible de définir la nature de ces composantes non linéaires. En effet, contrairement à son homologue linéaire (l'analyse en composantes principales, ACP), l'ACC, et par extension la plupart des méthodes de projection non linéaire, ne permet pas de lier de manière analytique, et donc significativement, les composantes non linéaires aux variables initiales. Comme le précise [Saporta, 2006], ce point est fondamental, car lier une composante aux variables initiales permet de trouver et de donner une signification à cette composante. C'est dans ce contexte que nous avons développé une méthodologie permettant d'enrichir les méthodes de réduction non linéaire, en leur donnant la capacité d'extraire la qualité de la représentation des variables dans les composantes non linéaires. Notre approche d'extraction d'informations est fondée sur une reproductibilité et une adaptation de la technique utilisée pour l'interprétation des composantes principales. Nous sommes partis du principe que l'ACC est une extension non linéaire de l'ACP, comme le souligne son auteur en la décrivant comme une ACP par parties [Dreyfus *et al.*, 2002]. D'autre part, le fait que la plupart des méthodes de projection non linéaire sont fondées sur une même notion, qui est la préservation locale de la topologie ou de la structure des données, nous a amené à penser que le processus d'estimation de la représentation des variables dans les composantes développé pour l'ACC peut s'adapter à d'autres techniques de projection non linéaire. Cela s'est vérifié lors d'une validation expérimentale sur des données synthétiques.

Organisation

Ce manuscrit est structuré en deux parties : la première concerne l'**état de l'art** et la seconde décrit nos **contributions**. L'état de l'art est organisé à l'image de la figure 1 où les différentes tâches nécessaires pour élaborer un outil de discrimination ont été détaillées. Ainsi, le **chapitre 1** fait référence à l'étape de classification/prédiction de la figure 1. Il présente différentes méthodes de discrimination capables d'être employées pour classer nos patients réalisant le *tilt-test*. Le **chapitre 2** fait référence aux étapes de prétraitement et de sélection/extraction de variables de la figure 1. Celui-ci présente les méthodes permettant de définir les entrées du modèle de discrimination et les techniques employées pour extraire et obtenir l'information pertinente liée à notre problématique. Enfin, le dernier chapitre de l'état de l'art (**chapitre 3**) donne des indications sur la méthodologie pour évaluer et comparer les modèles construits, afin de sélectionner le plus approprié pour résoudre notre problème.

La deuxième partie expose nos contributions. Elle est également divisée en plusieurs chapitres. Le **chapitre 4** présente avec soin la problématique étudiée, en donnant un bref état de l'art du domaine de la prédiction de la syncope lors de l'examen du *tilt-test*. Le **chapitre 5** aborde dans le détail les études réalisées pour prédire le résultat du *tilt-test*, en fonction des différentes approches abordées dans l'état de l'art. Le dernier chapitre (**chapitre 6**) présente une nouvelle méthode capable d'extraire l'information et d'interpréter le résultat obtenu par des méthodes de projection non linéaire.

La figure 2 propose une démarche de lecture de cette thèse et indique les différentes connexions entre les chapitres.

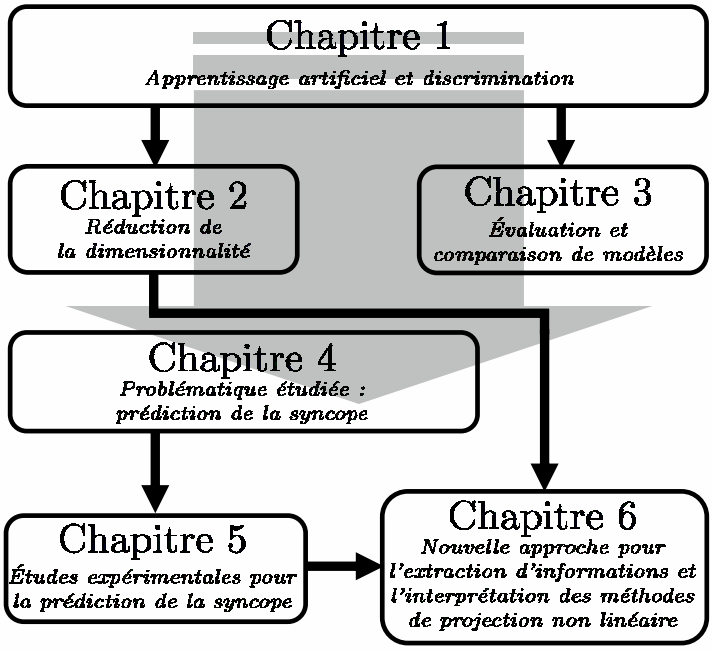


FIG. 2 – Organigramme de lecture possible de cette thèse.

Première partie

État de l'art

Chapitre 1

Apprentissage artificiel et discrimination

1.1 Introduction

Ce premier chapitre va permettre, dans un premier temps, d'apporter des précisions sur l'apprentissage artificiel, tout en faisant le lien avec notre problématique. Dans un second temps, aux sections 1.2, 1.3 et 1.4, les méthodes de discrimination nécessaires à la résolution de notre problème seront détaillées.

1.1.1 Apprentissage artificiel

Le diagnostic médical, ou plutôt, l'**aide au diagnostic** est réalisée par des outils fondés soit sur des modèles, soit sur des données. Concernant les outils fondés sur des modèles, la complexité de ces derniers peut être si importante que la modélisation devient impossible. La mise en œuvre des outils doit alors être réalisée à partir de données. La littérature nous invite, peut être de manière un peu abusive, à employer le terme de **modèle** pour définir ces outils. En effet, de nombreux ouvrages rattachent ce terme aussi bien au résultat d'une **modélisation**, qu'à l'aboutissement d'un **apprentissage**, allant jusqu'à employer ces deux termes comme des synonymes. On notera simplement que la modélisation suppose d'approcher une réalité de manière très précise, par exemple sous une forme « analytique », tandis que l'apprentissage implique de passer par une série d'ajustements et de tests à partir d'un ensemble de données. En dépit de ces remarques, nous emploierons dans ce manuscrit, le terme « modèle » comme le résultat d'un apprentissage.

Ce dernier point introduit alors l'apprentissage artificiel, défini par [Cornuéjols and Miclet, 2002], comme une notion englobant toute méthode permettant de construire un modèle réel à partir d'un ensemble de données soit en améliorant un modèle partiel (ou moins général), soit en créant complètement le modèle. La popularité croissante de l'apprentissage artificiel est certainement due à son approche multidisciplinaire. En effet, de part la diversité des outils produits et des problèmes traités, l'apprentissage artificiel se trouve au carrefour de nombreuses disciplines, comme le montre la figure 1.1 reprise du site internet de l'équipe de recherche « Équipe Apprentissage Machine[1] », du laboratoire LEIBNIZ[2].

Bien que la liste des domaines cités sur la figure 1.1 ne soit pas exhaustive, nous pouvons tout de même observer une diversité singulière, avec laquelle l'apprentissage artificiel a pu, depuis une cinquantaine d'années, se développer et s'exporter dans de très nombreuses disciplines. Les

[1] http ://www-leibniz.imag.fr/Apprentissage/Apprentissage-intro.html
[2] http ://www-leibniz.imag.fr/

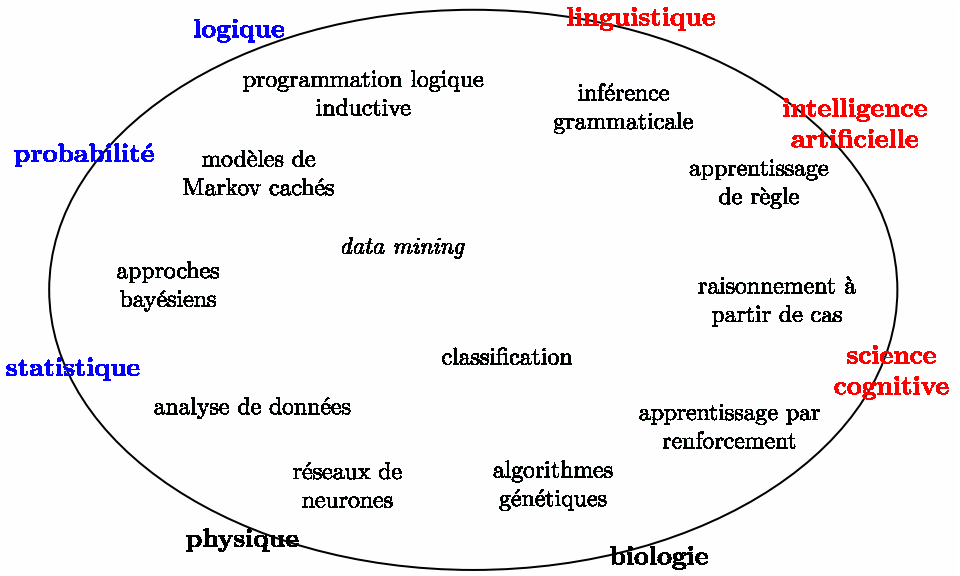


FIG. 1.1 – Illustration des domaines scientifiques apparentés à l'apprentissage artificiel.

méthodes issues de l'apprentissage artificiel se regroupent habituellement en deux thématiques : la **reconnaissance de formes** [Nagy, 1968; Jain *et al.*, 2000; Bishop, 2006] et la **fouille de données** [Witten and Frank, 2005; Han and Kamber, 2006; Tufféry, 2007].

Historiquement, il y a une soixantaine d'années, la reconnaissance de formes (RdF) est apparue grâce aux méthodes de l'apprentissage artificiel, dans des domaines tels que la reconnaissance d'images et de la parole. De nos jours, la variété des domaines d'application fait que cet axe est toujours aussi actif. Aussi, au début des années 1990, face à l'explosion croissante des capacités de calcul et de stockage des ordinateurs, de nouvelles problématiques se sont posées aux chercheurs et industriels. Ces derniers voyaient en effet un intérêt économique à analyser de grands ensembles de données[3]. Ainsi, est apparue la fouille de données (*data mining*), appelée aussi **extraction de connaissances** (KDD : *Knowledge Discovery Data*). Celle-ci consiste à rechercher et à extraire de l'information au sein de gros ensembles de données. Comme le définissent [Cornuéjols and Miclet, 2002], la fouille de données prend en charge l'ensemble du processus d'extraction de connaissances. Cependant, certains ouvrages [Witten and Frank, 2005; Han and Kamber, 2006], essentiellement anglo-saxons, proposent une autre vision de la fouille de données, en la considérant comme une étape essentielle du processus d'extraction de connaissances, ainsi que le montre le schéma de la figure 1.2.

Les processus de fouille de données et de reconnaissance de formes montre peu de différences, seul l'objectif diffère. En effet, la fouille de données est apparue, afin de permettre l'analyse de données de n'importe quelle nature, et d'offrir une visualisation interactive des données. Ainsi, dans un cadre médical, la fouille de données permet aux différents observatoires du GROG[4] (Groupes Régionaux d'Observation de la Grippe) de surveiller et de visualiser chaque hiver la progression de la grippe. Tandis qu'un outil développé suivant le processus de reconnaissance de formes permettrait quant à lui de définir si un patient est atteint ou non de la grippe. Cette dernière illustration du processus de reconnaissance de formes révèle parfaitement le cadre de notre étude, qui rappelons-le, doit permettre de déterminer pour un patient le résultat de l'examen du *tilt-test*.

[3]La littérature cite fréquemment l'exemple concernant les organismes de crédit, qui utilisent des « scores » de risque pour proposer le montant de crédit le plus adapté à chaque client, en fonction notamment du profil du demandeur et de sa demande.

[4]http ://www.grog.org

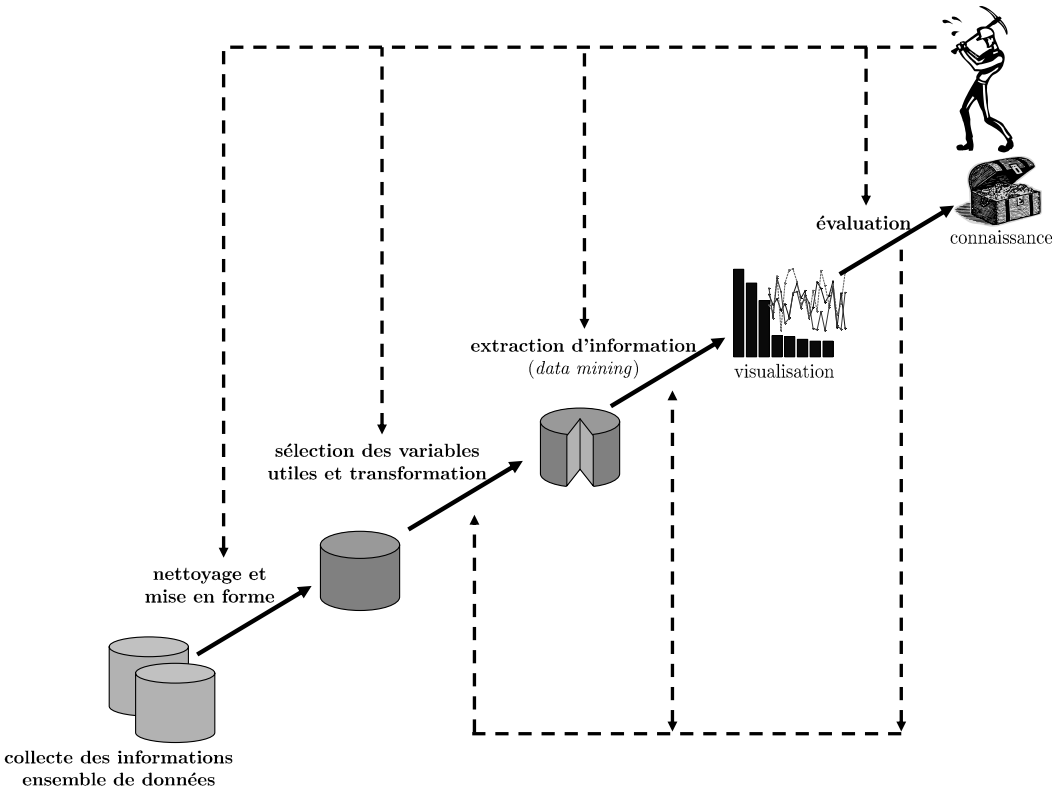


FIG. 1.2 – Schéma d'un processus de fouille de données [Han and Kamber, 2006].

1.1.2 Reconnaissance de formes

Parallèlement à l'apprentissage artificiel, la reconnaissance de formes s'est fortement développée au cours des vingt dernières années, en raison notamment des progrès de l'informatique [Bishop, 2006]. L'une des vocations de la RdF est le traitement de données. Elle se fonde sur des connaissances *a priori* du problème ou sur de l'information tirée d'un ensemble de mesures et d'observations. Ce principe introduit prématurément la notion d'induction, que nous verrons à la section 1.1.3. Une observation, appelée aussi exemple, correspond à la description d'un épisode relatif à un événement. Par exemple, dans le cadre d'un diagnostic médical, une observation peut être représentée par les différentes mesures et symptômes observés sur un patient, tels que le rythme cardiaque, la présence de fièvre. De manière générale, ces mesures sont appelées **variables**, ou encore attributs, caractéristiques ou descripteurs. Celles-ci sont donc les entrées du processus et la sortie correspond alors à l'événement. Ainsi, les données manipulées sont habituellement présentées sous forme de couples (**observation**, **étiquette**).

La figure 1.3 rappelle le schéma général d'un processus de reconnaissance de formes, dans lequel le **monde physique** est un espace de *dimension infini* où les observations (les formes) sont représentées par une multitude de propriétés. Ces caractéristiques sont, lorsque cela est possible, converties en données numériques lors de la phase de **capture/transcription numérique**. Le **prétraitement** consiste à éliminer l'information nuisible à l'analyse des données, de manière à obtenir une meilleure homogénéité inhérente à chaque variable et entre les variables. L'étape de **sélection/extraction de variables** a pour objectif de conserver l'information pertinente liée à la problématique traitée, impliquant une réduction de la dimensionnalité de l'espace représentant les observations. Jusque là, ces étapes semblent être communes à celles du processus de fouille de données présenté à la figure 1.2. Aussi, comme évoqué précédemment, c'est par l'objectif à

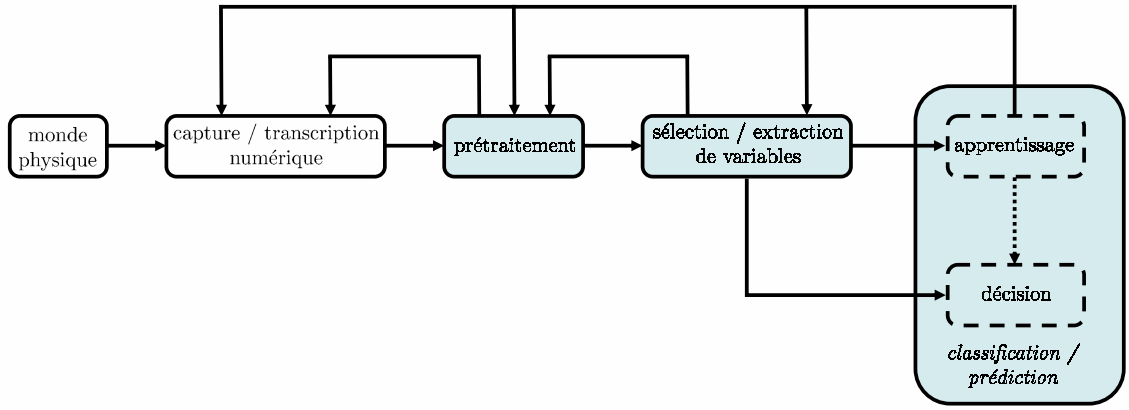


FIG. 1.3 – Schéma général d'un système de reconnaissance de formes [Belaïd and Belaïd, 1992].

atteindre que les deux processus se différencient. La RdF souhaite au final obtenir un outil de **régression**, de **classification** ou de **discrimination**. La nature de cet outil dépend avant tout de la forme du résultat à obtenir. En effet, dans le cas où la sortie est inconnue, l'objectif est de regrouper des sous-ensembles de données en classes, ou plus précisément en *clusters* : on parle alors de **classification**, ou de **classification automatique**. Lorsque la sortie est connue, le processus permet d'obtenir un outil de **régression** si la sortie est quantitative (numérique), ou un outil de **discrimination** si la sortie est qualitative. Notons que dans la littérature, le terme de « discrimination » est parfois remplacé par celui de « classification », et dans ce cas, celui de « classification » par « classification automatique ». Aussi, dans notre manuscrit, où nous manipulons des données étiquetées, les termes de discrimination et de classification seront employés sans distinction.

Ces différents outils aboutissent à la dernière phase (notée classification/prédiction sur la figure 1.3) qui regroupe les deux tâches d'**apprentissage**[5] et de **décision**. Celles-ci jouent des rôles assez proches. Dans le cadre de la discrimination, l'apprentissage et la décision tentent tous les deux d'attribuer une observation à une classe de référence. Ainsi, le résultat de l'apprentissage est soit la réorganisation ou le renforcement des classes existantes en tenant compte de l'apport de la nouvelle observation, soit la création d'une nouvelle classe représentant la nouvelle observation. L'apprentissage se charge alors d'acquérir la connaissance et de l'organiser en classes de référence. Tandis que la décision donne un « avis » sur l'appartenance ou non de l'observation aux classes préalablement définies lors de l'apprentissage et désigne celles qui sont les plus « proches ».

1.1.3 Introduction à la discrimination

Comme évoqué précédemment, la classification au sens de la discrimination fait partie d'une des réalisations de l'apprentissage supervisé. Nous pouvons alors définir son objectif comme la tâche d'un processus à affecter pour une nouvelle observation, la classe parmi toutes les classes possibles, en fonction des règles préalablement définies durant l'apprentissage.

Dès lors, les outils de classification doivent permettre d'établir des règles de classification et d'affectation, afin d'obtenir l'appartenance aux classes, de nouvelles observations. Prenons l'exemple d'un diagnostic médical où une augmentation de la température corporelle d'un individu peut définir un symptôme d'une maladie. En effet, nous savons que la température du corps oscille

[5] L'apprentissage peut être de type supervisé lorsque la sortie est connue, ou non supervisé lorsque la sortie est inconnue.

habituellement entre 36 et $37,2°\text{C}$ et que par expérience, certaines maladies (notamment en cas d'infection) font augmenter cette température. Cette réaction naturelle montre que l'organisme tente de se défendre face à des bactéries ou à des virus : on appelle cela la fièvre. Ainsi, la fièvre est caractérisée chez un individu, si sa température (T) dépasse le seuil de $37,5°\text{C}$. À partir de cette constatation, nous pouvons en retirer les règles suivantes :

$$\begin{array}{ll} \text{si} & T \geq 37,5\ °\text{C}, \quad \text{alors l'individu est potentiellement malade ;} \\ \text{si} & T < 37,5\ °\text{C}, \quad \text{alors l'individu n'est vraisemblablement pas malade.} \end{array}$$

Cet exemple simpliste, qui bien évidement ne suffit pas à diagnostiquer une maladie, montre parfaitement le principe d'une classification. Pour arriver à l'obtention de ce type de règles, la littérature fournit une diversité de méthodes de classification ; leur quantité rend l'établissement d'une bonne taxinomie difficile. Ainsi, à l'image de [Tufféry, 2007], nous pouvons dans un premier temps différencier les techniques **transductives** des techniques **inductives**.

Les techniques **transductives** déterminent directement l'appartenance d'une nouvelle observation sans passer par une phase d'apprentissage et donc sans créer de modèle. L'exemple le plus révélateur est la méthode des k-plus proches voisins (k-ppv) qui détermine la classe d'une nouvelle observation en regardant parmi les observations connues et déjà classées, les classes des k observations qui sont ses plus proches voisins. Cet algorithme a l'inconvénient de manipuler systématiquement l'ensemble des données disponibles pour chaque nouvelle affectation, ce qui nécessite si le nombre d'observations disponibles est important, une grande capacité de calcul et de stockage. Cet inconvénient s'étend à l'ensemble des techniques transductives, les rendant par conséquent peu attractives [Tufféry, 2007]. Il est alors préférable d'utiliser un modèle qui résume le contenu des données, afin de pouvoir obtenir rapidement l'appartenance de nouvelles observations. C'est dans ce contexte que nous retrouvons les techniques dites **inductives**, où celles-ci déterminent l'appartenance aux classes d'une nouvelle observation à partir d'un modèle élaboré durant une **phase d'apprentissage**. Comme explicité précédemment, cette phase définit les relations entre les variables d'entrées et les variables de sortie, à partir des observations disponibles. C'est ainsi qu'apparaît le principe de l'**induction**, qui comme le définit [Cornuéjols, 2005] consiste à produire une connaissance générale à partir de faits particuliers, afin de pouvoir faire des prédictions sur des événements futurs. Dans le cas de la classification, l'induction permet de déterminer la classe d'une nouvelle observation, sachant que ses variables obéissent aux mêmes lois que celles qui ont défini la connaissance.

Revenons sur l'exemple où un seuil sur la température corporelle d'un individu ($T = 37,5°\text{C}$), permettait d'aider le diagnostic et de relever la présence d'une maladie. En son temps, ce seuil a été déterminé par des médecins, lors de multiples observations sur des patients malades. Ces observations ont ainsi permis depuis, de pouvoir envisager une maladie en fonction de la présence de fièvre chez un patient. Cet exemple illustre parfaitement le principe de l'induction, qui rappelons-le, tente de tirer une loi générale à partir de faits particuliers.

Les approches inductives exploitent alors une phase d'apprentissage. [Bishop, 2006] les répartit suivant deux approches : génératives et discriminantes. Les approches génératives, abordées principalement dans le chapitre 1.2, créent des modèles en estimant des densités de probabilité afin d'obtenir les règles de classification et d'affectation. Ces approches recherchent alors à minimiser le recouvrement des densités de probabilité des classes. Les approches dites discriminantes, évoquées dans les chapitres 1.3 et 1.4, obtiennent directement l'appartenance d'une nouvelle observation aux classes, sans passer par les estimations de densités.

La tâche réalisée par un système d'apprentissage est donc de trouver des régularités à partir d'un ensemble d'observations et d'en tirer des règles de décision. Dans un contexte de classification par des techniques inductives, les régularités et les règles obtenues sont alors résumées dans un modèle de classification (par moment nous simplifierons son appellation par classifieur, ou juste par modèle), qui doit permettre de les généraliser, afin d'obtenir la classe de nouvelles observations.

Avant de poursuivre, nous noterons $\mathbf{X}$, la matrice contenant n observations $\mathbf{x}_1, ..., \mathbf{x}_n$, chacune décrite par p variables. Le vecteur $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$ donne alors les p éléments de l'observation i, et l'élément (i, j) de $\mathbf{X}$ (noté x_{ij}) correspond à la j-ème variable de la i-ème observation.

1.2 Approches probabilistes pour la classification

1.2.1 Introduction

Parmi les techniques inductives de classification évoquées précédemment, nous avons différencié deux types de méthodes : les approches génératives et les approches discriminantes. Cette première section porte sur les approches génératives qui sont fondées sur le calcul de plusieurs paramètres à partir de la distribution des observations. Aussi, comme nous le verrons, les méthodes basées sur ces approches sont fondées sur une vision bayésienne qui permet d'offrir un caractère théorique très intéressant pour la classification.

Ce type de méthodes repose sur l'idée que les variables et les classes peuvent être traitées comme des variables aléatoires. Ainsi, l'appartenance à la classe d'une observation tirée au hasard dans l'espace des observations de l'échantillon est une réalisation d'une variable aléatoire dont la valeur est le numéro de la classe (variable discrète).

1.2.2 Vision bayésienne

1.2.2.1 Règle de décision de Bayes

Dans les approches statistiques relatives aux problèmes de classification et plus généralement dans la théorie de la décision [Bishop, 2006], la règle de décision de Bayes est fondamentale, car elle fournit la limite théorique du taux d'erreur d'un modèle. [Dreyfus *et al.*, 2002] proposent la définition de la règle suivante :

Définition 1 *Pour affecter une observation à une classe, nous minimisons le risque d'erreur en prenant la décision d'affecter l'observation à la classe dont la probabilité a posteriori est la plus grande.*

Dans le cadre de la classification, cette limite est obtenue par la connaissance exacte de la distribution des observations pour chacune des classes ; notons que cette exactitude sera remise en question. Ainsi, les fonctions de densités de probabilité conditionnelles des K classes $p(\mathbf{x}|\mathcal{C}_k)$[6], dites aussi fonctions de vraisemblance, et les probabilités *a priori* $P(\mathcal{C}_k)$[7] sont connues. Dès lors, en manipulant les règles de la théorie des probabilités, [Bishop, 2006] montre que le théorème de Bayes (1.1) permet d'obtenir l'affectation d'une nouvelle observation aux classes, par l'intermédiaire des probabilités *a posteriori* $P(\mathcal{C}_k|\mathbf{x})$:

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{p(\mathbf{x})}\,, \tag{1.1}$$

[6] $p(\mathbf{x}|\mathcal{C}_k)$ indique la probabilité de l'observation $\mathbf{x}$, sachant que $\mathbf{x}$ appartient à la classe $\mathcal{C}_k$.
[7] $P(\mathcal{C}_k)$ indique la probabilité qu'une observation tirée au hasard dans l'échantillon appartienne à la classe $\mathcal{C}_k$.

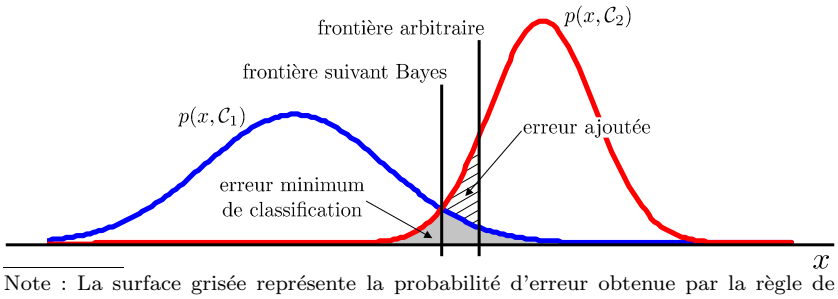
où $p(\mathbf{x})$ est le facteur de normalisation équivalent à $\sum_{k=1}^{K} p(\mathbf{x}|\mathcal{C}_k)P(\mathcal{C}_k)$. Ainsi, selon la règle de décision, l'observation $\mathbf{x}$ est affectée à la classe $\mathcal{C}_k$, où la probabilité *a posteriori* $P(\mathcal{C}_k|\mathbf{x})$ est maximum.

Ce principe est certainement très séduisant, mais ses performances sont fortement dépendantes de la qualité de l'échantillon qui est extrait de la population à étudier. En effet, l'estimation des probabilités *a posteriori* est correcte si les variables obéissent aux mêmes vraisemblances de l'échantillon utilisé pour les estimer.

Les expressions analytiques des vraisemblances et des probabilités *a priori* sont rarement connues exactement. C'est pourquoi, il est nécessaire de les estimer à partir des observations de l'échantillon. Les probabilités *a priori* des classes s'obtiennent naturellement en faisant le rapport entre le nombre d'observations de chaque classe sur le nombre total d'observations. En revanche, la difficulté majeure du théorème de Bayes demeure dans l'estimation des densités de probabilité. En effet, leur estimation est susceptible de se heurter au problème connu sous le nom de la « malédiction de la dimensionnalité » (*cf.* section 2.1). Les observations sont représentées par des vecteurs de variables dont leur dimension peut être très grande. Or, dès que la dimension devient importante, l'estimation des densités de probabilité est très difficile et peut devenir très approximative, dès lors que l'échantillon ne possède pas suffisamment d'observations. En effet, comme il le sera montré dans l'introduction de la deuxième partie (page 60), l'échantillon devrait croître de façon exponentielle avec le nombre de variables.

1.2.2.2 Classifieur bayésien

La combinaison du théorème de Bayes et de la règle de décision de Bayes constitue ce que nous pouvons appeler le classifieur de Bayes. Il présente théoriquement le meilleur outil de classification si les probabilités *a priori* et les vraisemblances sont connues exactement, comme le montre la figure 1.4 qui compare deux frontières de décision obtenues par la règle de Bayes et de manière arbitraire. Ainsi, pour la frontière arbitraire, nous pouvons observer l'erreur ajoutée à l'erreur de Bayes sur le recouvrement des densités de probabilité conjointe $p(x, \mathcal{C}_k) = p(x|\mathcal{C}_k)\, P(\mathcal{C}_k)$.



Note : La surface grisée représente la probabilité d'erreur obtenue par la règle de décision de Bayes. La surface hachurée représente la probabilité d'erreur ajoutée à la précédente par le choix d'une autre frontière de décision.

FIG. 1.4 – Règle de décision de Bayes pour un problème à deux classes.

Cependant, dans la pratique il est rare de connaître précisément les vraisemblances, ainsi le classifieur de Bayes présente un intérêt théorique plutôt que pratique [Dreyfus *et al.*, 2002].

À la section 1.2.4, nous détaillerons des méthodes de classification basées sur l'approche bayésienne, telle que le classifieur de Bayes naïf et l'analyse discriminante.

1.2.3 Estimation des densités de probabilité

Le théorème de Bayes (1.1) fait intervenir les densités de probabilité conditionnelles et les probabilités *a priori*, or nous avons évoqué qu'il est rare qu'elles soient connues exactement. C'est pourquoi, il est nécessaire de les estimer, et ce, à partir des observations de l'échantillon. Deux familles de méthodes permettent de les estimer : les approches paramétriques et non paramétriques. Elles se différencient par une connaissance sur la distribution des observations, permettant pour les approches paramétriques de faire certaines hypothèses sur les vraisemblances. [Stoppiglia, 1997; Dreyfus *et al.*, 2002] proposent une comparaison expérimentale entre les méthodes d'estimation paramétrique et non paramétrique.

1.2.3.1 Estimation paramétrique

L'estimation des densités de probabilité par des méthodes paramétriques consiste à faire une hypothèse sur la forme analytique de la densité. Les observations disponibles permettent alors d'estimer les paramètres liés à la densité avancée par l'hypothèse. Ainsi, par la connaissance de la nature analytique, la densité peut être déduite en tout point de l'espace des variables.

Dans le cas où l'hypothèse de la distribution des observations considère une loi gaussienne, hypothèse couramment utilisée, la connaissance de la répartition des observations permet d'obtenir le centre de gravité μ_k et la matrice de covariance Σ_k (variance dans le cas unidimensionnel) du groupe d'observations de la classe k. Ainsi, nous supposons que la densité de probabilité $p(\mathbf{x}|\mathcal{C}_k)$ suit une loi multinormale, telle que :

$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{2\pi^{p/2}\sqrt{\det(\Sigma_k)}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k)\right), \tag{1.2}$$

où p indique la dimension de l'espace des observations.

L'utilisation de cette loi apparaît dans les méthodes basées sur les fonctions discriminantes décrites à la section 1.2.4.1.

1.2.3.2 Estimation non paramétrique

Les distributions d'observations se présentent rarement par des lois simples. Ainsi, sans connaissance *a priori* de la distribution des observations, les méthodes d'estimation non paramétrique permettent d'éviter de faire des hypothèses quant à leur distribution.

Le principe de cette estimation des vraisemblances consiste à délimiter une région autour d'un vecteur $\mathbf{x}$ représenté dans l'espace des variables et de déterminer le nombre d'observations k contenu dans un volume V. Ainsi, [Bishop, 2006] montre, qu'à partir de (1.3), la densité de probabilité $p(\mathbf{x})$ peut être estimée, en connaissant n le nombre total d'observations par $\hat{p}$:

$$\hat{p}(\mathbf{x}) = \frac{k}{n\,V}\,. \tag{1.3}$$

Deux paramètres inconnus demeurent dans l'estimation non paramétrique : k et V. La méthodologie consiste à fixer un des deux paramètres et à calculer l'autre. Cela conduit donc à deux approches :

- méthode à base de noyau, connue sous le nom de noyau de *Parzen* [Parzen, 1962], qui fixe le volume V et compte le nombre d'observations k dans ce volume ;

- méthode des k-plus proches voisins, qui fixe le nombre d'observations k et calcule le volume V contenant les k plus proches observations.

À la figure 1.5, nous montrons $\hat{p}(x)$ et des estimations de densités de $p(x)$ pour différentes valeurs de k (2, 10 et 30). Sachant que le nombre d'observations n dans cet exemple est de 150, nous pouvons alors remarquer l'aplatissement de l'estimation de la densité avec l'augmentation du nombre des k voisins considérés.

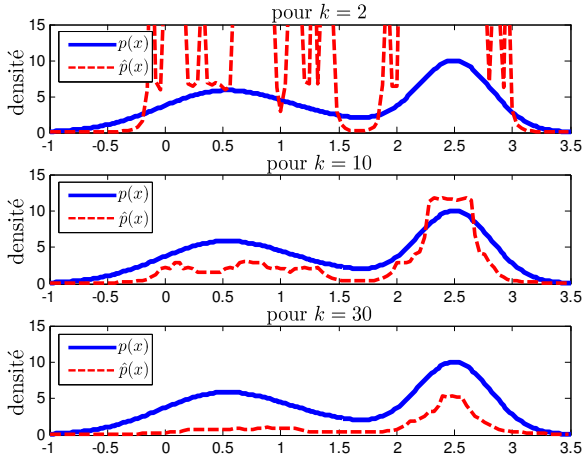


FIG. 1.5 – Estimations de densités par la méthode des k-plus proches voisins.

Dans l'exemple qui suit, nous souhaitons déterminer l'appartenance aux classes d'une observation $\mathbf{x}$, donc en déterminant les probabilités *a posteriori*. Ainsi, nous déterminons le volume V autour de $\mathbf{x}$, contenant les k observations. La relation (1.3) permet alors d'estimer la densité de probabilité associée à chaque classe par $p(\mathbf{x}|\mathcal{C}_k) = \frac{k_k}{n_k V}$, où n_k est le nombre total d'observations appartenant à la classe $\mathcal{C}_k$ et k_k donne le nombre d'observations appartenant à la classe $\mathcal{C}_k$ parmi les k plus proches observations de $\mathbf{x}$. D'autre part, la probabilité *a priori* s'obtient par $P(\mathcal{C}_k) = \frac{n_k}{n}$ et le facteur de normalisation est donné par $p(\mathbf{x}) = \frac{k}{n V}$, où n donne le nombre total d'observations. En combinant ces relations avec le théorème de Bayes, nous obtenons, une fois les simplifications faites, la probabilité *a posteriori* d'appartenance à la classe k suivante :

$$P(\mathcal{C}_k|\mathbf{x}) = \frac{k_k}{k}\,. \tag{1.4}$$

Les méthodes d'estimation non paramétrique ont besoin d'un nombre important d'observations. À la fin de la section 1.2.2.1, nous avons vu que ce nombre devrait croître exponentiellement avec la dimension afin d'obtenir une estimation convenable des densités. Malgré l'intérêt de ces estimations qui ne formulent pas d'hypothèses sur la distribution, la contrainte associée à la quantité d'observations nécessaire rend l'estimation non paramétrique très délicate. Cette remarque est valable dans une moindre mesure pour les estimations paramétriques, qui pour déterminer les paramètres des lois ont besoin également d'un nombre non négligeable d'observations.

1.2.4 Classifieurs basés sur la théorie de Bayes

À la section 1.2.2.2, nous avons introduit le classifieur de Bayes obtenu par l'association stricte du théorème et de la règle de décision de Bayes. Dans cette section, nous détaillons ce type de classifieurs et proposons deux approches : les fonctions discriminantes linéaires et quadratiques et le classifieur naïf.

1.2.4.1 Fonctions discriminantes

L'approche de classification la plus utilisée, fondée sur les relations de Bayes, est certainement celle basée sur la distribution normale. Ceci lie naturellement ce modèle de classification à la section 1.2.3.1 sur l'estimation paramétrique des densités. Ainsi, dans cette approche, nous faisons l'hypothèse que les densités de probabilité $p(\mathbf{x}|\mathcal{C}_k)$ suivent des lois multinormales telles que (1.2).

La règle de décision de Bayes affecte une observation à la classe $\mathcal{C}_k$ telle que $p(\mathcal{C}_k|\mathbf{x})$ est maximale. Nous cherchons alors la classe k qui maximise $p(\mathcal{C}_k|\mathbf{x})$, ou encore $\log(p(\mathcal{C}_k|\mathbf{x}))$. Ainsi, en reprenant le théorème de Bayes (1.1), cela revient à trouver la classe k qui maximise :

$$\log(p(\mathbf{x}|\mathcal{C}_k)\, P(\mathcal{C}_k))\,. \tag{1.5}$$

Le facteur de normalisation $p(\mathbf{x})$ disparaît de la relation, car il est indépendant des classes et la fonction discriminante $y_k(\mathbf{x})$ s'écrit donc :

$$\begin{aligned} y_k(\mathbf{x}) &= \log(p(\mathbf{x}|\mathcal{C}_k)) + \log(P(\mathcal{C}_k))\,, \\ &= -\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma_k^{-1}(\mathbf{x}-\mu_k) - \frac{1}{2}\log\left(\det(\Sigma_k)\right) + \log(P(\mathcal{C}_k))\,. \end{aligned} \tag{1.6}$$

Ce choix du critère de classification, en d'autres termes la règle de décision de Bayes, minimise la probabilité de fausse classification. Ainsi, l'observation $\mathbf{x}$ est assignée à la classe $\mathcal{C}_k$ si $y_k(\mathbf{x}) > y_j(\mathbf{x})$, $\forall j \neq k$. La frontière de décision est donnée pour $y_k(\mathbf{x}) = y_j(\mathbf{x})$ et elle est de forme quadratique. Cette technique de classification basée sur le théorème de Bayes est aussi appelée **analyse discriminante probabiliste** [Tufféry, 2007].

Pour simplifier la règle de décision quadratique et aboutir à une règle linéaire, nous pouvons faire l'hypothèse d'homoscédasticité [Tufféry, 2007] qui signifie que, pour chaque variable, les groupes d'observations de chaque classe ont la même variance, et donc que les matrices des covariances des variables sont égales : $\Sigma_k = \Sigma$. Comme Σ est désormais indépendant des classes, la fonction discriminante $y_k(\mathbf{x})$ devient :

$$y_k(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_k)^T\Sigma^{-1}(\mathbf{x}-\mu_k) + \log(P(\mathcal{C}_k))\,, \tag{1.7}$$

$$= \mathbf{x}^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(P(\mathcal{C}_k))\,, \tag{1.8}$$

$$= \mathbf{w}_k^T\mathbf{x} + w_{k0}\,. \tag{1.9}$$

où $\mathbf{w}_k = \Sigma^{-1}\mu_k$ et $w_{k0} = -\frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(P(\mathcal{C}_k))$. Les paramètres $\mathbf{w}_k$ et w_{k0} sont les coefficients de la fonction discriminante, ils seront appelés vecteurs poids et biais. Cette forme d'expression sera utilisée par la suite pour décrire les fonctions discriminantes, notamment dans la section 1.3.

Remarquons que le développement de (1.7) a vu le terme $\mathbf{x}^T\Sigma^{-1}\mathbf{x}$ disparaître dans (1.8), car ce dernier est indépendant des classes.

La figure 1.6 illustre la différence entre les frontières de décision obtenues par les fonctions discriminantes quadratique et linéaire. Dans cet exemple, trois classes sont représentées par deux variables.

Prenons le cas de deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$, avec lesquelles nous faisons toujours l'hypothèse d'homoscédasticité. Nous allons alors montrer que sous cette hypothèse, nous pouvons arriver

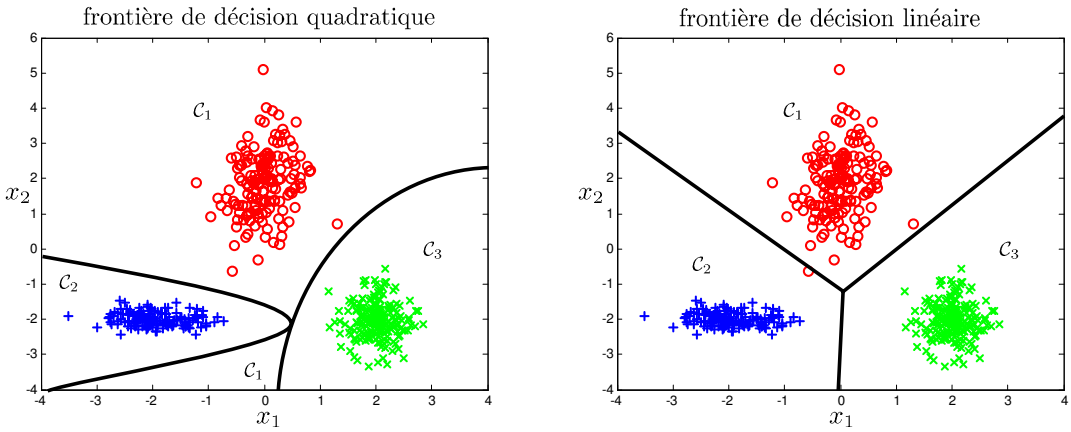


FIG. 1.6 – Illustrations de frontières de décision obtenues par les fonctions discriminantes quadratique et linéaire.

à une nouvelle fonction discriminante $y(\mathbf{x})$, correspondant à la fonction discriminante de Fisher [Fisher, 1936]. La construction de cette fonction sera détaillée à la section 1.3.2.1. Ainsi, dans ce problème à deux classes, l'appartenance de $\mathbf{x}$ à la classe $\mathcal{C}_1$, défini précédemment par $y_1(\mathbf{x}) > y_2(\mathbf{x})$ est équivalent à $y(\mathbf{x}) > 0$ pour $y(\mathbf{x}) = y_1(\mathbf{x}) - y_2(\mathbf{x})$. Dès lors, en reprenant la relation (1.8), nous obtenons l'expression de la fonction discriminante suivante :

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{x}^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log(p(\mathcal{C}_1)) - \mathbf{x}^T\Sigma^{-1}\mu_2 - \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2 + \log(p(\mathcal{C}_2))\,, \\ &= (\mu_1-\mu_2)^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)\,. \end{aligned} \tag{1.10}$$

Le terme $\log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)$ pourrait disparaître de (1.10), si l'hypothèse de l'équiprobabilité des classes était considérée. On retrouverait alors la fonction discriminante de Fisher :

$$y(\mathbf{x}) = (\mu_1-\mu_2)^T\Sigma^{-1}\mathbf{x} - \frac{1}{2}\left(\mu_1^T\Sigma^{-1}\mu_1 - \mu_2^T\Sigma^{-1}\mu_2\right)\,. \tag{1.11}$$

La considération des hypothèses d'homoscédasticité et d'équiprobabilité fait référence cette fois-ci à l'**analyse discriminante géométrique** [Tufféry, 2007]. Dans ce cas particulier, et toujours en considérant un problème à deux classes, on remarque que la règle d'affectation est liée à une notion de distance. En effet, l'affectation d'une observation $\mathbf{x}$ est obtenue par le calcul de la distance entre cette observation et les centres de gravité μ_1 et μ_2 des deux gaussiennes représentatives des deux groupes d'observations (classes $\mathcal{C}_1$ et $\mathcal{C}_2$). Cette distance s'appelle distance de Mahalonobis :

$$\Delta^2 = (\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\,. \tag{1.12}$$

L'interprétation des distances dans l'affectation des classes éloigne, quelque peu, cette approche de discrimination linéaire de la relation de Bayes. Moins efficace que la méthode de discrimination quadratique, cette approche a néanmoins l'intérêt d'être simple et rapide.

1.2.4.2 Classifieur de Bayes naïf

Nous avons comparé précédemment deux types d'analyses discriminantes, avec comme principal problème, l'estimation des densités de probabilité, notamment pour l'analyse quadratique qui fait l'hypothèse que la distribution des observations suit une loi multinormale. Le classifieur de

Bayes naïf permet d'éliminer ce problème. Cette méthode de classification probabiliste est simple, elle est liée au théorème et à la règle de décision de Bayes. Ce classifieur fait l'hypothèse d'indépendance entre les variables, permettant ainsi, de simplifier considérablement la détermination des densités de probabilité. En effet, $\hat{p}(\mathbf{x}|\mathcal{C}_k)$ la densité de probabilité conditionnée à chacune des classes est estimée à partir des densités univariées $p(x_j|\mathcal{C}_k)$, $j = 1, \ldots, p$. Ce point est intéressant, car il implique l'estimation individuelle des p variables, donc en une dimension, évitant ainsi la « malédiction de la dimensionnalité ». La densité associée à chaque classe $p(\mathbf{x}|\mathcal{C}_k)$ est alors estimée par :

$$\hat{p}(\mathbf{x}|\mathcal{C}_k) = \prod_{j=1}^{p} p(x_j|\mathcal{C}_k)\,. \tag{1.13}$$

Une fois les densités de probabilité estimées, il suffit de déterminer les probabilités *a posteriori* (1.1) pour obtenir l'affectation d'une observation aux classes en suivant la règle de Bayes.

1.2.5 Conclusions

Nous pouvons rappeler dans cette conclusion, que compte tenu des hypothèses qui peuvent être faites sur les distributions des observations, il est plus que nécessaire que l'échantillon soit représentatif de la population étudiée.

Nous avons pu noter que le principal problème de cette approche statistique est l'estimation des densités de probabilité. En effet, elles conditionnent les performances, car si les probabilités *a priori* sont égales, alors les probabilités *a posteriori* dépendent totalement des vraisemblances des classes. Et en outre, si les vraisemblances sont égales (les variables ne sont pas discriminantes), alors les probabilités *a posteriori* dépendent cette fois-ci uniquement des probabilités *a priori*.

Pour éviter le problème d'estimation des densités, d'autres méthodes de classification, connues sous les noms d'approches discriminantes, ou directes, proposent une alternative intéressante. En effet, ce type de méthodes, comme les réseaux de neurones, permettent une estimation directe des probabilités *a posteriori*, sans passer par le calcul des probabilités *a priori* et surtout sans estimer les vraisemblances.

1.3 Classification Linéaire

1.3.1 Introduction

Dans la section précédente, la séparation des groupes d'observations de chaque classe, donc l'obtention de frontières de décision, était obtenue suivant des approches probabilistes. Ces approches nécessitaient la détermination de paramètres liés à la distribution des observations ; cela sous-entendait de faire des hypothèses sur les lois régissant la répartition des observations de l'échantillon. Nous avons mis en évidence certains problèmes liés à ces approches, qui suggèrent l'utilisation d'autres méthodes pour effectuer la classification, sans faire d'hypothèses sur la distribution. Ainsi, cette deuxième section aborde le problème de classification sous un angle différent, les méthodes étudiées construisent des fonctions de discrimination, afin de séparer les observations de classes différentes. Ces méthodes, nommées approches discriminantes ou directes, vont permettre comme leurs homologues (approches génératives ou indirectes) d'obtenir des règles de décision permettant de classer de nouvelles observations.

Le lecteur pourra noter que les sections 1.3.2.2 et 1.3.2.3 peuvent faire office de préambule à la présentation des réseaux de neurones développés dans la section 1.4.2.

Afin de faciliter les démonstrations, nous commençons par expliciter les méthodes dans le cadre de problèmes à deux classes, puis nous proposerons différentes procédures pour généraliser et étendre ces méthodes pour des problèmes à classes multiples.

1.3.2 Problème à deux classes

La séparation des classes peut s'effectuer en déterminant une fonction séparant les observations de chaque classe. Considérons l'exemple donné à la figure 1.7, où deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$ sont représentées et caractérisées par deux variables x_1 et x_2. On peut remarquer qu'une droite suffit à les séparer. Dans ce cas bidimensionnel, la droite d'équation $y(\mathbf{x}) = w_1x_1 + w_2x_2 + w_0$, admettant comme paramètres w_j $(j = 0, 1, 2)$ sépare effectivement les deux classes. Ainsi, d'après la figure, toute forme $\mathbf{x}$ dans l'espace des variables appartenant à la classe $\mathcal{C}_1$ doit conduire à une valeur positive de $y(\mathbf{x})$ et négative pour la classe $\mathcal{C}_2$.

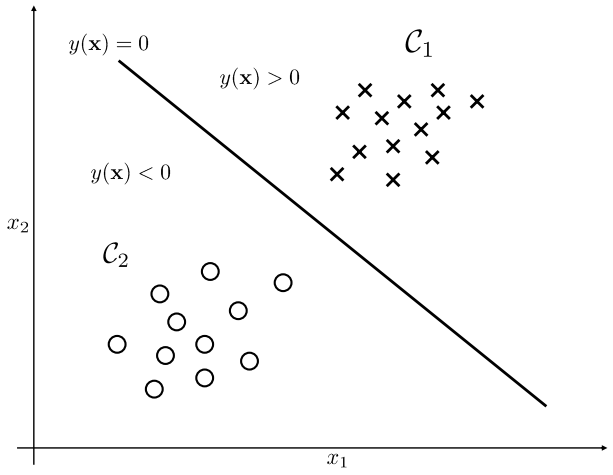


FIG. 1.7 – Illustration d'une séparation linéaire en deux classes d'un ensemble de données.

La fonction discriminante bidimensionnelle, en l'occurrence la droite séparatrice (figure 1.7), peut aisément être généralisée à p dimensions, elle est appelée dans ce cas **hyperplan séparateur**. Pour construire cette fonction et obtenir les règles de décision, nous avons besoin d'informations sur les deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$ définies *a priori*. Pour cela, nous utilisons un ensemble d'observations $\mathcal{X}$:

$$\mathcal{X} = \{(\mathbf{x}_1, t_1), ..., (\mathbf{x}_n, t_n)\},\ \mathbf{x}_i \in \mathbf{R}^p,\ t_i \in \{-1, 1\}\,, \tag{1.14}$$

où le vecteur $\mathbf{x}_i$ contient les valeurs prises par la i-ème observation sur les p variables et t_i indique l'appartenance à la classe de l'observation $\mathbf{x}_i$. Nous adoptons la notation suivante : une observation $\mathbf{x}_i$ appartient à la classe $\mathcal{C}_1$ si l'étiquette $t_i = 1$, et à la classe $\mathcal{C}_2$ pour $t_i = -1$.

La représentation la plus simple de la fonction discriminante est obtenue par :

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \sum_{j=1}^{p} w_j x_j + w_0\,, \tag{1.15}$$

où le vecteur $\mathbf{w}$ de dimension p et le scalaire w_0 sont appelés respectivement **vecteur des poids** et **biais**. Les règles de décision adoptées par cette fonction discriminante sont :

$$\begin{cases} \text{si} \quad \mathbf{w}^T\mathbf{x}_i + w_0 \geq 0 \quad \text{alors} \quad \mathbf{x}_i \in \mathcal{C}_1\,, \\ \\ \text{si} \quad \mathbf{w}^T\mathbf{x}_i + w_0 < 0 \quad \text{alors} \quad \mathbf{x}_i \in \mathcal{C}_2\,. \end{cases} \tag{1.16}$$

La frontière de décision correspondante est donc définie par la relation $y(\mathbf{x}) = 0$. Dès lors, si les deux classes sont linéairement séparables, pour chaque observation nous avons $t_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 0$.

La figure 1.8 donne quelques caractéristiques géométriques de l'hyperplan $y(\mathbf{x})$ en fonction de ses paramètres. Ainsi, la distance d'un point $\mathbf{x}_i$ à l'hyperplan est $\frac{|y(\mathbf{x}_i)|}{||\mathbf{w}||}$ et le décalage de l'hyperplan à l'origine est de $\frac{|w_0|}{||\mathbf{w}||}$.

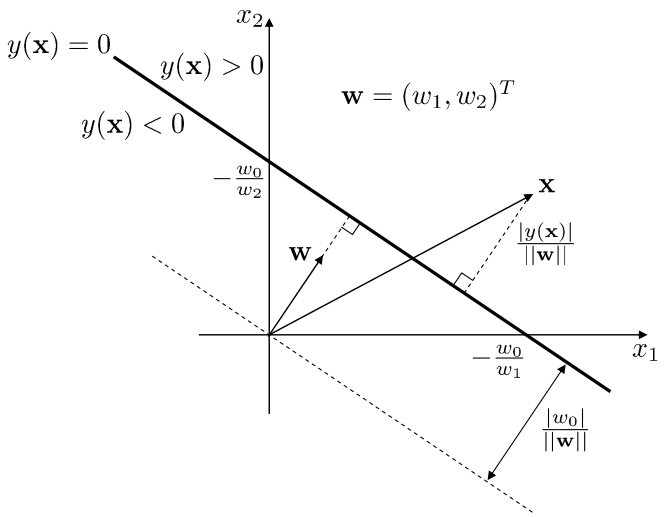


FIG. 1.8 – Géométrie d'une fonction de discrimination linéaire.

Maintenant que nous connaissons la géométrie de l'hyperplan, il va falloir déterminer ses paramètres w_j $(j = 0, \ldots, p)$, afin de séparer les deux classes. Plusieurs solutions le permettent, elles se différencient par l'objectif à atteindre et nous en présentons quelques unes.

1.3.2.1 Fonction discriminante de Fisher

Dans la présentation des approches probabilistes pour la classification, les fonctions discriminantes ont été abordées (*cf.* section 1.2.4.1). Ainsi, sous l'hypothèse d'homoscédasticité, nous avons pu aboutir à la fonction discriminante de Fisher. Nous allons aborder l'approche de Fisher sous un angle différent. En effet, cette approche aborde le problème de classification linéaire par une réduction de dimension ; c'est dans ce cadre de réduction que cette approche sera également abordée à la section 2.3.2.3. Ainsi, dans un premier temps, les observations sont projetées sur une droite passant par l'origine de vecteur directeur $\mathbf{w}$, maximisant la séparation des deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$. Comme le montre la figure 1.9, l'hyperplan discriminant est orthogonal à la droite de projection, optimisant la frontière de décision.

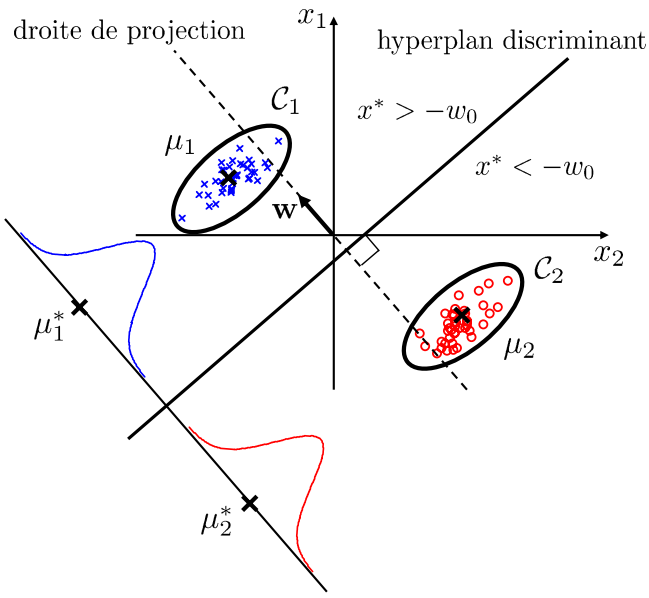


FIG. 1.9 – Hyperplan discriminant de Fisher.

L'observation $\mathbf{x}$ projetée sur la droite, notée x^* est obtenue par :

$$x^* = \mathbf{w}^T \mathbf{x}\,. \tag{1.17}$$

En se basant sur (1.15) et (1.16), on peut donc classer cette observation en considérant un seuil w_0. Ainsi, si $x^* > -w_0$, l'observation $\mathbf{x}$ appartient à la classe $\mathcal{C}_1$, tandis que si $x^* < -w_0$, l'observation appartient à la classe $\mathcal{C}_2$.

Notons μ_1 et μ_2 (1.18), les centres de gravité des observations appartenant aux classes $\mathcal{C}_1$ et $\mathcal{C}_2$, composées respectivement de n_1 et n_2 observations. Ainsi, les centroïdes des deux classes projetées sur la droite de projection sont obtenues respectivement par $\mu_1^* = \mathbf{w}^T\mu_1$ et $\mu_2^* = \mathbf{w}^T\mu_2$ (voir figure 1.9).

$$\mu_k = \frac{1}{n_k} \sum_{\mathbf{x}_i \in \mathcal{C}_k} \mathbf{x}_i\,. \tag{1.18}$$

L'ajustement des composantes du vecteur $\mathbf{w}$ optimise la séparation des classes. Dans cet objectif, Fisher propose de maximiser une fonction permettant une grande séparation entre les moyennes des classes projetées, tout en réduisant la variance de chacune des classes sur la droite de projection. Ainsi, le critère de Fisher est défini par le rapport entre la variance interclasse et la variance intraclasse totale. Il est donné par :

$$J(\mathbf{w}) = \frac{(\mu_2^* - \mu_1^*)^2}{s_1^2 + s_2^2}\,, \tag{1.19}$$

où s_k^2 représente la variance intraclasse de la classe k, définie par :

$$s_k^2 = \sum_{x_i^* \in \mathcal{C}_k} (x_i^* - \mu_k^*)^2\,. \tag{1.20}$$

En utilisant (1.17) et les équations des variances interclasse et intraclasse, on peut réécrire le critère de Fisher sous la forme suivante :

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}\,, \tag{1.21}$$

où $\mathbf{S}_B$ est la matrice de covariance interclasse, donnée par :

$$\mathbf{S}_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T\,, \tag{1.22}$$

et $\mathbf{S}_W$ est la matrice de covariance intraclasse, donnée par :

$$\mathbf{S}_W = \sum_{\mathbf{x}_i \in \mathcal{C}_1} (\mathbf{x}_i - \mu_1)(\mathbf{x}_i - \mu_1)^T + \sum_{\mathbf{x}_i \in \mathcal{C}_2} (\mathbf{x}_i - \mu_2)(\mathbf{x}_i - \mu_2)^T\,. \tag{1.23}$$

L'optimisation des paramètres est obtenue par la maximisation du critère de Fisher, donc pour $\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = 0$ et la démonstration de [Bishop, 2006] nous amène à obtenir $\mathbf{w} = \mathbf{S}_W^{-1}(\mu_2 - \mu_1)$. Ainsi, d'après les règles définies en (1.16) une observation $\mathbf{x}$ est attribuée à la classe $\mathcal{C}_1$, si $\mathbf{w}^T\mathbf{x} > w_0$. Dès lors, il reste à déterminer le biais introduit par w_0. Nous pouvons le déterminer en modélisant les densités de probabilité de chaque classe dans l'espace projeté par une gaussienne (*cf.* section 1.2.3.1). Nous pouvons trouver le seuil optimal conformément à la minimisation d'erreur introduit par le théorème de Bayes (1.1). En supposant, que les distributions gaussiennes ont des matrices de covariances égales [Webb, 2002], le biais optimal peut alors être déterminé par :

$$w_0 = -\frac{1}{2}(\mu_1 + \mu_2)^T \mathbf{S}_W^{-1}(\mu_2 - \mu_1) + \log\left(\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)}\right)\,, \tag{1.24}$$

L'association du vecteur paramètres $\mathbf{w}$ et du biais w_0 permet de retrouver la relation (1.11), qui était donnée comme la fonction discriminante de Fisher à l'issue d'hypothèses faites sur la nature des observations (*cf.* section 1.2.4.1).

1.3.2.2 Moindres carrés

Comme la technique précédente, cette estimation des paramètres de l'hyperplan peut être considérée comme une approche globale. Dans le sens où l'ensemble des observations est utilisé pour estimer les paramètres. Cette approche globale cherche à optimiser le critère (1.25), en minimisant la somme des carrés des erreurs entre les sorties réelles (t_i), correspondant à l'appartenance aux classes, et les sorties obtenues par la fonction discriminante.

$$J(\mathbf{w}) = \sum_{i=1}^{n} \left(t_i - \left(\mathbf{w}^T\mathbf{x}_i + w_0\right)\right)^2 = \sum_{i=1}^{n} e_i^2\,. \tag{1.25}$$

Pour faciliter la résolution de ce problème, nous allons adopter une nouvelle notation. L'ensemble des paramètres $\widetilde{\mathbf{w}}$ est appelé vecteur des poids élargi, composé du vecteur des poids $\mathbf{w}$ et du biais w_0. Ce nouveau vecteur est donc défini par $(w_0, \mathbf{w}^T)^T$. Pour obtenir les $p+1$ paramètres de la fonction discriminante, nous devons transformer également l'ensemble des observations $\mathbf{X}$ de dimension p en élargissant leur espace à $p+1$ dimensions. Alors, nous notons $\widetilde{\mathbf{X}}$, l'ensemble des observations dans l'espace élargi, tel que la i-ème ligne de $\widetilde{\mathbf{X}}$ correspond au vecteur $\widetilde{\mathbf{x}}_i^T$, représentant la i-ème observation dans cet espace, défini par $(1, \mathbf{x}^T)^T$. Ainsi, la fonction discriminante (1.15) devient :

$$y(\mathbf{x}) = \widetilde{\mathbf{w}}^T\widetilde{\mathbf{x}}\,, \tag{1.26}$$

et le critère (1.25) à minimiser devient :

$$J(\widetilde{\mathbf{w}}) = \sum_{i=1}^{n} \left(t_i - \widetilde{\mathbf{w}}^T\widetilde{\mathbf{x}}_i\right)^2\,. \tag{1.27}$$

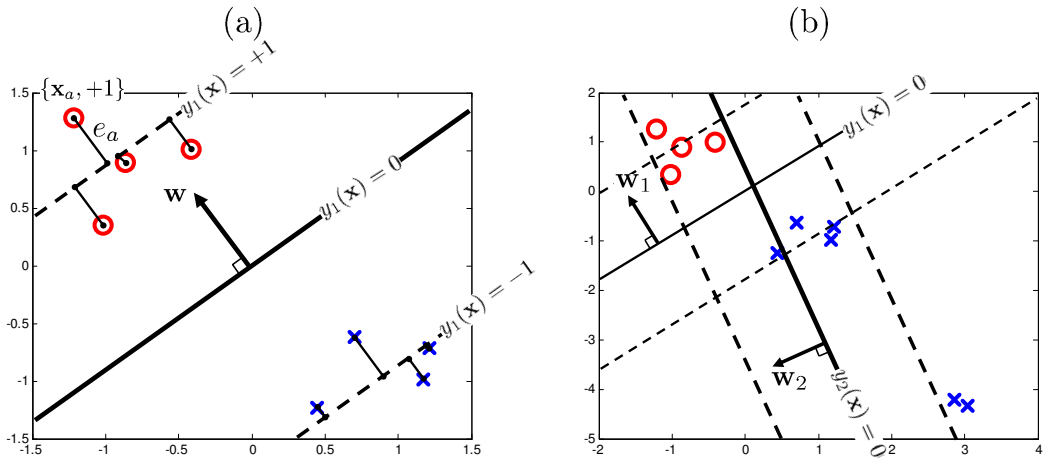
À partir de l'ensemble des observations disponibles, on définit le vecteur $\mathbf{t}$ réunissant l'appartenance des observations aux classes, tels que, la i-ème ligne de $\mathbf{t}$ correspond à t_i, renseignant l'appartenance de l'observation $\mathbf{x}_i$. L'utilisation conjointe de $\widetilde{\mathbf{X}}$ et $\mathbf{t}$ permet de déterminer les paramètres de $\widetilde{\mathbf{w}}$ en minimisant la somme des carrés de la fonction d'erreur (1.27), dont l'écriture matricielle est :

$$J(\widetilde{\mathbf{w}}) = (\mathbf{t} - \widetilde{\mathbf{X}}\widetilde{\mathbf{w}})^T(\mathbf{t} - \widetilde{\mathbf{X}}\widetilde{\mathbf{w}})\,. \tag{1.28}$$

Pour $\frac{\partial J(\widetilde{\mathbf{w}})}{\partial \widetilde{\mathbf{w}}} = 0$, on obtient la relation $(\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})\widetilde{\mathbf{w}} - \widetilde{\mathbf{X}}^T\mathbf{t} = 0$. Ce qui amène à trouver les paramètres $\widetilde{\mathbf{w}}$ par la relation suivante :

$$\widetilde{\mathbf{w}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{t}\,. \tag{1.29}$$

La figure 1.10(a) représente l'hyperplan $y_1(\mathbf{x})$ obtenu par la méthode des moindres carrés. Rappelons que cette méthode cherche à minimiser la somme des carrés des erreurs, où pour l'observation $\mathbf{x}_a$, étiquetée $t_a = +1$, l'erreur e_a est donnée par $t_a - (\mathbf{w}^T\mathbf{x_a} + w_0)$. Pour évaluer la robustesse de cette méthode, nous avons ajouté des observations et calculé les nouveaux paramètres de l'hyperplan séparateur $y_2(\mathbf{x})$. Comme le montre la figure 1.10(b), ces nouvelles observations peuvent être considérées comme aberrantes (*cf.* section 2.2.2). Ainsi, la comparaison des deux hyperplans $y_1(\mathbf{x})$ et $y_2(\mathbf{x})$montre la grande sensibilité de cette méthode en présence d'observations aberrantes, rendant par conséquent la méthode beaucoup moins efficace.



(a) Hyperplan discriminant $y_1(\mathbf{x})$ par la méthode des moindres carrés. (b) Comparaison de deux hyperplans $y_1(\mathbf{x})$ et $y_2(\mathbf{x})$ en présence d'observations aberrantes.

FIG. 1.10 – Influence d'observations aberrantes lors de la construction d'hyperplans discriminants par la méthode des moindres carrés.

Un autre inconvénient de cette méthode est d'ordre numérique. Il n'est pas rare que l'inversion de la matrice $\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}}$, présent dans (1.29) pose des problèmes. En outre, cette méthode globale utilise obligatoirement l'ensemble des observations pour estimer les paramètres de l'hyperplan. En effet, si de nouvelles observations étiquetées deviennent disponibles, cette méthode doit alors reprendre entièrement le processus d'estimation sans pouvoir s'adapter aux paramètres préalablement déterminés.

Compte tenu de ces dernières remarques, il serait alors agréable d'utiliser un algorithme itératif, où tout ne serait pas à refaire. Cela conduit à exploiter $\frac{\partial J(\widetilde{\mathbf{w}})}{\partial \widetilde{\mathbf{w}}}$, le gradient de $J(\widetilde{\mathbf{w}})$ (1.27) : $\nabla J(\widetilde{\mathbf{w}}) = \widetilde{X}^T(\widetilde{\mathbf{X}}\widetilde{\mathbf{w}} - \mathbf{t})$. À l'instant k, cette technique de gradient ajoute au vecteur $\widetilde{\mathbf{w}}$ un vecteur colinéaire et de sens opposé à $\nabla J(\widetilde{\mathbf{w}})$. On dit que l'on fait une descente de gradient sur la surface d'erreur. Le principe est simple, si l'on cherche un endroit situé plus bas que tous les autres,

alors il suffit de se déplacer systématiquement vers le bas, en suivant les plus grandes pentes. Ces pentes sont données par le gradient.

Ainsi, une fois l'initialisation aléatoire du vecteur des poids $\widetilde{\mathbf{w}}$ réalisée, l'algorithme d'adaptation modifie les poids par :

$$\widetilde{\mathbf{w}}(k+1) = \widetilde{\mathbf{w}}(k) - \rho \nabla J(\widetilde{\mathbf{w}})\,, \tag{1.30}$$

jusqu'à atteindre le critère désiré $J(\widetilde{\mathbf{w}}) < \epsilon$. Le paramètre ρ pondère l'adaptation des poids et il est défini comme le pas d'apprentissage. On peut simplifier en notant : $\Delta\widetilde{\mathbf{w}} = -\rho\nabla J(\widetilde{\mathbf{w}})$.

La version séquentielle de cette procédure, qui permet d'utiliser individuellement chaque observation, est donnée dans l'algorithme 1.1. Durant cette adaptation, la modification des poids est dépendante de la donnée $\{\mathbf{x}_i, t_i\}$ évaluée, et la modification est alors obtenue par :

$$\Delta\widetilde{\mathbf{w}} = -\rho\,\widetilde{\mathbf{x}}_i(\widetilde{\mathbf{w}}^T\widetilde{\mathbf{x}}_i - t_i)\,. \tag{1.31}$$

Algorithme 1.1 : Règle d'adaptation de *Widrow-Hoff*

Données : $\mathcal{X} = \{\mathbf{x}_i, t_i\}_{i=1,\ldots,n}$, $\mathbf{x}_i \in \mathbf{R}^p$, $t_i \in \{-1, 1\}$
Résultat : $\widetilde{\mathbf{w}}(k)$

```
début
    k ← 0;
    initialiser aléatoirement w̃(k);
    répéter
        sélectionner aléatoirement une observation {x_i, t_i};
        w̃(k+1) ← w̃(k) + Δw̃
        k ← k + 1;
    jusqu'à Σ_{i=1}^{n} (t_i − w̃(k)^T x̃_i)^2 < ε ;
fin
```

La règle *Widrow-Hoff* [Widrow and Hoff, 1960], connue aussi sous le nom de règle *Delta*, est très importante. Elle est considérée comme le fondement de la rétropropagation qui, comme la règle *Delta*, est une technique de descente de gradient. Comme nous le verrons dans la section 1.4.2, la rétropropagation a permis de faire évoluer considérablement les réseaux de neurones, permettant l'apprentissage d'architectures plus complexes (assemblées de plusieurs couches) que le perceptron, composé comme nous allons le voir, d'une seule couche de neurones.

1.3.2.3 Algorithme du perceptron

L'algorithme du perceptron [Rosenblatt, 1958] est une autre technique pour déterminer les paramètres de la fonction discriminante. Parmi les techniques de séparation linéaire, cet algorithme est probablement l'un des plus anciens et l'un des plus simples.

Comme la règle de *Widrow-Hoff*, cet algorithme d'apprentissage peut être interprété de manière connexionniste, où les coefficients de l'hyperplan sont alors vus comme les poids de connexions entre neurones. En effet, constitué d'un seul neurone (voir figure 1.11), le perceptron est le réseau le plus simple pour séparer les observations en deux classes.

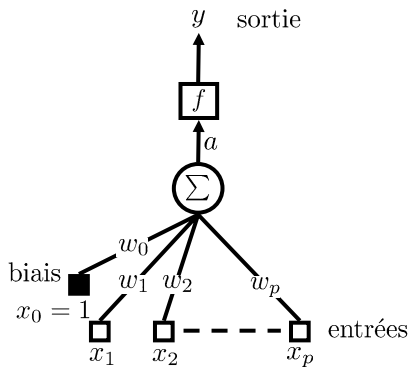


FIG. 1.11 – Représentation du perceptron de Rosenblatt [Rosenblatt, 1962].

La sortie du perceptron $y(\mathbf{x})$ dépend de la somme des entrées x_i pondérées par les poids w_i (avec $i = 0, ..., p$) ; la littérature qualifie cela comme l'activation ou le potentiel du neurone (a), tel que :

$$a = \sum_{i=1}^{p} w_i x_i + w_0 \,. \tag{1.32}$$

Le potentiel a est ensuite présenté à une fonction, nommée fonction d'activation f, afin d'obtenir la sortie du perceptron $y(\mathbf{x}) = f(a)$. Dans le perceptron original, cette fonction est un seuil tel que, si le potentiel du neurone a dépasse le seuil introduit par la fonction d'activation $f(\cdot)$, alors la sortie $y(\mathbf{x})$ vaut 1, sinon -1. Dans la pratique, le seuil habituellement utilisé est 0 :

$$y(\mathbf{x}) = f(a) = \begin{cases} +1 \text{ si } a \geq 0\,, \\ -1 \text{ si } a < 0\,. \end{cases} \tag{1.33}$$

Par cette fonction, la sortie du perceptron $y(\mathbf{x})$ revient simplement à prendre le signe de a :

$$\begin{aligned} y(\mathbf{x}) &= f\left(\sum_{i=1}^{p} w_i x_i + w_0\right), \\ &= \text{sign}\left(\sum_{i=1}^{p} w_i x_i + w_0\right). \end{aligned} \tag{1.34}$$

Comme avec la méthode des moindres carrés, l'espace des poids et des entrées peut être élargi, facilitant l'écriture : rappelons que $\widetilde{\mathbf{w}} = (w_0, \mathbf{w}^T)^T$ et $\widetilde{\mathbf{x}} = (1, \mathbf{x}^T)^T$. Ainsi, la formulation de la sortie du perceptron peut alors être donnée par :

$$y(\mathbf{x}) = \text{sign}\left(\widetilde{\mathbf{w}}^T \widetilde{\mathbf{x}}\right). \tag{1.35}$$

Nous avons vu que la méthode de Fisher cherche un hyperplan maximisant le ratio entre les variances interclasse et intraclasse dans l'espace de projection. La méthode des moindres carrés minimise une erreur globale, entre les sorties réelles et les sorties obtenues par le modèle. Contrairement à ces deux méthodes, l'algorithme du perceptron cherche à minimiser les observations mal classées, de manière à respecter $t_i(\widetilde{\mathbf{w}}^T \widetilde{\mathbf{x}}_i) > 0$ pour toutes les observations. Son critère peut s'écrire sous la forme suivante :

$$J(\widetilde{\mathbf{w}}) = -\sum_{\widetilde{\mathbf{x}}_i \in \mathcal{Y}} t_i(\widetilde{\mathbf{w}}^T \widetilde{\mathbf{x}}_i)\,, \tag{1.36}$$

où $\mathcal{Y}$ est l'ensemble des observations mal classées ne satisfaisant donc pas $t_i(\widetilde{\mathbf{w}}^T \widetilde{\mathbf{x}}_i) > 0$.

L'algorithme itératif modifie les poids associés aux entrées du neurone à chaque itération en fonction des données qui lui sont présentées, de manière à minimiser l'équation (1.36). Deux approches algorithmiques, que l'on peut retrouver dans [Bishop, 1995; Webb, 2002], permettent l'adaptation des poids. La première considère à chaque itération l'ensemble des données, c'est la version non stochastique de l'algorithme qui adapte les poids par :

$$\widetilde{\mathbf{w}}(k+1) = \widetilde{\mathbf{w}}(k) + \rho \sum_{\widetilde{\mathbf{x}}_i \in \mathcal{Y}} t_i \, \widetilde{\mathbf{x}}_i \, . \tag{1.37}$$

Cette version est appelée *batch update*, [Cornuéjols and Miclet, 2002] la décrivent précisément, où le terme ρ définit toujours le pas d'apprentissage, permettant de pondérer l'adaptation des poids.

La seconde approche, donc la version stochastique, adapte les poids à chaque itération, en fonction d'une donnée choisie aléatoirement, telle que :

$$\widetilde{\mathbf{w}}(k+1) = \widetilde{\mathbf{w}}(k) + \rho \, t_i \, \widetilde{\mathbf{x}}_i \, . \tag{1.38}$$

Cette version est également détaillée par [Dreyfus *et al.*, 2002; Cornuéjols and Miclet, 2002]. L'algorithme 1.2 décrit son fonctionnement.

Algorithme 1.2 : Version stochastique de l'algorithme du perceptron

Données : $\mathcal{X} = \{\mathbf{x}_i, t_i\}_{i=1,\dots,n}$, $\mathbf{x}_i \in \mathbf{R}^p$, $t_i \in \{-1, 1\}$
Résultat : $\widetilde{\mathbf{w}}(k)$

```
début
    k ← 0;
    initialiser aléatoirement w̃(k);
    répéter
        sélectionner aléatoirement une observation {x_i, t_i};
        si t_i(w̃(k)^T x̃_i) ≤ 0 alors
            w̃(k+1) ← w̃(k) + ρ t_i x̃_i ;
            k ← k + 1;
        fin
    jusqu'à t_i(w̃(k)^T x̃_i) > 0, ∀ i = 1, ..., n ;
fin
```

L'algorithme du perceptron itère la modification des poids tant qu'il reste des observations mal classées. Nous avons fait l'hypothèse que les données sont linéairement séparables, ainsi, l'algorithme du perceptron converge vers une solution, qui est cependant loin d'être optimale. [Belaïd and Belaïd, 1992] font remarquer que l'hyperplan obtenu est souvent « tangent » aux classes à séparer. Si les classes n'étaient pas linéairement séparables, l'algorithme 1.2 ne s'arrêterait jamais. Dès lors, un critère d'arrêt complémentaire à $t_i(\widetilde{\mathbf{w}}(k)^T\widetilde{\mathbf{x}}_i) > 0, \forall\, i = 1, \dots, n$ pourrait être ajouté, tenant compte d'un nombre maximal d'itérations à ne pas dépasser : $k \leq k_{max}$. Ces observations ont été mises en évidence pour la première fois par [Minsky and Papert, 1969] en montrant les limitations du perceptron, sur son incapacité à traiter des problèmes non linéairement séparables.

La figure 1.12 montre un exemple de convergence du perceptron. À l'instant k, l'hyperplan discriminant de paramètres $\widetilde{\mathbf{w}}(k)$ attribue des erreurs pour les observations $\mathbf{x}_a$ et $\mathbf{x}_b$. Selon l'algorithme stochastique 1.2, l'une de ces deux observations intervient pour modifier les poids. La donnée $\mathbf{x}_a$ est choisie, entraînant l'adaptation suivante : $\widetilde{\mathbf{w}}(k+1) = \widetilde{\mathbf{w}}(k) + t_a \, \widetilde{\mathbf{x}}_a$. Cette adaptation permet à l'instant $k+1$ de séparer correctement les données.

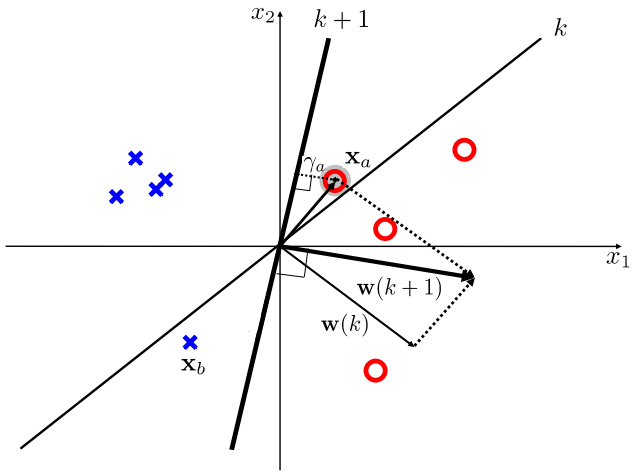


FIG. 1.12 – Algorithme du perceptron : évolution de l'hyperplan discriminant de l'itération k à l'itération $k+1$.

L'algorithme présenté précédemment est une des possibilités pour l'apprentissage du perceptron. Plusieurs variantes existent, concernant notamment la règle d'adaptation des poids [Bishop, 1995; Theodoridis and Koutroumbas, 2006; Webb, 2002]. Mentionné dans (1.37) et (1.38), le pas d'apprentissage ρ influence le processus. [Webb, 2002] propose plusieurs règles pour fixer sa valeur. Il introduit également la notion de marge, permettant de définir une distance minimale entre les observations et l'hyperplan séparateur. Ainsi, pour une marge $b > 0$, le vecteur de poids est modifié chaque fois que $t_i(\widetilde{\mathbf{w}}^T\widetilde{\mathbf{x}}_i) \leq b$, garantissant une marge de $\frac{b}{||\mathbf{w}||}$.

[Dreyfus *et al.*, 2002; Cristianini and Shawe-Taylor, 2000] abordent également la notion de marge, qui sans l'utiliser dans le processus d'apprentissage, proposent son utilisation une fois l'adaptation terminée. Ainsi, ils ont pu introduire la notion de stabilité des observations, permettant d'obtenir une idée de leur confiance quant à leur classification. Elle est obtenue par la quantité suivante :

$$\gamma_i = \frac{t_i(\mathbf{w}^T\mathbf{x}_i + w_0)}{||\mathbf{w}||}\,. \tag{1.39}$$

Nous pouvons noter que $\gamma_i > 0$ implique une classification correcte de l'observation $\mathbf{x}_i$ et que $|\gamma_i|$ donne la distance séparant l'observation de l'hyperplan. Ainsi, la plus petite distance, notée γ permet de définir la marge du perceptron. Dans l'exemple de la figure 1.12, la marge de l'ensemble de données d'apprentissage γ est $|\gamma_a|$, où $\mathbf{x}_a$ est l'observation la plus proche de l'hyperplan. Ainsi, la région autour et centrée sur l'hyperplan, d'épaisseur $2\,\gamma$, ne contient aucune observation d'apprentissage [Dreyfus *et al.*, 2002].

L'hyperplan de marge maximale est connu pour être plus robuste aux perturbations des observations, et, permet d'améliorer la généralisation [Cristianini and Shawe-Taylor, 2000; Dreyfus *et al.*, 2002]. C'est sur ce concept que sont fondées les *support vector machines*.

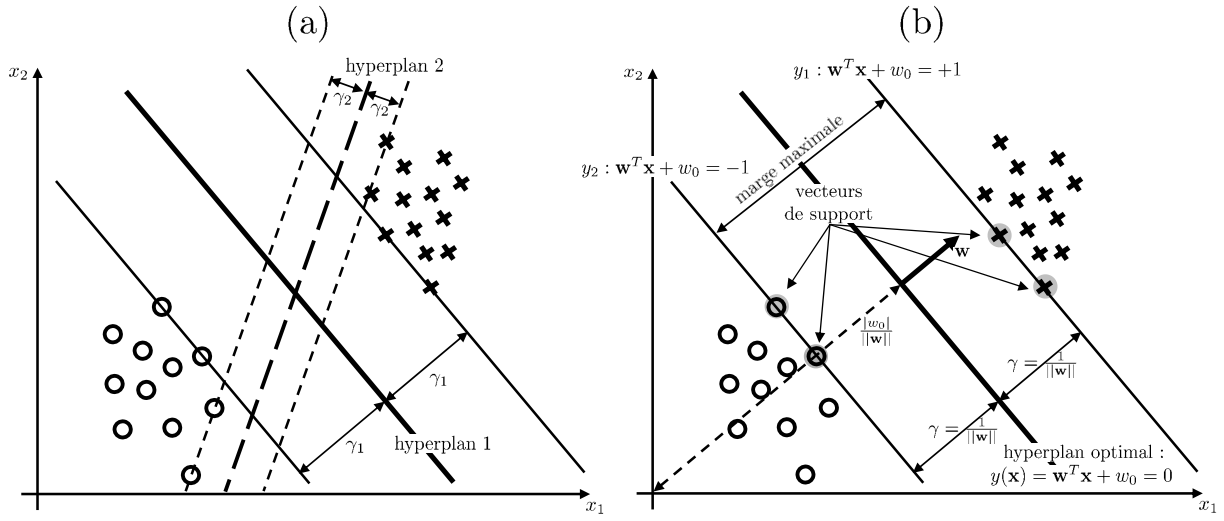
1.3.2.4 *Support Vector Machines*

Inspirées de la théorie statistique de l'apprentissage [Vapnik, 1995; Vapnik, 1998], les *Support Vector Machines*[8] constituent la forme la plus connue des méthodes à noyaux. Les SVM sont des

[8]Il n'y a pas de consensus sur la traduction française des SVM, nous pouvons ainsi trouver : séparateurs à vastes marges ou encore machines à vecteurs de support.

méthodes de classification binaire par apprentissage supervisé, elles sont basées sur l'utilisation de fonctions noyaux permettant une séparation optimale des données.

Nous avons pu observer qu'il existe une infinité d'hyperplans séparant les observations de deux classes. Nous avons également introduit la notion de marge dans l'algorithme du perceptron, car il révélait un aspect important dans la classification, en permettant d'améliorer la généralisation. Ces remarques amènent naturellement aux SVM, ayant comme propriété majeure d'obtenir un hyperplan unique et optimal (figure 1.13). Cet hyperplan est déterminé de manière à ce que la marge, la distance minimale de l'hyperplan aux observations d'apprentissage, soit maximale. Ainsi, pour une marge maximale, l'hyperplan séparateur est défini comme optimal. Cependant, malgré cette intéressante propriété, la solution unique n'est pas synonyme de généralisation optimale [Dreyfus *et al.*, 2002].



Note : (a) Comparaison entre deux droites séparatrices. Malgré la bonne séparation pour les deux hyperplans, l'hyperplan 1 est rendu optimal en maximisant la marge γ. (b) Géométrie de l'hyperplan optimal : cet hyperplan est perpendiculaire au segment de droite le plus court joignant une donnée d'apprentissage à l'hyperplan (marge). La marge est égale à $\frac{1}{||\mathbf{w}||}$ lorsque les paramètres de l'hyperplan sont normalisés.

FIG. 1.13 – Hyperplan discriminant maximisant les marges.

L'approche des SVM est simple, elle peut être décomposée en deux étapes. La première étape réalise une transformation des données faisant passer leur représentation de l'espace d'origine à un espace de plus grande dimension, appelé **espace de redescription**. Cela sera abordé lorsque nous évoquerons le cas où les observations ne sont pas linéairement séparables. Enfin, la seconde étape sépare les classes en rendant la marge maximale. Pour résumer, les SVM consistent à choisir l'hyperplan possédant une marge maximale dans l'espace de redescription.

Cas où les données sont linéairement séparables

On se place toujours dans le cadre de classification binaire, la fonction discriminante est donc toujours sous la forme de :

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0\,. \tag{1.40}$$

En respectant les mêmes règles de décision (1.16), toutes les observations sont correctement classées si $t_i(\mathbf{w}^T\mathbf{x}_i + w_0) > 0\,,\ \forall\, i$.

Si l'on reprend l'exemple d'apprentissage du perceptron de la figure 1.12, une fois les paramètres de l'hyperplan déterminés, la marge de l'apprentissage γ est calculée. Cette marge correspond à la distance de l'observation $\mathbf{x}_a$ à l'hyperplan : $\gamma = |\gamma_a| = \frac{|y(\mathbf{x}_a)|}{||\mathbf{w}||}$. Ainsi, les observations d'apprentissage sont à une distance supérieure à γ et satisfont l'expression $t_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq \gamma$. Avec les SVM, nous cherchons l'hyperplan qui maximise cette marge. Ainsi pour simplifier l'expression, nous pouvons normaliser les paramètres $\mathbf{w}$ et w_0 tels que la valeur de $y(\mathbf{x})$ aux points les plus proches des classes soit égale à 1 pour $\mathbf{x} \in \mathcal{C}_1$ et -1 pour $\mathbf{x} \in \mathcal{C}_2$. Ainsi, la marge γ devient $\frac{1}{||\mathbf{w}||}$, donc la marge maximale $\frac{2}{||\mathbf{w}||}$, satisfaisant cette fois-ci $t_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1\,,\ \forall\, i$.

Par conséquent, nous souhaitons optimiser les paramètres $\mathbf{w}$ et w_0 maximisant la distance $\frac{2}{||\mathbf{w}||}$, ce qui revient à minimiser $\frac{1}{2}||\mathbf{w}||^2$:

$$\begin{cases} \text{minimiser} & \frac{1}{2}||\mathbf{w}||^2\,, \\ \\ \text{tel que} & t_i(\mathbf{w}^T\mathbf{x}_i + w_0) \geq 1\,,\ \ i = 1, \ldots, n\,. \end{cases} \tag{1.41}$$

Ce problème de minimisation fait apparaître $p + 1$ paramètres, rendant la résolution difficile pour des valeurs de p importantes. Ainsi, nous pouvons transformer la **formulation primale** du problème (1.41) en une **formulation duale**, en faisant intervenir une fonction appelée lagrangien, qui pour ce problème est définie par :

$$\mathcal{L}(\mathbf{w}, w_0, \alpha) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{n} \alpha_i (t_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1)\,. \tag{1.42}$$

Le passage à la formulation duale se fait en introduisant des multiplicateurs de Lagrange $\alpha_i \geq 0$, $i = 1, \ldots, n$ pour chaque contrainte, soit pour chaque observation.

L'optimalité du problème (1.41) revient à déterminer le point-selle du lagrangien. La tâche est donc de minimiser par rapport aux variables $\mathbf{w}$ et w_0 et de maximiser par rapport aux variables duales α_i. La forme quadratique de $\mathcal{L}$ amène à trouver une solution unique. Au point-selle, l'annulation des dérivées partielles de $\mathcal{L}$ est :

$$\begin{cases} \frac{\partial}{\partial \mathbf{w}}\mathcal{L}(\mathbf{w}, w_0, \alpha) = 0 \\ \\ \frac{\partial}{\partial w_0}\mathcal{L}(\mathbf{w}, w_0, \alpha) = 0 \end{cases} \quad \text{ce qui conduit à} \quad \begin{cases} \mathbf{w} = \sum\limits_{i=1}^{n} \alpha_i t_i \mathbf{x}_i \\ \\ \sum\limits_{i=1}^{n} \alpha_i t_i = 0\,. \end{cases} \tag{1.43}$$

Selon les conditions de Karush-Kuhn Tucker [Theodoridis and Koutroumbas, 2006], au point-selle, pour chaque α_i la solution optimale satisfait :

$$\alpha_i(t_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1) = 0, i = 1, \ldots, n\,. \tag{1.44}$$

Ce dernier résultat implique que les observations $\mathbf{x}_i$ sur les hyperplans $y_1(\mathbf{x})$ et $y_2(\mathbf{x})$, telles que $\mathbf{w}^T\mathbf{x}_i + w_0 = \pm 1$, sont uniquement considérées et les multiplicateurs de Lagrange α_i correspondants sont non nuls. Alors, tous les autres α_i valent 0. Ces observations sont connues comme étant les **vecteurs de support**, nous noterons l'ensemble de ces observations par $\mathcal{SV}$.

Dès lors, en éliminant $\mathbf{w}$ et w_0 de $\mathcal{L}(\mathbf{w}, w_0, \alpha)$, en substituant (1.43) dans (1.42), nous atteignons la forme duale du problème d'optimisation :

$$\begin{cases} \max\limits_{\alpha} \left\{ \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j t_i t_j \mathbf{x_i}^T \mathbf{x_j} \right\} , \\ \alpha_i \geq 0, i = 1, \ldots, n \,, \\ \sum\limits_{i=1}^{n} \alpha_i t_i = 0 \,. \end{cases} \tag{1.45}$$

Ainsi, en résolvant ce problème d'optimisation quadratique (1.45) nous obtenons les coefficients α_i et le vecteur des poids (1.43) réalise l'hyperplan de marge maximale $\frac{2}{||\mathbf{w}||}$. On a donc la fonction de l'hyperplan :

$$\begin{aligned} y(\mathbf{x}) &= \sum_{i=1}^{n} \alpha_i t_i \mathbf{x}_i^T \mathbf{x} + w_0 \,, \\ &= \sum_{\mathbf{x}_i \in \mathcal{SV}} \alpha_i t_i \mathbf{x}_i^T \mathbf{x} + w_0 \,. \end{aligned} \tag{1.46}$$

La valeur de w_0 peut être trouvée en utilisant des exemples de $\mathcal{SV}$ dans (1.44).

Cas où les données sont non linéairement séparables

Nous avons pu introduire les SVM et leur résolution dans un cas idéal où les données sont linéairement séparables. Cependant, dans de nombreux problèmes réels, ce cas idyllique ne se présente pas, les classes ne peuvent donc pas être séparées linéairement (figure 1.14), et par conséquent, la recherche d'un hyperplan optimal n'a pas de sens. Ainsi, pour surmonter cette nouvelle contrainte, nous allons introduire une notion de « tolérance » faisant appel à une technique dite des **variables ressort** (*slack variables*). L'autre approche, qui néanmoins est complémentaire, concerne la transformation des observations d'entrées dans un espace de redescription de plus grande dimension. Cette question sera abordée dans la section 1.4 sur la classification non linéaire.

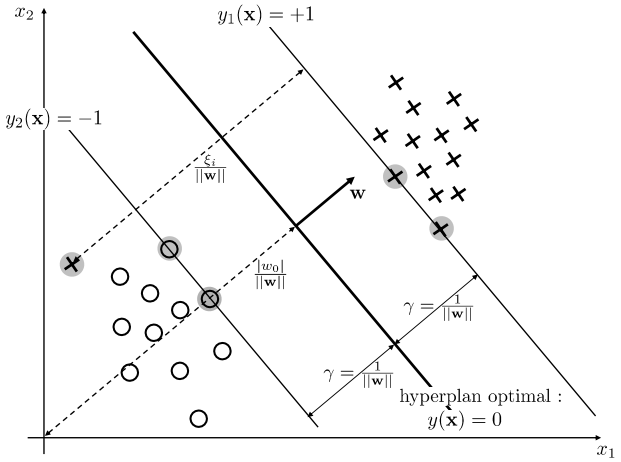


FIG. 1.14 – Hyperplan discriminant obtenu par les *support vector machines* en présence de données non linéairement séparables.

La technique des variables ressort permet de construire un hyperplan en admettant des erreurs, mais en les minimisant, ce qui amène à assouplir les contraintes en introduisant les variables ressort $\xi \geq 0$:

$$t_i(\mathbf{w}^T \mathbf{x}_i + w_0) \geq 1 - \xi_i \,, \;\; i = 1, \ldots, n \,. \tag{1.47}$$

La minimisation de $\frac{1}{2}||\mathbf{w}||^2$ donnée par la formulation primale en (1.41) devient :

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i\,, \tag{1.48}$$

avec $C > 0$. Ce coefficient est défini comme un paramètre de régularisation, il donne un compromis entre la marge et le nombre d'erreurs admissibles. L'ajout du terme $C\sum_i \xi_i$ peut être considéré comme une mesure d'une certaine quantité de mauvaise classification. Ainsi, une faible valeur de C entraîne une faible tolérance. D'autres formulations existent, comme $C\sum_i \xi_i^2$ [Vapnik, 1998].

On peut alors déduire la formulation duale, proche de (1.45), où contrairement au cas linéairement séparable, les α_i ont une borne supérieure :

$$\begin{cases} \max\limits_{\alpha}\left\{\sum\limits_{i=1}^{n}\alpha_i - \frac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_j t_i t_j \mathbf{x_i}^T\mathbf{x_j}\right\}\,, \\ 0 \leq \alpha_i \leq C, i = 1,\ldots,n\,, \\ \sum\limits_{i=1}^{n}\alpha_i t_i = 0\,. \end{cases} \tag{1.49}$$

Comme précédemment, les observations, pour lesquelles $\alpha_i > 0$, sont des vecteurs de support. Aussi, les observations satisfaisant $0 < \alpha_i < C$ sont sur la marge. Enfin, les observations mal classées apparaissent lorsque $\xi_i > 1$.

1.3.3 Extension de la discrimination à plus de deux classes

Il existe plusieurs approches pour étendre les procédures de discrimination de deux classes à plusieurs classes. La plupart des ouvrages sur la reconnaissance de formes traitent de cette généralisation, cependant le lecteur pourra trouver dans [Belaïd and Belaïd, 1992] des informations complémentaires agrémentées d'exemples précis et de démonstrations basées sur le travail de [Tou and Gonzalez, 1974].

La plupart des méthodes de classification linéaire présentées dans cette section peuvent se généraliser en suivant les procédures développées par la suite, cependant on peut trouver dans [Webb, 2002] des descriptions spécifiques pour chacune de ces méthodes, moindres carrés, fonction discriminante de Fisher et SVM.

Posons K le nombre de classes et $\mathcal{C}_1,\ldots,\mathcal{C}_K$ les classes.

1.3.3.1 Une classe contre toutes les autres

Dans ce processus, chaque classe est séparée de toutes les autres par une simple fonction discriminante. Ainsi, les K hyperplans sont construits suivant la propriété suivante :

$$\forall k = 1,\ldots,K\;\; y_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x} + w_{k0} \begin{cases} > 0 & \text{si} \quad \mathbf{x} \in \mathcal{C}_k\,, \\ < 0 & \text{sinon}\,, \end{cases} \tag{1.50}$$

où le vecteur $\mathbf{w}_k$ et le biais w_{k0} sont les paramètres du k-ième hyperplan.

Par exemple, une généralisation de la méthode des moindres carrés (section 1.3.2.2) peut amener intuitivement à cette procédure. En effet, en reprenant l'équation des poids (1.29) et en

codant différemment le vecteur $\mathbf{t}$ définissant l'appartenance aux classes, nous pouvons obtenir directement les K hyperplans par :

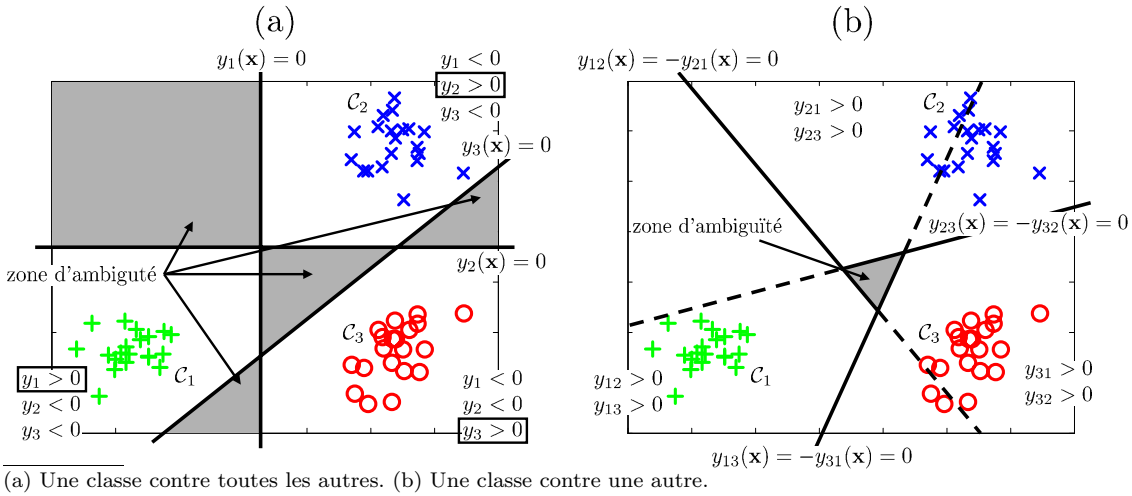
$$\widetilde{\mathbf{W}} = (\widetilde{\mathbf{X}}^T\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}^T\mathbf{T}\,, \tag{1.51}$$

où d'une part, $\mathbf{W}$ caractérise une matrice telle que la k-ième colonne correspond au vecteur des poids élargi $\widetilde{\mathbf{w}}_k = (w_{0k}, \mathbf{w}_k^T)^T$ de la classe $\mathcal{C}_k$. Et d'autre part, $\mathbf{T}$ définit la matrice d'appartenance des observations aux classes, telle que la i-ème ligne de $\mathbf{T}$ correspond le vecteur $\mathbf{t}_i^T$. Par exemple, en présence de trois classes, si l'observation $\mathbf{x}_i$ appartient à la classe $\mathcal{C}_3$ alors $\mathbf{t}_i = (-1, -1, +1)^T$. [Webb, 2002; Bishop, 2006] proposent d'autres procédures de généralisation des moindres carrés.

1.3.3.2 Une classe contre une autre

Dans ce processus, chaque classe est séparée de chaque autre classe par une fonction discriminante différente. Séparés deux à deux, $K(K-1)/2$ hyperplans sont nécessaires pour discriminer les K classes, telles que $y_{ij}(\mathbf{x}) = -y_{ji}(\mathbf{x})$, où y_{ij} définit l'hyperplan séparant la classe $\mathcal{C}_i$ de la classe $\mathcal{C}_j$ sans considérer les autres classes. Ainsi, une observation $\mathbf{x}$ est assignée à la classe $\mathcal{C}_i$ si $y_{ij}(\mathbf{x}) > 0 \;\forall\, j \neq i$.

La figure 1.15 donne un exemple de séparation à trois classes, suivant les deux approches décrites précédemment. La méthode des moindres carrés a été choisie pour déterminer les hyperplans.



(a) Une classe contre toutes les autres. (b) Une classe contre une autre.

FIG. 1.15 – Comparaison d'approches pour séparer les observations d'un problème à trois classes : une classe contre toutes les autres et une classe contre une autre.

1.3.3.3 K fonctions discriminantes

Cette méthode s'approprie les deux approches précédentes, où l'on considère une nouvelle fois K hyperplans $y_k(\mathbf{x})$, $k = 1, \ldots, K$. Une observation $\mathbf{x}$ est assignée à la classe $\mathcal{C}_i$ si $y_i(\mathbf{x}) > y_j(\mathbf{x}) \;\forall\, j \neq i$. On peut alors faire le parallèle avec l'approche « une classe contre une autre », car l'hyperplan séparant deux classes $\mathcal{C}_i$ et $\mathcal{C}_j$ nommé précédemment par $y_{ij}(\mathbf{x})$ peut dans ce cas être défini par :

$$\begin{aligned} y_{ij}(\mathbf{x}) &= y_i(\mathbf{x}) - y_j(\mathbf{x})\,, \\ &= (\mathbf{w}_i - \mathbf{w}_j)^T\mathbf{x} + (w_{i0} - w_{j0})\,. \end{aligned} \tag{1.52}$$

La frontière entre ces deux classes est donc donnée pour $y_i(\mathbf{x}) = y_j(\mathbf{x})$, soit $y_i(\mathbf{x}) - y_j(\mathbf{x}) = 0$.

Sur l'exemple donné à la figure 1.16, les K hyperplans $(y_i(\mathbf{x}) - y_j(\mathbf{x}),\ \forall\, i \neq j,\ i, j = 1, 2, 3)$ sont construits en fonction de ceux obtenus par la méthode « une classe contre une autre » ($y_k(\mathbf{x})$, $k = 1, 2, 3$, voir figure 1.15).

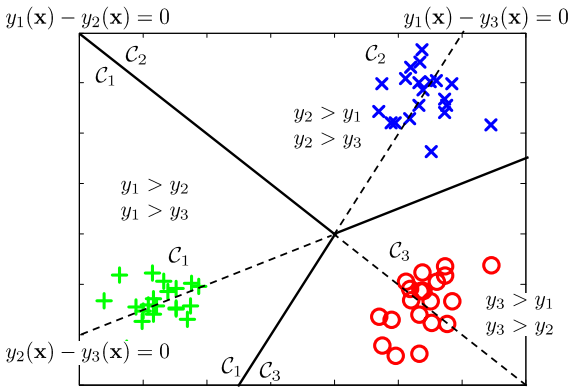


FIG. 1.16 – Séparation linéaire de trois classes, fondée sur un cas particulier de l'approche « une classe contre une autre ».

Cette dernière approche élimine les zones d'ambiguïté. Cependant d'une manière générale, les problèmes d'ambiguïté peuvent se résoudre en prenant en compte d'autres paramètres. Ainsi, en présence d'incertitude sur l'appartenance d'une donnée à une certaine classe, on peut utiliser les probabilités *a priori* des classes pour prendre la décision. Ou encore, on peut attribuer l'observation à la classe dont l'hyperplan est le plus proche.

1.3.4 Conclusions

Cette partie, consacrée à la classification linéaire par des approches discriminantes, a permis d'éviter les difficultés liées aux estimations des densités de probabilité, soulevées dans la conclusion de la section 1.2. Les différentes approches détaillées se différencient par la nature de leur critère à optimiser. La fonction discriminante de Fisher cherche à maximiser la séparation des classes ; cette méthode est un cas particulier des approches probabilistes, détaillées à la section 1.2. La technique des moindres, quant à elle, cherche à minimiser la somme des carrés de l'erreur de chaque observation. L'inconvénient de cette approche est d'obtenir une fonction discriminante très rigide, et fortement dépendante à la présence d'observations aberrantes. Cependant, cette méthode avait permis d'introduire la règle *Delta*, qui apporte plus de souplesse dans la détermination de la fonction discriminante. Cette règle est une technique d'optimisation basée sur une descente de gradient, elle peut, à l'image de l'algorithme du perceptron, s'interpréter de manière connexionniste. Comme expliqué précédemment, l'hyperplan séparateur obtenu par ces différentes, n'est pas optimal. L'apprentissage juge bon de s'arrêter une fois l'objectif atteint. Ainsi, les marges apparaissant entre les classes ne sont pas optimales, et rendent par conséquent la classification peu robuste. L'introduction des *support vector machine* a permis d'optimiser ces marges. Néanmoins, malgré l'amélioration de la classification par ces approches, des limitations apparaissent également. En effet, pour certains problèmes, une séparation linéaire ne peut pas être suffisante car elle comporte le risque de rendre le modèle peu efficace. C'est dans ce contexte qu'il peut être préférable d'employer des méthodes plus complexes, afin d'obtenir des fonctions discriminantes non linéaires.

1.4 Classification non linéaire

1.4.1 Introduction

Pour des problèmes complexes, une séparation linéaire est rarement suffisante pour obtenir de bonnes performances de classification. Des fonctions de décision plus riches sont alors nécessaires, afin d'obtenir des frontières de décision non linéaires et parfois discontinues. On parle alors de régions de décision [Bishop, 1995].

Aux sections 1.4.2 et 1.4.3, nous détaillerons deux des approches décrites précédemment, respectivement les réseaux de neurones et les *support vector machines*, afin de montrer leur adaptation dans des problèmes où les données sont non linéairement séparables. La section 1.4.2.5 apportera des réponses aux problèmes de convergences évoqués sur l'algorithme du perceptron (*cf.* section 1.3.2.3).

1.4.2 Réseaux de neurones artificiels

1.4.2.1 Introduction

Les réseaux de neurones artificiels (RNAs), également appelés réseaux connexionnistes, sont des structures qui prennent leur inspiration dans le fonctionnement élémentaire du système nerveux [Asselin de Beauville and Kettaf, 2005]. Le lecteur, soucieux et intéressé par l'analogie entre l'approche biologique des neurones et le développement théorique des réseaux de neurones artificiels, pourra se référer à [Jodouin, 1994a] qui propose tout au long de son ouvrage de judicieux parallèles. Les paragraphes suivants présentent un bref historique des réseaux de neurones, et peut être complété par les ouvrages de [Jodouin, 1994a; Abdi, 1994; Rennard, 2006].

Traditionnellement, on attribue la naissance des réseaux de neurones aux travaux de [McCulloch and Pitts, 1943]. Ils ont montré que les RNAs peuvent réaliser des fonctions logiques et arithmétiques. La structure employée était fondée sur des neurones logiques ou binaires interconnectés, qui ne connaissent que les réponses 0 ou 1. Ce modèle du neurone inspira la création du perceptron de [Rosenblatt, 1958] et de l'Adaline [Widrow and Hoff, 1960] (*cf.* sections 1.3.2.2 et 1.3.2.3).

Le Perceptron de [Rosenblatt, 1958] fut le premier réseau apprenant, il fut donc une étape importante dans l'histoire des réseaux de neurones. À partir d'observations étiquetées (contexte d'apprentissage supervisé), il est capable d'ajuster les poids d'un neurone, afin de converger vers une configuration apte à réaliser des opérations de classification ou de généralisation. L'Adaline de [Widrow and Hoff, 1960] (pour ADAptive LInear NEuron, devenu plus tard ADAptive LINear Element) est vu comme une extension du perceptron et de sa loi d'apprentissage. On peut noter que la première règle d'apprentissage fut néanmoins donnée par [Hebb, 1949][9], dans un contexte non supervisé.

Ces deux types de réseaux structurés autour d'une seule couche de neurones ne peuvent, dans un contexte de classification, partitionner l'espace d'entrée que de manière linéaire, comme montré précédemment. La seule alternative pour dépasser les limites de ces réseaux monocouches est d'augmenter le nombre de couches de neurones entre les entrées et la couche de neurones de

[9] La règle de Hebb consiste à renforcer la connexion entre deux neurones lorsque ceux-ci sont actifs au même moment. Cette règle ne tient pas compte de l'écart entre les sorties réelles et désirées, tout comme la règle du perceptron qui est une variante de la règle d'apprentissage de Hebb.

sortie. Or, à la fin des années 1960, on ne connaissait pas d'algorithmes permettant de réaliser l'apprentissage de perceptrons constitués de plusieurs couches (perceptrons multicouches). L'analyse et les démonstrations de [Minsky and Papert, 1969] sur les limitations du perceptron ont fortement ralenti les recherches dans ce domaine, en désintéressant la communauté scientifique. Il fallut attendre les années 1980, pour voir réapparaître un regain d'activité sur les réseaux de neurones, notamment par l'intermédiaire du modèle de [Hopfield, 1982] basé sur la notion de mémoire auto-associative.

L'autre moment clé fut la découverte de nouveaux modèles capables de dépasser les limites du perceptron : le plus célèbre est le perceptron multicouches. Plus précisément, la découverte ne fut pas le modèle, mais l'algorithme d'apprentissage, qui permit l'utilisation de modèles plus complexes. En effet, F. Rosenblatt avait déjà pris conscience de la nécessité de couches cachées dans un réseau, afin d'élargir les capacités du perceptron. Ainsi, [Rumelhart *et al.*, 1986] ont publié un nouvel algorithme d'apprentissage, appelé l'algorithme de rétropropagation de l'erreur (*backpropagation*), qui permit l'apprentissage et donc l'optimisation des paramètres de réseaux de neurones à plusieurs couches.

Depuis la recrudescence des travaux réalisés autour des réseaux de neurones, de nombreux modèles sont apparus. Cependant, ils ont tous comme origine le modèle fondateur : le neurone de McCulloch-Pitts. Seuls, les neurones sont capables de faire des opérations élémentaires, pouvant aller jusqu'à la réalisation de fonctions non linéaires simples. Cependant, leur intérêt demeure dans leur association constituant un réseau [Dreyfus *et al.*, 2002]. Cela amène naturellement à définir les réseaux de neurones artificiels, comme des modèles mathématiques, pouvant être vus comme des graphes dont les nœuds sont les neurones, et les arêtes, orientées et pondérées, sont les liens synaptiques (connexions). C'est dans ces liens, appelés poids, que se trouve la connaissance. Ainsi, lorsque les réseaux apprennent, ils construisent dans leurs poids une description numérique du domaine.

L'importance des travaux réalisés sur les réseaux a entraîné le développement de nombreux types de réseaux. Ils peuvent ainsi être classés en deux grandes catégories : les réseaux **non bouclés** (dit aussi réseaux statiques) et les **réseaux bouclés** (réseaux dynamiques), ceux-ci seront décrits dans la prochaine section. D'autre part, les réseaux de neurones peuvent aussi être caractérisés par :

- leur **architecture** ou **topologie**, comprenant la structure d'interconnexion des neurones et les fonctions d'activation utilisées ;
- leur **dynamique** de propagation, propagation vers l'avant (*feedforward*) ou propagation avec retour en arrière (*feedback*) ;
- leur **apprentissage**, supervisé ou non supervisé.

Notons, que nous pouvons également dissocier les réseaux conçus d'un côté pour la description des données (*cf.* section 2.3.3), et d'un autre côté pour la classification automatique (*clustering*, contexte non supervisé) et pour la classification et la prédiction (contexte supervisé). On peut associer ces deux orientations respectivement aux réseaux descriptifs et prédictifs.

1.4.2.2 Architecture

L'architecture décrit la structure d'interconnexion entre les neurones qui composent le réseau : on parle aussi de topologie. Contrairement à ce que nous avons vu à la section 1.3.2.3, les neurones peuvent être arrangés par couches. On distingue alors les couches cachées de neurones et la couche de neurones de sortie. Dans les **réseaux non bouclés** (figure 1.17a-b-c), les entrées des neurones d'une couche sont les sorties des neurones de la couche précédente et les entrées des neurones de la première sont reliées au « monde extérieur ». Dans les **réseaux à connexions locales** (figure 1.17c-d-e), les neurones d'une couche ne sont pas tous forcément reliés à la couche suivante.

Les réseaux non bouclés propagent donc l'information de l'entrée vers la sortie (*feedforward*). Dans les **réseaux bouclés** appelés aussi réseaux récurrents (figure 1.17d-e), on retrouve le même type d'interconnexion entre neurones, mais avec des possibilités de retours en arrière (*feedback*). Ces réseaux sont ainsi utilisés pour la modélisation et la commande dynamique de processus, les rétroactions permettent à ce type de réseaux de présenter un comportement temporel.

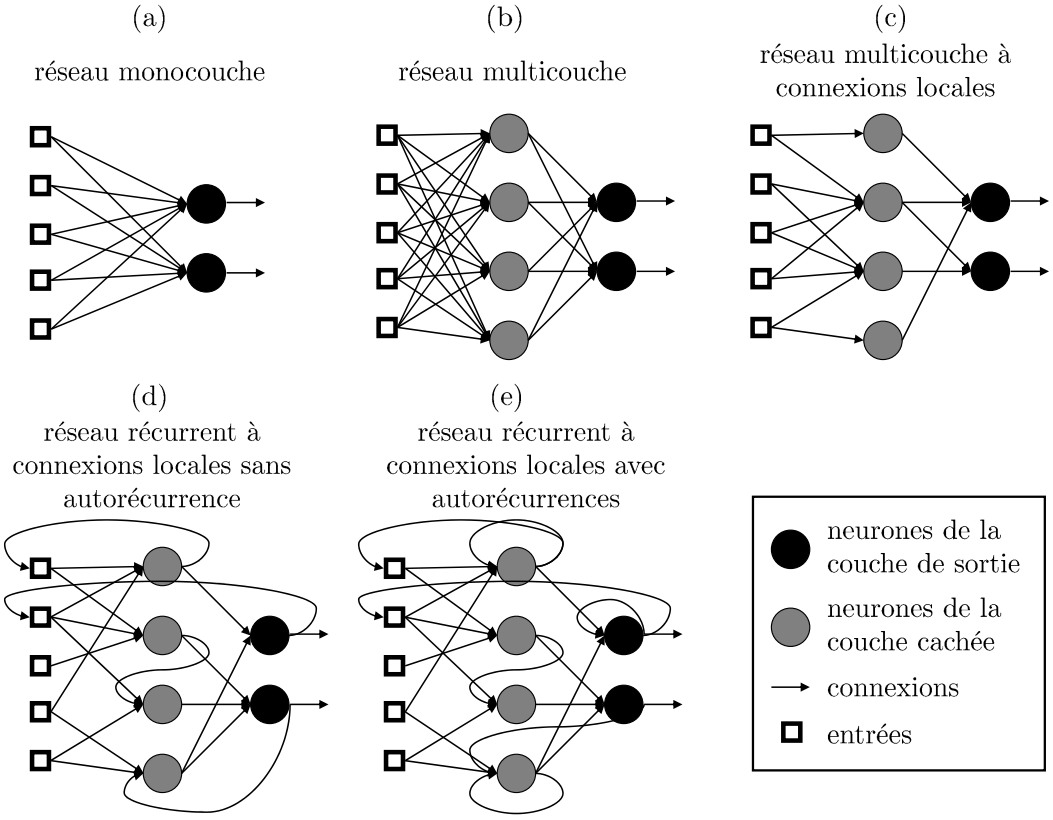
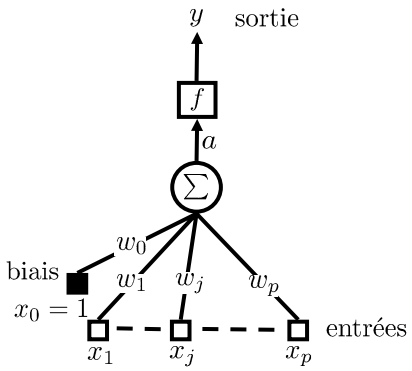


FIG. 1.17 – Principaux types d'architectures et de structures d'interconnexions des réseaux de neurones.

Les réseaux liés à notre problématique sont les réseaux non bouclés, c'est pourquoi nous concentrerons notre attention sur ce type de réseaux et plus particulièrement sur les perceptrons multicouches, qui ont été largement utilisés dans nos travaux.

1.4.2.3 Le neurone

À la section 1.3.2.3, nous avons abordé le neurone lors de la description du perceptron. Ce dernier était constitué d'un seul neurone, la figure 1.18 en redonne une illustration.



- x_j $(j = 1, \ldots, p)$ sont les entrées,
- y est la sortie,
- w_j $(j = 1, \ldots, p)$ sont les poids,
- w_0 est le biais,
- a est est appelé activation ou potentiel du neurone,
- f est la fonction d'activation.

FIG. 1.18 – Représentation détaillée du perceptron [Rosenblatt, 1962].

Précédemment, on a montré qu'un neurone est une fonction algébrique paramétrable, qui réalise une fonction ayant comme argument une combinaison linéaire des entrées x_j pondérées par les poids w_j, avec $j = 1, \ldots, p$. Cette combinaison complétée d'un biais w_0 est appelée **activation** ou **potentiel**, elle est obtenue par la relation suivante :

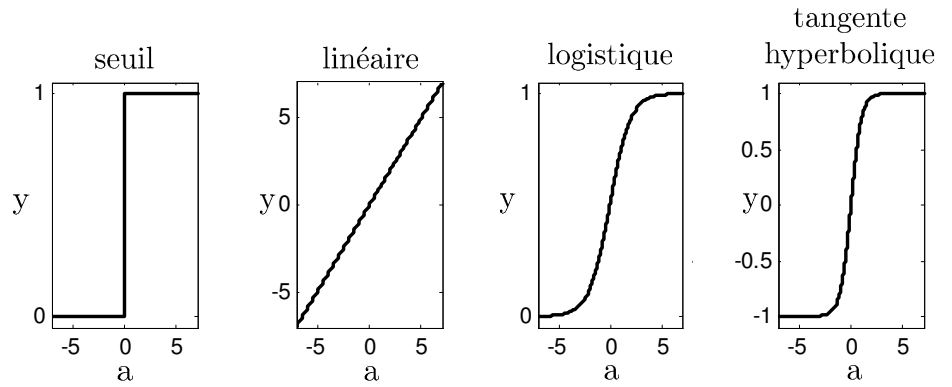
$$a = \sum_{j=1}^{p} w_j x_j + w_0 \,. \tag{1.53}$$

La valeur de sortie du neurone y est transformée par une fonction d'activation $f(\cdot)$, elle est donnée par :

$$y = f\,(a) = f\left(\sum_{i=j}^{p} w_j x_j + w_0\right) \,. \tag{1.54}$$

En reprenant la forme matricielle, on obtient : $y = f\left(\mathbf{w}^T\mathbf{x} + w_0\right)$.

Dans le cas du perceptron (*cf.* section 1.3.2.3), on a vu que le neurone s'active lorsque la somme des entrées pondérées par les poids dépasse une valeur de seuil donnée. La fonction d'activation $f(\cdot)$ dans ce cas correspond à la fonction seuil de la figure 1.19. Cette même figure propose trois autres fonctions d'activation, dont deux non linéaires : logistique et tangente hyperbolique, celles-ci sont appelées fonctions sigmoïdes.



$y = f(a) = \begin{cases} 0 & \text{si } a < 0 \\ 1 & \text{si } a > 0 \end{cases}$ $\quad y = f(a) = a$ $\quad y = f(a) = \dfrac{1}{1+e^{-a}}$ $\quad y = f(a) = \dfrac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$

FIG. 1.19 – Fonctions d'activation usuelles pour les réseaux de neurones.

Remarque 1 *Notons que le remplacement de la fonction d'activation seuil pour le perceptron par une fonction non linéaire, ne change pas la nature linéaire de la discrimination. L'avantage de ce changement donnerait une nature probabiliste de la sortie du réseau.*

1.4.2.4 Apprentissage

L'apprentissage des réseaux de neurones était donné en introduction comme une caractéristique importante. Il s'effectue en modifiant l'intensité des connexions entre les neurones, [Haykin, 1999] définit l'apprentissage comme :

Définition 2 *le mécanisme pour lequel les paramètres libres d'un réseau de neurones (poids) sont adaptés à travers un processus de stimulation par l'environnement dans lequel le réseau est intégré. Le type d'apprentissage est déterminé par la façon dont les changements de paramètres sont mis en œuvre.*

La figure 1.20 illustre les deux types d'apprentissage : supervisé et non supervisé.

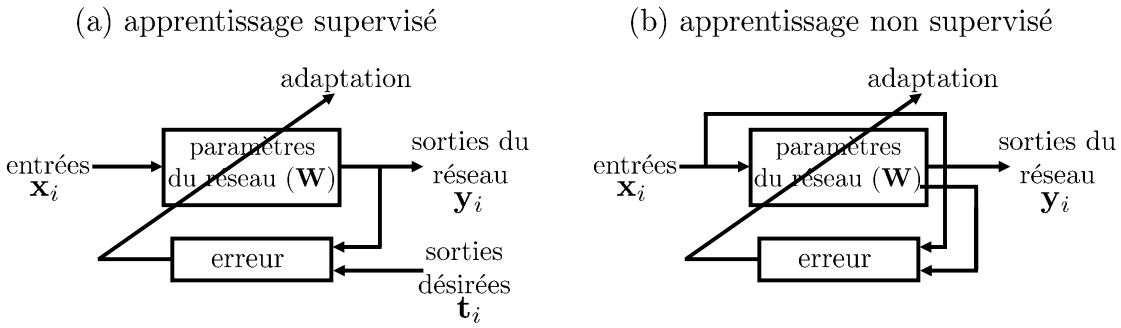


FIG. 1.20 – Synoptique de l'apprentissage supervisé et non supervisé d'un réseau de neurones [Amat and Yahiaoui, 2002].

Dans l'apprentissage non supervisé, on ne connaît pas quelle doit être la sortie du réseau. Le réseau cherche à détecter des points communs entre les observations présentées et adapte les poids afin de donner une sortie équivalente pour des observations d'entrées proches. Cette approche est utilisée notamment dans des applications de *clustering*, où l'on cherche à regrouper des observations. À la section 2.3.3.2, un réseau de neurones auto-organisant, l'algorithme des cartes auto-organisatrices de Kohonen est présenté. Parfaite illustration de l'apprentissage non supervisé des réseaux de neurones, cette technique classique d'agrégation regroupe des données en fonction de régularités statistiques.

L'apprentissage supervisé, qui correspond plus précisément à notre problématique, utilise un « professeur » pour guider le réseau vers la solution recherchée. Pour une entrée propagée dans le réseau, le professeur compare la réponse désirée à celle obtenue par le réseau. À la section 1.3, nous avons vu deux règles d'apprentissage basées sur ce principe : la règle *Delta* et la règle du perceptron (*cf.* sections 1.3.2.2 et 1.3.2.3). Par la règle *Delta*, la modification des poids est proportionnelle à l'erreur commise en sortie, contrairement à la règle du perceptron. Cette différence s'explique par la nature des fonctions d'activation des neurones de sortie, où pour le perceptron la fonction d'activation est un seuil. Le seuillage de l'activation du neurone « biaise » la comparaison entre la sortie désirée et la sortie du neurone, car pour des observations situées d'un même côté de l'hyperplan, la valeur renvoyée par la sortie sera toujours identique. Cette différence amène à distinguer deux cas : **apprentissage par correction d'erreur** , où les poids sont modifiés proportionnellement à la valeur de l'erreur et l'**apprentissage par renforcement**, où l'on ne tient pas compte de l'ampleur de l'erreur.

1.4.2.5 Perceptron multicouches

Lors de l'introduction sur les réseaux de neurones, nous avons relevé l'intérêt du perceptron multicouches dans l'évolution des réseaux de neurones. Rappelons, que l'origine de leur utilisation fut la découverte de règles d'apprentissage adaptées à leur architecture. En effet, l'architecture des perceptrons multicouches (PMC et en anglais *MLP*, pour *multi-layer perceptron*) est fondée sur le perceptron, où des couches de neurones, dites couches cachées, sont ajoutées entre les entrées et la couche de sortie (figure 1.21). L'intérêt de cette évolution est de surpasser les limitations du perceptron, afin d'optimiser la capacité d'apprendre des réseaux à effectuer une tâche.

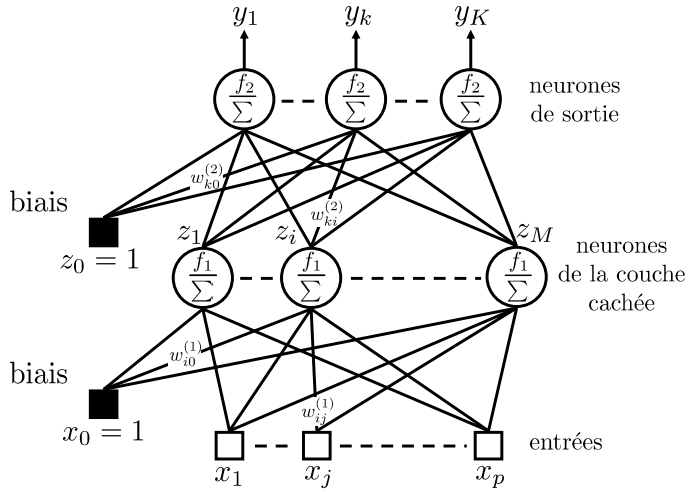


FIG. 1.21 – Illustration d'un réseau de neurones à une couche cachée (PMC) [Bishop, 1995].

La sortie d'un neurone caché (a_i) est obtenue par une combinaison linéaire des p variables d'entrées et d'un biais, toujours pondérées par des poids, telle que :

$$\begin{aligned} a_i &= \sum_{j=1}^{p} w_{ij}^{(1)} x_j + w_{i0}^{(1)} \,, \\ &= \sum_{j=0}^{p} w_{ij}^{(1)} x_j \,, \end{aligned} \tag{1.55}$$

où $w_{ij}^{(1)}$ représente un poids de la première couche (la couche cachée) associant l'entrée x_j au neurone i.

Comme précédemment, la sortie de chaque unité de la couche cachée z_i est vue par une fonction d'activation $f_1(\cdot)$: $z_i = f_1(a_i)$. Dans cette écriture, on considère que l'ensemble des neurones d'une couche ont la même fonction d'activation. Sur le même principe, les sorties du réseau ne considèrent plus directement les entrées mais les sorties des M neurones cachés z_i. On retrouve alors la formulation suivante :

$$\begin{aligned} a_k &= \sum_{i=1}^{M} w_{ki}^{(2)} z_i + w_{k0}^{(2)} \,, \\ &= \sum_{i=0}^{M} w_{ki}^{(2)} z_i \,. \end{aligned} \tag{1.56}$$

Vue par leur fonction d'activation $f_2(\cdot)$, l'expression de la sortie est $y_k = f_2(a_k)$. Ainsi, en combinant les relations (1.55) et (1.56), nous obtenons directement les sorties du réseau de la figure 1.21 en fonction des entrées :

$$y_k = f_2 \left(\sum_{i=0}^{M} w_{ki}^{(2)} f_1 \left(\sum_{j=0}^{p} w_{ij}^{(1)} x_j \right) \right) \,. \tag{1.57}$$

Dans ce type de réseau, nous remarquons que l'information circule des entrées vers les sorties, d'où l'appellation *feedforward*.

Apprentissage

Le regain d'intérêt pour les réseaux de neurones fut en partie lié à la capacité d'apprentissage des réseaux multicouches, grâce à la procédure de rétropropagation. Cette procédure est basée sur une extension de la règle *Delta* : faisant donc intervenir une descente de gradient (*cf.* section 1.3.2.2). Rappelons que la règle *Delta* consiste à propager une observation de l'entrée du réseau à travers la couche de neurones, afin d'obtenir les valeurs de sorties. Celles-ci comparées aux sorties désirées. Elles fournissent alors les erreurs permettant d'adapter les poids des neurones de sortie. Sans couche cachée, la connaissance de l'erreur des neurones de sortie permet un calcul direct du gradient et rend l'adaptation des poids de ces uniques neurones aisée, comme cela avait été montré par la règle *Delta*. Cependant, dans un réseau à couches cachées, ne connaissant pas les sorties désirées des neurones cachés, il demeure alors impossible de connaître les erreurs de ces neurones. Par conséquent, en l'état, ce processus ne peut pas être utilisé pour l'adaptation des poids des neurones cachés.

L'intuition qui permit de résoudre cette difficulté et qui donna naissance à la rétropropagation, fut la suivante : l'activité d'un neurone est liée aux neurones de la couche précédente. Ainsi, l'erreur d'un neurone de sortie est due aux neurones cachés de la couche précédente proportionnellement à leur influence ; donc en fonction de leur activation et des poids qui relient les neurones cachés au neurone de sortie. Dès lors, conformément à la figure 1.22, on cherche à obtenir les contributions des M neurones cachés qui ont donné l'erreur du neurone de sortie k.

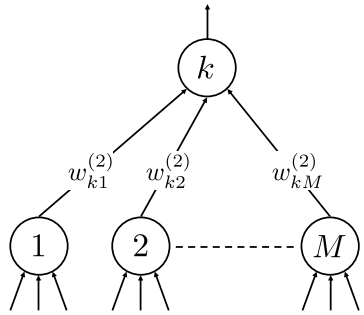


FIG. 1.22 – Relations entre le neurone de sortie k et les M neurones cachés.

La suite de la procédure de la rétropropagation consiste alors à propager dans le réseau le gradient de l'erreur (erreur obtenue lors de la propagation d'une entrée). Cette fois-ci, la propagation de l'erreur d'un neurone de sortie part de la couche de sortie vers les neurones cachés ; rappelant ainsi, le terme de rétropropagation.

Dès lors, pour connaître l'erreur de tous les neurones du réseau, il suffit alors de retracer à l'envers le cheminement de l'activation originale, en partant des erreurs des neurones de sortie. Une fois l'erreur de chaque neurone connue, les relations de modification des poids peuvent être obtenues. La démonstration de la rétropropagation, proposée en annexe A.1, donne l'adaptation des poids du neurone k de la couche de sortie par :

$$\Delta w_{ki}^{(2)} = -\rho \frac{\partial E}{\partial w_{ki}^{(2)}} = -\rho \, (t_k - y_k) \cdot f_2'(v_k) \cdot z_i \, , \tag{1.58}$$

où t_k et y_k sont respectivement la sortie désirée et la sortie réelle du neurone k correspondant à l'entrée $\mathbf{x}$ propagée et v_k donne l'activation de ce neurone. Au terme de la démonstration de la rétropropagation, on obtient l'adaptation des poids du neurone i de la couche cachée par :

$$\Delta w_{ij}^{(1)} = -\rho \, f_1'(u_i) \sum_k \delta_k w_{ki}^{(2)} \cdot x_j \, , \tag{1.59}$$

avec, $\delta_k = (t_k - y_k) \cdot f_2'(v_k)$ et u_i donne l'activation du neurone i de la couche cachée.

L'adaptation des poids ne s'effectue qu'une fois la rétropropagation terminée et l'opération est réitérée jusqu'à ce que le critère d'arrêt soit atteint. Cependant, comme pour les règles d'apprentissage *Delta* et du perceptron, il y a deux façons d'appliquer la rétropropagation :

- une **version séquentielle** (ou **stochastique**), où après chaque présentation d'une entrée, les poids sont modifiés ;

- une **version *batch***, où les dérivées sont accumulées au fur et à mesure que les entrées sont propagées, et les poids sont modifiés une fois le passage de toutes les entrées effectué.

La version *batch* de l'algorithme permet de converger à chaque itération vers des solutions précises, mais elle induit un temps de calcul d'adaptation des poids proportionnel au nombre d'observations d'apprentissage. La version séquentielle n'est pas dépendante du nombre d'observations d'apprentissage, elle est aussi adaptée lorsque toutes ces observations ne sont pas disponibles au début de l'apprentissage. Elle implique un choix aléatoire des observations d'apprentissage, ce qui permet une exploration plus vaste de la fonction d'erreur, et dans certains cas permet également d'éviter des minimums locaux [Rennard, 2006].

La règle de rétropropagation est simple, elle est basée sur la plus forte pente, où le gradient donne la direction à suivre sur la surface d'erreur et le pas d'apprentissage ρ indique la distance à parcourir. Cet algorithme fonctionne correctement si la fonction d'erreur est parfaitement convexe. Cependant, ce cas idéal est rare, et l'utilisation d'un simple gradient ne permet pas toujours d'obtenir le minimum global. Ainsi, la rétropropagation, étant une méthode de gradient, peut être immobilisée dans un minimum local, où les performances du réseau sont nettement sous-optimales. Minsky et Papert dans la réédition en 1988 de leur précédent ouvrage [Minsky and Papert, 1969] ont insisté sur ce problème, montrant que dans sa version originale, la rétropropagation ne permettait pas de pallier cette limitation.

Amélioration de l'apprentissage

Cet algorithme est ainsi très sensible à l'initialisation des poids, qui peut l'entraîner à rester bloqué dans des minimums locaux. C'est pourquoi en pratique, il est courant de lancer plusieurs apprentissages partant chacun d'une initialisation différente, et de conserver le réseau le plus performant. Cette approche est certainement efficace, mais engendre un temps de calcul très important. On notera que ce problème peut également être résolu par différentes heuristiques. Par exemple, on peut ajouter du bruit dans les observations d'apprentissage ou dans la procédure de modification des poids. Bruiter les observations d'apprentissage permet également d'améliorer la robustesse du réseau en le rendant moins sensible aux variations des observations d'entrée [Grandvalet *et al.*, 1997]. En outre, [Franzini *et al.*, 1990; Pearlmutter, 1992] proposent de réduire à chaque itération tous les poids d'une petite quantité afin d'éviter la saturation de l'activation des neurones quand leurs liens possèdent des poids trop grands.

Les difficultés de convergence, liées à la rétropropagation qui dépendent en partie du choix de la valeur du pas d'apprentissage, ont amené, dans les années qui ont suivi l'apparition de cette règle d'apprentissage, aux développements de nombreuses variantes. Celles-ci ont comme objectifs d'accélérer la convergence du processus d'apprentissage et d'améliorer la capacité de généralisa-

tion. En effet, un pas d'apprentissage faible entraîne une convergence lente et l'algorithme, s'il se dirige vers un minimum local, ne peut s'en extirper. [Jodouin, 1994b] évoque alors l'augmentation épisodique et brusque de la taille du pas pour forcer la descente hors de minimums locaux. Cette approche peut paraître approximative, mais elle montre bien la difficulté dans le choix de la valeur du pas d'apprentissage, qui, trop faible, entraîne une convergence lente. En revanche, si le pas est trop élevé cela conduit à des oscillations dues à la présence de vallées et de plateaux sur la surface d'erreur.

Une technique souvent employée pour stabiliser la descente est d'introduire un terme au gradient, appelé terme d'inertie, ou aussi « moment » :

$$\Delta w_{ij}(t) = -\rho \frac{\partial E(t)}{\partial w_{ij}} + \alpha \, \Delta w_{ij}(t-1) \, . \tag{1.60}$$

Cette variante, appelée *backpropagation with momentum*, atténue les effets d'un pas d'apprentissage trop grand, en donnant une inertie à l'adaptation de chaque poids. La modification étant moins brutale permet de réduire les oscillations et donc d'accélérer la convergence.

La littérature propose également d'agir directement sur le pas d'apprentissage, en le diminuant progressivement ou en adaptant sa valeur en fonction de la forme de la surface d'erreur. En effet, une valeur du pas peut être adaptée au début du processus de modification et ne plus l'être par la suite. Ainsi, [Pearlmutter, 1992] propose de commencer avec une faible valeur et de l'augmenter durant l'apprentissage.

D'autres heuristiques permettent une adaptation automatique de la valeur du pas d'apprentissage ou du terme d'inertie. Ainsi, idéalement le pas doit être élevé dans les descentes. Aussi, lorsque le signe de la dérivée de la fonction d'erreur est constant, il doit se réduire en cas d'oscillations, correspondant à des changements de signes [Rennard, 2006]. Suivant ce principe, on trouve plusieurs règles qui adaptent le pas d'apprentissage en fonction du changement de signe de la dérivée ou de la constance. Pour cela, les méthodes comme *resilient backpropagation* (notée *Rprop*) de [Riedmiller and Braun, 1993], *delta-bar-delta* [Jacobs, 1988] ou encore la méthode *Quickprop* de [Fahlman, 1988], considèrent à chaque itération le gradient précédent. Ces techniques abordées par [Jodouin, 1994b; Rennard, 2006] reposent sur l'estimation locale de la forme de la surface d'erreur et adaptent le pas en fonction du signe du gradient.

Malgré ces règles d'optimisation de la valeur du pas, la rétropropagation et ces proches variantes utilisent toujours comme seule information la pente locale de la fonction d'erreur. Ainsi, pour affiner la recherche, la courbure de la fonction d'erreur, donc les variations de la pente, peuvent être considérées en évaluant les dérivées partielles secondes. La nouvelle classe de méthodes plus sophistiquées est fondée sur les algorithmes du second ordre. Ils sont basés sur la méthode de Newton qui adapte les poids par la relation (1.61), où la matrice hessienne $\mathbf{H}$ donne les dérivées secondes de la fonction d'erreur par rapport aux poids : $\mathbf{H}_{ij} = \frac{\partial^2 E}{\partial w_i \partial w_j}$.

$$\Delta w = -\mathbf{H}^{-1} \nabla E \, . \tag{1.61}$$

Comme les méthodes basées uniquement sur le gradient, les méthodes du second ordre déterminent le gradient par l'algorithme de rétropropagation et font généralement une approximation de la matrice hessienne ou de son inverse [Bishop, 1995], car le coût de son calcul peut rapidement devenir prohibitif. Les méthodes connues sous le nom de quasi-Newton généralisent de manière itérative la méthode de Newton en construisant à chaque itération des approximations de plus en plus précises de l'inverse de la matrice hessienne. La méthode la plus utilisée pour la

mise à jour de $\mathbf{H}$ est certainement celle de Broyden-Fletcher-Goldfarb-Shanno (BFGS) [Broyden, 1970]. Cette approche permet de réduire très significativement le nombre d'itérations avant la convergence, cependant l'approximation de $\mathbf{H}^{-1}$ est correcte lorsque la méthode est proche d'un minimum de la fonction d'erreur. Ainsi, comme le conseille [Dreyfus *et al.*, 2002], il est préférable de commencer la minimisation par une méthode de gradient simple, puis d'utiliser la méthode BFGS lorsqu'on estime être proche d'un minimum. Ne disposant pas de règles théoriques pour connaître le moment du passage du gradient à BFGS, l'utilisateur doit par conséquent procéder par tâtonnements.

La méthode de Levenberg-Marquardt [Levenberg, 1944; Marquardt, 1963] est une autre méthode du second ordre. Elle propose une alternative intéressante en modifiant les poids par la relation suivante :

$$\Delta w_{ij} = -\left[\mathbf{H} + \mu\mathbf{I}\right]^{-1} \nabla E\,, \tag{1.62}$$

où $\mathbf{I}$ est une matrice identité. Cette méthode a la particularité de s'adapter à la forme de la surface d'erreur, en faisant un compromis entre la direction du gradient et la méthode de Newton. En effet, pour de petites valeurs de μ, la méthode de Levenberg-Marquardt s'approche de la méthode de Newton, et pour de grandes valeurs de μ, l'algorithme est tout simplement fonction du gradient. Notons que le pas μ est adapté automatiquement par l'algorithme. Malgré les propriétés intéressantes de cette méthode, elle nécessite le calcul de l'inverse de $[\mathbf{H} + \mu\mathbf{I}]$, rendant son utilisation délicate pour des réseaux possédant beaucoup de poids. Ainsi, [Dreyfus *et al.*, 2002] proposent dans la pratique d'utiliser la méthode BFGS si le réseau possède beaucoup de poids et, la méthode Levenberg-Marquardt dans le cas contraire. Les méthodes du second ordre diminuent considérablement le nombre d'itérations pour arriver à un optimum, mais elles augmentent le temps de calcul. [Fletcher, 1987; Press *et al.*, 1992; Bishop, 1995] détaillent dans leur ouvrage ces méthodes d'optimisation.

Pour résumer les méthodes d'adaptation des poids d'un perceptron multicouches, nous avons donc évoqué : les méthodes de gradient, avec les variantes à pas variable et les méthodes du second ordre.

Dans la remarque 1 (page 41), nous avons noté l'inutilité d'utiliser des fonctions d'activation non linéaires pour le perceptron afin d'obtenir une frontière de décision non linéaire. Cette limitation des réseaux simple couche montre l'importance des couches cachées de neurones. Cependant si les fonctions d'activation dans les couches cachées sont linéaires, alors l'ajout de couches n'a pas d'effet. En effet, sans couche cachée la sortie est : $\mathbf{y} = \mathbf{W}^T\,\mathbf{x}$. Avec une couche cachée composée de fonctions d'activation linéaires, la sortie devient :

$$\mathbf{y} = \mathbf{W}_2(\mathbf{W}_1^T\,\mathbf{x}) = (\mathbf{W}'^T\,\mathbf{x})\,,\ \text{ avec } \mathbf{W}' = \mathbf{W}_1^T\,\mathbf{W}_2\,. \tag{1.63}$$

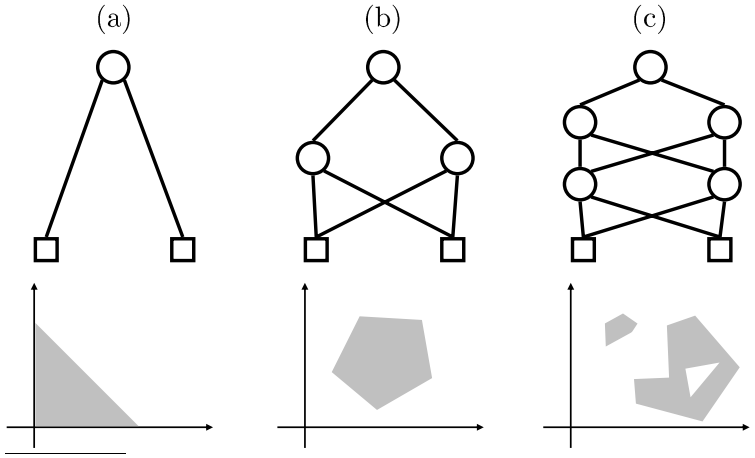
On peut alors remarquer l'inefficacité de la couche cachée, où la relation entre l'entrée $\mathbf{x}$ et la sortie $\mathbf{y}$ est toujours linéaire. En revanche, pour des fonctions d'activation non linéaires, la sortie devient alors :

$$\mathbf{y} = f\left(\mathbf{W}_2\ f\left(\mathbf{W}_1^T\,\mathbf{x}\right)\right) \neq f\left(\mathbf{W}'^T\,\mathbf{x}\right)\ ,\ \text{ avec } \mathbf{W}' = \mathbf{W}_1^T\,\mathbf{W}_2\,. \tag{1.64}$$

L'inégalité obtenue montre le caractère non linéaire d'un réseau possédant des fonctions d'activation non linéaires au sein des couches cachées.

Complexité de la topologie

Les frontières de décision séparant les classes dans un réseau simple couche sont linéaires. L'augmentation de couches, par l'insertion de couches cachées implique une augmentation de la complexité des régions identifiant les différentes classes, comme montré à la figure 1.23. Ainsi, trois couches de neurones (figure 1.23c) sont nécessaires pour représenter des régions disjointes.

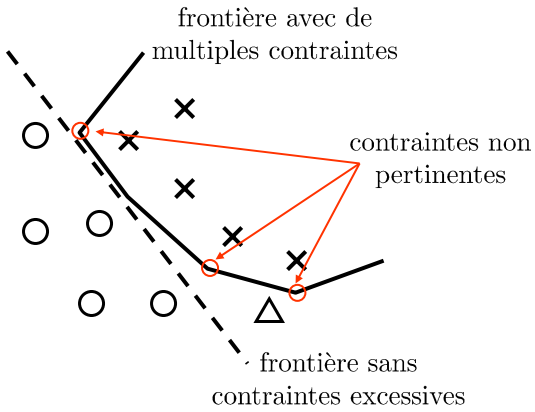


Note : (a) Un perceptron avec un seul neurone réalise une séparation linéaire. (b) Un réseau avec une couche cachée peut réaliser une région de décision suivant une forme convexe. (c) Un réseau avec deux couches cachées, donc avec trois couches de poids peut générer des régions de décisions arbitraires.

FIG. 1.23 – Influence de l'architecture d'un réseau de neurones sur les frontières de décision [Bishop, 1995].

Théoriquement, il n'y a pas de limitation dans le nombre de couches cachées que peut comporter un réseau. Cependant en pratique, il est rare d'en avoir plus de deux, et bien souvent, une seule couche cachée suffit. L'une des raisons est le temps de calcul, car l'ajout de couches entraîne une augmentation du nombre de poids, rendant l'apprentissage et la convergence excessivement longs.

En outre, l'ajout de couches entraîne une augmentation de contraintes, pouvant être non pertinentes (voir figure 1.24) et provoquant dans ce cas un risque de surapprentissage. Le phénomène du surapprentissage (appelé *overfitting* ou *overtraining* par les anglo-saxons) sera abordé précisément à la section 3.3.1.3, mais nous pouvons toutefois signaler que ce phénomène s'apparente à une intense mémorisation des observations d'apprentissage, à l'image d'un élève apprenant ses cours « par cœur ». Le problème résultant de ce genre d'apprentissage endommage fortement les capacités de généralisation.



Note : La classification de l'observation « Δ » se voit par la frontière issue d'un modèle complexe, associée à la classe « ○ », alors qu'il paraît plus judicieux de la lier aux observations « × ».

FIG. 1.24 – Influence de contraintes non pertinentes, résultant d'un modèle trop complexe (ajout de neurones cachés et surapprentissage), sur les frontières de décision.

Sans connaissance *a priori* du problème, la difficulté majeure dans la conception d'un réseau est de déterminer son architecture. En effet, il n'existe pas d'outils analytiques permettant de connaître le nombre idéal de couches cachées et de neurones par couche ; le choix délicat de ces paramètres revient donc à l'utilisateur. Cependant, des auteurs comme [Baum and Haussler, 1988; Murata *et al.*, 1994; Kohavi, 1995; Rudolph, 1997] ont proposé des heuristiques permettant de trouver le nombre de neurones nécessaires à la résolution d'un problème. Par exemple, [Baum and Haussler, 1988] proposent une règle empirique pour trouver le nombre de neurones cachés, en fonction du nombre et de la dimension des observations d'apprentissage, du nombre de neurones de sortie et de l'erreur empirique acceptable. Cette règle, comme le souligne [Rennard, 2006], permet de privilégier la généralisation au détriment de la mémorisation. Elle ne permet pas d'obtenir le nombre de couches et [Rennard, 2006] observe également le caractère aléatoire de cette règle, la rendant peu fiable. En effet, l'ensemble des propositions faites par la littérature reposent sur des expérimentations, ne s'appliquant par conséquent qu'à des cas particuliers. Par ailleurs, afin de trouver la meilleure architecture, les utilisateurs doivent procéder par essais. Ainsi, dans la pratique, des méthodes plus générales sont possibles, mais ont l'inconvénient d'être extrêmement lourdes. En effet, pour trouver l'architecture idéale, il est conseillé de procéder pas à pas. Ainsi, [Fahlman and Lebiere, 1990] proposent, par la méthode *cascade-correlation*, de commencer avec un réseau simple et d'augmenter progressivement la taille du réseau, en ajoutant si nécessaire des neurones ou des couches cachées afin d'améliorer les performances du réseau. À l'inverse, les méthodes OBD (*Optimal Brain Damage*) et OBS (*Optimal Brain Surgeon*) introduites respectivement par [Le Cun *et al.*, 1990] et [Hassibi *et al.*, 1993; Hassibi and Stork, 1993] sont des méthodes dites d'élagage (*pruning*). Elles limitent la complexité du réseau en supprimant, au terme de l'apprentissage, des connexions entre neurones, donc des poids qui ont peu d'influence sur l'erreur de sortie ou qui sont nuisibles au bon fonctionnement du réseau.

L'apprentissage avec régularisation est une autre famille de méthodes très utilisées. Il se fonde sur des techniques de **pénalisation** qui consistent à modifier la fonction d'erreur afin de pénaliser les poids peu utiles au réseau. Ces approches s'apparentent aux méthodes d'élagage qui supposent, qu'initialement, le réseau a été surdimensionné. La méthode de modération des poids (*weight decay*) réduit ainsi progressivement les connexions inutiles, et neutralise le phénomène de surapprentissage.

À l'image de l'apprentissage avec bruit, une autre technique très utilisée est l'« arrêt prématuré » (*early stopping*). En effet, contrairement aux techniques précédentes, cette technique ne modifie pas et n'améliore pas la topologie du réseau, mais arrête l'apprentissage du réseau avant que celui-ci ne mémorise par cœur les exemples d'apprentissage. Elle repose sur l'évaluation périodique de la performance de généralisation du réseau de neurones durant sa phase d'apprentissage et sur des observations non utilisées durant l'apprentissage. Ainsi lors de l'apprentissage, les observations de validation permettent d'observer l'erreur de généralisation en même temps que l'erreur d'apprentissage, comme montré à la figure 1.25. Ainsi, si l'erreur de prédiction de validation augmente pendant que l'erreur sur la base d'apprentissage continue de diminuer, on peut alors considérer que le réseau entre dans la zone de surapprentissage. Afin de l'éviter, il suffit d'arrêter l'apprentissage au moment où les deux erreurs divergent (voir figure 1.25). En pratique cela n'est pas aussi simple, car malgré l'allure générale donnée à la figure 1.25, l'erreur de validation fluctue autour de cette allure et possède donc plusieurs minima locaux. Le problème se pose alors de savoir à quel moment arrêter l'apprentissage. [Prechelt, 1998] a étudié et comparé plusieurs critères permettant de définir ce moment sur une étude concernant l'utilisation de perceptrons multicouches. D'autres détails sont donnés à la section 3.3.1.3.

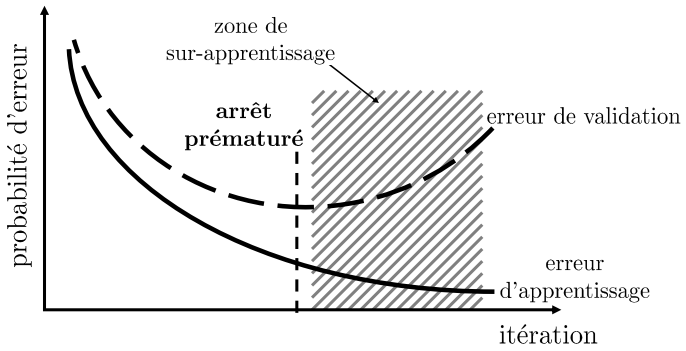


FIG. 1.25 – Arrêt prématuré de l'apprentissage afin d'éviter le surapprentissage.

Toutes les approches présentées ont un but commun, le contrôle de la complexité du réseau, afin d'améliorer les performances de généralisation et la rapidité du processus d'apprentissage. Nous retrouvons ces heuristiques accompagnées d'autres techniques dans le chapitre 9 (*Learning and Generalization*) de l'ouvrage de [Bishop, 1995] ou encore dans le chapitre 14 (Heuristique pour la généralisation) de [Thiria *et al.*, 1997].

En outre, comme il le sera montré à la section 3.2, la qualité d'un modèle est fortement dépendante de ces paramètres et leur nombre est un critère important dans le choix du modèle. Celui-ci est alors d'autant plus efficace qu'il possède le moins de paramètres possible : donc pour les réseaux de neurones moins de couches et moins de neurones. Dans cette perspective, [Dreyfus *et al.*, 2002] cherchent la modélisation la plus **parcimonieuse**, afin d'améliorer la robustesse et la capacité de généralisation des modèles. [Dreyfus *et al.*, 2002] évoquent alors les travaux de [Barron, 1993] qui montre que le nombre de paramètres croît exponentiellement avec le nombre de variables pour des modèles linéaires et que ce même nombre croît linéairement pour des modèles non linéaires. Par conséquent, une modélisation dépendant de paramètres non linéaires est plus parcimonieuse qu'une modélisation résultant de paramètres linéaires.

Ainsi, pour que le réseau respecte cette caractéristique fondamentale, il faut que les fonctions d'activation des neurones cachés soient non linéaires, par exemple de type sigmoïde. Aussi, [Hornik *et al.*, 1989] montrent qu'un réseau composé d'une seule couche cachée en nombre fini et à fonctions d'activation non linéaire est capable d'approximer toute fonction bornée. Ces remarques amènent à caractériser certains réseaux de neurones comme des modèles parcimonieux et des approximateurs universels. Ainsi, l'une des architectures privilégiées d'un réseau de neurones pourrait alors considérer une seule couche cachée de neurones possédant des fonctions d'activation de type sigmoïde, et une couche de sortie. Il est toutefois nécessaire de relativiser ces remarques qui ont été démontrées de manière générale et qui peuvent se révéler fausses pour un problème particulier.

1.4.2.6 Réseaux RBF

Les réseaux RBF (*Radial Basis Function*) sont également considérés comme des réseaux à couches. Ces réseaux approximent un comportement par une collection de fonctions, appelées fonctions noyaux. Ces réseaux utilisent des fonctions locales qui donnent des réponses uniquement dans un domaine restreint et qui est défini par leur champ récepteur. Généralement, ce champ est circulaire de centre $\mathbf{c}_i$, il permet d'obtenir la distance d_i entre une observation $\mathbf{x}$ et le centre de la fonction i :

$$d_i = ||\mathbf{x} - \mathbf{c}_i|| \,. \tag{1.65}$$

La réponse de la fonction noyau est donc maximale en son centre et décroît généralement de façon monotone avec la distance. La fonction la plus utilisée est la gaussienne qui est décrite par deux paramètres : la position de son centre $\mathbf{c}_i$ et la taille σ_i de son champ récepteur.

Architecture

Nous avons précédemment décrit les réseaux RBF comme des réseaux à une couche cachée, cependant, contrairement aux PMC, les réseaux RBF possèdent uniquement des poids reliant l'unique couche cachée à la couche de sortie, comme le montre la figure suivante 1.26.

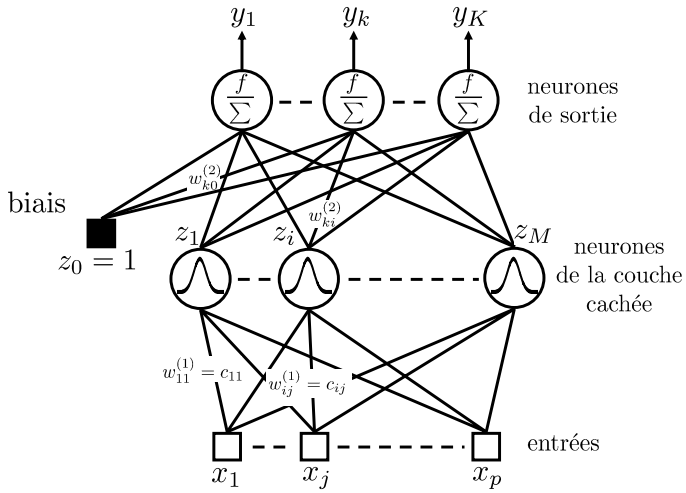


FIG. 1.26 – Réseau RBF.

Ainsi, les paramètres liant les entrées aux neurones cachés ne sont plus des poids ordinaires, mais définissent les centres des fonctions noyaux. Le paramètre du neurone est désormais le champ récepteur. Avec ce type de réseaux, chaque neurone caché détermine la distance (1.65) séparant le centre de son noyau à l'observation d'entrée et la réponse d'un neurone i. Son activation z_i est, donc, proportionnelle à cette distance :

$$z_i = \exp\left(-\frac{d_i^2}{2\sigma_i^2}\right) . \tag{1.66}$$

La fonction d'activation f des neurones de la couche de sortie est choisie comme pour les PMC et la relation d'un neurone de sortie k peut alors être donnée par :

$$y_k(\mathbf{x}) = f\left(\sum_i (w_{ki}^{(2)} \exp\left(-\frac{d_i^2}{2\sigma_i^2}\right) + w_{k0}^{(2)}\right) . \tag{1.67}$$

Dans un contexte de classification, l'attribution de l'appartenance à la classe d'une observation est obtenue par la couche des neurones de sortie en fonction des activations des neurones cachés. L'illustration proposée à la figure 1.27 montre la représentation des classes obtenues par ce type de réseau. L'activation des neurones à l'observation d'entrée Δ est plus forte pour le neurone i. Enfin, on notera que de part leur architecture, les réseaux RBF ne produisent plus d'hyperplans séparateurs pour réaliser la classification.

Apprentissage

L'apprentissage des réseaux RBF est réalisé généralement en deux tâches distinctes :

- la détermination des paramètres des fonctions noyaux (couche cachée de neurones) ;

- l'adaptation des poids de la couche de neurones de sortie.

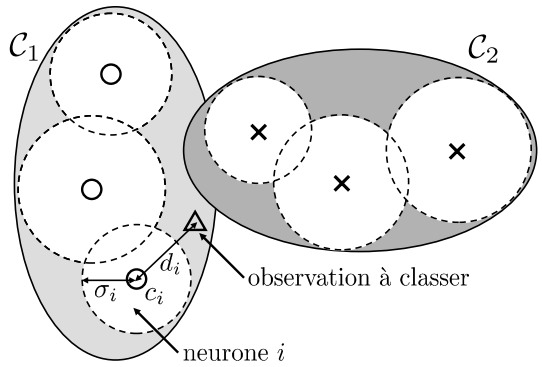


FIG. 1.27 – Illustration de la représentation des classes par un réseau RBF.

Le nombre et la position des neurones de la couche cachée sont souvent déterminés par des techniques d'agglomération (*clustering*. Par exemple, [Mustawi *et al.*, 1992], s'inspirant de l'algorithme *K-means*, proposent dans un premier temps d'associer un neurone à chaque observation d'apprentissage. Puis ils déterminent le nombre et la position des neurones en regroupant itérativement des neurones proches. La taille du champ récepteur doit être choisie judicieusement afin d'obtenir un bon compromis entre la performance d'apprentissage et la capacité de généralisation. En effet, si les champs récepteurs sont trop petits, le réseau peut alors avoir des difficultés à généraliser. En outre, si les champs sont trop grands, ils risquent de se chevaucher et entraîner des ambiguïtés. Le compromis privilégie souvent la taille du champ récepteur d'un neurone i en fonction de la distance séparant le centre du neurone $\mathbf{c}_i$ à une observation d'apprentissage, que nous notons $\mathbf{x}_r$. Respectant la contrainte suivante $||\mathbf{x}_r - \mathbf{c}_i|| < ||\mathbf{x}_r - \mathbf{c}_l|| \ \forall\, \mathbf{c}_l$, cette observation est choisie telle que :

$$\mathbf{x}_r = \underset{\forall \mathbf{x}_j}{\operatorname{argmax}} \{||\mathbf{x}_j - \mathbf{c}_i||\} \ . \tag{1.68}$$

La dernière étape de l'apprentissage est donc l'adaptation des poids de la couche de sortie. Comme nous l'avons évoqué, les réseaux RBF possèdent une seule couche de poids, permettant ainsi d'utiliser des règles simples d'adaptation des poids, telles que la règle *Delta* (*cf.* section 1.3.2.2).

1.4.2.7 Conclusions

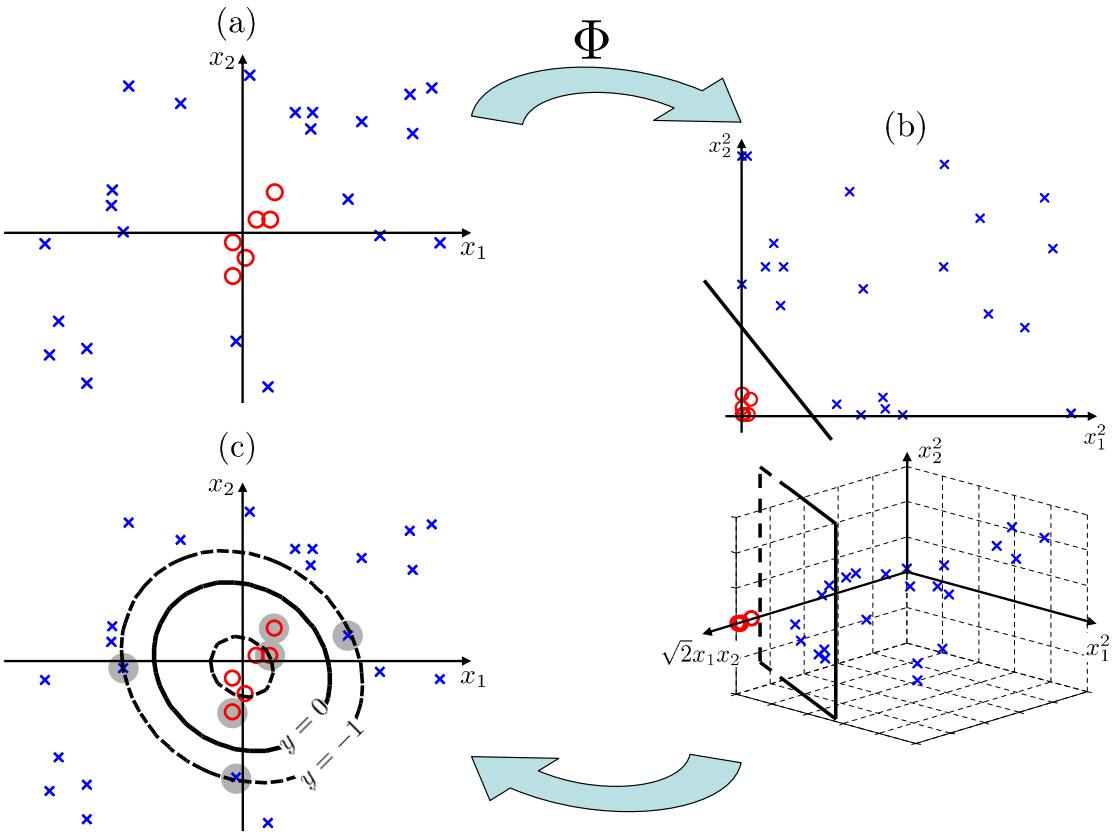
Les réseaux RBF sont à l'image des PMC à une couche cachée, considérés comme des approximateurs universels [Cybenko, 1989]. L'avantage principal des réseaux RBF par rapport aux PMC réside dans leur apprentissage, qui est beaucoup plus simple et rapide. Cependant, la facilité d'utilisation des réseaux RBF est contrebalancée par des performances moins bonnes que celles des PMC, notamment lorsque les observations sont bruitées ou lorsque celles-ci sont représentées dans de grandes dimensions.

[Rennard, 2006] nous fait remarquer que les réseaux RBF permettent de déformer l'espace des variables afin de rendre linéairement séparables des problèmes qui ne l'étaient pas au départ. Cette particularité est démontrée en annexe B, où comme avec les perceptrons multicouches, les réseaux RBF permettent de résoudre le problème du *XOR*.

1.4.3 *Support vector machines* non linéaires

Les SVM ont été introduits dans le cadre de classification linéaire à deux classes, et nous avons pu montrer leur capacité à trouver une frontière de décision maximisant la marge entre les observations de chaque classe. En présence de données non linéairement séparables, l'utilisation de variables ressorts permet de tolérer des erreurs de manière à trouver l'hyperplan optimal. La résolution de ce type de problème peut également être effectuée par une transformation des observations d'entrée dans un espace de plus grande dimension, appelé espace de redescription. Cette transformation a pour but de révéler de nouvelles caractéristiques permettant de séparer linéairement les données dans ce nouvel espace. En effet, on peut penser qu'intuitivement plus la dimension de l'espace de description est grande, plus la probabilité d'y trouver un hyperplan optimal est élevée. C'est donc dans ce nouvel espace que les SVM vont être utilisés.

Ce principe est illustré par l'exemple proposé à la figure 1.28(a), où le problème consiste à discriminer idéalement les deux classes (« × » et « ○ ») qui sont non linéairement séparables. Les observations sont caractérisées par deux variables x_1 et x_2. Ainsi, à la figure 1.28(b), on peut observer que la transformation Φ des observations par une projection en deux et trois dimensions, permet dans cette nouvelle représentation une séparation linéaire. La figure 1.28(c) montre la séparation des données dans l'espace original. Dans la suite de cette section, nous apporterons des explications supplémentaires sur la projection des observations par la transformation Φ.



Note : (a) Observations représentées dans l'espace d'origine. (b) Observations et hyperplan représentés dans l'espace de redescription (deux et trois dimensions). (c) Retour à l'espace d'origine et expression de la fonction discriminante.

FIG. 1.28 – Illustration de la transformation de l'espace par une fonction noyau (dans le cadre des *support vector machines*) sur un exemple de discrimination non linéaire.

Rappelons que les SVM déterminent l'hyperplan séparateur par le calcul des produits scalaires entre les vecteurs d'observation de l'espace d'entrée. Aussi, nous avons pu également remarquer que ce même calcul de l'hyperplan est fonction uniquement du nombre de ces observations d'entrées (plus précisément des vecteurs de support ($\mathbf{x} \in \mathcal{SV}$), souvent bien inférieurs aux nombre d'observations), n'attachant donc pas d'importance à la dimension de l'espace. Par conséquent, la transformation de ces observations dans un espace de plus grande dimension ne devrait pas poser de problèmes si le calcul du produit scalaire des données transformées est toujours possible.

La transformation de l'espace d'entrée peut être réalisée par Φ, ainsi la transformation $\Phi(\mathbf{x})$ d'un vecteur d'entrée $\mathbf{x}$ est donnée par :

$$\mathbf{x} = (x_1, \ldots, x_p)^T \mapsto \Phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_p(\mathbf{x}), \ldots, \phi_q)^T \,,\ q > p \,. \tag{1.69}$$

La dimension q du vecteur $\Phi(\mathbf{x})$ est beaucoup plus grande que la dimension p du vecteur d'entrée original, de manière à favoriser la découverte d'un hyperplan séparateur linéaire. Ainsi, l'hyperplan séparateur peut s'exprimer comme suit :

$$y(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) + w_0 \,. \tag{1.70}$$

Dès lors, le problème d'optimisation intégrant la transformation peut se réécrire de la façon suivante :

$$\begin{cases} \max\limits_{\alpha} \left\{ \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j t_i t_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \right\} , \\ 0 \leq \alpha_i \leq C, i = 1, \ldots, n \,, \\ \sum\limits_{i=1}^{n} \alpha_i t_i = 0 \,. \end{cases} \tag{1.71}$$

Comme précédemment, les observations pour lesquelles $\alpha_i > 0$ sont des vecteurs de support. Les observations satisfaisant $0 < \alpha_i < C$ sont sur la marge. De plus, les observations pour lesquelles $\xi_i > 1$ sont mal classées.

Rappelons que l'hyperplan séparateur est obtenu par le calcul des produits scalaires entre les vecteurs des observations de l'espace d'entrée. Cependant, l'apparition de la transformation Φ dans la nouvelle expression du problème peut montrer des limites lorsque cette nouvelle représentation, issue de la transformation, est exprimée dans une très grande dimension, voire dans une dimension infinie. Il devient alors difficile, et même impossible de calculer le produit scalaire $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle$.

Ce problème peut être contourné en calculant les produits scalaires à partir des vecteurs des observations d'entrée initiale, et non à partir des vecteurs transformés (comme cela a été proposé en 1.71). Cette simplification est réalisée par des **fonctions noyaux** [Scholkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004] qui seront notées $K(\mathbf{x}, \mathbf{x}')$. Ainsi, la transformation opérée par Φ est donc réalisée par une fonction noyau et afin d'obtenir l'équivalence : $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle \equiv K(\mathbf{x}, \mathbf{x}')$, la fonction noyau doit être symétrique et satisfaire la condition de Mercer :

$$\int K(\mathbf{x}, \mathbf{x}') g(\mathbf{x}) g(\mathbf{x}') d\mathbf{x} d\mathbf{x}' \geq 0 \,, \tag{1.72}$$

pour toutes fonctions g satisfaisant :

$$\int g^2(\mathbf{x}) d\mathbf{x} < +\infty \,. \tag{1.73}$$

Dans ces conditions, le produit scalaire est remplacé par la fonction noyau telle que :

$$\sum_j \phi_j(\mathbf{x})\phi_j(\mathbf{x}') = K(\mathbf{x}, \mathbf{x}')\,, \tag{1.74}$$

et par conséquent, l'espace dit de redescription n'est jamais explicité.

Les conditions décrites précédemment visant à valider l'utilisation d'un noyau ne sont pas vérifiables facilement. C'est pourquoi, il est d'usage d'utiliser des familles de fonctions noyaux conventionnelles, comme :

- $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}' + c)^d$, fonction polynomiale, où d et c définissent respectivement le degré du polynôme et une constante ;
- $K(\mathbf{x}, \mathbf{x}') = e^{-\frac{||\mathbf{x}^T\mathbf{x}'||^2}{2\sigma^2}}$, fonction gaussienne, appelée aussi fonction à base radiale (*RBF*, *cf.* section 1.4.2.6), où σ^2 définit l'écart-type ;
- $K(\mathbf{x}, \mathbf{x}') = tanh(a(\mathbf{x}^T\mathbf{x}' - b))$, fonction sigmoïde (cf. section 1.4.2.3).

Une alternative est possible, combiner des fonctions noyaux existantes, de manière à en créer de nouvelles [Cornuéjols and Miclet, 2002; Shawe-Taylor and Cristianini, 2004]. De nombreux ouvrages, comme ceux de [Scholkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004], traitent spécifiquement des méthodes à base de noyaux. En outre, la plupart des ouvrages traitant de l'apprentissage machine ou de la reconnaissance de formes, abordent désormais les méthodes noyaux, comme [Webb, 2002; Bishop, 2006; Theodoridis and Koutroumbas, 2006].

Prenons un exemple simple utilisé notamment à la figure 1.28, et considérons la fonction noyau suivante (fonction polynomiale) $K(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T\mathbf{x}')^2$, ainsi qu'un espace d'entrée de deux dimensions $\mathbf{x} = (x_1, x_2)$. Les nouvelles caractéristiques provenant de la transformation Φ sont alors obtenues par :

$$\begin{aligned} K(\mathbf{x}, \mathbf{x}') &= (\mathbf{x}^T\mathbf{x}')^2 = (x_1x'_1 + x_2x'_2)^2\,, \\ &= x_1^2x_1'^2 + 2x_1x'_1x_2x'_2 + x_2^2 + x_2'^2\,, \\ &= (x_1^2, \sqrt{2}x_1x_2, x_2^2)^T(x_1'^2, \sqrt{2}x'_1x'_2, x_2'^2)\,, \\ &= \Phi(\mathbf{x})^T\Phi(\mathbf{x}')\,. \end{aligned} \tag{1.75}$$

La transformation Φ correspond donc à $[x_1^2, \sqrt{2}x_1x_2, x_2^2]$ et illustre la projection des données d'un espace de deux dimensions en un espace à trois dimensions.

Comme le montre la démonstration précédente, on s'aperçoit que le produit scalaire de deux observations appartenant à l'espace engendré par la transformation Φ revient à calculer la valeur de la fonction noyau pour ces mêmes observations. Par conséquent, la connaissance de l'espace de redescription n'est pas nécessaire pour déterminer le produit scalaire de deux observations décrit dans cet espace.

Désormais, en connaissant la fonction noyau et l'égalité $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$, nous pouvons redéfinir la formulation duale, telle que :

$$\begin{cases} \max\limits_{\alpha} \left\{ \sum\limits_{i=1}^{n} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{n} \alpha_i \alpha_j t_i t_j K(\mathbf{x}_i, \mathbf{x}_j) \right\}, \\ 0 \leq \alpha_i \leq C, i = 1, \ldots, n, \\ \sum\limits_{i=1}^{n} \alpha_i t_i = 0. \end{cases} \tag{1.76}$$

On peut remarquer que la formulation du problème avec ou sans noyau (section 1.3.2.4), est peu différente, nous avons juste remplacé le produit scalaire $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ par $K(\mathbf{x}_i, \mathbf{x}_j)$. Ainsi, l'hyperplan séparateur est donné simplement par l'équation :

$$y(\mathbf{x}) = \sum_{\mathbf{x}_i \in \mathcal{SV}} \alpha_i t_i K(\mathbf{x}_i, \mathbf{x}) + w_0 . \tag{1.77}$$

1.4.4 Conclusions

Dans cette partie consacrée à la classification non linéaire, la description des deux principales approches, réseaux de neurones et *support vector machines*, a permis d'apprécier leurs grandes capacités d'adaptation à la configuration des données. D'un point de vu théorique, ces méthodes paraissent très efficaces, en dépit du nombre de paramètres à déterminer ; particulièrement pour les réseaux de neurones. En effet, comme nous le verrons dans la description de la partie expérimentale, l'utilisation des réseaux de neurones en pratique nécessite une mise en œuvre rigoureuse.

1.5 Résumé et discussions

L'étendue de ce chapitre montre la diversité des approches et des méthodes qui peuvent être utilisées pour résoudre un problème de classification. Fondées sur des approches génératives ou discriminantes, les méthodes s'emploient en fonction de la configuration des données, et de la nature de la sortie voulue (directe ou probabiliste). Notons que pour de nombreuses méthodes discriminantes, leurs sorties sont exploitées de manière à obtenir une probabilité d'appartenance, comme le présentent par exemple [Suykens *et al.*, 2002; Lu, 2005; Bishop, 2006].

La capacité des méthodes de classification non linéaire semble être particulièrement intéressante et performante. Cependant, avant de mettre en œuvre de tels modèles, il faut s'assurer de la nécessité des propriétés non linéaires du modèle. En effet, si la classification ne nécessite pas une complexité particulière, l'utilisation de méthodes linéaires serait plus judicieuse. Si l'utilisation de méthodes non linéaires est nécessaire, le choix entre les *support vector machines* ou les réseaux de neurones ne peut *a priori* être décidé avant de les avoir comparés.

En introduction de ce chapitre, nous avons évoqué les techniques transductives qui, à l'inverse des techniques inductives, classifient et affectent de nouvelles observations sans passer par une étape d'apprentissage. En tant que principale représentante des techniques transductives, la méthode des k-plus proches voisins illustre parfaitement ce procédé. Elle doit manipuler toutes les données disponibles pour trouver l'affectation d'une seule observation, nous obligeant ainsi à disposer d'une grande capacité de stockage et de calcul. Il est généralement préférable de disposer d'un modèle résumant le contenu des données, afin de pouvoir contrôler et appliquer rapidement le modèle à de nouvelles observations. Le désintéressement pour les techniques transductives les a rendues beaucoup moins répandues, ce qui a eu comme effet de concentrer les recherches sur les techniques inductives.

À la section 1.4.2.5, nous avons évoqué la complexité de la topologie obtenue par les réseaux de neurones. Celle-ci est liée à une surabondance de neurones, pouvant entraîner un surapprentissage. Dès lors, en s'adaptant trop parfaitement aux données d'apprentissage, le modèle (obtenu par les réseaux de neurones ou par d'autres méthodes) est devenu si complexe qu'il en a perdu ses capacités de généralisation. Ce phénomène peut par des techniques particulières, voir ses effets réduits. Aussi, il est important de noter que ce phénomène est accentué lorsque les données sont représentées dans de grandes dimensions, puisque la complexité du modèle est généralement dépendante de la dimension de l'espace des variables. Ainsi, l'apprentissage et la qualité des modèles pourraient être améliorés, en réduisant le nombre d'entrées. Cette tâche de réduction est identifiée dans le schéma global d'un processus de reconnaissance de formes (voir figure 1.3 à la page 14), comme la « sélection/extraction de variables ». Nous l'évoquerons dans le chapitre suivant concernant la réduction de la dimensionnalité.

Notons enfin, que dans la seconde partie de ce manuscrit, consacrée aux contributions, nous nous sommes attachés à utiliser et à comparer les méthodes de classification linéaire et non linéaire évoquées dans ce chapitre.

Chapitre 2

Réduction de la dimensionnalité

2.1 Introduction

Nous avons pu observer, sur le schéma de principe d'un système de reconnaissance de formes (figure 1.3, page 14), que la conversion, ou plutôt la transcription numérique des données provenant du monde physique, génère une description brute de l'information pouvant conduire à une représentation des observations dans de grandes dimensions (*cf.* section 1.1.2). Aussi, les capacités de stockage n'étant plus un problème aujourd'hui, les données fournies sont de plus en plus nombreuses et exprimées dans des dimensions de plus en plus grandes. Ainsi, les descriptions des formes, donc les variables, peuvent être très nombreuses, et en présence d'un nombre important de variables disponibles, nous devrions y trouver l'information utile. Cependant dans ce contexte, où il n'y pas de tri, du bruit, des variables « nuisibles » ou redondantes peuvent s'ajouter aux données. Or, des analyses théoriques et des études expérimentales ont montré la faiblesse de nombreux algorithmes en présence de variables non pertinentes ou redondantes [Langley, 1996]. D'un point de vue algorithmique, [Gutierrez-Osuna, 2002] précise également qu'en présence de variables redondantes, de nombreuses techniques statistiques sont confrontées à des difficultés numériques[1].

Au-delà des aspects de complexité de calcul et de capacité de stockage, la réduction de la dimensionnalité peut permettre également d'améliorer les performances de classification. De là, [Bellman, 1961] a introduit l'expression « malédiction de la dimensionnalité » (*curse of dimensionality*), révélant le problème causé par des formes représentées dans de grandes dimensions, lorsque le nombre d'observations est limité. Il spécifiait notamment, que le nombre d'échantillons nécessaires pour estimer précisément une distribution de données augmente exponentiellement avec la dimension. Ainsi, plus le nombre de variables augmente, plus le nombre d'échantillons devrait être important. Cela constitue un obstacle majeur dans l'apprentissage artificiel, qui est souvent contraint pour des raisons de coût ou de contexte, à manipuler des ensembles d'observations réduits ; le nombre d'observations est souvent très limité et il est parfois impossible de pouvoir en obtenir autant que nécessaire. L'expression de la malédiction de la dimensionnalité est connue aussi sous le nom du « phénomène de l'espace vide » (*the empty space phenomenon*). [Thiria *et al.*, 1997; Lee *et al.*, 2004] le caractérisent en énonçant quelques propriétés inattendues dans des espaces de grandes dimensions, et proposent l'illustration suivante : le volume d'une sphère inscrit dans un cube tend vers zéro quand la dimension augmente.

La diminution de l'influence de ces difficultés peut être alors obtenue par une réduction de la dimensionnalité de l'ensemble de données ; en d'autres termes, des variables caractérisant le

[1] [Gutierrez-Osuna, 2002] spécifie que des variables redondantes conduisent à obtenir la matrice de covariance de l'ensemble de données singulières, la rendant non inversible.

problème sont éliminées. La réduction ne peut se faire au hasard, elle doit évidemment se concentrer sur des variables inutiles et non pertinentes. Cette tâche apparaît dans le schéma global d'un processus de reconnaissance de formes (figure 1.3), comme la « sélection/extraction de variables ». Cette phase est importante et incontournable, car elle conditionne le processus de classification. Ainsi, deux approches, illustrées à la figure 2.1, sont principalement utilisées pour réduire la dimension :

- l'**extraction de caractéristiques**, qui consiste à réduire la dimensionnalité de l'espace d'entrée en appliquant une transformation sur les variables initiales, afin d'obtenir une nouvelle représentation plus synthétique. Cette transformation peut être linéaire ou non linéaire ;

- la **sélection de variables**, qui sélectionne parmi les variables d'entrée, les plus pertinentes de manière à former un sous-ensemble de variables préservant l'information utile.

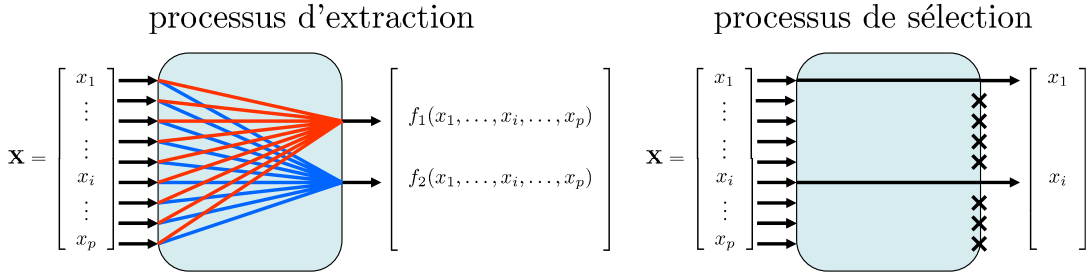


FIG. 2.1 – Extraction de caractéristiques et sélection de variables [Webb, 2002].

Ces deux approches seront détaillées dans les sections 2.3 et 2.4. Cependant, avant de les aborder, une analyse préliminaire est nécessaire : il s'agit du **prétraitement**. La littérature occulte un peu cette étape. Elle la considère uniquement comme une « préparation » (sous-entendant une tâche peu importante) des données aux tâches de réduction et de classification, et évoque essentiellement la normalisation des données. Nous ne pouvons pas contredire cette vision, cependant elle est un peu réductrice. En effet, à l'image du livre de [Pyle, 1999] qui traite exclusivement de la préparation des données, nous verrons à la section 2.2 l'étendue des méthodes qui permettent de préparer les données de manière optimale avant leurs manipulations.

Avant de poursuivre, rappelons que la matrice des observations $\mathbf{X}$ contient n observations $\mathbf{x}_1, ..., \mathbf{x}_n$, chacune décrite par p variables. Le vecteur $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$ donne alors les p éléments de l'observation i, et l'élément (i, j) de $\mathbf{X}$ (noté x_{ij}) correspond à la j-ème variable de la i-ème observation.

2.2 Prétraitement

2.2.1 Introduction

La phase de prétraitement est aussi importante que celle de la réduction de dimension ; elles sont profondément liées. C'est pourquoi, avant de chercher à obtenir de bonnes performances de classification, les données doivent subir un prétraitement afin d'éliminer toute incertitude sur leur légitimité à apparaître dans la base de données. Les ouvrages de [Pyle, 1999; Theodoridis and Koutroumbas, 2006] abordent dans le détail les étapes de préparation des données avant leur analyse.

2.2.2 Données aberrantes (*outliers*)

Une donnée aberrante (*outlier*) est une donnée qui diffère de manière importante des autres données. Elle peut être vue comme une donnée éloignée (d'un point de vue de la distance) de la valeur moyenne de la distribution de la variable correspondante (figure 2.2) [Theodoridis and Koutroumbas, 2006].

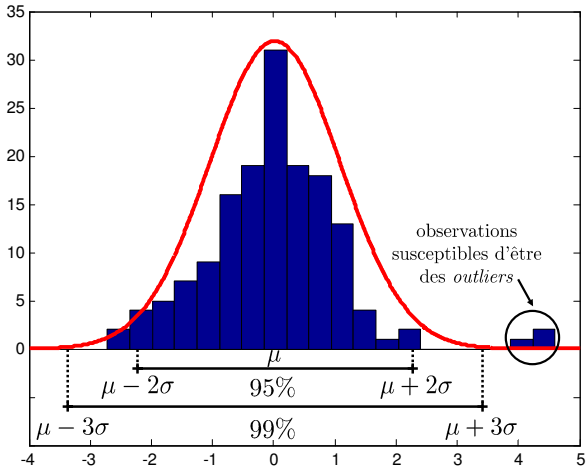
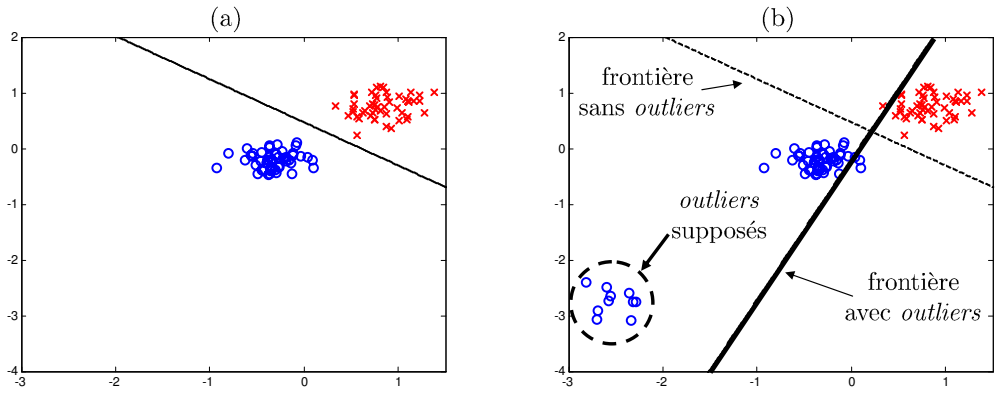


FIG. 2.2 – Détection d'observations aberrantes sur une variable aléatoire de distribution normale.

Ce type de données soulève un inconvénient majeur dans l'utilisation de la mesure de la moyenne, car ces données ont tendance à entraîner la valeur moyenne dans leur direction. Nous avons déjà pu voir leur effet sur la détermination de l'hyperplan discriminant par la méthode des moindres carrés. Ces données peuvent influencer fortement la distribution de la variable et donc les performances des outils de classification durant l'apprentissage, engendrant un biais important, comme montré à la figure 2.3. Cet exemple, issu de [Bishop, 2006], montre la forte sensibilité du critère des moindres carrés aux valeurs aberrantes dans la détermination de la frontière de décision [Bishop, 2006; Theodoridis and Koutroumbas, 2006]. Ce type de données demande donc un traitement approprié.



Note : (a) Hyperplan obtenu par les moindres carrés. (b) Comparaison des hyperplans en fonction de la présence de valeurs aberrantes.

FIG. 2.3 – Influence de la présence de valeurs aberrantes sur la frontière de décision obtenue par les moindres carrés.

Dans le paragraphe précédent, nous avons évoqué l'éloignement de la donnée en termes de distance, qui peut être évaluée en fonction d'un seuil, soulevant ou non la présence de valeurs aberrantes. Plusieurs techniques permettent de déterminer ce seuil, notamment en fonction de l'écart

type. Pour une variable aléatoire de distribution normale (de moyenne μ et d'écart type σ), l'intervalle $[\mu-2\sigma\,,\,\mu+2\sigma]$ couvre 95% des données et l'intervalle $[\mu-3\sigma\,,\,\mu+3\sigma]$ en couvre 99%. Ainsi, pour une variable dont on estime qu'elle suit une loi normale (vérifiable par le test Kolmogorov-Smirnov [Tufféry, 2007]), nous pouvons estimer qu'une observation supérieure à $\mu+2\sigma$ ou inférieure à $\mu-2\sigma$ serait susceptible d'être une valeur aberrante, comme l'illustre la figure 2.2.

Dans le cas où les données ne suivent pas une loi normale, cas que l'on peut être amené à rencontrer souvent, il faut effectuer un traitement que l'on peut considérer comme non paramétrique, (si l'on veut faire un parallèle avec l'estimation de densité de probabilité). Ainsi, plutôt que d'utiliser la valeur de l'écart type pour déterminer le seuil, nous pouvons nous servir du premier quartile ($Q1$) et du troisième quartile ($Q3$) pour évaluer la dispersion des données et définir le seuil. Ainsi, on peut considérer qu'une observation est aberrante si sa valeur se trouve à l'extérieur de l'intervalle $[Q1-1,5(Q3-Q1)\,,\,Q3+1,5(Q3-Q1)]$. Ces seuils sont représentatifs des valeurs utilisées par les diagrammes en boîtes à moustaches (*boxplot*) [Tufféry, 2007]. En outre, le seuil désignant la présence de valeurs aberrantes peut être, tout aussi bien, défini soit arbitrairement, soit en suivant des informations extérieures à la distribution des données.

Cependant, une fois la détection des observations aberrantes effectuée, que doit-on en faire ? On pourrait simplement considérer ces valeurs comme erronées et donc les supprimer. Cependant, [Pyle, 1999; Tufféry, 2007] proposent de travailler comme en présence de valeurs manquantes. Le traitement de ce type de valeurs sera développé dans la section 2.2.4.

[Tufféry, 2007] différencie une valeur extrême d'une valeur aberrante, même si ces deux types de valeurs ont des propriétés similaires. En effet, il peut être judicieux de penser qu'une valeur, considérée comme aberrante, peut tout à fait être représentative d'un état rare ou particulier. Dès lors, sa suppression pourrait appauvrir l'échantillon des données. Il serait donc intéressant de la conserver dans l'état.

2.2.3 Normalisation des données

La normalisation des données permet de s'affranchir des différences de « normes » des variables. En effet, des variables avec des grandes valeurs peuvent avoir une plus grande influence que des variables avec des petites valeurs, sans pour autant être plus significatives. La technique la plus simple et la plus utilisée traite chaque variable indépendamment, et calcule pour chaque variable x_j sa valeur moyenne $\overline{x}_j$ et son écart type σ_j. Ainsi, la normalisation de l'observation $\mathbf{x}_i$ de cette variable, identifiée par l'élément x_{ij} de $\mathbf{X}$, est normalisée ($\widetilde{x}_{ij}$) par l'expression suivante :

$$\widetilde{x}_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j}\,. \tag{2.1}$$

Le résultat de cette normalisation sur l'ensemble des observations x_{ij}, $i = 1, ..., n$ de la variable permet d'obtenir une distribution de cette variable ayant comme propriétés une valeur moyenne nulle et une variance de un. Dans la littérature, on peut observer cette transformation sous le nom de transformation centrée réduite ou encore normalisation statistique. D'autres techniques de normalisation linéaire limitent les valeurs des variables entre $[0, 1]$ ou encore $[-1, 1]$. La normalisation appropriée dépend bien évidemment du traitement qui sera effectué sur les données. D'autres types de normalisations sont possibles, fondées sur des fonctions non linéaires, comme des fonctions de type sigmoïde ou logarithmiques [Pyle, 1999; Theodoridis and Koutroumbas, 2006].

2.2.4 Données manquantes

La présence de valeurs manquantes [Barnett and Lewis, 1994] dans une base de données est un handicap bien souvent insurmontable pour la plupart des algorithmes. Une approche simpliste serait d'éliminer les observations où des variables sont manquantes ou inversement, d'éliminer les variables où des observations sont manquantes. Cependant, il n'est pas recommandé de les ignorer. L'élimination de données disponibles est un luxe que peu d'applications peuvent s'offrir, engendrant une perte d'information substantielle ; notamment dans le domaine médical (*cf.* introduction) [Dupont, 2002].

Le remplacement des valeurs manquantes doit préserver les aspects importants des distributions et les relations entre les variables. On ne doit donc pas chercher à les prédire précisément. Aussi, lorsque les valeurs manquantes excèdent 10 à 20% pour une même variable, il est d'usage de ne pas chercher à les remplacer [Pyle, 1999]. Plusieurs techniques permettent de les remplacer et elles peuvent être également utilisées dans le traitement des données aberrantes (*cf.* section 2.2.2) [Pyle, 1999].

Pour illustrer et montrer l'impact du remplacement de valeurs manquantes dans la distribution d'une variable, nous évaluons quatre techniques couramment utilisées, sur un exemple donné aux figures 2.4−2.7. Les données sont caractérisées par deux variables x_1 et x_2, où deux observations de la variable x_2 sont manquantes. Ainsi, pour $x_1 = 3$ et $x_1 = 4,5$, nous cherchons à remplacer les valeurs manquantes de x_2 par les quatre techniques présentées et nous comparons et évaluons dans le tableau 2.1 l'impact sur la distribution de la variable x_2.

- le remplacement **par la valeur moyenne** de la variable observée. Cette technique est relativement simple et très utilisée, elle a l'avantage de préserver la valeur moyenne. Cependant, en remplaçant pour une même variable, toutes les données par la même valeur, cette approche réduit la variance de la variable et les corrélations avec les autres variables. Toutefois, en présence de peu de valeurs manquantes, cette technique peut être intéressante ;

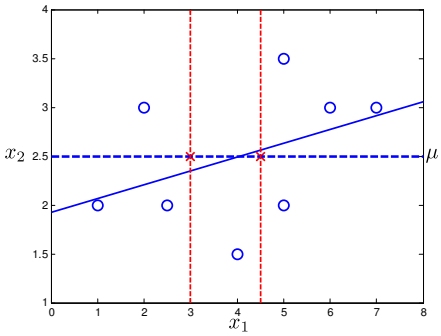


FIG. 2.4 – Remplacement de valeurs manquantes par la valeur moyenne.

- le remplacement **par une valeur aléatoire**, mais représentative de la distribution de la variable. Prenons une nouvelle fois le cas où une variable suit une loi normale. À partir des caractéristiques de cette distribution (moyenne μ et écart type σ), on tire aléatoirement une valeur dans une fourchette dite non déviante, appartenant à l'intervalle $[\mu - 2\sigma\,,\ \mu + 2\sigma]$. Cette technique réduit les effets sur la variance de la variable ;

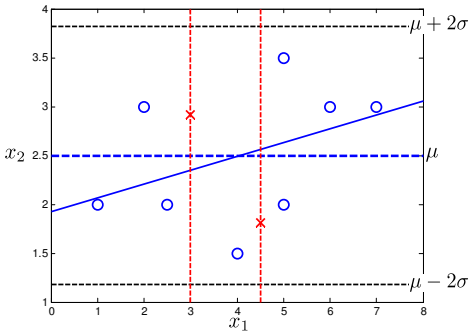


FIG. 2.5 – Remplacement de valeurs manquantes par une valeur aléatoire.

- le remplacement **par le plus proche voisin**. Cette approche remplace la valeur manquante par une valeur ayant les caractéristiques les plus proches, le nombre de variables à considérer n'est pas défini *a priori* ;

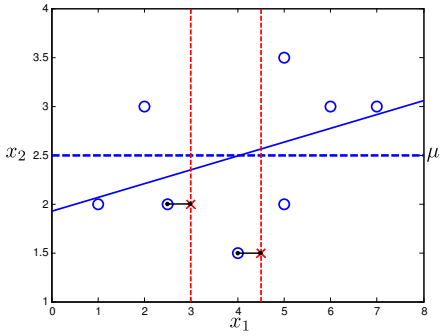


FIG. 2.6 – Remplacement de valeurs manquantes par le plus proche voisin.

- le remplacement **par une valeur prédite par un modèle de régression** réalisé à partir des données disponibles.

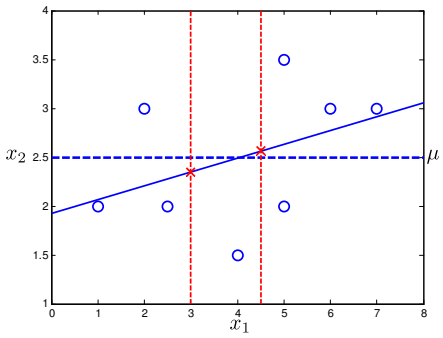


FIG. 2.7 – Remplacement de valeurs manquantes par une valeur prédite.

Comme le montre cet exemple, le remplacement par la valeur moyenne fait chuter la variance en conservant la moyenne. Le remplacement par une valeur aléatoire, quant à elle, change la moyenne tout en réduisant les effets sur la variance. Or, nous pouvons noter que dans l'analyse des données la variance est essentielle [Tufféry, 2007].

technique de remplacement	**moyenne de x_2**	**variance de x_2**	**corrélation entre x_1 et x_2**
sans remplacement	2,50	0,707	0,413
par la valeur moyenne (figure 2.4)	2,50	0,624	0,405
par une valeur aléatoire (figure 2.5)	2,47	0,678	0,305
par le plus proche voisin (figure 2.6)	2,35	0,709	0,356
par une valeur prédite par un modèle de régression (figure 2.7)	2,49	0,626	0,420

TAB. 2.1 – Comparaison et évaluation de l'impact du remplacement de valeurs manquantes sur la distribution d'un ensemble de données.

2.2.5 Conclusions

Dans cette première partie du chapitre consacré à **la réduction de la dimensionnalité**, nous avons présenté les premières analyses et transformations nécessaires au traitement des données. Les différentes étapes et techniques abordées lors du prétraitement montrent les aberrations que peuvent introduire les systèmes d'acquisition. Il est alors nécessaire de s'en prémunir avant d'entreprendre toute action d'analyse, aux risques de biaiser les interprétations. Malgré les nombreuses techniques statistiques disponibles pour prétraiter les données, l'intervention humaine est nécessaire, bien que celle-ci soit souvent négligée dans la pratique [Jermyn *et al.*, 1999].

Une bonne préparation des données est donc un prérequis incontournable au succès de l'analyse de données. Aussi, selon [Jermyn *et al.*, 1999], 60 à 80% du temps employé à l'analyse des données devrait être consacré à la phase de préparation. Bien que la littérature manifeste un intérêt modéré à la préparation des données, le lecteur pourra néanmoins trouver des informations supplémentaires dans [Famili *et al.*, 1997; Hernández and Stolfo, 1998; Pyle, 1999].

2.3 Extraction de caractéristiques

2.3.1 Introduction

Dans cette section, nous allons aborder l'extraction de caractéristiques, qui permet d'extraire et de construire de nouvelles caractéristiques fondées sur une transformation des variables originales. Ce principe peut rappeler la classification non linéaire par les SVM (*cf.* section 1.4.3), où les observations dans l'espace d'entrée sont transformées dans l'espace des **représentations** afin de permettre une classification linéaire de ces observations dans ce nouvel espace. Cependant, la transformation des observations dans le cadre de l'extraction de caractéristiques ne cherche pas spontanément à optimiser la séparation des classes, mais à faire apparaître des structures et des relations dans les données en définissant un sous-espace. Cette transformation peut être, comme dit précédemment, linéaire ou non linéaire. Elle part du principe que l'information, contenue dans un grand nombre de variables, peut être représentée par quelques caractéristiques non visibles directement.

2.3.2 Approches linéaires

Dans cette partie, nous abordons les approches linéaires pour la réduction de dimension, en particulier, l'analyse en composantes principales (ACP) [Jolliffe, 2002], qui fait partie des techniques incontournables pour ce type d'analyse. Cependant, les limitations de cette méthode nous amèneront à envisager d'autres techniques, comme l'analyse factorielle discriminante à la section 2.3.2.3. Celle-ci offre l'avantage de considérer l'appartenance des observations aux classes.

2.3.2.1 Analyse en composantes principales

Aborder la réduction de dimension impose naturellement d'évoquer l'analyse en composantes principales introduite par [Pearson, 1901]. C'est une technique très ancienne, et encore très largement utilisée : elle est à la base de beaucoup d'autres techniques. À partir d'une représentation des données de type « observations − variables », où les variables sont numériques et continues, l'ACP cherche une représentation dans des sous-espaces vectoriels de plus faible dimension préservant au mieux la distribution des observations. [Hotteling, 1933], qui participa au développement de l'ACP, la définit comme une projection orthogonale des observations dans un espace de plus faible dimension telle que la variance, des observations projetées, est maximisée. Les caractéristiques de la nouvelle représentation ne sont donc pas corrélées et permettent d'apporter une réponse aux problèmes des variables redondantes. De nombreux ouvrages décrivent précisément cette analyse et quelques extensions, comme [Jolliffe, 2002; Saporta, 2006; Bishop, 2006].

Trivialement, l'ACP détermine le premier axe principal maximisant la variance des observations, puis le deuxième axe principal, orthogonal au premier, maximisant toujours la variance des observations. La construction des autres axes se déroule suivant le même processus. L'objectif est donc de trouver les axes orthogonaux qui permettent de maximiser la variance des observations projetées. Les coordonnées de ces nouveaux axes dans l'espace des variables d'origines sont obtenues par le calcul des vecteurs propres de la matrice des covariances. Les vecteurs propres **u**

associés aux axes sont ordonnés suivant la variance restituée sur chacun d'eux, qui elle est obtenue par les valeurs propres λ. Par conséquent, ces valeurs propres nous donnent l'information sur la contribution d'inertie de chacun des p axes principaux. Ainsi, le pourcentage d'inertie expliquée I_q par les q premiers axes est donné par :

$$I_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{i=1}^{p} \lambda_i} \ , \ q \leq p \, . \tag{2.2}$$

La projection d'une observation $\mathbf{x}_i$ de dimension p sur les q nouveaux axes, permet d'obtenir les nouvelles coordonnées $\mathbf{y}_i$ de cette observation, telles que chacun des q éléments du vecteur $\mathbf{y}_i$ est obtenu par $y_j = \mathbf{x}_i \, \mathbf{u}_j$, avec $j = 1, \ldots, q$ et $\mathbf{u}_j$ donne le vecteur propre associé à la j-ème composante principale. Dès lors, la projection de toutes les observations disponibles donne les nouvelles caractéristiques appelées composantes principales (CPs). Chacune des q composantes principales est une combinaison linéaire des p variables initiales, telle que la k-ième CP est définie par :

$$cp_k = \mathbf{X}\mathbf{u}_k \, . \tag{2.3}$$

Les CPs sont des vecteurs indépendants et sont donc non corrélées linéairement entre elles, évitant la présence de variables redondantes. Géométriquement, la première composante principale, notée cp_1, donne la direction du nuage des observations qui suit l'axe d'étirement maximal du nuage, comme le montre la figure 2.8.

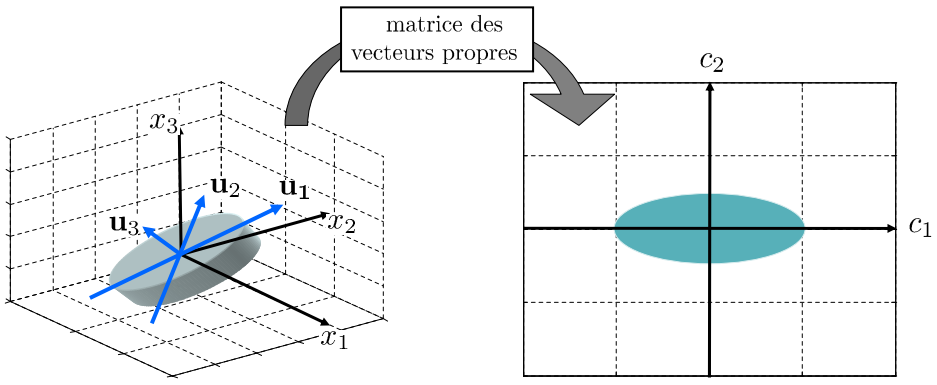


FIG. 2.8 – Illustration de la réduction de dimension par l'analyse en composantes principales.

Cette transformation linéaire respecte la topologie globale des données originales, mais dans laquelle les axes n'ont pas forcément de signification par rapport à la structure des données. Cependant, une bonne interprétation de l'ACP permet de dégager un certain nombre d'informations.

Choix du nombre de composantes principales à conserver

La variance de chacune des CPs est donnée dans l'ordre décroissant des valeurs propres, par : $\lambda_1, \lambda_2, \ldots, \lambda_p$. Ainsi, en éliminant les axes où la variance est faible, on obtient la réduction de la dimension. L'information sur la variance, associée à chaque axe, ne permet pas de choisir formellement le nombre de composantes à conserver. D'autre part, une décroissance régulière des valeurs propres indique que les données sont peu structurées et rend par conséquent difficile le choix de la dimension de l'espace de projection. Plusieurs critères permettent d'apporter une

solution à ce problème, notamment la règle de Kaiser [Kaiser, 1961]. Cette règle est certainement la plus utilisée : on conserve les CPs correspondant aux valeurs propres supérieures à la moyenne des valeurs propres. Ainsi, si les observations sont centrées réduites (*cf.* section 2.2.3), on retient uniquement les CPs correspondant à des valeurs propres supérieures à 1. La règle de Kaiser appliquée à l'exemple donné à la figure 2.9 conserverait les quatre premières composantes principales. Une autre approche, fondée sur une analyse graphique, est le « test de l'éboulis » (*scree test* [Cattell, 1966]). Ce critère consiste à tracer les valeurs propres dans l'ordre décroissant et conserver les CPs jusqu'à la première rupture de la pente des valeurs propres (voir figure 2.9, où l'on conserverait les cinq premières composantes principales). Plus récemment, [Karlis *et al.*, 2003] proposent de tenir compte de la dispersion des valeurs propres ; ainsi, ils conservent les valeurs propres supérieures au critère suivant :

$$\lambda > 1 + 2\sqrt{\frac{p-1}{n-1}}\,. \tag{2.4}$$

Rappelons que p et n définissent respectivement le nombre de variables et le nombre d'observations. Enfin, un dernier type d'approche se fonde sur le pourcentage d'inertie à conserver, généralement de l'ordre de 80 à 90%. Cette approche, critiquée par [Saporta, 2006], est cependant souvent employée. L'auteur condamne son utilisation en indiquant que l'on ne peut pas donner un seuil universel sans tenir compte de la taille des observations (matrice $\mathbf{X}$) à analyser et des corrélations entre les variables.

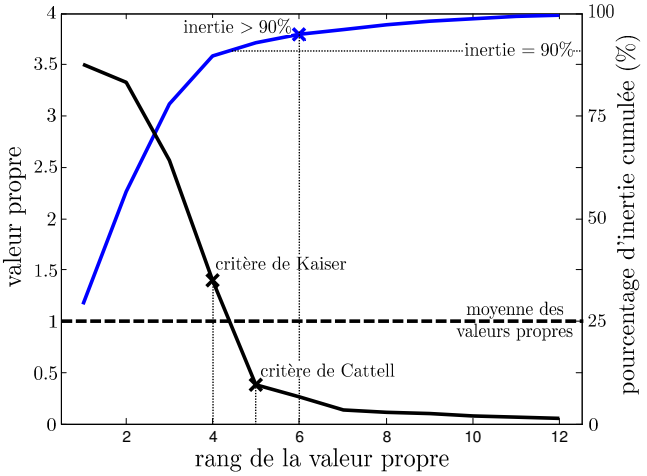


FIG. 2.9 – Comparaison de critères pour le choix du nombre de composantes principales à conserver.

Interprétation

Plusieurs interprétations peuvent être faites à partir de l'ACP, notamment sur la visualisation des relations entre les variables, ainsi que sur les groupements d'observations et des variables. Pour plus de détails, le lecteur pourra se référer aux ouvrages de [Saporta, 2006; Lebart *et al.*, 2006], mais aussi celui de [Georgin, 2002]. Ce dernier offre de nombreux exemples très détaillés et exploitables à l'aide du logiciel Microsoft® Excel.

Nous avons déjà abordé le pourcentage d'inertie expliquée pour chaque axe principal, qui permet d'observer la quantité d'informations restituées sur chaque axe. L'ACP est une technique qui autorise de faire une synthèse sur les données initiales, en apportant de nombreuses informations

sur leur structure. Dans les paragraphes suivants, nous nous intéresserons particulièrement à deux interprétations.

La première est **la qualité de la représentation des variables dans les composantes principales**. Celle-ci permet de déterminer la signification à donner aux nouvelles caractéristiques, soit les composantes principales, en les reliant aux variables originales. Tout d'abord, il faut projeter les variables sur les axes principaux, ainsi, les coordonnées factorielles des p « points − variables » sur l'axe j sont obtenues par $\mathbf{u}_j\sqrt{\lambda_j}$. Par conséquent, en appliquant ce calcul à toutes les composantes principales, nous pouvons obtenir la qualité de la représentation Q_j^i (2.5) d'une variable i dans la composante j. Sachant que u_j^i représente le i-ème élément du vecteur propre associé à la j-ème composante principale.

$$Q_j^i = \frac{\left(\sqrt{\lambda_j}\, u_j^i\right)^2}{\sum_{l=1}^{p}\left(\sqrt{\lambda_l}\, u_l^i\right)^2} \tag{2.5}$$

Cette approche, évoquée dans les ouvrages exposant les aspects théoriques de l'ACP, est cependant très peu utilisée dans les applications faisant intervenir une ACP. Une explication plus détaillée est proposée au chapitre 6, relatant les contributions apportées sur l'extraction d'information et l'interprétation des méthodes de projection.

La seconde interprétation, fondée sur **la représentation simultanée des observations et des variables**, est obtenue par le diagramme de double projection, nommé plus couramment *biplot* [Gabriel, 1971; Smith and Cornell, 1993; Gower and Hand, 1996]. Cette représentation est réalisée dans l'espace réduit et son objectif est d'interpréter directement sur le diagramme les deux éléments suivants : les projections des observations sur les axes principaux et les corrélations entre les variables et les CPs. Ces corrélations sont obtenues à partir des « points − variables », les détails de la construction du *biplot* peuvent être également trouvés dans [Georgin, 2002; Lebart *et al.*, 2006]. Le graphique obtenu identifie alors les relations entre des variables et des groupes d'observations dans l'espace réduit. Ainsi, si une CP a une forte corrélation avec une variable initiale, alors une grande valeur de cette CP pour un groupe d'observations sera associée à une grande valeur de la variable pour ces mêmes observations.

2.3.2.2 Multidimensional scaling

Le *multidimensional scaling* (MDS) est une autre technique très populaire. Elle peut être traduite par « mise à l'échelle multidimensionnelle » ou encore « positionnement multidimensionnel » [Torgeson, 1952; Shepard, 1962; Borg and Groenen, 2005]. Cette technique consiste à trouver une projection dans un espace de faible dimension en préservant au mieux les distances entre chaque paire d'observations. L'objectif reste identique à l'ACP, mais la présentation des données est différente. Dans le cas de MDS, nous n'avons plus la représentation « observations − variables », mais une matrice contenant les **distances** ou les similarités (ou dissimilarités) entre les observations (comme dans l'exemple proposé au tableau 2.2). Si ces distances sont euclidiennes, le résultat du MDS sera similaire à celui obtenu par l'ACP. Dans l'autre cas, en présence de mesures de proximité (similarités ou dissimilarités), où l'information est de nature ordinale, nous pouvons retrouver une version étendue de l'algorithme nommé *nonmetric* MDS.

Son objectif peut être résumé, par un désir de reconstituer une « carte » des observations à partir d'une matrice de proximité, en recherchant une représentation des observations dans un espace euclidien.

L'exemple le plus populaire est le repositionnement des villes sur une carte, à partir de l'information sur les distances les séparant. Ainsi, on dispose d'une matrice donnant les distances entre chaque ville (tableau 2.2), le MDS doit restituer le positionnement des villes sur la carte.

	AMIENS	ANGERS	⋯	LE MANS	⋯	STRASBOURG	TOULOUSE
AMIENS	0	342	⋯	238	⋯	440	690
ANGERS	342	0	⋯	104	⋯	679	468
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
LE MANS	238	104	⋯	0	⋯	596	524
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
STRASBOURG	440	679	⋯	596	⋯	0	765
TOULOUSE	690	468	⋯	524	⋯	765	0

TAB. 2.2 – Matrice représentative de l'ensemble de données employé par la méthode MDS pour reconstituer le positionnement des villes sur une carte.

Le résultat obtenu est montré à la figure 2.10(a). Nous pouvons comparer ce résultat au positionnement des villes à partir de leurs latitudes et longitudes donné à la figure 2.10(b). On remarque une forte similitude, validant par conséquent le repositionnement MDS.

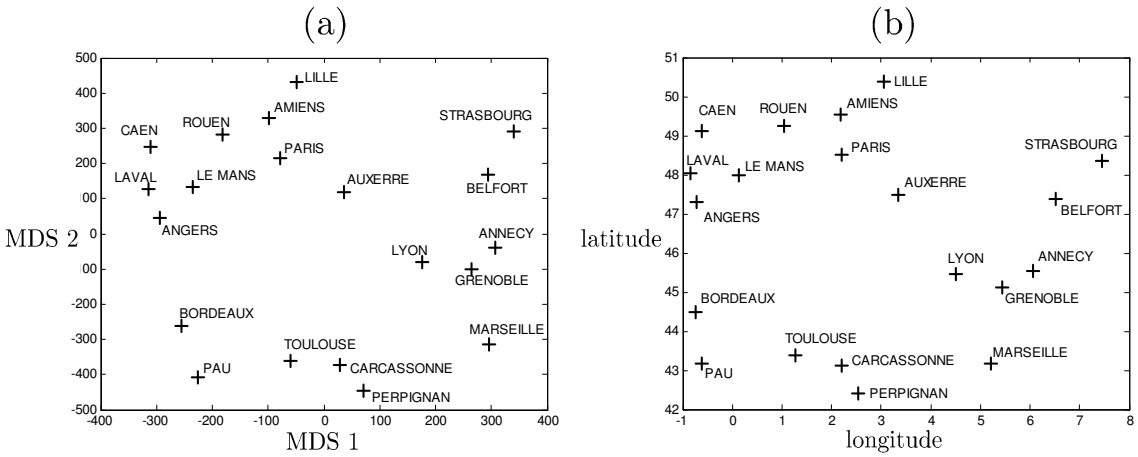


FIG. 2.10 – Projection par la méthode *multidimensional scaling* pour reconstituer le positionnement des villes sur une carte.

2.3.2.3 Analyse factorielle discriminante

Dans l'analyse de données étiquetées, où l'appartenance des observations aux classes est connue, l'ACP ne garantit pas une projection permettant de faciliter la séparation des classes. La figure 2.11(a) illustre parfaitement cette remarque [Theodoridis and Koutroumbas, 2006] : les deux classes (« x » et « o ») suivent une distribution gaussienne de même matrice de covariance (le vecteur propre $\mathbf{u}_1$ correspond à la plus grande valeur propre). Ainsi, nous pouvons remarquer que la projection des données sur ce premier axe entraîne un recouvrement des deux classes et ne permet donc pas de les discriminer correctement.

Dans un contexte de classification supervisée, l'appartenance des observations aux classes est une information utile et qui, si elle le peut, doit être considérée tout au long du processus de construction du modèle. Ainsi, contrairement à l'ACP, l'analyse factorielle discriminante (AFD) recherche de nouvelles directions (ou caractéristiques) sur lesquelles les projections des classes

sont bien séparées ; ces directions sont appelées axes factoriels discriminants. Le critère de projection est, cette fois-ci, la maximisation du rapport entre la variance inter-classe et la variance intra-classe.

Comme évoqué à la section 1.3.2.1, dans le cadre de classification linéaire, nous avons cherché une fonction discriminante par une réduction de dimension. Pour cela, nous avons abordé la fonction discriminante de Fisher [Fisher, 1936], donnant l'hyperplan discriminant comme une fonction orthogonale à la droite de projection. Cette droite de projection sépare aux mieux la moyenne de chaque classe tout en réduisant leur variance, et elle correspond à l'axe obtenu par l'AFD, représentatif de la nouvelle caractéristique. C'est dans ce contexte que nous pouvons voir que l'approche de Fisher déborde le simple cadre de la discrimination [Bishop, 2006]. Le lecteur pourra donc se référer à la section 1.3.2.1 et aux nombreux ouvrages décrivant cette méthode, comme [Confais, 2003; Saporta, 2006; Tufféry, 2007].

L'illustration 2.11 compare les projections obtenues sur le premier axe de l'ACP et de l'AFD. Ainsi, nous pouvons apercevoir que la première caractéristique résultante de l'AFD semble plus apte à conserver la séparation des données, contrairement à l'ACP qui, sur sa première composante principale, obtient un recouvrement des classes.

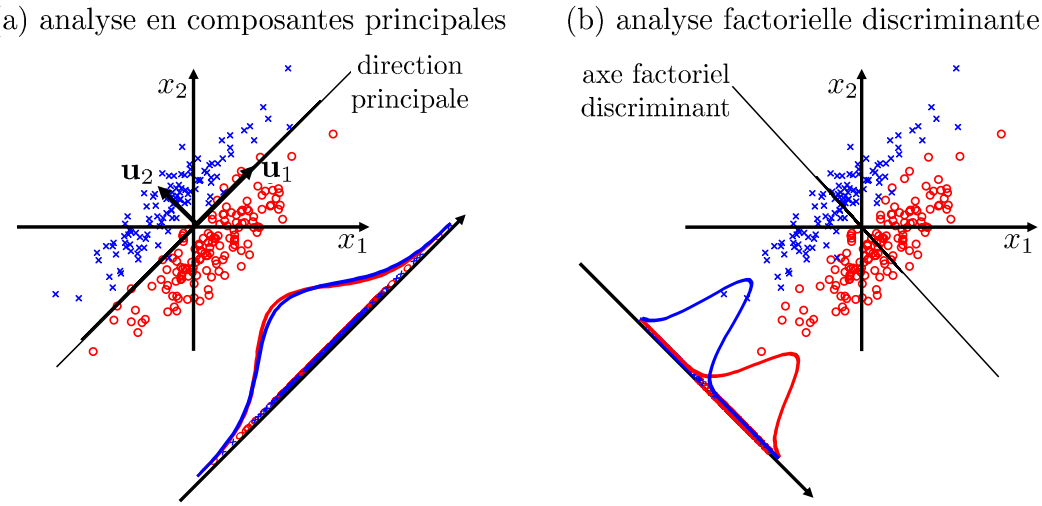


FIG. 2.11 – Comparaison de la projection par une analyse en composantes principales et par une analyse factorielle discriminante [Theodoridis and Koutroumbas, 2006].

2.3.2.4 Conclusions

Nous avons décrit l'ACP comme une méthode de réduction de dimension, autorisant notamment une visualisation des données dans leur globalité. Cette méthode permet de décorréler les variables de l'espace d'origine dans un nouvel espace et de débruiter les données par l'élimination des axes considérés comme insignifiants. En maximisant la variance des observations projetées, l'ACP demeure une méthode très sensible aux valeurs extrêmes. Aussi, l'ACP ne traduit que des liaisons linéaires entre les variables.

La facilité d'implémentation rend néanmoins ces méthodes de projection linéaire très populaires et justifie leur large utilisation. Cependant, comme indiqué auparavant, ces méthodes ne peuvent pas détecter des structures ou des relations non linéaires présentes dans les données, ce qui oblige selon les applications, à utiliser d'autres approches.

2.3.3 Approches non linéaires pour la réduction de la dimensionnalité

Les méthodes de réduction non linéaire ont l'avantage de considérer les relations non linéaires, contrairement à leurs homologues linéaires. En effet, les critères d'optimisation de la projection sont, dans le cas des méthodes non linéaires, fondés sur des notions de topologie et de voisinage, de manière à préserver localement la structure des données. Ainsi, ces techniques de préservation de structures ont pu notamment évoluer par l'utilisation d'une mesure plus complexe que la distance euclidienne, en l'occurrence la distance géodésique [Lee *et al.*, 2004] (voir figure 2.12).

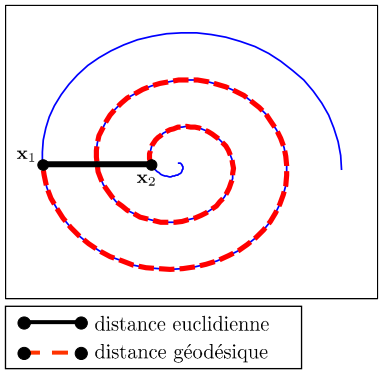


FIG. 2.12 – Comparaison des mesures de distances euclidienne et géodésique entre deux points $\mathbf{x}_1$ et $\mathbf{x}_2$ [Lee *et al.*, 2004].

Il est à noter que l'ACP est à la base de nombreuses méthodes de réduction non linéaire, les courbes principales [Hastie and Stuetzle, 1989] sont sans doute l'exemple le plus frappant. Ce type d'approches cherche à remplacer les axes principaux de l'ACP par des courbes. Nous pouvons également citer la méthode de projection *kernel*-PCA [Scholkopf *et al.*, 1998], fondée sur la technique des noyaux. Les méthodes à base de noyaux peuvent faire penser naturellement aux SVM (*cf.* section 1.4.3), où, les observations de l'espace d'origine sont projetées par une transformation non linéaire dans l'espace des représentations. La méthode *kernel*-PCA réalise une ACP dans cet espace. [Moerland, 2000; Nasser, 2007] ont pu montré l'efficacité et l'intérêt de cette approche, notamment des caractéristiques résultantes de ce traitement lors de l'entraînement de modèles de classification. Cette technique a l'inconvénient d'être très gourmande en termes de calculs, ce qui a poussé [Moerland, 2000] à proposer des algorithmes qui accélèrent le processus.

Intuitivement, la découverte et la modélisation des structures non linéaires dans l'ensemble des observations originales pourraient être réalisées en combinant plusieurs transformations linéaires. Dans ce contexte, [Bishop, 2006] propose la méthodologie suivante : après le partitionnement des observations par un algorithme de classification automatique, comme la technique des *K-means*, on pourrait appliquer une ACP à chaque groupe d'observations. Cependant, ce type d'approche ne permet pas de considérer les observations dans leur globalité, et associe donc difficilement les projections de chaque partition [Bishop, 2006].

Globalement, deux approches permettent de réaliser une réduction non linéaire de la dimension : les approches algébriques et les approches neuronales.

Nous adopterons la notation suivante, les vecteurs d'entrées $\mathbf{x}_i$ $(i = 1, ..., N)$ sont définis dans l'espace d'origine de dimension p, les vecteurs de sorties $\mathbf{y}_i$ $(i = 1, ..., N)$ sont projetés dans l'espace réduit de dimension q, avec $q \leq p$. Les distances, entre deux échantillons i et j dans l'espace d'origine et dans l'espace réduit, sont notées respectivement d^*_{ij} et d_{ij}.

2.3.3.1 Approches algébriques pour la réduction de dimension non linéaire

Locally Linear Embedding (LLE) est une méthode de réduction non linéaire basée sur des aspects géométriques [Roweis and Sam, 2000; Saul and Roweis, 2003][2]. Cette méthode algébrique tente de projeter les observations d'entrée dans un espace de plus faible dimension, en considérant que les observations, globalement non linéaires, sont localement linéaires, amenant ainsi à conserver les configurations locales.

L'algorithme, illustré à la figure 2.13, commence par chercher les k plus proches voisins autour de chaque observation d'entrée $\mathbf{x}_i$. Puis, il exprime leurs relations en calculant les vecteurs poids de reconstruction ($\mathbf{w}$) en minimisant la fonction de coût suivante :

$$E(w) = \sum_i \left| \mathbf{x}_i - \sum_j w_{ij} \mathbf{x}_j \right|^2 . \tag{2.6}$$

Sachant que le poids w_{ij} est associé au couple d'observations $(\mathbf{x}_i, \mathbf{x}_j)$, où $\mathbf{x}_j$ appartient au voisinage de $\mathbf{x}_i$ (parmi les k plus proches voisins). Aussi, la minimisation impose de respecter deux contraintes suivantes :

$$\begin{cases} w_{ij} = 0, \text{si } \mathbf{x}_j \text{ n'est pas un voisin proche de } \mathbf{x}_i, \\ \sum_j w_{ij} = 1 . \end{cases} \tag{2.7}$$

Les poids de reconstruction w_{ij} reflètent les propriétés géométriques de l'espace initial, soit les structures locales. La projection $\mathbf{y}_i$ de l'observation $\mathbf{x}_i$ est réalisée en minimisant ce nouveau critère :

$$E(y) = \sum_i \left| \mathbf{y}_i - \sum_j w_{ij} \mathbf{y}_j \right|^2 . \tag{2.8}$$

Similaires à la fonction de coût précédente (2.6), les poids du critère de projection sont dorénavant fixes de manière à préserver les structures locales propre à l'espace initial.

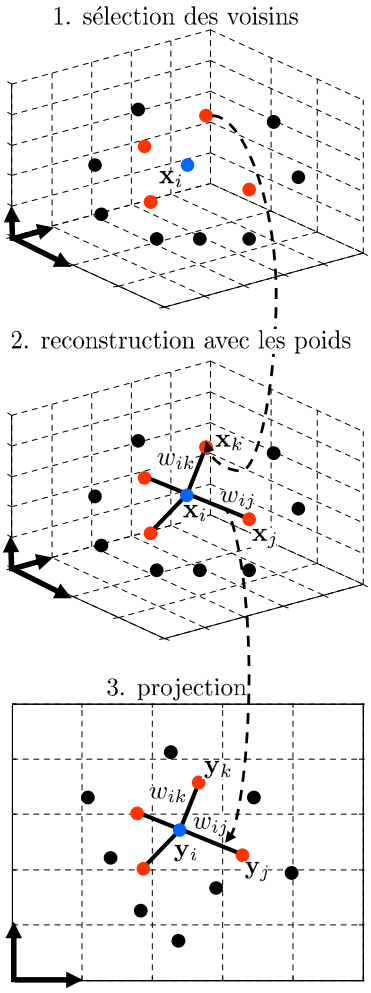


FIG. 2.13 – Algorithme de projection LLE [Roweis and Sam, 2000].

L'algorithme *isometric feature mapping* ou encore *isomap*[3] [Tenenbaum *et al.*, 2000] est une technique de réduction de dimension qui, à l'image de MDS, utilise également une matrice de dissimilarités. Cependant, dans le cas d'*isomap*, la mesure de dissimilarité entre deux observations est définie en termes de distance géodésique [Lee *et al.*, 2004]. Elle est obtenue par le plus court chemin entre deux observations passant par d'autres observations. Dans [Tenenbaum *et al.*, 2000], ce chemin est obtenu à l'aide d'un « graphe » liant chaque observation à ses k plus proches voisins (voir figure 2.14). Une fois la matrice de dissimilarité construite, il reste à la traiter par MDS.

[2]Des informations complémentaires sont disponibles sur http ://www.cs.toronto.edu/~roweis/lle/

[3]Des informations complémentaires, notamment algorithmiques, sont disponibles sur le site de Josh Tenenbaum http ://isomap.stanford.edu/ .

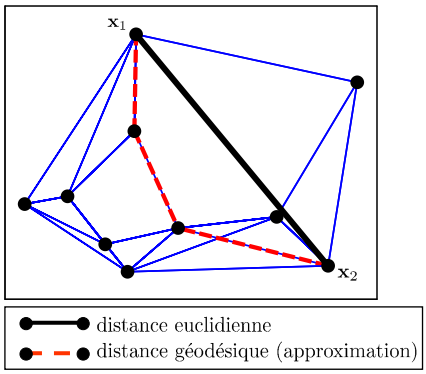


FIG. 2.14 – Illustration du « graphe » et de la distance obtenue par la méthode *isomap* entre deux points $\mathbf{x}_1$ et $\mathbf{x}_2$ (pour $k = 3$).

Une autre variante de MDS est la réduction non linéaire de Sammon [Sammon Jr, 1969] (Sammon*'s nonlinear mapping* − NLM) qui, comme MDS, préserve les distances entre les observations dans l'espace de dimension réduit. Cet algorithme effectue la réduction en utilisant la fonction de coût ci-dessous :

$$E = \frac{1}{\sum_i \sum_{j \neq i} d^*_{ij}} \sum_i \sum_{j \neq i} \frac{\left(d^*_{ij} - d_{ij}\right)^2}{d^*_{ij}} \,. \tag{2.9}$$

2.3.3.2 Approches neuronales pour la réduction de dimension non linéaire

Dans son ouvrage, [Bishop, 2006] présente des approches neuronales pour la réduction de la dimensionnalité d'un problème. Il évoque notamment les réseaux de neurones auto-organisants qui, à la section 1.4.2.4, avaient été abordés afin d'illustrer l'apprentissage non supervisé. Parmi ce type de réseaux de neurones, les cartes auto-organisatrices de Kohonen (SOM, *self-organizing map*) [Kohonen, 1982; Kohonen, 1995] sont certainement les plus utilisées. L'algorithme SOM effectue un partitionnement de l'espace en plusieurs *clusters* (appelé **quantification vectorielle**) et une projection non linéaire des observations originales dans un espace discret de très faible dimension, appelé « carte » ou « grille ». La grille est prédéfinie, généralement rectangulaire ou hexagonale. Elle doit aboutir à devenir une représentation discrète de l'espace d'entrée. Chaque neurone de l'espace de projection, appartenant donc à la grille, est lié à l'espace des observations par un vecteur référent. L'apprentissage s'efforce d'adapter ces vecteurs référents à la distribution des observations, en conservant la topologie de la carte. Ainsi, deux neurones proches sur la carte doivent avoir leur vecteur référent proche dans l'espace des observations.

Dans un but d'améliorer l'algorithme SOM, la méthode nommée par son auteur *generative topographic mapping* (GTM) [Bishop *et al.*, 1997; Bishop *et al.*, 1998] permet de s'affranchir de quelques faiblesses de SOM. Soulignées dans [Kohonen, 1995], ces faiblesses font référence à l'absence d'une fonction de coût ou à la difficulté d'ajuster les paramètres de l'apprentissage. De plus, l'optimisation de l'adaptation de la grille passe par une connaissance *a priori* sur la forme de la structure, ce qui limite encore l'utilisation de cette méthode.

Apparue dans les années quatre-vingt-dix, l'analyse en composantes curvilignes [Demartines, 1994; Demartines and Hérault, 1997; Herault *et al.*, 1999] (ACC, *curvilinear component analysis*) a été proposée également comme une amélioration de SOM, où l'espace de projection n'est plus fixé *a priori* par une grille. L'ACC peut être vue comme une extension neuronale de la méthode de Sammon. [Dreyfus *et al.*, 2002] interprètent cette méthode comme une extension non linéaire de l'ACP. Ils évoquent une ACP « par parties », rappelant le principe introduit par [Bishop, 2006], qui

combine plusieurs ACP. Cette fois-ci, les observations sont « étirées » et projetées dans un espace de plus petite dimension de manière à respecter localement la topologie des observations d'entrée.

Pour optimiser le résultat de la projection en termes de préservation de la topologie, les distances de l'espace de sortie (d_{ij}) doivent être proportionnelles aux distances de l'espace d'entrée (d^*_{ij}). Il est évident que la correspondance des distances ne peut pas être parfaite, pour cause de réduction de la dimension. Dès lors, une fonction de pondération $F(d_{ij}, \lambda_y)$ est introduite dans la fonction de coût à minimiser (2.10) et permet de favoriser et de conserver la topologie locale. Le paramètre de voisinage λ_y peut évoluer avec le temps [Demartines and Hérault, 1997].

$$E = \frac{1}{2} \sum_i \sum_{j \neq i} (d^*_{ij} - d_{ij})^2 F(d_{ij}, \lambda_y) \tag{2.10}$$

La fonction $F(d_{ij}, \lambda_y)$ est définie comme monotone, positive et décroissante par rapport à d_{ij}. Dans leurs simulations, [Demartines and Hérault, 1997] utilisent l'expression de F suivante :

$$F(d_{ij}, \lambda_y) = \begin{cases} 1 & \text{si } d_{ij} \leq \lambda_y \,, \\ 0 & \text{si } d_{ij} > \lambda_y \,. \end{cases} \tag{2.11}$$

La minimisation de la fonction de coût (2.10) est réalisée par un algorithme de descente de gradient stochastique :

$$\forall i \neq j \,, \quad \Delta \mathbf{y}_i = \alpha(t) F(d_{ij}, \lambda_y) \frac{d^*_{ij} - d_{ij}}{d_{ij}} (\mathbf{y}_i - \mathbf{y}_j) \,, \tag{2.12}$$

où le taux d'apprentissage $\alpha(t)$ et le paramètre de voisinage λ_y, tous deux compris entre $[0, 1]$, décroissent en fonction du temps. Dès lors, chaque itération de la descente de gradient a un coût calculatoire proportionnel à n^2 (où, n représente le nombre d'observations), limitant la méthode à de petites bases de données. Le problème de coût est résolu en effectuant une quantification vectorielle (partitionnement de l'espace en plusieurs *clusters*, *cf.* l'algorithme SOM) avant l'ACC, afin de fournir un sous-ensemble de vecteurs, appelés centroïdes, représentant au mieux la distribution des observations d'origine. On peut alors faire intervenir dans (2.10) les distances entre les centroïdes à la place des distances entre les observations. Le coût calculatoire de l'algorithme devient proportionnel au nombre de centroïdes.

Les trois étapes de l'ACC, la quantification vectorielle, la projection des centroïdes et la projection des observations en fonction de ces centroïdes, permettent à cet algorithme d'obtenir de bonnes performances en un temps raisonnable. [Lee *et al.*, 2000; Lee *et al.*, 2004] proposent une amélioration de l'ACC, en permettant d'automatiser le choix des paramètres. Par ailleurs, ils utilisent la distance curviligne (ADC − *Curvilinear Distance Analysis*) à la place de la distance euclidienne. La distance curviligne est, en fait, équivalente à la distance géodésique (figure 2.12). À l'image de la méthode *isomap*, où la distance entre deux observations est obtenue par le chemin passant par les plus proches voisins (voir la figure 2.14), la méthode ADC remplace les plus proches voisins par les plus proches centroïdes.

Un des intérêts de ces approches est la possibilité de projeter facilement une autre observation dans le nouvel espace réduit. Ceci n'est pas forcément le cas pour d'autres méthodes, qui limitent donc leur utilisation à la visualisation.

Dans cette section, nous avons abordé différentes approches pour réaliser la réduction de dimension en tenant compte de relations non linéaires. Bien évidemment, la liste des méthodes citées n'est pas exhaustive. Nous pouvons aussi trouver plusieurs variantes des méthodes présentées, comme la méthode *hessian-based* LLE proposée par [Donoho and Grimes, 2003], qui se fonde notamment sur le calcul de la matrice Hessienne.

2.3.3.3 Interprétation

L'erreur de projection pourrait être obtenue en examinant la fonction de coût, telle que celle de l'ACC (2.10) obtenue pour chaque paire d'observations après la projection. Cependant, cette approche distingue difficilement les erreurs des « petites » et des « grandes » distances. Or la conservation de la topologie locale est l'objectif de la plupart des méthodes de projection non linéaire. L'analyse de l'erreur de la projection ne permet donc pas d'observer rigoureusement la qualité de la projection, au sens de la conservation de la topologie.

[Demartines, 1992] a proposé une représentation, appelée « $dy - dx$ », afin de vérifier la préservation de la topologie obtenue par l'algorithme SOM. Dans son utilisation originale, cette représentation consiste à tracer pour chaque paire de neurones un point $[dy, dx]$ et comparer les distances des neurones sur la grille (dy) avec les distances des vecteurs poids (dx). Ainsi, une projection préservant correctement la topologie s'observe lorsque les dy sont proportionnelles aux dx, au moins pour des petites distances de dy.

Dans le cadre de l'ACC, [Demartines and Hérault, 1997] ont adapté cette représentation à leur algorithme. Dès lors, l'axe nommé précédemment dx représente désormais les distances d^*_{ij} et l'axe dy est remplacé par d_{ij}. On voit aisément que l'on compare, cette fois-ci, les distances de l'espace d'origine à celles de l'espace réduit. Ainsi, de la même manière, la topologie initiale est respectée si on obtient une bonne corrélation entre les distances de ces deux espaces. Des détails supplémentaires seront donnés à la section 6.4.2.

Comme l'ont fait [Demartines and Hérault, 1997] pour l'ACC, on peut généraliser cette représentation à beaucoup d'autres méthodes de réduction. En effet, comme évoqué en introduction de cette partie, la plupart des approches de réduction non linéaire sont fondées sur des notions de topologie et de voisinage, similaires à l'ACC.

2.3.3.4 Illustrations

La figure 2.15 propose deux exemples traditionnellement employés dans la démonstration et l'évaluation des méthodes de projection non linéaire : le « petit suisse » (*swiss role*) et les deux anneaux imbriqués.

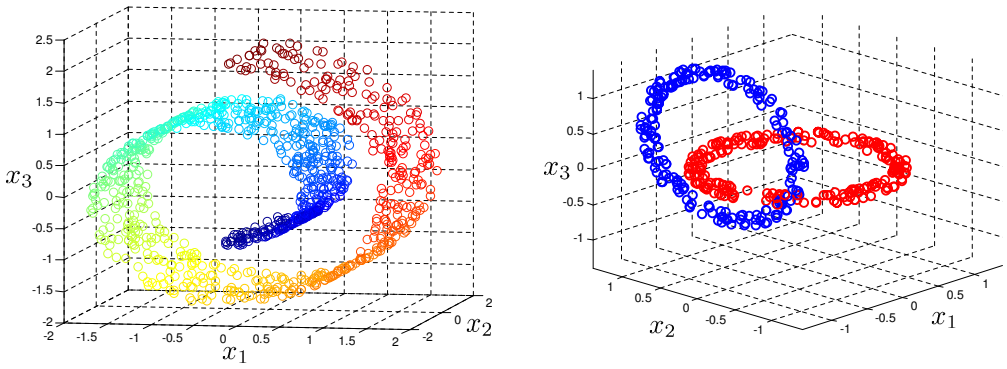


FIG. 2.15 – Ensembles de données, du « petit suisse » (*swiss role*) et de deux anneaux imbriqués, utilisés pour comparer les méthodes de projection.

Trois méthodes (ACP, ACC et LLE) sont comparées pour chaque exemple (voir figure 2.16). Pour chaque méthode, la représentation $dy - dx$ est proposée afin d'évaluer, grâce aux distances entre les observations, la déformation obtenue lors du passage à une dimension plus réduite.

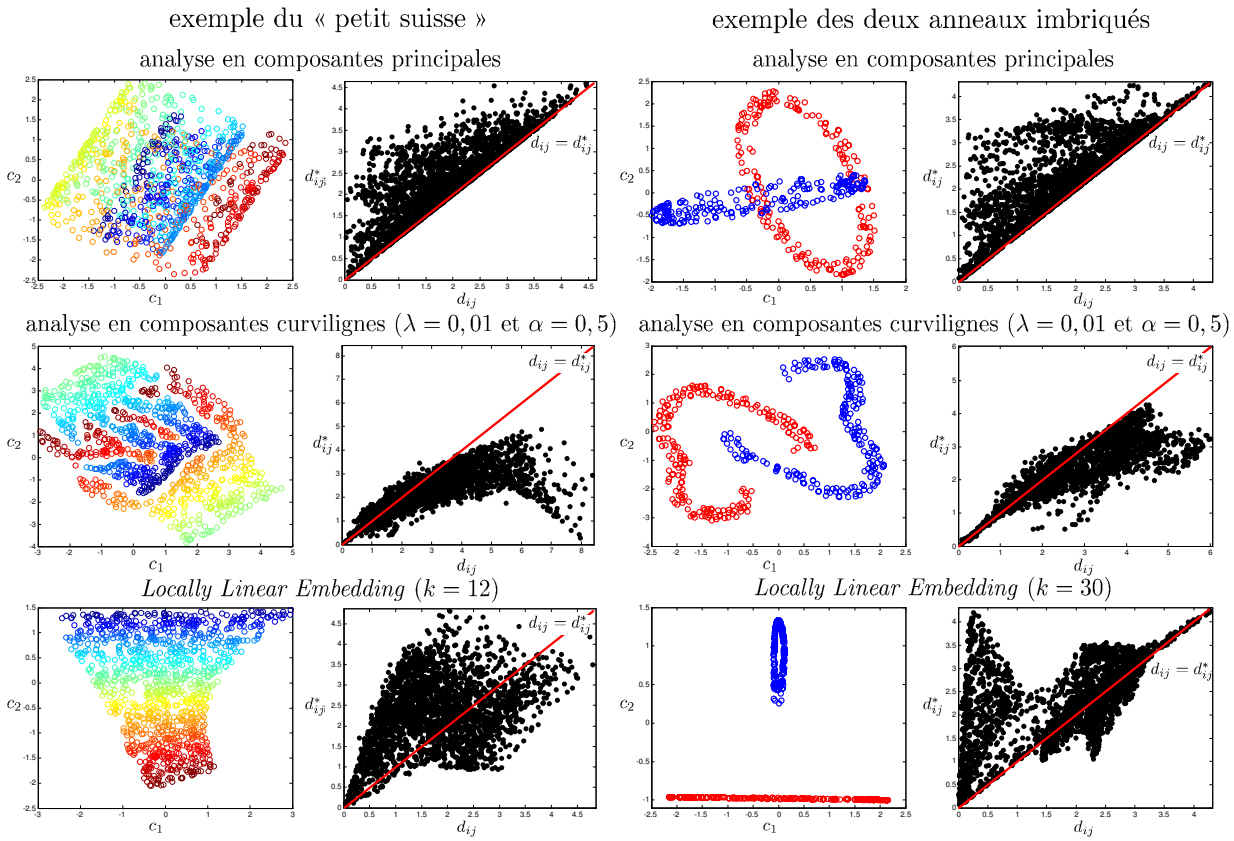


FIG. 2.16 – Comparaison de projections linéaires et non linéaires sur l'exemple du « petit suisse » (*swiss role*) et de deux anneaux imbriqués.

On peut observer le pouvoir des méthodes non linéaires à « étirer » et « déplier » les données, facilitant l'extraction et la visualisation des structures non linéaires. De plus, pour l'ACC, la représentation $dy - dx$ montre dans l'espace de projection une préservation des petites distances et un allongement des grandes distances, ce qui implique l'étirement des observations.

2.3.4 Conclusions

Nous avons pu observer et comparer l'efficacité de la projection par des méthodes non linéaires. Cependant, un inconvénient majeur subsiste. En effet, [Illouz and Jardino, 2001] relèvent que la méthode NLM ne donne aucune information sur le rôle et la présence des variables dans l'espace réduit. Cette remarque peut naturellement être généralisée à la plupart des méthodes de projection [Guérif, 2006], particulièrement pour celles opérant suivant une approche non linéaire. En effet, il n'existe pas de moyen analytique pour extraire la représentation des variables dans les nouvelles composantes non linéaires, contrairement à l'ACP (*cf.* sections 2.3.2.1 et 6.3, sur la qualité de la représentation des variables dans les composantes principales). Cependant, au chapitre 6, nous proposerons une approche qui généralise la méthodologie employée par l'ACP. Elle concerne l'extraction de la qualité de la représentation des variables dans les nouvelles composantes, pour les méthodes de réduction linéaire et non linéaire.

Tout comme avec l'ACP, les méthodes non linéaires projettent les observations sans tenir compte de l'appartenance des observations aux classes.

Au regard des deux dernières remarques, la manière qui peut sembler la plus simple pour réduire la dimension, tout en conservant une identification claire des caractéristiques issues du prétraitement, est de rester dans l'espace initial et d'y sélectionner les variables les plus pertinentes.

2.4 Sélection de variables

2.4.1 Introduction

En introduction de ce chapitre, nous avons souligné la nécessité de réduire la dimension d'un problème, notamment pour améliorer les performances de classification et de prédiction [Langley, 1996]. Par ailleurs, la création d'un outil de classification revient à créer un modèle, et il paraît légitime que ce modèle soit le plus simple possible. Ainsi, tout en renforçant la classification, la simplicité du modèle améliore la vitesse d'exécution, le pouvoir de généralisation, ainsi que la compréhension des entrées du modèle (les variables ou les caractéristiques nécessaires à la modélisation [Dreyfus *et al.*, 2002]). Pour atteindre ce but, nous avons abordé auparavant la réduction de dimension par des méthodes de projection. Ces dernières permettent d'obtenir une nouvelle représentation, par la transformation des variables originales en caractéristiques. Cependant, malgré leur efficacité, deux inconvénients subsistent. Le premier concerne la difficulté d'obtenir des informations précises sur la constitution des nouvelles caractéristiques, diminuant par conséquent la compréhension du modèle. Le second, et non le moindre, concerne un aspect plus pratique relatif au processus global des systèmes de reconnaissance de formes. En effet, avec les méthodes d'extraction, la nouvelle représentation des données nécessite l'ensemble des variables originales lors de la projection. En effet, nous avons préalablement montré, que les nouvelles caractéristiques sont des combinaisons linéaires ou non linéaires des variables. Par conséquent, les entrées du modèle et donc la dimension de l'espace d'apprentissage est bien réduite, mais le nombre de variables à recueillir reste inchangé (matérialisé par le bloc fonctionnel **capture/transcription numérique** de la figure 1.3, page 14). Cela peut être contraignant pour beaucoup d'applications, pour lesquelles on aimerait réduire le nombre de variables à acquérir, afin de réduire les temps d'acquisition, le nombre de capteurs et la capacité nécessaire au stockage.

Dès lors, l'objectif de la sélection est de choisir, parmi toutes les variables originales, un sous-ensemble de variables pertinentes. En réduisant le nombre de variables à recueillir, ces méthodes réduisent implicitement l'espace d'apprentissage.

Toujours en introduction de ce chapitre, nous avons évoqué les problèmes causés par des variables redondantes, notamment d'un point de vue algorithmique. De plus, ces variables, repérables par des corrélations élevées, peuvent être considérées comme superflues, dans le sens qu'aucune information supplémentaire n'apparaît en les ajoutant. Dès lors, spontanément, on pourrait analyser la corrélation entre toutes les paires de variables, et réaliser un tri. Cependant, comme illustré parfaitement par [Guyon and Elisseeff, 2003], une corrélation très élevée entre deux variables ne signifie pas nécessairement une absence de complémentarité entre ces variables.

[Guyon and Elisseeff, 2003] illustrent la complémentarité des variables par le célèbre problème *XOR* (« OU-exclusif », évoqué en annexe B). Cet exemple, montré à la figure 2.17, cherche à établir en deux classes ($\mathcal{C}_1$ et $\mathcal{C}_2$), quatre groupes d'observations, caractérisés par deux variables (x_1 et x_2). En observant les densités de probabilité de chaque variable pour chaque classe ($p(x_j|\mathcal{C}_k)$, avec $j, k = 1, 2$), on remarque qu'individuellement, les variables n'ont pas de pouvoir discriminant. En effet, chaque variable voit les densités de probabilité associées aux deux classes se chevauchent quasiment totalement. Tandis que combinées, elles produisent une bonne séparation des classes, ce qui a permis d'obtenir en annexe B, une résolution efficace de ce problème de discrimination par des méthodes linéaires et non linéaires.

Illustré par ce même exemple, [Bishop, 1995] soulève la nécessité de considérer toutes les combinaisons possibles de variables, de manière à évaluer leur complémentarité afin d'optimiser les performances de classification. Dès lors, un autre problème apparaît, cette fois-ci, de nature

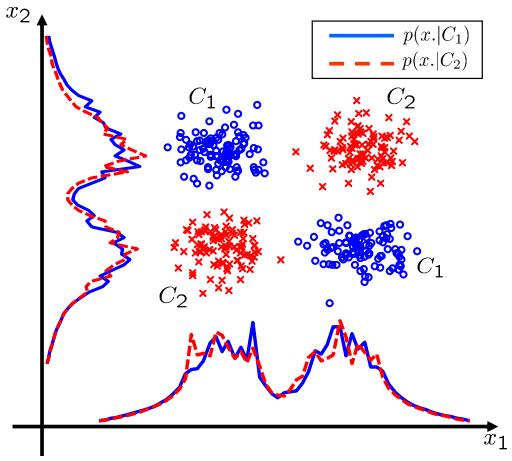


FIG. 2.17 – Illustration détaillée du problème *XOR* en deux dimensions.

combinatoire . En effet, la présence de p variables entraîne 2^p combinaisons possibles. Augmentant exponentiellement avec p, ce nombre peut être réduit en fixant q, un nombre désiré de variables constituant le sous-ensemble final. Considérant ce nouveau paramètre, une recherche exhaustive demanderait désormais l'évaluation d'un nombre de combinaisons égal à :

$$\mathrm{C}_p^q = \frac{p!}{q!\,(p-q)!}\,. \tag{2.13}$$

Cependant, comme le montre la figure 2.18, le nombre de possibilités peut encore augmenter très rapidement, même pour des valeurs modérées de q, rendant l'utilisation d'une recherche exhaustive toujours inconcevable [Bishop, 1995; Jain and Zongker, 1997]. Par exemple, afin de garantir l'optimalité d'un sous-ensemble composé de 12 variables (q), choisies parmi 24 disponibles (p), nous trouvons encore 2,7 millions de combinaisons à évaluer. En outre, le paramètre q n'est rarement connu *a priori*. Pour échapper à l'absence d'information sur le nombre de variables finales et de l'explosion combinatoire engendrée par des recherches exhaustives, il devient nécessaire de définir une procédure de recherche, permettant de guider la sélection et l'élimination des variables.

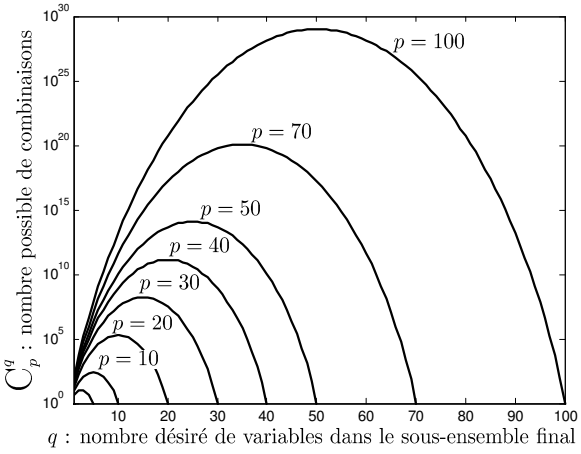


FIG. 2.18 – Évaluation du nombre de combinaisons de variables possibles, en fonction du nombre de variables disponibles (p) et du nombre de variables à sélectionner (q).

L'organigramme de la figure 2.19, proposé par [Liu *et al.*, 1998; Liu and Yu, 2002], illustre la procédure traditionnelle pour la sélection d'un sous-ensemble de variables. Cette procédure repose sur deux éléments principaux :

- le **critère d'évaluation**, qui doit permettre d'estimer si un sous-ensemble de variables est meilleur qu'un autre ;

- la **procédure de recherche**, qui doit permettre de chercher les sous-ensembles candidats de variables.

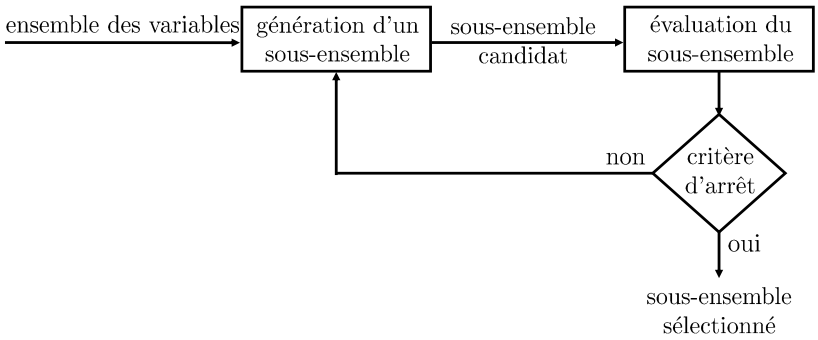


FIG. 2.19 – Procédure traditionnelle de recherche d'un sous-ensemble de variables [Liu *et al.*, 1998; Liu and Yu, 2002].

À l'instar de l'extraction de caractéristiques, l'objectif de la sélection de variables est de réduire la dimension d'un problème. Dans le cadre de la sélection, la procédure, définie par la figure 2.19, montre que l'objectif est établi par le choix d'un certain nombre de variables pertinentes suivant un critère. Cependant, cet objectif peut encore être précisé. En effet, [Kudo and Sklansky, 2000] apportent un regard légèrement différent et surtout plus minutieux sur l'objectif à atteindre. Ainsi, en tenant compte des capacités algorithmiques des méthodes de sélection, ils sont parvenus à extraire les trois objectifs suivants :

- l'algorithme trouve un sous-ensemble de taille donné pour lequel le critère est optimal ;

- l'algorithme trouve le plus petit sous-ensemble pour lequel son critère est supérieur à un seuil donné ;

- l'algorithme trouve un sous-ensemble en réalisant un compromis des deux précédents objectifs.

Nous verrons, par la suite, que ces objectifs sont accessibles en fonction des procédures de recherche employées.

L'ensemble des remarques exposées montre l'enjeu et la complexité de la tâche de sélection de variables. Cette tâche peut s'exprimer comme un problème d'optimisation combinatoire, où l'on cherche parmi un ensemble fini de solutions admissibles, la ou les solutions qui optimise un critère et donc la résolution du problème.

2.4.2 Critères d'évaluation

Rappelons que nous sommes dans le cadre de l'amélioration des performances d'un système de classification par la sélection de variables. À ce titre, il est nécessaire de définir une mesure de pertinence, faisant état de la qualité de la variable ou du sous-ensemble de variables sélectionnées. [Bennani, 2001] définit une variable pertinente telle que sa suppression entraîne une détérioration des performances du système de classification.

Idéalement, en classification supervisée, le critère d'évaluation d'un sous-ensemble de variables pourrait être fondé sur le taux de classification. Ce dernier serait obtenu par l'évaluation des performances de généralisation du modèle, une fois l'apprentissage réalisé ; les entrées de ce modèle

seraient composées des variables pré-sélectionnées. Les méthodologies d'évaluation des performances d'un modèle seront au chapitre 3. Cependant, comme le note judicieusement [Bishop, 1995], les procédures d'apprentissage peuvent être très coûteuses, notamment avec les réseaux de neurones, et répéter le processus d'évaluation pour chaque sous-ensemble pourrait devenir excessivement long. [Bishop, 1995] suggère alors d'utiliser des méthodes de classification plus simples et plus rapides, telles que des techniques linéaires (*cf.* section 1.3), pour sélectionner les variables et, ainsi, générer le modèle « final » avec des méthodes de classification plus sophistiquées à partir du sous-ensemble préalablement déterminé. Cependant, par cette approche, le sous-ensemble de variables, obtenu durant la sélection, ne sera pas forcément optimal pour la conception du modèle. En effet, [Liu and Yu, 2002] soulignent judicieusement qu'un sous-ensemble de variables peut être optimal suivant un certain critère et peut ne plus l'être pour un autre. Ils notent ainsi l'importance et l'influence du critère d'évaluation dans le processus de sélection. Ainsi, pour un outil de classification donné, la sélection du sous-ensemble peut être biaisée par le taux de classification obtenu, et par conséquent, ce même sous-ensemble peut donner des taux de classification bien moins optimaux pour d'autres classifieurs. Ce processus de recherche appartient à une catégorie de méthodes nommée *wrapper*, qui suggère l'utilisation d'un algorithme d'apprentissage dans la phase de sélection de variables [Kohavi and John, 1997].

Un autre ensemble de méthodes permet de rechercher des sous-ensembles de variables, sans utiliser un algorithme d'apprentissage : cette catégorie de méthodes est nommée *filter* [Kohavi and John, 1997; Hall, 2000; Yu and Liu, 2003]. En effet, en présence d'observations étiquetées, le choix d'un sous-ensemble de variables peut se faire en considérant l'habilité du sous-ensemble à discriminer les classes. Dans ce cas, la pertinence d'une variable pourrait être définie par une mesure de séparabilité des classes, ou encore, par une évaluation du recouvrement entre les classes [Theodoridis and Koutroumbas, 2006]. Cette pertinence s'obtiendrait indépendamment d'un algorithme d'apprentissage.

Les deux catégories, *filter* et *wrapper*, se distinguent donc en fonction de la participation de l'algorithme d'apprentissage dans la sélection du sous-ensemble de variables ; la figure 2.20 illustre ces deux approches.

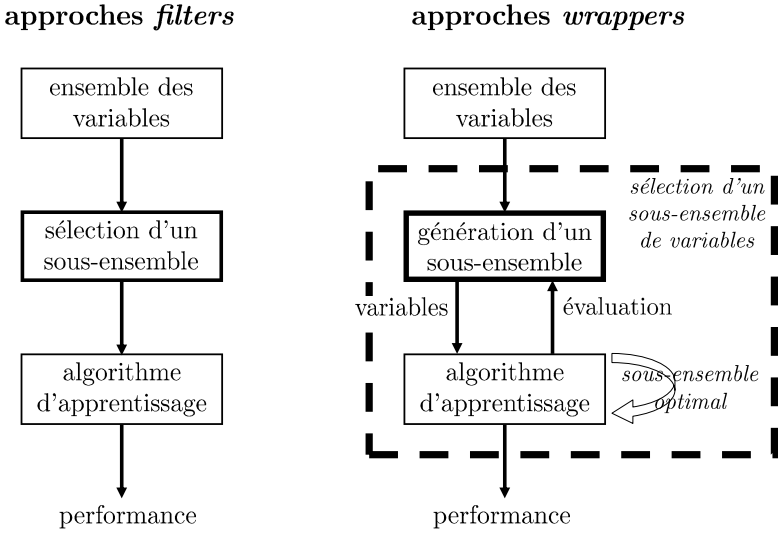


FIG. 2.20 – Approches à la sélection de sous-ensembles de variables (*filter* et *wrapper*) fondées sur l'intégration d'un algorithme d'apprentissage [Yang and Honavar, 1997].

Des auteurs, comme [Blum and Langley, 1997], font référence à une autre catégorie, appelée *embedded*, qui réalise la sélection de variables parallèlement à l'apprentissage. Plus couramment, on joindra aux deux catégories principales (*filter* et *wrapper*), une dernière catégorie nommée hybride [Das, 2001; Xing *et al.*, 2001; Sebban and Nock, 2002]. Celle-ci tente de tirer avantage des précédentes approches, en exploitant leurs différents critères d'évaluation dans plusieurs étapes de la recherche du sous-ensemble [Liu and Yu, 2005]. Cette dernière approche peut être privilégiée en présence d'un nombre de variables très important. Dans la pratique, on pourrait dans une première étape faire une pré-sélection par des méthodes de type *filter*, afin de réduire le nombre de variables. Puis, pour optimiser la sélection, une deuxième étape fondée sur une approche de type *wrapper* pourrait être réalisée afin d'obtenir la sélection du sous-ensemble final.

Dans la présentation générale des approches évaluant la pertinence d'un sous-ensemble, nous avons abordé l'utilisation du taux de classification comme critère d'évaluation des approches de type *wrapper*. Ce critère est obtenu nécessairement par un algorithme d'apprentissage. Concernant les approches de type *filter*, [Webb, 2002; Theodoridis and Koutroumbas, 2006] décrivent rigoureusement plusieurs mesures d'évaluation, afin d'estimer la capacité d'une variable ou d'un sous-ensemble à séparer les classes. Nous pouvons les regrouper en plusieurs catégories.

2.4.2.1 Mesures de distances probabilistes

Les mesures de distances probabilistes sont parfois appelées mesures de discrimination, ou encore, mesures de divergence [Theodoridis and Koutroumbas, 2006].

À la section 1.2.2, nous avons évoqué la règle de décision de Bayes, où pour deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$, nous attribuons le vecteur d'entrée $\mathbf{x}$ à la classe $\mathcal{C}_1$, si $P(\mathcal{C}_1|\mathbf{x}) > P(\mathcal{C}_2|\mathbf{x})$. Par conséquent, l'erreur de classification, et donc la capacité de discrimination, s'identifie par l'écart entre les deux probabilités *a posteriori* $P(\mathcal{C}_1|\mathbf{x})$ et $P(\mathcal{C}_2|\mathbf{x})$. Connaissant la relation liant ces probabilités aux densités de probabilité, nous pouvons désormais retrouver cette même information dans le rapport entre les densités de probabilité $p(\mathbf{x}|\mathcal{C}_1)$ et $p(\mathbf{x}|\mathcal{C}_2)$. Dans leur démonstration, [Theodoridis and Koutroumbas, 2006] nous permettent d'aboutir à la relation (2.14). Ainsi, pour évaluer la capacité d'une variable x à discriminer deux classes, nous pouvons utiliser le calcul de la distance probabiliste, donné par la relation suivante, connue aussi sous le nom de divergence :

$$J_D = \int_{-\infty}^{+\infty} [p(x|\mathcal{C}_1) - p(x|\mathcal{C}_2)] \log \left(\frac{p(x|\mathcal{C}_1)}{p(x|\mathcal{C}_2)} \right) dx \,. \tag{2.14}$$

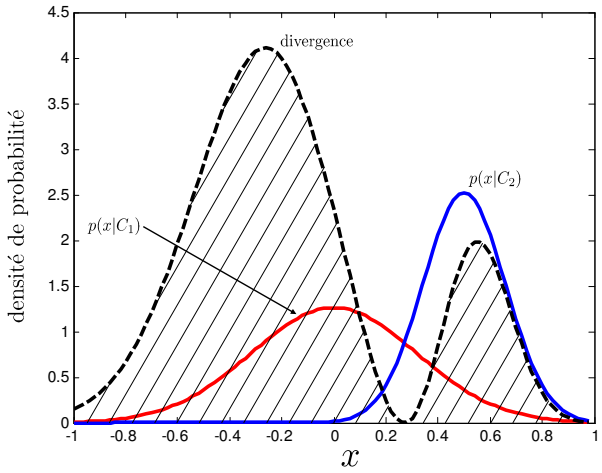
La figure 2.21 montre l'évolution de cette mesure en fonction du recouvrement entre les deux classes. La valeur du critère J_D est l'aire sous la courbe en pointillés (notée divergence), soit la zone hachurée. Ainsi, nous pouvons observer que cette mesure est maximale lorsque le recouvrement des classes est minimal.

Sous certaines conditions, il est fréquent d'estimer les densités de probabilité par des gaussiennes, soit μ_1 et μ_2 les moyennes et Σ_1 et Σ_2 les matrices de covariances respectivement pour les deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$.

$$J_D = \frac{1}{2}\text{trace}\{\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I\} + \frac{1}{2}(\mu_1 - \mu_2)^T \left(\Sigma_1^{-1}\Sigma_2^{-1}\right)(\mu_1 - \mu_2)\,, \tag{2.15}$$

où I est la matrice identité. D'autre part, si nous supposons que les matrices de covariances sont égales, soit $\Sigma_1 = \Sigma_2 = \Sigma$, alors la mesure de divergence devient :

$$J_D = (\mu_1 - \mu_2)\Sigma^{-1}(\mu_1 - \mu_2)\,. \tag{2.16}$$



Note : Le critère J_D est l'aire sous la courbe en pointillés (notée divergence), il correspond donc à la zone hachurée.

FIG. 2.21 – Observation de la mesure de divergence (distance probabiliste) en fonction du niveau de recouvrement des classes.

Cette relation est tout simplement la distance probabiliste la plus connue, nommée distance de Mahalanobis. Ces relations illustrent un cas à deux classes, mais nous pouvons facilement les généraliser aux cas multiclasses [Webb, 2002; Theodoridis and Koutroumbas, 2006]. [Webb, 2002] propose, en annexe de son ouvrage, un nombre important d'autres mesures de distances probabilistes, toujours fondées sur les densités de probabilité.

2.4.2.2 Mesures de distances fondées sur les matrices de covariances

Conformément à la section 1.3.2.1, où nous avons abordé la classification linéaire, nous notons respectivement $\mathbf{S}_B$ et $\mathbf{S}_W$, les matrices de covariances interclasse et intraclasse. [Webb, 2002; Theodoridis and Koutroumbas, 2006] proposent plusieurs critères basés sur ces matrices, afin d'évaluer la capacité d'une variable ou d'un sous-ensemble de variables à séparer des classes. Le critère le plus populaire est certainement J_1 :

$$J_1 = \text{trace}\{\mathbf{S}_W^{-1}\mathbf{S}_M\}\,, \tag{2.17}$$

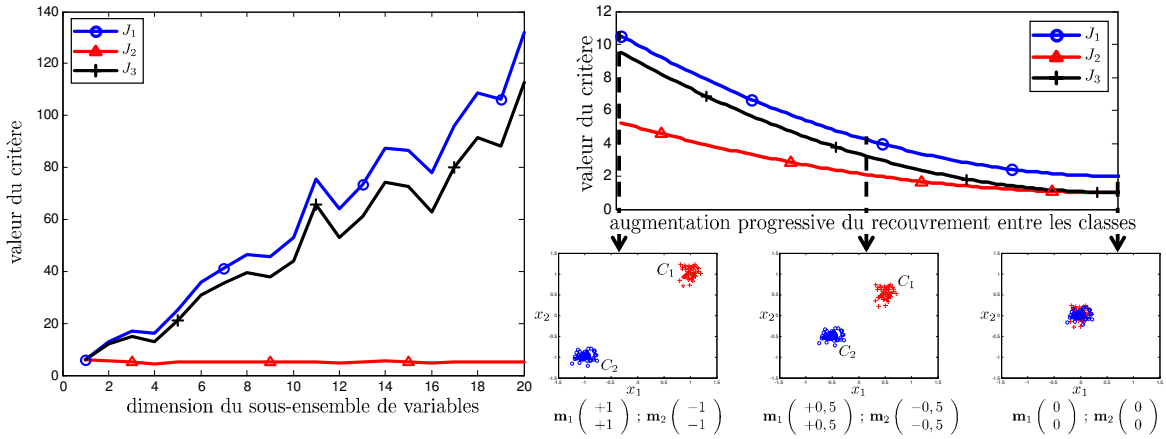
notons que $\mathbf{S}_M = \mathbf{S}_W + \mathbf{S}_B$. Dans quelques applications, la matrice $\mathbf{S}_M$ de J_1 est remplacée par $\mathbf{S}_B$, tout comme pour le critère J_2 :

$$J_2 = \frac{\text{trace}\{\mathbf{S}_M\}}{\text{trace}\{\mathbf{S}_W\}}\,. \tag{2.18}$$

Aussi, l'opérateur « trace » est parfois remplacé par le « déterminant », comme dans le critère J_3 :

$$J_3 = \frac{\det\{\mathbf{S}_M\}}{\det\{\mathbf{S}_W\}}\,. \tag{2.19}$$

La figure 2.22(a) montre que le critère J_2 est indépendant de l'élargissement de la dimension, soit à l'augmentation du nombre de variables, contrairement au critère J_3. La figure 2.22(b) illustre l'évolution des critères en fonction de l'augmentation du recouvrement entre les classes, lorsque la variance interclasse diminue.



Note : (à gauche) Évolution des critères en fonction de l'augmentation du nombre de variables considérées. (à droite) Évolution des critères en fonction de l'augmentation du recouvrement.

FIG. 2.22 – Observation de l'évolution des critères basés sur les matrices de covariances en fonction de différentes distributions des observations.

Comme nous l'avons montré, les trois critères permettent d'évaluer la pertinence d'un sous-ensemble de variables. Cependant, en considérant uniquement l'évaluation d'une variable, nous pouvons observer une simplification des expressions des matrices de covariance interclasse et intraclasse, respectivement par $\mathbf{S}_B = (\mu_1 - \mu_2)^2$ et par $\mathbf{S}_W = \sigma_1^2 + \sigma_2^2$. Cette nouvelle formulation nous permet de faire le parallèle avec la fonction discriminante de Fisher (*cf.* section 1.3.2.1), où, μ_1 et σ_1 sont respectivement la moyenne et la variance des observations liées à la classe $\mathcal{C}_1$. Ainsi, en combinant $\mathbf{S}_B$ et $\mathbf{S}_W$, nous retrouvons la relation appelée critère de Fisher. Rappelons que ce critère, que l'on notera FDR (pour *Fisher Discriminant Ratio*), minimise le ratio entre la variance interclasse et la variance intraclasse :

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\,. \tag{2.20}$$

[Theodoridis and Koutroumbas, 2006] généralisent ce critère au traitement de K classes :

$$FDR_M = \sum_{i}^{K} \sum_{j \neq i}^{K} \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2}\,. \tag{2.21}$$

2.4.2.3 Mesures de dépendance

Les mesures de dépendance évaluent la capacité d'une variable à en prédire une autre. En d'autres termes, elles cherchent à mesurer la redondance d'information dans les variables [Yu and Liu, 2004].

[Yu and Liu, 2003] indiquent qu'une variable est pertinente, si elle est corrélée à la variable de sortie (classe) et si elle n'est pas redondante avec d'autres variables. Cette définition incite à mesurer une certaine corrélation entre les variables originales et la « variable − classe ». Notons que, si les variables étaient continues, nous pourrions utiliser le coefficient de corrélation linéaire de Pearson, comme le proposent [Guyon and Elisseeff, 2003; Stoppiglia *et al.*, 2003]. Cependant, dans notre contexte de classification, la « variable − classe[4] » est composée d'étiquettes liant les

[4]Nous simplifierons par la suite le terme « variable − classe » par classe.

observations aux classes. Ainsi, [Guyon and Elisseeff, 2003] suggèrent d'utiliser des critères mesurant la séparabilité, comme le critère de Fisher (FDR), utilisé notamment par [Golub *et al.*, 1999; Furey *et al.*, 2000] dans des applications liées à la bio-informatique. Dans cette même direction, où l'on cherche à évaluer la corrélation des variables avec la classe, la méthode ANOVA (*ANalysis Of VAriance*) est parfois utilisée [Sahai, 2000; Guyon and Elisseeff, 2003].

Évaluer uniquement la corrélation ou la dépendance entre les variables et la classe, amène un inconvénient majeur. En effet, d'une part, cette approche ne peut pas écarter les variables redondantes, et d'autre part, comme évoqué par [Guyon and Elisseeff, 2003], elle peut éliminer des variables peu corrélées avec la classe sans prendre garde à d'hypothétiques complémentarités avec d'autres variables. [Stoppiglia *et al.*, 2003; Theodoridis and Koutroumbas, 2006] proposent une procédure intégrant cette remarque, elle sera discutée à la section 2.4.3.1.

Remarquons que ce type de critère d'évaluation, appartenant à la catégorie *filter*, est parfaitement adapté pour traiter des ensembles de données contenant un nombre important de variables. Ainsi dans le domaine de la bio-informatique, l'analyse de l'expression de gènes entraîne une représentation du problème pouvant atteindre plusieurs milliers de variables. Par exemple, [Golub *et al.*, 1999; Furey *et al.*, 2000] ont utilisé le critère de Fisher pour faire la sélection. Autre exemple, [Mercier *et al.*, 2004] ont analysé un ensemble contenant pas moins de 6 135 gènes. Raisonnablement, avant d'effectuer l'apprentissage du modèle, ils ont trié et classé les gènes par pertinence. Dans leur étude [Mercier *et al.*, 2004] ont utilisé, pour la sélection, des méthodes fondées sur l'analyse d'indépendance, comme ANOVA et RELIEF.

L'algorithme RELIEF est fondé sur un processus aléatoire qui estime la qualité de chaque variable pour un problème de classification, en assignant un poids de pertinence. Ainsi, pour une observation choisie aléatoirement, l'algorithme recherche deux observations parmi ses plus proches voisins : la première appartenant à sa classe (appelée *nearest hit :* $\mathbf{x}_h$) et la seconde étant de classe différente (appelée *nearest miss :* $\mathbf{x}_m$). [Kononenko, 1994] estime le poids w_j de la variable j par la différence entre deux probabilités :

$$w_j = p(x_{ij}|x_{mj}) - p(x_{ij}|x_{hj})\,, \tag{2.22}$$

où x_{ij} représente la i-ème valeur de l'observation de la variable j, x_{hj} et x_{mj} sont les deux observations sélectionnées. Ainsi, une variable pertinente se distingue par une différence importante entre ces deux probabilités.

[Kononenko, 1994; Robnik-Sikonja and Kononenko, 2003] proposent et récapitulent un nombre important d'évolutions de l'algorithme RELIEF, comme le remplacement des deux observations les plus proches (de même classe et de classe différente) par deux sous-ensembles d'observations. Cela permet d'obtenir une plus grande résistance au bruit. Ainsi, l'algorithme recherche deux sous-ensembles d'observations parmi les plus proches voisins de l'observation choisie, nous noterons $\mathcal{X}_h$ et $\mathcal{X}_m$, les sous-ensembles contenant les observations les plus proches respectivement de même classe et de classes différentes à l'observation choisie.

L'algorithme 2.1 montre une estimation des poids w faite par [Kononenko, 1994].

Algorithme 2.1 : Pseudo-code de l'algorithme d'évaluation RELIEF.

Données :
$\mathcal{X} = \{\mathbf{x}_i, t_i\}_{i=1,\ldots,n}$, $\mathbf{x}_i \in \mathbf{R}^p$, $t_i \in \{-1, 1\}$
m : nombre d'observations considéré dans l'évaluation des p poids w
Résultat : w_j : poids des variables $j = 1, \ldots, p$

début
  pour chaque $j = 1$ *à* p **faire**
    $w_j \leftarrow 0$;
    pour chaque $i = 1$ *à* m **faire**
      sélectionner une observation x_{ij} par tirage aléatoire de son indice i ;
      trouver le sous-ensemble d'observations $\mathcal{X}_h$ plus proche de x_{ij} et de même classe ;
      trouver le sous-ensemble d'observations $\mathcal{X}_m$ plus proche de x_{ij} et de classe différente ;
      $w_j \leftarrow w_j - \left(\sum_{x_{rj} \in \mathcal{X}_h} |x_{ij} - x_{rj}| - \sum_{x_{rj} \in \mathcal{X}_m} |x_{ij} - x_{rj}| \right)$;
    fin
    $w_j \leftarrow w_j / m$;
  fin
fin

L'intérêt des méthodes d'évaluation de type *filter* réside dans leur rapidité d'exécution, grâce notamment à la non-utilisation d'outils de classification. Cependant, comme souligné auparavant par [Liu and Yu, 2002][5], le sous-ensemble optimal pourrait se révéler inefficace une fois appliqué à un outil de classification. Cet inconvénient n'apparaît pas dans les méthodes de type *wrapper*, où l'évaluation est fondée directement sur l'outil de classification. En contre-partie, cette démarche rend les méthodes *wrapper* fortement dépendantes du classifieur utilisé. En d'autres termes, le sous-ensemble sélectionné peut être rendu inexploitable pour d'autres classifieurs : interdisant une certaine généralisation. Le choix du type d'évaluation peut donc dépendre de l'importance que l'on souhaite donner aux outils de classification.

2.4.3 Génération de sous-ensembles : procédures de recherche

Dans la section précédente, nous avons mis en évidence les deux principales approches inhérentes à l'évaluation d'un sous-ensemble ; nous allons maintenant aborder les stratégies de recherche. Rappelons que q définit le nombre de variables sélectionnées, sachant que cette valeur n'est pas censée être connue *a priori*, nous la considérons alors égale à p.

2.4.3.1 Classement des variables

L'obtention d'un classement, indiquant la qualité de discrimination des variables, peut être réalisé indépendamment d'un outil de classification. Sans classifieur, il est donc nécessaire de définir un « score » pour évaluer la pertinence de chaque variable. L'évaluation des variables peut se faire par des approches de type *filter*, en employant un critère mesurant la séparabilité, comme FDR par exemple. La valeur du critère J est alors calculée pour chacune des variables $J(x_j)$, $j = 1, \ldots, p$. Les variables sont ensuite triées afin d'obtenir un classement allant de la variable la plus pertinente à la moins pertinente. Cette démarche est également appelée dans la littérature anglo-saxonne *variable/feature ranking* [Guyon and Elisseeff, 2003; Stoppiglia *et al.*, 2003] ou encore *scalar feature selection* [Theodoridis and Koutroumbas, 2006]. Ainsi, la sélection de q variables correspondrait aux q meilleures valeurs de $J(x_j)$. Cette approche fait référence à une méthode de sélection dite « naïve », comme le montre l'exemple suivant.

[5] Au début de la section 2.4.2, nous avons évoqué une remarque de [Liu and Yu, 2002] dans laquelle les auteurs soulignaient qu'un sous-ensemble de variables peut être optimal suivant un certain critère et peut ne plus l'être pour un autre critère.

La figure 2.23 illustre un problème de classification binaire, dans lequel trois variables sont disponibles (x_1, x_2 et x_3). L'observation indépendante de chaque variable, permet de classer les variables de la plus discriminante vers la moins discriminante : x_3, x_1 et x_2. En effet, la représentation des densités de probabilité indique que la variable x_3 sépare les classes avec un minimum de recouvrement.

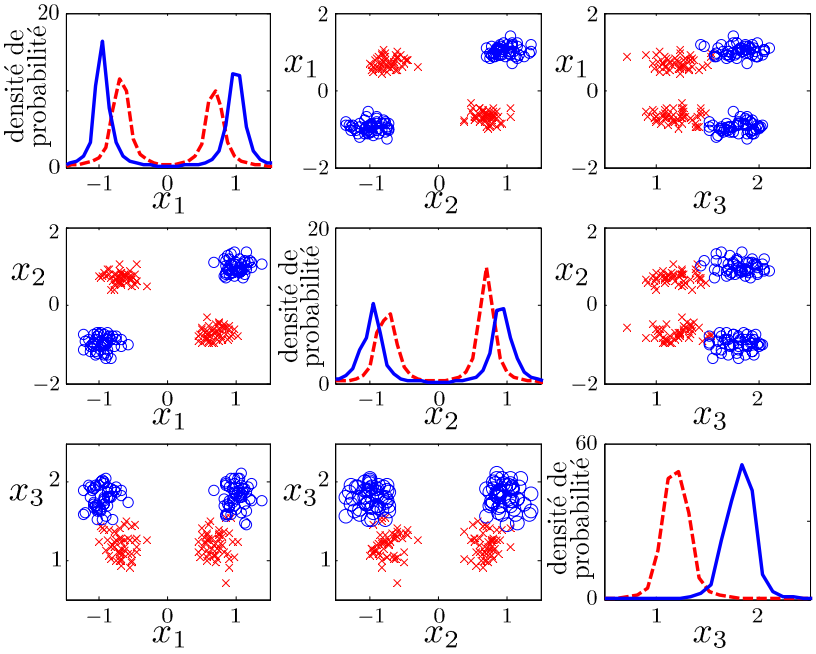


FIG. 2.23 – Illustration détaillée d'un problème binaire à trois variables dans le cadre d'une sélection naïve de variables.

Parmi ces trois variables, nous désirons en sélectionner deux. Dès lors, l'approche naïve serait de sélectionner les deux premières variables les plus discriminantes, soit x_3 et x_1. Cependant, dans cette représentation, comme le montre la figure 2.23, aucune frontière (aussi bien linéaire que non linéaire) ne permet de séparer les classes sans erreur. Cette même figure nous montre alors que l'utilisation des variables apparaissant individuellement peu pertinente (x_1 et x_2), permet de séparer très distinctement les deux classes.

Un peu simpliste, cet exemple démontre néanmoins le besoin de d'employer une procédure de recherche plus élaborée afin d'évaluer la complémentarité dans les combinaisons de variables. En outre, la nécessité d'améliorer la procédure est également motivée par le besoin de considérer la dépendance entre les variables : [Stoppiglia *et al.*, 2003; Guyon and Elisseeff, 2003; Theodoridis and Koutroumbas, 2006] suggèrent alors une mesure de corrélation. Ces auteurs proposent le calcul de la corrélation entre deux variables x_j et x_l par la relation :

$$\rho_{jl} = \frac{\sum_{i=1}^{n} x_{ij}\, x_{il}}{\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{il}^2}, \tag{2.23}$$

où $|\rho_{jl}| \leq 1$. Rappelons que x_{ij} ($i = 1, \ldots, n$ et $j = 1, \ldots, p$) donne la i-ème valeur observée de la variable j. La relation (2.23) présuppose que les variables sont centrées (*cf.* section 2.2.3, sur la normalisation).

Afin de considérer judicieusement ce nouveau paramètre ρ_{jl} dans le processus de sélection, [Theodoridis and Koutroumbas, 2006] proposent alors la procédure donnée dans l'algorithme 2.2.

Algorithme 2.2 : Pseudo-code d'une variante de l'algorithme de sélection « naïve » de variables.

Données :
$\langle FS \rangle = \{x_1, \ldots, x_p\}$: ensemble des variables initiales
α_1 : pondère l'importance relative de la capacité de discrimination de la variable
α_2 : pondère l'importance relative de la corrélation entre les variables
q_{max} : nombre de variables à sélectionner
Résultat : $\langle SS_{q_{max}} \rangle$: sous-ensemble de q_{max} variables sélectionnées

début
  $\langle SS_1 \rangle \leftarrow \emptyset$;
  $x_+ \leftarrow \underset{x_j \in \langle FS \rangle}{\text{argmax}} \{J(x_j)\}$;
  $\langle SS_1 \rangle \leftarrow \{x_+\}$;
  pour chaque $q = 2$ *à* q_{max} **faire**
    $x_+ \leftarrow \underset{x_j \in \{\langle FS \rangle - \langle SS_{q-1} \rangle\}}{\text{argmax}} \left\{ \alpha_1 J(x_j) - \frac{\alpha_2}{q-1} \sum_{x_{jr} \in \langle SS_{q-1} \rangle} |\rho_{j_r j}| \right\}$;
    $\langle SS_q \rangle \leftarrow \langle SS_{q-1} \rangle \cup \{x_+\}$;
  fin
fin

Notons que cet algorithme ajoute à chaque itération la variable la plus pertinente (en fonction du critère J) parmi celles qui n'ont pas été sélectionnées. Ainsi, les q_{max} variables du sous-ensemble final $\langle SS_{q_{max}} \rangle$ sont triées par ordre décroissant de leur pertinence. Dans de nombreuses applications, et lorsque le nombre de variables initiales n'est pas trop important, ou encore lorsque q_{max} ne peut pas être défini *a priori*, il est courant de choisir $q_{max} = p$, de comparer les p sous-ensembles ($\langle SS_1 \rangle$, $\langle SS_2 \rangle$, $\cdots$, $\langle SS_q \rangle$) et de sélectionner le meilleur.

2.4.3.2 Sélection d'un sous-ensemble de variables

Les procédures de recherche s'emploient à générer des sous-ensembles dans l'espace des variables. Rappelons que la présence de p variables entraîne un nombre de 2^p combinaisons possibles. Ce nombre de combinaisons peut être réduit à C_p^q, si q est fixé. Cependant, cette simplification ne permet toujours pas de réduire significativement le nombre de combinaisons à évaluer (voir figure 2.18, page 78). Dans ces conditions, une recherche exhaustive n'est pas envisageable. Néanmoins, l'utilisation d'une méthode optimale, connue sous le nom de *Branch and Bound* (B & B) [Narendra and Fukunaga, 1977] peut être envisagée, si le critère d'évaluation J est monotone avec le nombre de variables sélectionnées, tel que :

$$J(x_{j_1}) \leq J(x_{j_1}, \ldots, x_{j_l}) \leq \cdots \leq J(x_{j_1}, \ldots, x_{j_{l+1}}) \,, \tag{2.24}$$

où x_{j_l} indique la l-ième variable sélectionnée. Cependant, trouver un critère d'évaluation monotone n'est pas une chose aisée [Bennani, 2001]. Dès lors, pour éviter l'explosion combinatoire d'une recherche exhaustive, il est nécessaire de changer de stratégie de recherche, en se basant sur des procédures dites sous-optimales qui s'occupent de gérer la sélection et l'élimination des variables. Dans cet objectif, il peut être judicieux d'utiliser des **heuristiques**, afin de s'approcher de la solution optimale.

Comme nous le verrons par la suite, les approches heuristiques suivent rigoureusement une direction de recherche, qui peut les amener à converger vers des solutions non optimales. Ainsi, afin d'optimiser la recherche et les solutions fournies par des méthodes heuristiques, d'autres méthodes

ont fait leur apparition. Celles-ci permettent d'assurer une meilleure diversité et efficacité dans la recherche de sous-ensembles de variables. Ainsi, ces nouveaux algorithmes s'autorisent à prospecter dans des zones jugées peu intéressantes, tout en évitant de s'y enfermer, mais intensifient également les recherches dans des zones paraissant plus pertinentes. Appelés **métaheuristiques**, ces algorithmes sont connus pour la qualité des solutions obtenues et pour leur adaptabilité à bon nombre de problèmes. [Dréo *et al.*, 2003] caractérisent ces algorithmes par :

- un processus stochastique (au moins pour la majorité), permettant de faire face à l'explosion combinatoire ;

- une analogie aux phénomènes de la physique (recuit simulé), de la biologie (algorithmes génétiques) et de l'éthologie (colonies de fourmis) ;

- des difficultés de réglage des paramètres et le temps de calcul élevé.

La lecture des précédents paragraphes nous révèle l'existence de trois catégories dans la génération de sous-ensembles de variables : exhaustive, déterministe et non déterministe (stochastique). Cependant, il n'y a pas réellement de consensus sur l'appellation de ces catégories. En effet, l'auteur Anil K. Jain, très actif dans le domaine de la reconnaissance de formes, propose dans [Jain and Zongker, 1997] une hiérarchisation des approches de sélection légèrement différente de celles que nous trouvons classiquement dans la littérature. Dans un premier temps, il dissocie les méthodes optimales des méthodes sous-optimales, puis à partir des méthodes sous-optimales, il différencie les méthodes apportant une seule solution au problème de celles en apportant plusieurs. Pour finir, dans chacun des deux derniers groupes, il présente les approches déterministes et non déterministes.

Dans ce manuscrit, nous ferons un choix plus classique dans la présentation des stratégies de recherche, en suivant la classification de [Liu *et al.*, 1998] qui considèrent les trois catégories suivantes : exhaustive, heuristique et non déterministe. Cet excellent ouvrage donne un aperçu global des méthodes de sélection, où, après la décision de la stratégie à adopter, il sollicite le choix de la direction de recherche à prendre. On dénombre trois types de stratégies :

- **stratégie ascendante**, qui commence à partir d'un ensemble vide de variables, puis ajoute progressivement une variable ;

- **stratégie descendante**, qui commence avec toutes les variables, puis élimine progressivement une variable ;

- **stratégie aléatoire**, qui choisit aléatoirement un sous-ensemble de variables, puis ajoute ou retire progressivement des variables.

2.4.3.3 Approches heuristiques

Les approches heuristiques répondent à une contrainte calculatoire, notamment lorsque le nombre de variables empêche une évaluation exhaustive. Ces approches permettent alors de faire un compromis entre le nombre de combinaisons à évaluer et le coût global de l'évaluation. Les algorithmes opèrent par recherche séquentielle. Cette stratégie réduit le nombre de combinaisons à évaluer, en appliquant des recherches locales suivant une direction qu'il reste à définir : ascendante, descendante ou encore aléatoire.

Parmi les méthodes utilisant une évaluation de type *wrapper*, les techniques heuristiques les plus connues sont fondées sur des sélections séquentielles ascendante et descendante, respectivement nommées *sequential forward selection* (SFS) et *sequential backward selection* (SBS). Ces méthodes identifient le meilleur sous-ensemble de variables en ajoutant ou en éliminant progressivement des variables. Ainsi, SFS commence avec un ensemble vide de variables, et choisit à chaque itération la variable améliorant le plus la valeur du critère. Son implémentation est décrite dans le pseudo-code de l'algorithme 2.3 adapté de [Domingos, 1997]. Dans cette procédure, le critère d'évaluation J permet d'obtenir les performances des sous-ensembles de variables, il est généralement fondé sur les performances de classification. Dans notre exemple, nous chercherons à le maximiser. Inversement, SBS commence avec toutes les variables, et choisit d'éliminer à chaque itération la variable qui améliore le plus le critère.

Algorithme 2.3 : Pseudo-code de l'algorithme de sélection de variables SFS.

Données :
$\langle FS \rangle = \{x_1, \ldots, x_p\}$: ensemble des variables initiales
q_{max} : nombre de variables à sélectionner
Résultat : $\langle SS_{q_{max}} \rangle$: sous-ensemble de q_{max} variables sélectionnées

```
1 début
2 |  ⟨SS_0⟩ ← ∅ ;
3 |  pour chaque q = 1 à q_max faire
  |  |  x_+ ← argmax_{x_j ∈ {⟨FS⟩ − ⟨SS_{q−1}⟩}} {J(⟨SS_{q−1}⟩ ∪ {x_j})} ;
4 |  |
5 |  |  ⟨SS_q⟩ ← ⟨SS_{q−1}⟩ ∪ {x_+} ;
6 |  fin
7 fin
```

Comme pour l'algorithme 2.2 (variante de la sélection naïve), l'algorithme SFS ajoute à chaque itération la variable la plus pertinente parmi celles non sélectionnées. La différence entre les deux approches demeure dans le critère, où SFS utilise le résultat d'un algorithme d'apprentissage dans l'évaluation des sous-ensembles. Cela n'interdit pas d'exécuter l'algorithme pour $q_{max} = p$, afin de considérer le maximum de combinaisons que permet cette méthode. Dans ce cas, au terme du processus (lorsque $q = p$), l'algorithme aura évalué $1 + p \cdot (p + 1)/2$ sous-ensembles.

Ces deux méthodes (SFS et SBS) ont l'avantage de réduire considérablement le nombre de sous-ensembles de variables à évaluer. Cependant, leur principale limitation est leur incapacité durant leur processus à éliminer une variable sélectionnée pour SFS et, à sélectionner une variable éliminée pour SBS. Pour remédier à cela, d'autres méthodes sont apparues en proposant des possibilités de « retours en arrière ».

Dans cette perspective, la méthode *Plus-L Minus-R Selection* (LRS) [Stearns, 1976] modifie les processus précédents afin d'offrir plus de souplesse. Deux cheminements sont possibles :

- si $L > R$, LRS commence avec un ensemble vide de variables, ajoute L variables, puis en retire R, jusqu'à la sélection des p variables ;

- si $L < R$, LRS commence avec toutes les variables, élimine R variables, puis en ajoute L, jusqu'à l'élimination des p variables.

Comme précédemment, le choix de la variable à sélectionner ou à éliminer est suggéré par un critère. En améliorant les méthodes précédentes, LRS demande cependant à l'utilisateur de définir les paramètres L et R, rendant la méthode et, donc, le résultat dépendant des choix effectués.

Dérivées de la méthode LRS, et toujours fondées sur les méthodes SFS et SBS, les méthodes de sélection séquentielle ascendante/descendante flottante (SFFS et SFBS − *sequential floating forward/backward selection*) offrent un « retour en arrière » plus flexible que LRS. En effet, contrairement à LRS, le nombre de variables à supprimer (SFFS) et à ajouter (SFBS) n'est plus fixé par des paramètres constants, mais est défini en fonction de l'évolution du critère. Par conséquent, l'intervention de l'utilisateur dans le réglage des paramètres, qui nuisait tant à LRS, disparaît avec les méthodes flottantes. Une fois de plus, deux parcours sont possibles :

- comme son homologue, SFFS commence avec un ensemble vide de variables. Après chaque sélection d'une variable, l'algorithme élimine une à une des variables déjà sélectionnées, jusqu'à ne plus pouvoir améliorer le critère. Alors, il répète une nouvelle séquence, en recommençant par ajouter une nouvelle variable optimisant toujours le critère. Le processus se termine lorsque les p variables sont sélectionnées. Une schématisation de ce processus fondée sur [Semani *et al.*, 2004] est donnée par l'algorithme 2.4 ;

- SFBS commence avec toutes les variables. Après chaque élimination d'une variable, l'algorithme ajoute une à une des variables éliminées auparavant, jusqu'à ne plus obtenir d'amélioration du critère. Alors, il répète une nouvelle séquence, en recommençant par éliminer la variable optimisant le plus le critère, jusqu'à l'élimination totale des variables.

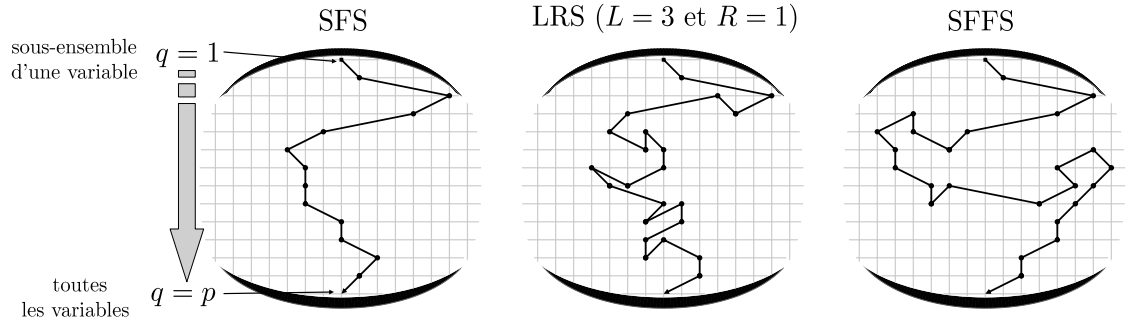
Algorithme 2.4 : Pseudo-code de l'algorithme de sélection de variables SFFS.

Données :
$\langle FS \rangle = \{x_1, \ldots, x_p\}$: ensemble des variables initiales
q_{max} : nombre de variables à sélectionner
Résultat : $\langle SS_{q_{max}} \rangle$: sous-ensemble de q_{max} variables sélectionnées

```
début
    ⟨SS_0⟩ := ∅ ;
    pour chaque q = 1 à q_max faire
        x_+ ← argmax_{x_j ∈ {⟨FS⟩ − ⟨SS_{q−1}⟩}} {J(⟨SS_{q−1}⟩ ∪ {x_j})} ;
        ⟨SS_q⟩ ← ⟨SS_{q−1}⟩ ∪ {x_+} ;
        STOP ← faux
        répéter
            x_− ← argmax_{x_j ∈ ⟨SS_q⟩} {J(⟨SS_q⟩ − {x_j})} ;
            si J(⟨SS_q⟩ − x_−) > J(⟨SS_q⟩) alors
                ⟨SS_{q−1}⟩ ← ⟨SS_q⟩ − {x_−} ;
                q ← q − 1 ;
            sinon
                STOP ← vrai ;
            fin
        jusqu'à STOP ;
    fin
fin
```

Malgré l'évolution des méthodes séquentielles par la flexibilité du « retour en arrière », ces approches peuvent engendrer rapidement un sous-ensemble de variables sous-optimal, dû à la convergence du processus d'optimisation dans un optimum local. En effet, dans l'algorithme SFFS (tout comme pour SFBS), si juste après son élimination, la variable éliminée devient la nouvelle variable ajoutée, alors l'algorithme, dans la forme présentée ci-dessus, ne pourra pas améliorer le sous-ensemble de variables. Il resterait bloqué sur ce sous-ensemble. De plus, avec ce processus, il n'est plus possible de connaître avant son exécution le nombre de sous-ensembles qui sera évalué. Cependant, [Kudo and Sklansky, 2000] estiment ce nombre à $p^{2,40}$, en considérant

que l'algorithme ne se bloque pas et qu'il obtient dans le cas de SFFS, un sous-ensemble final de p variables ($q_{max} = p$). La figure 2.24 illustre les descriptions faites des méthodes de sélection séquentielle ascendante, en comparant leur progression dans l'espace des variables.



Note : Sur les trois graphiques, chaque intersection de la q-ième ligne correspond une combinaison de q variables sélectionnées.

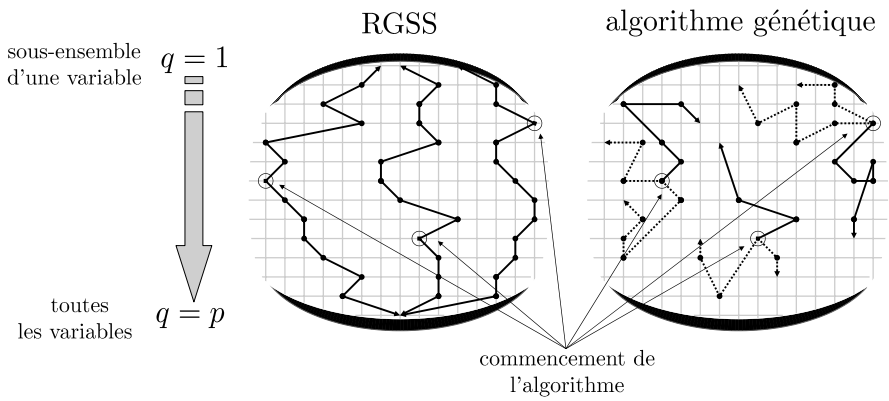
FIG. 2.24 – Comparaison de la progression dans l'espace des variables de méthodes de sélection séquentielle ascendante.

Avec l'intérêt des heuristiques, de nombreuses approches sont apparues, comme la stratégie bidirectionnelle (SBiS – *sequential bidirectional selection*). Fondée sur les deux méthodes principales de la sélection séquentielle, soit SFS et SBS, cette approche réalise simultanément ces deux processus et force la convergence des recherches ascendante et descendante au même sous-ensemble de variables. Afin de garantir cette convergence particulière, chaque variable sélectionnée par SFS ne sera jamais éliminée par SBS et chaque variable éliminée par SBS ne sera jamais sélectionnée par SFS. Cette approche permet de privilégier les variables sélectionnées ou conservées au début des deux progressions, permettant ainsi de réduire le risque de s'enfermer trop rapidement dans un optimum local.

Globalement, les méthodes heuristiques sont rapides, mais elles sont loin de garantir une solution optimale. Étant déterministes, ces méthodes ne pourront jamais s'extraire d'optimums locaux, même lors d'exécutions répétées de plusieurs processus. Une solution possible pour remédier à ce problème était de conjuguer à leurs processus une « dose » de hasard, comme sélectionner aléatoirement un sous-ensemble de variables de départ. Ainsi, ce type d'approches serait à mi-chemin entre les heuristiques précédentes et les approches non déterministes. Dans cette perspective, nous pouvons évoquer la méthode RGSS[6] (*random generation plus sequential selection*). En effet, RGSS est une méthode de sélection pouvant être qualifiée de bidirectionnelle, qui utilise un processus aléatoire pour déterminer le commencement de la recherche. Ainsi, une fois le sous-ensemble de variables de départ déterminé aléatoirement, RGSS réalise les deux processus SFS et SBS. Alors, à partir de ce sous-ensemble de r variables, les deux processus progressent jusqu'à aboutir à la sélection de $p - r$ variables et l'élimination de r variables, respectivement pour SFS et SBS. Après l'exécution de plusieurs séquences, soit après le tirage aléatoire de plusieurs sous-ensembles de variables de départ, le meilleur sous-ensemble est conservé. Le choix aléatoire rend l'algorithme bien évidemment non déterministe et lui permet d'échapper aux optimums locaux ; ce qui le rend relativement efficace. À l'image de la figure 2.24, la progression dans l'espace des variables de l'algorithme non déterministe est donnée à la figure 2.25.

Notons que si la valeur du nombre de variables à sélectionner (q) était connu, l'ensemble des algorithmes présentés arrêteraient leur progression une fois le sous-ensemble « rempli » de q variables.

[6] Compte tenu de sa forte relation avec les méthodes séquentielles, nous avons préféré la présenter avec les autres approches heuristiques, même si cette méthode est considérée comme non déterministe.



Note : Sur les trois graphiques, chaque intersection de la q-ième ligne correspond une combinaison de q variables sélectionnées.

FIG. 2.25 – Comparaison de la progression dans l'espace des variables de méthodes de sélection non déterministes.

Pour finir, [Kudo and Sklansky, 2000; Webb, 2002] présentent, chacun des généralisations des méthodes de sélection séquentielle qui ajoutent ou éliminent à chaque itération du processus, non plus une variable mais plusieurs variables simultanément.

[Kudo and Sklansky, 2000] proposent une étude comparative très poussée entre plusieurs méthodes de sélection : approches optimales, séquentielles et non déterministes, en l'occurrence des algorithmes génétiques. Au terme de leur étude, ils recommandent lors de la manipulation d'ensembles de données contenant plus de 100 variables, l'utilisation exclusive d'algorithmes génétiques. En effet, ils démontrent que pour $p = 100$ variables, les méthodes SFS[7] et SFFS[8] demandent respectivement d'évaluer approximativement 5 050 et 63 000 combinaisons de variables, contre 10 000 pour les algorithmes génétiques[9]. Et pour $p = 300$, les méthodes SFS et SFFS évaluent respectivement 45 150 et 880 000 combinaisons de variables, contre 30 000 pour les algorithmes génétiques. [Kudo and Sklansky, 2000] concluent leur étude en associant le choix du processus de sélection de variables en fonction du nombre de variables initiales. Ainsi, lorsque ce nombre est compris entre 50 et 100 des méthodes heuristiques peuvent être employées. Au-delà de 100 variables initiales, ils suggèrent l'utilisation de méthodes non déterministes. Enfin, rappelons que si l'on estime que le nombre de variables est beaucoup trop important, l'alternative la plus efficace est l'utilisation de méthodes de type de *filter*, telles que RELIEF.

2.4.3.4 Approches non déterministes : les algorithmes génétiques

Par opposition aux approches heuristiques, qui donnaient des résultats similaires à chaque exécution, des approches non déterministes, telles que les métaheuristiques, produisent des résultats différents à chaque exécution.

À l'image de [Kudo and Sklansky, 2000], des auteurs comme [Collette and Siarry, 2002; Dréo *et al.*, 2003] suggèrent l'utilisation de techniques métaheuristiques dans l'optimisation de problèmes combinatoires difficiles, comme la sélection de variables. Dérivées des heuristiques qui s'appuient sur la connaissance du problème pour trouver une solution, les métaheuristiques présentent l'avantage d'une certaine indépendance par rapport au problème à résoudre (*cf.* section 2.4.3.2). Cet atout rend les métaheuristiques adaptables à bon nombre de problèmes. On peut

[7] les auteurs estiment le nombre de combinaisons obtenu par la méthode SFS à $p \cdot (p+1)/2$
[8] rappel : les auteurs estiment le nombre de combinaisons à évaluer pour SFFS par $p^{2,40}$
[9] les auteurs fixent la taille de la population à $2 \cdot p$, évoluant durant 50 générations

citer, comme méthodes, le recuit simulé [Kirkpatrick *et al.*, 1983], la recherche tabou [Glover, 1989; Glover, 1990] ou encore les algorithmes génétiques [Holland, 1975; Goldberg, 1991][10].

Les algorithmes génétiques (AG) constituent les techniques les plus connues parmi les algorithmes évolutionnaires. Ces derniers impliquent la simulation et la modélisation par ordinateur du processus de l'évolution naturelle. Les algorithmes génétiques sont fondés sur les mécanismes de la sélection naturelle et de la génétique, s'inspirant de la théorie de l'évolution des espèces [Darwin, 1859]. Cette théorie se base sur l'idée que, sous l'influence de contraintes extérieures, les espèces se sont modifiées de manière autonome au travers de processus de reproduction et de mutation.

[Amat and Yahiaoui, 2002] montrent judicieusement comment et pourquoi l'aspect biologique de l'évolution a attiré les ingénieurs et les informaticiens. Les systèmes utilisés dans l'industrie ont été et sont, pour la plupart, des systèmes figés, C'est-à-dire que ces systèmes sont destinés à des applications très spécifiques. Par conséquent, ils sont exploités dans des conditions connues, censées ne pas évoluer. Cette approche de la conception requiert alors une connaissance parfaite de l'environnement dans lequel le système est utilisé, de manière à anticiper d'éventuelles évolutions des conditions d'utilisation. On remarque rapidement les limitations de ce type de systèmes, qui ne sont pas plus évolutifs qu'adaptables au moindre changement de conditions extérieures. Face à ce manque de souplesse, des travaux ont été réalisés, de manière à permettre aux systèmes d'évoluer spontanément et de manière autonome en fonction des changements de conditions d'utilisation. L'initiateur de ces travaux fut John Holland, en développant la notion de parallélisme [Holland, 1962] et en formalisant par la suite les algorithmes génétiques [Holland, 1975], comme nous les connaissons encore aujourd'hui.

Les algorithmes génétiques emploient la notion de « population d'individus », où chaque individu représente une solution de l'espace de recherche. Ainsi, l'idée fondamentale repose sur l'hypothèse qu'une population d'individus contient potentiellement la solution, ou une bonne solution au problème. C'est par la combinaison de ces individus au cours des générations que la solution pourra émerger. La réussite d'un tel algorithme est fondée sur l'hypothèse que l'action des opérateurs génétiques sur des individus sélectionnés produit statistiquement des individus de plus en plus proches de la solution recherchée. En d'autres termes, le processus stochastique sous-jacent, doit permettre aux populations successives de converger vers ce qui est souhaité, soit l'optimum global de la fonction d'évaluation.

L'engouement pour les algorithmes génétiques est certainement dû à leur simplicité algorithmique et à leur capacité d'adaptation à de nombreuses problématiques. En effet, étant des algorithmes d'exploration, ils cherchent simplement à optimiser une fonction d'évaluation, sans tenir compte d'une quelconque connaissance *a priori* du problème à traiter. Cette vision des algorithmes génétiques a néanmoins montré ses limites [Janikow, 1993], lorsque qu'il a été montré que l'apport d'une connaissance du problème dans l'exploitation des AG permettait d'améliorer considérablement les performances [Janikow, 1993; Venturini, 1994; Ratle and Sebag, 2000]. Néanmoins, cette simplification du contexte, dans lequel les AG s'exécutent, évite de considérer certaines hypothèses contraignantes du domaine à explorer, comme des hypothèses concernant la continuité, l'existence de dérivées, l'unimodalité. Cependant cette qualité est la cause de leur principal défaut : le coût calculatoire. En effet, en se basant uniquement sur la fonction d'évaluation, l'optimisation nécessite un grand nombre d'évaluations pour aboutir à une solution « optimale ».

[10] L'ouvrage de [Goldberg, 1991] a été traduit en français par V. Corruble dans une édition de 1994.

Une illustration de la progression dans l'espace des variables des algorithmes génétiques est donnée à la figure 2.25. Son principe de fonctionnement est présenté à la figure 2.26. Comme nous pouvons le constater, le processus demeure extrêmement simple. Ainsi, à chaque itération, l'algorithme est amené à créer une nouvelle population possédant le même nombre d'individus. Par conséquent, seuls certains individus survivent dans la nouvelle génération. Enfin, pour converger vers de meilleures combinaisons au cours des générations, des opérateurs de **sélection**, de **croisement** et de **mutation** sont appliqués aux individus des populations successives.

La **sélection** des individus pour la reproduction et pour le remplacement, est l'opération qui conditionne l'évolution dans la recherche de solutions. En effet, cette étape sélectionne les individus qui vont se reproduire, survivre et disparaître. Dans le but de converger vers de meilleures solutions, les individus sont sélectionnés suivant leur adaptation au problème. C'est pourquoi, il est très important de bien établir la fonction d'évaluation, appelée aussi **fonction d'adaptation** (*fitness function*) : c'est elle qui évalue la qualité des individus. Ainsi, les individus ayant une bonne qualité auront plus de chances de participer à la reproduction et verront leur patrimoine génétique conservé à la génération suivante.

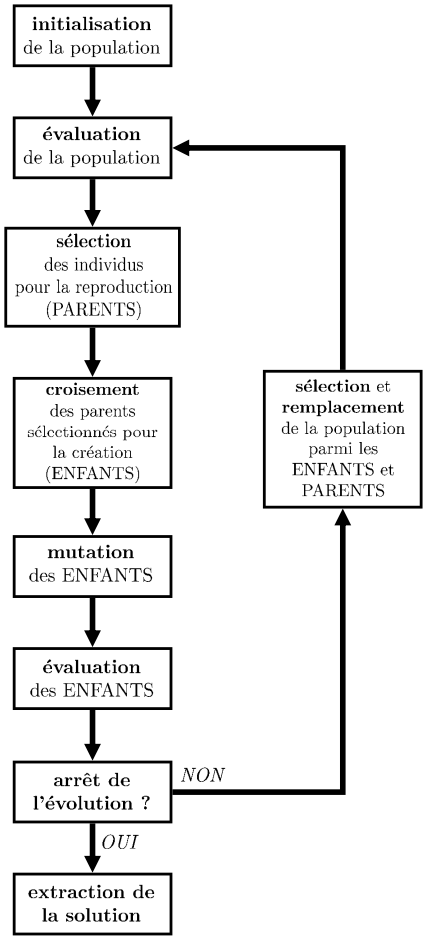


FIG. 2.26 – Organigramme du fonctionnement classique d'un algorithme génétique.

L'arrêt du processus est généralement défini après avoir atteint un nombre maximal d'itérations. L'utilisation d'un seuil de la performance à atteindre peut permettre également une fois ce seuil dépassé par un individu, de stopper l'évolution.

Sans rentrer pleinement dans les détails, nous allons présenter quelques étapes du processus des algorithmes génétiques. Pour plus d'informations, le lecteur pourra se référer aux ouvrages suivants [Holland, 1975; Goldberg, 1991; Dréo *et al.*, 2003].

Dans la présentation et la description de l'organigramme de fonctionnement (figure 2.26), nous avons abordé les opérateurs génétiques, comme étant prépondérants dans le bon fonctionnement du processus. On peut distinguer l'opérateur de sélection de l'opérateur de variations (croisement et mutation). Le choix des individus pour la reproduction et pour le remplacement est assuré par l'opérateur de croisement. [Goldberg, 1991; Dréo *et al.*, 2003] référencent plusieurs processus de sélection :

- la sélection proportionnelle, qui définit la probabilité de sélection d'un individu comme étant proportionnelle à sa performance ;

- la sélection par rang, qui organise la sélection suivant le principe proportionnel. Mais, elle calcule la probabilité de sélection par rapport au rang des individus et non plus sur leur

seule performance. Cette évolution de la sélection proportionnelle a pour but d'éviter de sélectionner en trop grand nombre l'individu ayant une performance bien supérieure aux autres individus de la population ;

- la sélection par tournoi, qui consiste à faire concourir deux à deux des individus choisis aléatoirement dans la population et à conserver pour la reproduction celui ayant la meilleure performance.

Avant de parler des opérateurs de variation, il est nécessaire d'apporter quelques précisions sur la représentation des individus. Traditionnellement, les algorithmes génétiques utilisent un codage binaire comme représentation des solutions, et particulièrement dans le cadre de la sélection de variables. Ainsi, un vecteur binaire de taille égale au nombre de variables initiales, caractérise chaque individu. Les éléments de ce vecteur, appelé aussi « chaîne de bits », prennent comme valeur « 0 » ou « 1 » ; les « 1 » désignent les variables sélectionnées. Prenons l'exemple, de l'évaluation du vecteur, ou plutôt, de l'individu suivant : 10010110. Le sous-ensemble correspondant à cet individu est donc composé des quatre variables suivantes : $\{x_1, x_4, x_6, x_7\}$ et son évaluation par le critère J se noterait $J(\{x_1, x_4, x_6, x_7\})$.

Les opérateurs de variation, le croisement et la mutation (figure 2.27) ont un rôle tout aussi important que la sélection, car ils produisent les nouveaux individus : les ENFANTS. En effet, ce sont ces opérateurs qui assurent l'exploration de la recherche, et donc la convergence, par la découverte de nouvelles régions et l'exploration plus minutieuse des régions considérées comme intéressantes. Notons que le croisement montré à la figure 2.27 est effectué sur un point, mais plusieurs peuvent tout aussi bien être considérés. Ces points sont bien évidemment choisis aléatoirement, tout comme le « bit » à muter, qui remplace un « 0 » par un « 1 », ou réciproquement un « 1 » par un « 0 ».

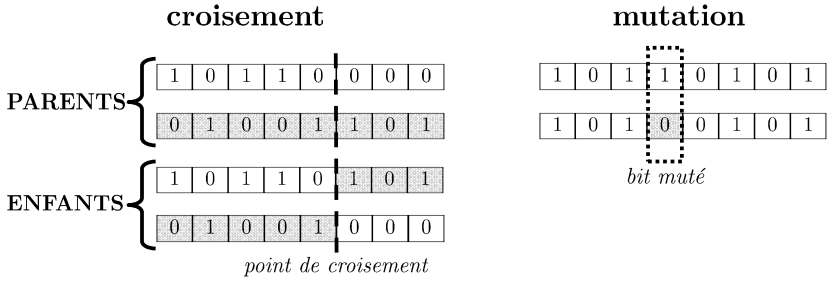


FIG. 2.27 – Opérateurs génétiques (croisement et mutation).

Le remplacement détermine quels individus parmi les PARENTS et les ENFANTS de la génération courante vont survivre à la génération suivante. Contrairement à la sélection pour la reproduction, les individus choisis pour le remplacement ne peuvent l'être qu'une seule fois et ceux qui ne sont pas choisis, disparaissent. Plusieurs techniques sont possibles, les plus courantes sont :

- le remplacement générationnel : les ENFANTS constituent entièrement la nouvelle population, indépendamment de la performance des PARENTS ;

- le remplacement élitiste : la nouvelle population conserve un certain nombre de PARENTS choisis parmi les meilleurs, et complète la population par des ENFANTS.

Le choix de la fonction d'évaluation conditionne l'efficacité de l'algorithme et par conséquent, influence fortement la résolution du problème. Dans certaines applications de sélection de variables

par des algorithmes génétiques, le critère utilisé (séparabilité des classes, taux de classification) est parfois pondéré par un coût. [Yang and Honavar, 1998] définissent le coût d'une solution candidate simplement par le nombre de variables sélectionnées. Ainsi, la fonction d'adaptation de l'algorithme génétique privilégierait ? à critère égal, une solution nécessitant moins de variables. Le coût peut être adapté individuellement à chaque variable, de manière à considérer dans la sélection les difficultés matérielles d'acquisition de certaines variables. Et dans ce cas, à critère sensiblement équivalent, pourquoi ne pas privilégier des variables plus simples ou moins coûteuses à obtenir.

Dans la sélection de variables, les algorithmes génétiques surpassent les méthodes heuristiques grâce à une plus grande diversité dans l'exploration de l'espace de recherche [Kudo and Sklansky, 2000]. Ceci fait des algorithmes génétiques, et plus généralement des approches non déterministes, des méthodes d'optimisation globale. Elles peuvent être adaptées lorsque l'espace de recherche devient trop vaste [Kudo and Sklansky, 2000]. En effet, dans ce cas, les méthodes « standards », certes plus rapides du point de vue du calcul, ne sont plus applicables du fait qu'elles se trouvent trop rapidement piégées dans des optimums locaux. En revanche, les algorithmes génétiques ont un coût calculatoire qui peut devenir très important. Cependant, basés sur une population, il peut être relativement aisé de paralléliser leurs exécutions : en déployant par exemple un individu par machine [Do, 2006].

Nous avons préalablement évoqué la facilité d'adaptation des algorithmes génétiques à de nombreux problèmes sans avoir à considérer une quelconque connaissance du problème. Cette particularité a fortement contribué à l'engouement de cette technique et son utilisation s'est retrouvée très répandue. Cependant, ces algorithmes ne garantissent pas la qualité des solutions obtenues, et sont aussi de gros consommateurs de temps de calcul. Aussi, des études, comme celles de [Janikow, 1993; Venturini, 1994; Ratle and Sebag, 2000], ont permis de montrer que ces algorithmes sont beaucoup plus efficaces si des connaissances du domaine sont intégrées dans leur processus (population initiale, croisements intelligents, mutations, etc.) : les connaissances du domaine introduites, par le biais d'un expert, permettent de guider, à travers un espace très vaste, la recherche des algorithmes dans des régions préférentielles. Aussi, même si la phase d'initialisation a théoriquement peu d'importance, car les AG sont censés converger vers un optimum global, il n'en demeure pas moins que cette phase influence énormément les résultats et la rapidité d'exécution ; c'est donc par le biais de l'initialisation qu'est souvent privilégiée l'intégration du maximum de connaissances du problème. D'autre part, pour améliorer leurs performances, il peut être intéressant de les associer avec d'autres méthodes d'optimisation, afin de combiner par exemple des algorithmes de recherche locale et globale [Lardeux *et al.*, 2006].

2.4.4 Conclusions

Avec le développement conséquent d'applications suscitant une optimisation combinatoire, de nombreux algorithmes ont vu le jour. Pour preuve, [Liu and Yu, 2002] proposent une présentation de 45 méthodes de sélection de variables. Celles-ci sont présentées en fonction du choix de la stratégie de recherche et du critère d'évaluation.

D'autre part, d'importants travaux ont été réalisés autour des méthodes de sélection, afin de les adapter aux différents types de problèmes pouvant être rencontrés. Ainsi, la communauté scientifique s'emploie désormais à « fouiller » des ensembles de données très spécifiques, caractérisés par un grand ou un petit nombre d'observations expliquées dans de grandes ou de faibles dimensions. Par exemple, [Jain and Zongker, 1997] se sont attachés à évaluer des méthodes de sélection dans des cas où peu d'observations sont disponibles. Aussi, on retrouve régulièrement

dans des conférences, des sessions spéciales où des compétitions sont organisées sur la sélection de variables et plus généralement sur l'extraction de connaissances (*data mining*). Nous pouvons citer, par exemple, le *World Congress on Computational Intelligence*[11] (WCCI) de 2006, qui avait, sous la direction de I. Guyon, organisé un remarquable concours. Annuellement, nous retrouvons une coupe organisée autour de la conférence *Knowledge Discovery and Data Mining*[12] (KDD). En 2006, la KDD Cup[13] s'était orientée sur le traitement de données médicales. Très active dans ce domaine, le lecteur pourra trouver de nombreuses informations auprès d'I. Guyon[14]. Toutes ces activités ont engendré de nouvelles méthodes, dont beaucoup sont fondées sur des combinaisons de méthodes existantes.

Compte tenu du nombre de combinaisons possibles dans le choix de l'approche à employer, entre le critère d'évaluation et la procédure de recherche, on aperçoit rapidement la difficulté pour choisir le processus de sélection. On a vu que le choix du critère d'évaluation pouvait dépendre du nombre de variables composant l'ensemble de départ. Plus ce nombre serait grand, plus on aurait tendance à choisir une méthode d'évaluation de type *filter*. L'utilisation d'outils de classification pour les méthodes de type *wrapper* rend la sélection plus performante (du point de vue du classifieur), mais la rend dépendante du classifieur utilisé lors de la sélection. Ce point donne, dans un contexte où l'outil de classification est susceptible de changer, un avantage indéniable aux méthodes *filters*.

Tout comme pour les critères d'évaluation, [Kudo and Sklansky, 2000] proposent des procédures de recherche à employer en fonction du nombre de variables disponibles au départ. Leurs suggestions sont fondées sur une étude expérimentale très poussée, qui a confirmé la nécessité d'utiliser exclusivement des méthodes non déterministes lorsque le nombre de variables devient relativement grand ($p > 100$). Lorsque p est compris entre 50 et 100 et compte tenu de l'objectif à atteindre, l'utilisation de méthodes heuristiques et non déterministes est préférable. En dessous de cette valeur, le choix de la méthode est dépendante de l'objectif à atteindre.

2.5 Résumé et discussions

L'étendue de ce chapitre montre la diversité des approches et des méthodes utilisables pour réduire la dimension d'un problème. Les méthodes de projection font une sorte de compression des données, et permettent alors d'obtenir une représentation de l'information initiale dans une dimension plus faible. Les méthodes de sélection, autorisent quant à elles, de réduire la dimension en sélectionnant uniquement les variables pertinentes. Le choix des méthodes à utiliser n'est pas aisé, il se fait généralement en fonction du nombre de variables disponibles et de l'ordre de grandeur de la réduction à obtenir. L'organigramme présenté à la figure 2.28 récapitule l'ensemble des méthodes présentées ; la liste n'est pas exhaustive, mais permet néanmoins de clarifier l'organisation des méthodes.

Comme nous l'avons évoqué dans ce chapitre, les méthodes de sélection ont un avantage important face aux méthodes de projection. En effet, au terme d'une projection, la nature ou la description physique des nouvelles composantes n'est pas clairement définie. Néanmoins, les méthodes de projection ont l'avantage d'être moins coûteuses en temps de calcul.

[11] http ://www.compsys.ia.ac.cn/wcci2006/index.html
[12] http ://www.sigkdd.org/kddcup/index.php
[13] http ://www.cs.unm.edu/kdd_cup_2006
[14] http ://www.clopinet.com/isabelle/Projects/modelselect/

FIG. 2.28 – Organigramme récapitulatif des approches et des méthodes de réduction de la dimension.

Avant d'achever ce chapitre, nous pouvons aborder une dernière approche, nommée **construction de caractéristiques** [Matheus, 1991; Wnek and Michalski, 1994]. Comme l'extraction de caractéristiques, cette approche est une variante de ce que certains auteurs nomment généralement transformation de caractéristiques [Liu and Motoda, 1998]. Contrairement à l'extraction, la construction de caractéristiques n'implique pas forcément une réduction de la dimension. En réalisant des opérations entre les variables initiales, de nouvelles variables peuvent être générées de manière à découvrir des informations manquantes. [Liu and Motoda, 1998] proposent un exemple considérant un problème à deux dimensions avec x_1 et x_2 définissant respectivement la longueur et la largeur. Ainsi, une troisième variable ($x_3 = x_1 \times x_2$), caractérisant l'aire, peut être découverte. Le problème peut donc passer à trois dimensions ou être réduit à une dimension caractérisée par la nouvelle variable x_3. Ce type d'approche est peu utilisé, car il oblige à connaître précisément les variables [Liu and Motoda, 1998].

L'objectif de ce chapitre peut être décrit brièvement comme la préparation des données et la sélection des entrées pour le modèle de classification. Cependant, il est difficile d'envisager la réalisation de ces tâches indépendamment de la tâche de classification. En effet, comme nous allons le voir dans le chapitre suivant (3), la validité du modèle, et donc celle des entrées choisies, se mesure principalement par les performances obtenues lors de la classification. Aussi, comme le montre la figure 1.3 (page 14), représentant le schéma général d'un système RdF, des connexions (montantes et descendantes) apparaissent entre les différentes tâches du processus. Les méthodes de sélection de type *wrapper* illustrent parfaitement ce fonctionnement, au sens où le processus de sélection des variables utilise les performances des algorithmes de classification pour optimiser la sélection.

Notons enfin, que dans la seconde partie de ce manuscrit, consacrée aux contributions, nous nous sommes attachés à employer et à comparer la plupart des méthodes de projection et de sélection de variables évoquées dans ce chapitre.

Chapitre 3

Évaluation et comparaison de modèles

3.1 Introduction

La résolution d'un problème de classification et plus généralement d'un problème de modélisation s'effectue en comparant des modèles afin de choisir le plus apte à résoudre le problème posé. L'évaluation des modèles est donc un préalable inévitable à la sélection. Elle est nécessaire pour connaître les performances d'un modèle et déterminer s'il est globalement significatif. Dès lors, deux objectifs se dégagent et seront abordés tout au long de ce chapitre : l'**évaluation** et la **comparaison** de modèles en vue de la sélection.

L'état de l'art proposé dans les chapitres précédents montre le nombre important d'approches, tant pour la classification que pour la sélection des variables d'entrée. Une recherche exhaustive d'un modèle optimal peut, en considérant rigoureusement le choix du modèle et de ses variables d'entrée, entraîner un nombre considérable de modèles. La sélection du modèle idéal peut être envisagée :

- en comparant différentes méthodes de classification (RNA, SVM), pour un même sous-ensemble de variables ;

- en comparant différentes méthodes de sélection de variables (SFS, SBS, AG), pour une même méthode de classification ;

- en comparant simultanément des méthodes de classification et des méthodes de sélection de variables.

On peut réduire la définition d'un modèle par les relations liant les variables d'entrée à celles de sortie. Ces relations, établies durant la phase d'apprentissage, définissent les règles de classification et doivent donc être déterminées avec pour objectif principal de réaliser la meilleure performance de classification pour une nouvelle observation : on parle alors de **généralisation**. Gallinari dans [Thiria *et al.*, 1997] définit la généralisation, comme la tâche accomplie par un modèle une fois son apprentissage effectué. Comme nous le montrerons à la section 3.3, l'estimation de la généralisation et globalement les mesures de performance d'un modèle sont dépendantes de l'échantillon sur lequel le problème est analysé. Ainsi dans ce chapitre, des méthodes, des techniques et des indices de mesures seront proposés afin d'évaluer précisément les performances des modèles de classification. Nous terminerons le chapitre en abordant le cas particulier d'analyses

médicales, où l'objectif principal du médecin est d'obtenir un verdict exact sur le diagnostic de la maladie d'un patient. Ainsi, compte tenu des nombreux examens possibles, il est pour lui utile de pouvoir comparer la performance de chaque examen et de chaque outil de diagnostic.

3.2 Mesures de la qualité d'un modèle

La qualité d'un modèle de classification est souvent définie par son taux de classification, indiquant sa capacité et ses performances de discrimination. Obtenue par le taux d'observations bien ou mal classées (erreur de classification), cette mesure peut être remplacée ou associée à d'autres indices plus pertinents, tels que l'aire sous la courbe de ROC que nous détaillerons à la section 3.4.2. Cependant, d'autres éléments moins quantitatifs, peuvent contribuer à rendre un modèle intéressant [Tufféry, 2007], comme :

- la **robustesse**, qui donne une information sur la performance de généralisation du modèle, donc sa sensibilité aux variations des observations. En d'autres termes, le modèle doit être le moins dépendant possible des observations d'apprentissage afin d'éviter le phénomène de surapprentissage. En outre, un modèle robuste est forcément dépendant des variables d'entrée qu'il observe. Par conséquent, la légitimité de la présence de ces variables doit être vérifiée et celles-ci doivent pouvoir être recueillies sans difficulté. Dans le cas contraire, le modèle doit pouvoir s'adapter à la présence de valeurs manquantes (*cf.* section 2.2.4). Selon [Tufféry, 2007], un modèle robuste doit pouvoir s'adapter également aux variables qui évolueraient dans le temps. Néanmoins, il précise légitimement que cette durée doit être raisonnable ;

- la **parcimonie**, déjà évoquée à la section 1.4.2.5, suggère de réaliser un modèle le plus simple possible. En effet, nous pouvons rappeler qu'avec des règles simples et explicites, le modèle améliore sa robustesse et sa capacité de généralisation. Aussi, la simplicité du modèle facilite sa compréhension et sa lisibilité, par exemple la lecture de la relation liant ses entrées à ses sorties ;

- le **coût calculatoire**, peut être un paramètre important dans l'évaluation d'un modèle, selon l'utilisation que l'on souhaite faire de ce dernier. En effet, si l'apprentissage du modèle doit être réalisé « *on-line*[1] » (temps réel), de manière à procéder à des ajustements en situation, il peut être souhaitable que l'apprentissage puisse être réalisé relativement rapidement. D'autre part, même en utilisation « *off-line* », plus l'apprentissage est rapide, plus le nombre de tests et d'ajustements peut être effectué afin d'affiner le modèle. Cette remarque qui, avait déjà été évoquée par [Bishop, 1995] et rappelée dans la section 2.4.2, permet de relier le coût calculatoire à la parcimonie.

Ces trois éléments montrent que la qualité d'un modèle est fortement liée à son pouvoir de **généralisation**. Plus sommairement, [Webb, 2002] décompose l'évaluation de la performance des modèles de classification en deux facteurs : le **pouvoir discriminant** de sa règle de classification et sa **fiabilité**.

[1] Nous mettons en garde le lecteur sur le terme *on-line*, où pour certains auteurs celui-ci signifie un apprentissage de type séquentiel.

3.3 Évaluation de la performance

Le choix d'un modèle requiert souvent une démarche empirique qui consiste à tester plusieurs solutions et à sélectionner la meilleure. Les performances des modèles, ou plutôt leur estimation, apparaissent naturellement comme critère de comparaison et deviennent alors le point sensible dans la démarche de sélection de modèles. Dans un premier temps, nous associons la mesure de performance à l'erreur du taux de classification.

Pour évaluer l'erreur d'un modèle, la première solution pourrait être l'utilisation des observations d'apprentissage pour estimer l'erreur de généralisation : les observations de test sont alors les mêmes que celles d'apprentissage. Nommée **resubstitution**, cette méthode est bien naturellement à proscrire. En effet, la simplicité de cette approche entraîne un biais important sur l'estimation de la généralisation, en produisant une estimation très optimiste de la probabilité d'erreur.

L'évaluation de modèles engendre une amélioration de ces derniers, due à l'élimination des mauvais modèles. Il est donc important d'avoir confiance dans les résultats des évaluations ; cela passe par l'utilisation de techniques appropriées et non biaisées. De nombreux outils statistiques sont disponibles, ils permettent d'estimer les capacités de généralisation et d'évaluer la confiance à donner aux estimations.

Avant de présenter les principales approches statistiques, nous allons revenir plus précisément sur la **généralisation**.

3.3.1 Facteurs influençant la généralisation

En introduction de ce chapitre, la généralisation avait été définie, comme la tâche accomplie par le modèle une fois l'apprentissage effectué [Thiria *et al.*, 1997]. [Dreyfus *et al.*, 2002] définissent l'erreur de généralisation par la probabilité du modèle à commettre une erreur de classification sur une nouvelle observation. La généralisation apparaît donc comme l'information principale ; malheureusement, elle est influencée par différents facteurs plus ou moins complexes : le problème étudié, l'échantillon de la population sur lequel l'analyse est effectuée, l'algorithme d'apprentissage et le modèle [Bishop, 1995]. Les deux derniers facteurs peuvent être plus spécifiques aux réseaux de neurones.

Rappelons que le modèle définit des règles liant ses variables d'entrée aux variables de sortie. Si des variables d'entrée ne sont pas pertinentes, alors le modèle risque d'avoir des paramètres superflus, impliquant une complexité inutile. Ainsi, l'allégement de la complexité du problème en réduisant sa dimensionnalité (*cf.* sections 2.3 et 2.4, extraction de caractéristiques et sélection de variables) peut diminuer la complexité du modèle. Précédemment, nous avons noté que l'apprentissage des modèles se fait sur un nombre fini d'observations disponibles. Ainsi, pour un problème donné, il peut exister un très grand nombre d'observations, défini à la figure 3.1 par la **population** : ensemble des couples « entrées-sorties » possibles. Sur cette population, toutes les observations ne sont pas disponibles, celles qui le sont servent à établir le modèle. Ces observations constituent alors l'**échantillon** : ensemble des couples « entrées-sorties » disponibles. Cependant, pour que l'analyse ait un sens et que les résultats soient exploitables, les échantillons d'observations utilisées pour définir et évaluer les règles, appelés respectivement groupe d'apprentissage et groupe de test, doivent être homogènes à la population.

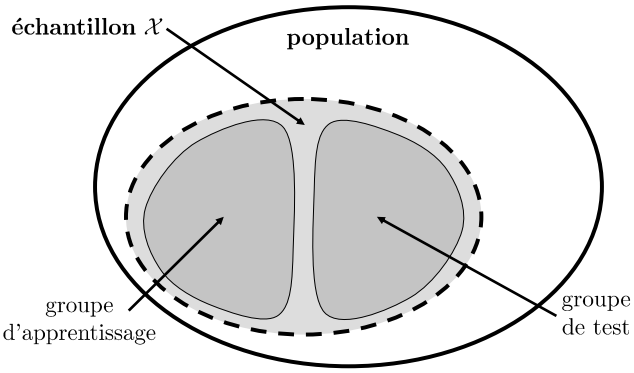


FIG. 3.1 – Illustration de l'analyse d'un problème [Denker *et al.*, 1987].

Ces remarques amènent à se poser un certain nombre de questions : pour un modèle donné, combien d'observations sont nécessaires afin d'obtenir une bonne généralisation ? Pour un modèle et un échantillon donnés, quel peut être l'écart entre la capacité de généralisation estimée sur l'échantillon et la capacité de généralisation réelle liée à la population (*cf.* section 3.3.3 sur les intervalles de confiance) ? Vapnik fut l'une des personnes ayant œuvré afin de répondre à ces questions. En effet, ses travaux [Vapnik, 1995; Vapnik, 1998], résumés remarquablement dans [Hastie *et al.*, 2001], ont permis de clarifier et d'expliquer l'apprentissage d'un point de vue statistique, en établissant des liens théoriques entre la complexité de l'échantillon et celle du modèle. Ces deux facteurs nuisent à l'évaluation de l'erreur de généralisation du modèle, on retrouvera ces deux facteurs respectivement dans les sections 3.3.1.1 et 3.3.1.2, en évoquant notamment le célèbre dilemme « biais-variance ».

3.3.1.1 Complexité de l'échantillon : approche *holdout* pour l'estimation de l'erreur de généralisation

L'alternative la plus classique à la méthode de resubstitution pour estimer la performance de généralisation d'un modèle est donnée par le processus illustré à la figure 3.2. Il consiste à diviser toutes les observations disponibles de l'échantillon en deux sous-ensembles :

- un **sous-ensemble d'apprentissage** qui détermine les paramètres du modèle (comme les paramètres de l'hyperplan ou les poids des réseaux de neurones) afin de fournir les règles de classification ;

- un **sous-ensemble de test**[2] qui évalue la généralisation. Les observations de ce sous-ensemble ne sont pas utilisées durant l'apprentissage.

Ainsi, les n observations qui composent l'ensemble $\mathcal{X}$ sont partagées aléatoirement[3] en deux sous-ensembles notées $\mathcal{X}_A$ (groupe d'apprentissage) et $\mathcal{X}_T$ (groupe de test), composés respectivement de $n_{_A}$ et $n_{_T}$ observations. Dès lors, pour m erreurs de classification parmi les $n_{_T}$ observations de test, la probabilité d'erreur est donc estimée par :

$$\hat{p_e} = \frac{m}{n_{_T}} \,. \tag{3.1}$$

[2]Dans la littérature le sous-ensemble de test est parfois appelé ou confondu avec le sous-ensemble de validation. Nous incitons le lecteur à prendre garde à la terminologie employée dans ce manuscrit, car nous distinguons le sous-ensemble de validation de celui de test.

[3]Il est courant de prendre deux tiers des observations pour le sous-ensemble d'apprentissage et un tier pour le sous-ensemble de test.

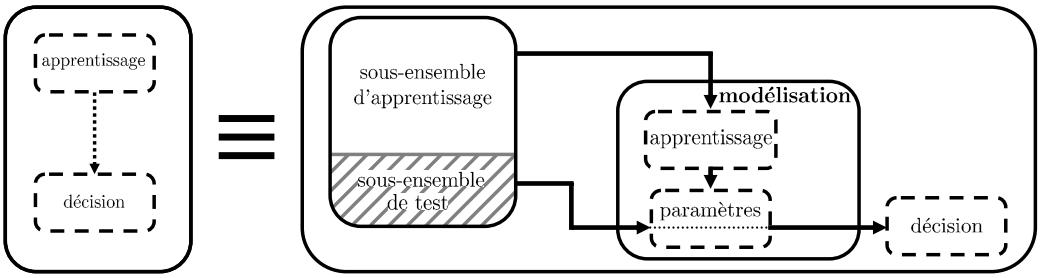


FIG. 3.2 – Partitionnement de l'ensemble des observations en deux sous-ensembles pour effectuer les tâches d'apprentissage et de test.

Cette méthode de découpage, nommée *holdout* dans la littérature anglo-saxonne, apporte rapidité et simplicité. Cependant, par son processus de division, cette méthode réduit le nombre d'observations dans chacun des deux sous-ensembles. Dans une configuration où peu d'observations sont disponibles, la séparation arbitraire en deux sous-ensembles peut engendrer des problèmes, tels qu'une sous-estimation ou une surestimation de l'erreur. En effet, il n'est pas illusoire de penser que l'appauvrissement d'observations dans le sous-ensemble d'apprentissage entraîne l'absence d'exemples particuliers qui, par conséquent, ne sont pas appris par le modèle durant la phase d'apprentissage. Si ces mêmes observations sont présentes dans le sous-ensemble de test, alors leur évaluation entraînera un biais important sur l'estimation des performances du modèle. Ces propos sont illustrés à la figure 3.3 [Tufféry, 2007], qui suggère qu'un petit échantillon d'apprentissage permet d'obtenir aisément une faible probabilité d'erreur en apprentissage, mais conduit à une forte probabilité d'erreur en test : le modèle ne peut pas généraliser ce qu'il n'a pas appris. En outre, l'augmentation de la taille du sous-ensemble d'apprentissage conduit à augmenter la probabilité d'erreur d'apprentissage et à réduire celle de test : le modèle a des difficultés à apprendre des exemples très particuliers noyés dans de nombreuses observations, évitant alors le surapprentissage. Dans ces conditions, le modèle s'avère être plus généralisable.

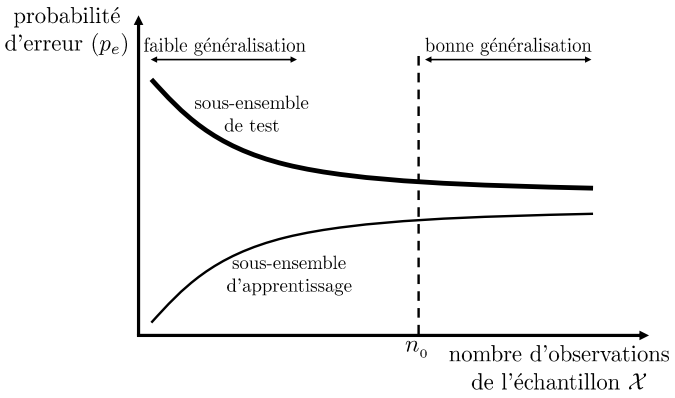


FIG. 3.3 – Influence du nombre d'observations de l'échantillon sur la probabilité d'erreur (p_e) de l'apprentissage et de test [Tufféry, 2007].

Si la convergence observée sur la figure 3.3 ne reflète pas forcément ce que nous pouvons observer dans la pratique (dans certains cas nous pouvons obtenir un taux d'erreur plus faible en test qu'en apprentissage), on peut cependant noter l'importance du nombre d'observations disponibles de l'échantillon $\mathcal{X}$. En effet, à partir d'un seuil noté n_0 sur la figure 3.3, le nombre d'observations peut devenir suffisant pour former les deux sous-ensembles et estimer avec un minimum de biais la capacité de généralisation du modèle. Notons, qu'il est difficile de chiffrer précisément la taille des sous-ensembles, car ils dépendent bien évidemment du problème à traiter.

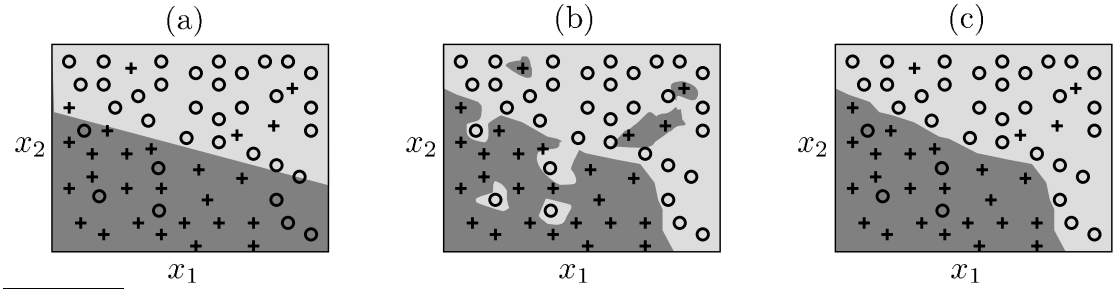
3.3.1.2 Complexité du modèle : dilemme biais-variance

Précédemment, nous avons introduit la notion de biais dans l'évaluation de l'erreur, ce biais est l'**erreur d'approximation** propre au modèle. À cette erreur, est associée l'**erreur d'estimation** due à la variance des prédictions du modèle ayant appris sur différents échantillons d'apprentissage. L'analyse de ces deux paramètres fait référence au compromis biais-variance [Geman *et al.*, 1992], ainsi deux cas extrêmes se dégagent en fonction de la complexité du modèle :

- un modèle peu complexe, possédant donc peu de paramètres, ne s'ajuste pas aux observations d'apprentissage et par conséquent, possède des erreurs de prédiction importantes sur ces mêmes observations, synonyme d'un **biais important**. Le modèle étant peu dépendant du sous-ensemble d'apprentissage, le changement des observations de ce sous-ensemble affecte peu les performances, ce qui engendre alors une **variance faible** dans les prédictions ;

- un modèle très complexe, possédant donc de nombreux paramètres, s'ajuste trop finement aux observations d'apprentissage et a donc peu d'erreurs de prédiction sur l'échantillon d'apprentissage : le **biais** est donc **faible**. Cependant, la forte dépendance du modèle aux observations d'apprentissage conduit lors d'un changement de ces observations, à des prédictions totalement différentes, introduisant alors une **variance élevée** dans les prédictions.

Ces deux cas font respectivement référence aux phénomènes de sous-apprentissage et de surapprentissage (voir figure 3.4) qui, comme nous l'avons dit précédemment, dépendent de la complexité du modèle et de l'échantillon. Suite à ces observations, une bonne estimation des performances se caractérise donc par un biais et une variance faible. Or ces deux objectifs sont contradictoires, et leur optimisation ne peut se faire que par un compromis [Thiria *et al.*, 1997; Cornuéjols and Miclet, 2002]. [Cornuéjols and Miclet, 2002] assimilent justement le compromis « biais-variance » au compromis « erreur d'approximation-erreur d'estimation ». Ainsi, pour extraire parmi les différents modèles de classification le plus apte à résoudre un problème, il est nécessaire de tenir compte de ces deux paramètres dans l'évaluation de l'erreur. Des méthodes statistiques permettent d'agir sur l'estimation, elles sont présentées à la section 3.3.2.



Note : (a) Sous-apprentissage. (b) Surapprentissage. (c) Compromis entre le sous-apprentissage et le surapprentissage.

FIG. 3.4 – Illustration des frontières de décision après un sous-apprentissage et un surapprentissage.

3.3.1.3 Surapprentissage

La qualité d'un modèle avait préalablement été définie en terme de robustesse et de parcimonie, qui permettent notamment d'améliorer les capacités de généralisation du modèle de classification. Nous avons noté à la section 3.3.1.1 lors de l'analyse de la figure 3.3, l'importance du nombre d'observations pour réaliser un modèle. En effet, un échantillon pourvu de peu d'observations risque d'entraîner un surapprentissage du modèle, avec le risque de nuire considérablement à la qualité de la généralisation.

Le phénomène du surapprentissage peut être expliqué lorsque le modèle s'ajuste trop finement aux observations d'apprentissage. Ce phénomène est généralement provoqué par un mauvais dimensionnement de la structure du modèle (comme par un nombre trop important de paramètres en regard du nombre d'observations d'apprentissage). Ainsi, la capacité du modèle à stocker beaucoup d'informations entraîne la structure du modèle dans une situation de surapprentissage, diminuant sa capacité de généralisation.

Ce phénomène affecte particulièrement les réseaux de neurones. En effet, à la section 1.4.2.5, nous avons présenté la technique de l'arrêt prématuré de l'apprentissage afin de neutraliser le surapprentissage. Cette technique sépare les observations réservées à l'apprentissage en deux sous-ensembles : un groupe pour l'apprentissage et un groupe pour la validation. Les observations de validation n'étant pas utilisées pour l'apprentissage permettent de vérifier périodiquement durant l'apprentissage, les capacités de généralisation du modèle. La figure 3.5 illustre la nouvelle répartition des observations, où l'on retrouve comme précédemment (voir figure 3.2) le sous-ensemble de test. Ce dernier utilisé après la phase d'apprentissage, estime la performance de généralisation indépendamment de l'apprentissage. Les observations n'étant pas infinies, deux tiers d'entre elles sont généralement réservées à l'apprentissage (sous-ensembles d'apprentissage et de validation).

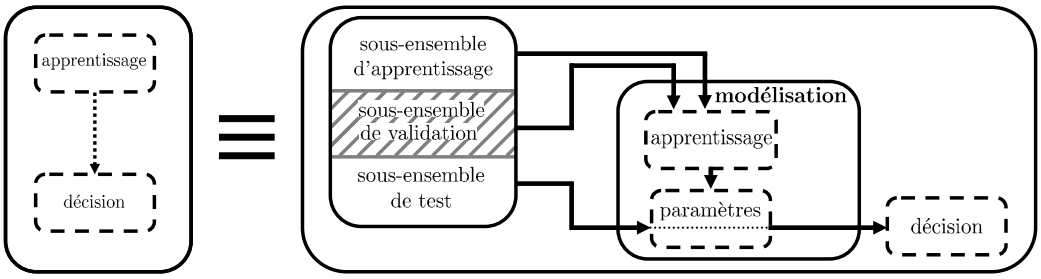


Fig. 3.5 – Partitionnement de l'ensemble des observations en trois sous-ensembles pour effectuer les tâches d'apprentissage, de test et de validation (pour gérer par exemple l'arrêt prématuré).

3.3.2 Méthodes d'estimation

Les remarques, apportées lors de l'analyse de la figure 3.3, amènent à envisager l'utilisation de la méthode *holdout* lorsque qu'un grand nombre d'observations est disponible. Cependant dans certaines applications, le nombre d'observations est souvent très limité et il est parfois impossible de pouvoir en obtenir autant que nécessaire. Ainsi, en présence d'échantillons dépourvus de nombreuses observations, d'autres techniques d'estimation peuvent et doivent être utilisées. Les méthodes citées ci-dessous ne sont pas exhaustives, elles sont néanmoins les plus employées. Elles se différencient par leurs procédures de découpage et de rééchantillonnage à partir des observations de l'échantillon initial $\mathcal{X}$.

3.3.2.1 Validation croisée

L'estimation par **validation croisée** (*cross-validation*) [Stone, 1974] est l'une des approches les plus connues pour évaluer les performances de généralisation en diminuant le biais de l'estimation. Comme pour l'approche *holdout*, cette approche partitionne l'ensemble des observations de $\mathcal{X}$ et estime l'erreur à partir des observations qui n'ont pas été utilisées pour l'apprentissage du modèle.

La technique des *K-fold cross-validation* est une méthode de validation croisée, où l'ensemble $\mathcal{X}$ est divisé aléatoirement en K sous-ensembles ($K > 1$) de taille approximativement égale. Le modèle apprend sur $K-1$ sous-ensembles et son erreur est évaluée sur le sous-ensemble n'ayant pas participé à l'apprentissage (sous-ensembles hachurés sur la figure 3.6). Répété K fois, ce processus permet d'offrir une meilleure estimation statistique, en diminuant le biais de l'erreur. L'erreur de généralisation est alors obtenue en effectuant la moyenne des K erreurs mesurées.

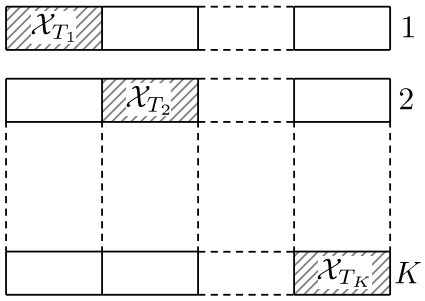


FIG. 3.6 – Illustration du partitionnement d'un ensemble d'observations pour une validation croisée.

La méthode *leave-one-out* [Lachenbruch and Mickey, 1968] est un cas particulier de la validation croisée et représente la limite statistique du nombre de sous-ensembles exploitables. En effet, avec cette méthode, l'échantillon initial, composé des n observations, est divisé en n sous-ensembles ($K = n$). Très coûteuse en calcul, cette méthode est très peu utilisée malgré ses grandes performances d'estimation.

3.3.2.2 Rééchantillonnage

Le *bootstrap* [Efron, 1979] est une technique de rééchantillonnage qui simule de nouveaux échantillons d'observations à partir de l'échantillon initial $\mathcal{X}$. Chaque échantillon est obtenu artificiellement par n tirages aléatoires avec remise des observations de l'échantillon initial. Ainsi, chaque observation a une probabilité de $1/n$ d'être intégrée dans le nouvel échantillon et par le mode de tirage avec remise, les observations qui seront les plus présentes dans le nouvel échantillon auront le plus de poids. Comme pour la validation croisée, cette opération est effectuée plusieurs fois, afin d'obtenir l'estimation de l'erreur de généralisation par la moyenne des résultats obtenus sur chacun des échantillons construits.

[Kohavi, 1995] a comparé ces deux types de méthodes (par découpage et par rééchantillonnage) et à l'issue de ses analyses expérimentales, il recommande l'utilisation de la validation croisée pour la sélection de modèles. L'auteur observe que cette méthode possède un meilleur compromis entre le biais et la variance que la méthode *bootstrap*, qui selon [Kohavi, 1995], bien que possédant une faible variance, le biais de la méthode de rééchantillonnage apparaît toutefois trop important.

Remarquons que ces deux approches sont efficaces mais coûteuses en temps et en calcul. En effet, elles répètent l'évaluation de chaque modèle sur plusieurs ensembles d'observations. C'est pourquoi lorsque l'on juge que le nombre d'observations est suffisamment important, on peut utiliser la méthode *holdout*, en considérant uniquement les deux ensembles d'observations : apprentissage et test [Goutte, 1997].

3.3.3 Intervalle de confiance

Les $n_{_T}$ observations de test utilisées pour l'évaluation de l'erreur d'un modèle proviennent généralement d'un échantillon de la population, et non pas de toute la population. Il est donc nécessaire de définir la confiance à accorder à l'estimation de la probabilité de l'erreur $\hat{p_e}$ (3.1), afin de définir le risque d'erreur de cette estimation. Ainsi, le risque définit l'écart entre la valeur estimée $\hat{p_e}$ sur l'échantillon de $n_{_T}$ observations et la valeur réelle p_e (inconnue) de la population totale. Cet écart peut être causé par la variabilité du phénomène à modéliser, ou encore par

l'imprécision du modèle. Un écart important, donc un risque d'erreur conséquent, suggérerait de porter attention à la validité à donner à la capacité de **prédiction** du modèle. Ainsi, à partir des deux paramètres ($\hat{p_e}$ et $n_{_T}$), nous pouvons déterminer l'intervalle confiance (IC) de la probabilité d'erreur p_e [Duda *et al.*, 2001] :

$$\text{IC à 0,95 de } p_e = \hat{p_e} \pm 1,96\sqrt{\frac{\hat{p_e}(1-\hat{p_e})}{n_{_T}}}\,. \tag{3.2}$$

Habituellement, l'IC est déterminé à 95% et désigne l'intervalle dans lequel la vraie valeur de l'erreur de la population a 95% de chance de s'y trouver. Cet intervalle donne une information importante dans l'interprétation des résultats, et permet de relativiser considérablement le poids à donner aux performances. Globalement, l'IC permet de préciser la confiance et la fiabilité de la classification d'une nouvelle observation par un modèle. À l'image de [Personnaz and Rivals, 2003], nous proposons quelques exemples :

- pour $n_{_T} = 50$, si $\hat{p_e} = 0\%$, alors IC à $0,95$ de p_e est $[0\% ; 8\%]$;
- pour $n_{_T} = 250$, si $\hat{p_e} = 0\%$, alors IC à $0,95$ de p_e est $[0\% ; 2\%]$;
- pour $n_{_T} = 50$, si $\hat{p_e} = 2\%$, alors IC à $0,95$ de p_e est $[0\% ; 11\%]$;
- pour $n_{_T} = 50$, si $\hat{p_e} = 5\%$, alors IC à $0,95$ de p_e est $[1\% ; 15\%]$;
- pour $n_{_T} = 250$, si $\hat{p_e} = 5\%$, alors IC à $0,95$ de p_e est $[2,5\% ; 9\%]$.

Précédemment, nous avons remarqué que l'estimation des performances d'un modèle passe par un partage de l'ensemble des observations (apprentissage/test) afin de tester l'efficacité de généralisation du modèle. Cependant, comme le nombre d'observations n'est pas infini, l'évaluation du modèle peut être biaisée si le nombre d'observations de test ($n_{_T}$) est trop faible. Ces remarques associées aux exemples de calcul de l'IC à $0,95$ montrent l'importance et la nécessité de posséder un nombre appréciable d'observations. Comme le montre la figure 3.7, plus le nombre d'observations augmente, plus l'intervalle de confiance diminue, rendant l'estimation de la probabilité d'erreur plus pertinente.

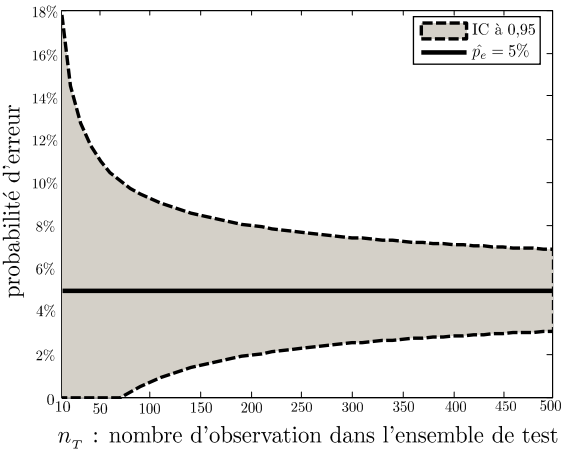


FIG. 3.7 – Influence du nombre d'observations de test sur l'intervalle de confiance à $0,95$ pour une probabilité d'erreur de 5%.

Pour améliorer la précision et donc réduire l'intervalle de confiance, le nombre d'observations doit être augmenté. Cependant, plutôt que de grossir arbitrairement l'échantillon de test, si toutefois cela est possible, il peut être plus judicieux de déterminer le nombre d'observations nécessaires afin de définir les bornes de l'intervalle de confiance[4]. Pour cela, il suffit de définir l'erreur acceptable de l'estimation (β) et d'adapter la relation (3.2) pour obtenir le nombre minimum d'observations :

$$n_{_T} \geq \frac{1,96^2 \cdot \hat{p_e}(1-\hat{p_e})}{\beta^2}\,. \tag{3.3}$$

[4]La question de la taille de l'échantillon est posée dans beaucoup de problèmes. [Tufféry, 2007] l'illustre dans les techniques de sondage, où dans ce domaine, on cherche à déterminer pour quelle taille de l'échantillon les résultats sont significatifs. Obtenir la taille minimale ou optimale permet dans cette application de réduire les coûts administratifs.

Par exemple, nous espérons obtenir à l'issue de l'analyse une erreur $\hat{p_e}$ de 15%, une limite inférieure et supérieure de IC à $0,95$ de 10% et 20%, donc $\beta = 5\%$. Pour atteindre ces objectifs, il faut que l'ensemble de test comporte au moins 196 observations. Dans le cas où, la valeur de $\hat{p_e}$ ne peut pas être définie, nous pouvons lui affecter la valeur la plus défavorable qui est de $0,5$ (erreur de 50%). Ainsi, en considérant le même écart ($\beta = 5\%$), le nombre d'observations nécessaires pour obtenir la précision voulue est maintenant de 384.

3.4 Mesures de performance d'un test diagnostique

Dans cette section, la présentation des mesures de performance est orientée dans le cadre de travaux médicaux. Ainsi, la notion de modèle employée jusqu'à présent sera associée aux tests cliniques.

3.4.1 Indices de performance

Jusqu'à présent dans ce manuscrit, les performances des modèles de classification étaient principalement obtenues par leur précision de classification sur le sous-ensemble de test : plus précisément, par la probabilité d'erreur de classification $\hat{p_e}$ donnant la proportion totale des observations mal classées. Cette unique information peut être insuffisante. En effet dans un problème à deux classes, appelé également problème binaire, où le résultat du test est soit **positif** soit **négatif**, l'indice du taux d'erreur ne distingue pas les erreurs de chacune des classes : les **faux positifs** et les **faux négatifs**. Cette mesure de performance considère alors les éléments de ces deux groupes comme des erreurs de même nature. Or lorsque les classes sont déséquilibrées, le taux d'erreur donne une fausse idée de la qualité du modèle de classification. Il n'est pas rare de trouver ce type de configuration dans certains problèmes, et notamment dans le domaine médical. En effet, dans le cadre de l'étude d'une pathologie rare, il est parfois difficile d'intégrer dans l'étude un grand nombre de patients atteints d'une pathologie particulière. Un exemple est donné en annexe C.

D'autre part dans certaines applications, il peut être nécessaire de faire la distinction entre ces deux erreurs de prédiction. Notamment dans l'évaluation de tests cliniques, où l'on cherche à déterminer les capacités d'un test à prédire une maladie. Prenons l'exemple d'une maladie très contagieuse, où le test clinique cherche à découvrir la présence de cette maladie chez des patients. Il est alors primordial que le test détecte chez tous les patients infectés la présence de la maladie, pour les traiter et éviter ainsi une épidémie. On peut alors imaginer que la fausse détection de la maladie chez des patients non infectés aurait une incidence moins importante qu'une non détection chez des patients infectés. Ainsi, ce test clinique doit avoir un taux de faux négatifs nul et un taux de faux positifs le plus faible possible. En outre, pour des pathologies qui nécessitent des soins très invasifs, le test clinique peut dans ce cas chercher à réduire au maximum le taux de faux positifs, pour éviter aux patients de subir des soins douloureux et inutiles. Ces remarques montrent la faiblesse de l'usage unique du taux d'erreur de prédiction en écartant des informations importantes. En effet, ces informations sont nécessaires aux cliniciens pour adapter les tests en fonction des exigences de prédiction.

La performance d'un test clinique peut alors être décrite en terme de fiabilité de diagnostic distinguant les patients pathologiques de ceux qui ne le sont pas. Ainsi, les résultats de tests médicaux sont utilisés afin de déterminer la probabilité d'un patient à souffrir d'une pathologie : on parle généralement de **prédiction positive** quand le résultat du test sur un patient indique une **présence de signes** pathologiques et une **prédiction négative** en **absence de signes**. La distinction entre positif et négatif dans un test n'étant pas toujours clairement définie (représentée par le chevauchement des deux densités de probabilité à la figure 3.8), il devient nécessaire

de déterminer un **seuil de décision** afin de lever l'incertitude lors de la prédiction du test, et de trancher la décision. La figure 3.8 illustre un exemple simple, où l'on cherche à définir la validité d'un test à prédire la présence d'une maladie chez un patient. Dans cet exemple, le test ou le modèle de classification est fondé sur la température corporelle (T) du patient (*cf.* exemple de la section 1.1.3). Ainsi, compte tenu de la distribution des deux groupes de patients de l'échantillon (sains et malades), le seuil de décision du modèle est donné pour $T = 37,5$ °C. Ce seuil permet alors de diviser l'échantillon de patients en quatre groupes :

- Vrai Positif (VP) représente le résultat **positif** du test en présence de signes pathologiques ;
- Vrai Négatif (VN) représente le résultat **négatif** du test en absence de signes pathologiques ;
- Faux Positif (FP) représente le résultat **positif** du test en absence de signes pathologiques ;
- Faux Négatif (FN) représente le résultat **négatif** du test en présence de signes pathologiques.

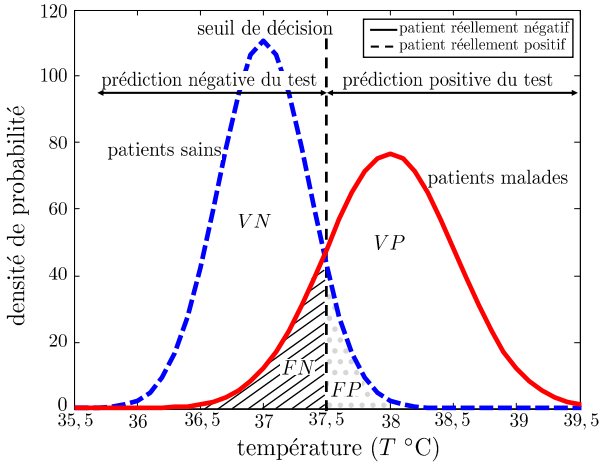


FIG. 3.8 – Illustration des quatre états possibles lors de la prédiction d'un modèle de classification binaire : vrai/faux positif et vrai/faux négatif.

Les quatre variables (VP, VN, FP et FN) peuvent se mettre sous la forme d'une matrice de confusion (tableau 3.1), où les éléments de la diagonale donnent les observations bien classées pour chacune des classes. Ainsi, à partir de cette matrice, plusieurs indices récapitulés dans le tableau 3.2 vont permettre de rendre l'interprétation des résultats du test diagnostique plus pertinente.

		observation réelle de l'état du patient	
		positive (malade)	négative (sain)
prédiction du test (modèle)	positive ($T \geq 37,5$°C)	VP	FP
	négative ($T < 37,5$°C)	FN	VN

TAB. 3.1 – Tableau/matrice de confusion issu de la prédiction d'un modèle de classification binaire : vrai/faux positif et vrai/faux négatif.

Parmi tous les indices disponibles à partir de la matrice de confusion, les plus connus sont la sensibilité et la spécificité. La **sensibilité** (S_e) exprime le pourcentage de patients détecter pathologique par le test parmi ceux présentant réellement la maladie, tandis que la **spécificité** (S_p) mesure le pourcentage de patients chez lesquels la pathologie n'est pas identifiée parmi ceux qui n'ont pas la maladie. La qualité du test dépend donc du meilleur compromis entre ceux deux

critères. Ce compromis peut être évalué par l'indice de ROC *(Receiver Operating Characteristic)*, qu'on cherche à minimiser afin d'obtenir conjointement les meilleures sensibilité et spécificité, réduisant par conséquent, les faux négatifs et les faux positifs. Cet indice est donné par la relation suivante :

$$ROC = \sqrt{(1 - spécificité)^2 + (1 - sensibilité)^2}\,. \tag{3.4}$$

L'augmentation d'un des deux paramètres occasionne une réduction de l'autre, c'est pourquoi dans l'évaluation d'un test, il est nécessaire de ne jamais considérer la sensibilité sans la spécificité et réciproquement.

Cependant, une question subsiste : si le résultat de l'examen d'un patient s'avère positif, donc pathologique, quelle est la probabilité que ce patient ait réellement la maladie ? De même, si le résultat d'un patient est négatif, quelle est la probabilité qu'il n'ait pas la maladie ? Pour le clinicien, ces deux questions, ou plutôt, leurs réponses sont fondamentales. En effet, elles permettent de connaître la fiabilité du test à mesurer d'une part, la probabilité de la présence de la maladie chez les patients qui ont un signe et d'autre part, la probabilité de l'absence de la maladie chez des patients qui n'ont pas de signe. Ces deux probabilités sont obtenues respectivement, par la **valeur prédictive positive** (VPP) et par la **valeur prédictive négative** (VPN). Cependant, comme montré en annexe C, ces deux indices sont fortement influencés par la prévalence de la maladie[5], rendant leur utilisation efficace si la prévalence de l'échantillon de patients analysé est identique à celle de la population.

Afin d'améliorer l'interprétation des résultats des tests et des modèles, nous pouvons utiliser les **rapports de vraisemblance**. Le **rapport de vraisemblance positif** (RV^+) exprime dans quelle proportion la présence de la maladie est vraisemblable chez un patient ayant obtenu un résultat positif au test. Par exemple, si la valeur de $RV^+ = 5$ alors le test sera cinq fois plus souvent positif chez les patients malades que chez les patients qui n'ont pas la maladie. Ainsi, l'outil de diagnostic est d'autant plus informatif que le RV^+ tend vers l'infini. Le **rapport de vraisemblance négatif** (RV^-) exprime la vraisemblance que le patient ne soit pas malade quand le test est négatif. De même, si $RV^- = 0,5$ alors le test sera deux fois moins souvent négatif chez les patients malades que chez ceux qui ne le sont pas. Par conséquent, plus ce nombre tend vers 0, plus le test est informatif. Les rapports de vraisemblance se déterminent comme suit :

$$RV^+ = \frac{S_e}{(1 - S_p)}\,, \tag{3.5}$$

$$RV^- = \frac{(1 - S_e)}{S_p}\,. \tag{3.6}$$

L'association de ces deux rapports permet de juger de l'utilité d'un test ou d'un modèle comme diagnostic. Ainsi, le test serait d'autant plus utile cliniquement s'il possédait une grande valeur de RV^+ et une petite valeur RV^-. Contrairement aux valeurs prédictives, les rapports de vraisemblance sont indépendants de la prévalence de la maladie dans la population. Ces rapports sont comme pour le taux de classification considérés comme des indices globaux ou des indices de synthèse.

[5]La prévalence de la maladie dans un échantillon donne le rapport du nombre de patients malades sur l'ensemble des patients de l'échantillon.

indice	**description**	**relation**
sensibilité (S_e)	probabilité de présence d'un signe chez les patients malades	$S_e = \frac{VP}{VP+FN}$
spécificité (S_p)	probabilité d'absence d'un signe chez les patients sains	$S_p = \frac{VN}{VN+FP}$
valeur prédictive positive (VPP)	probabilité de présence de maladie quand le signe est présent	$VPP = \frac{VP}{VP+FP}$
valeur prédictive négative (VPN)	probabilité d'absence de maladie quand le signe est absent	$VPN = \frac{VN}{VN+FN}$
taux de bonne classification (p_p)	probabilité de bonne classification	$p_p = \frac{VP+VN}{VP+VN+FP+FN}$
taux d'erreur de classification (p_e)	probabilité de fausse classification	$p_e = 1 - p_p$

TAB. 3.2 – Descriptions des indices d'évaluation obtenus à partir de la matrice de confusion (voir tableau 3.1).

Les indices, décrits dans le tableau 3.2, sont toujours obtenus à partir d'un échantillon prélevé sur la population que l'on souhaite étudier. Ainsi, pour évaluer la confiance à donner à ces indices, nous pouvons calculer les interfaces de confiance. Basé sur la relation (3.2), nous donnons ci-dessous l'intervalle de confiance de la sensibilité et de la spécificité :

$$\text{IC à 0,95 de } S_e = \hat{S}_e \pm 1,96\sqrt{\frac{\hat{S}_e(1-\hat{S}_e)}{VP+FN}}\,, \tag{3.7}$$

$$\text{IC à 0,95 de } S_p = \hat{S}_p \pm 1,96\sqrt{\frac{\hat{S}_p(1-\hat{S}_p)}{FP+VN}}\,. \tag{3.8}$$

Il est toujours souhaitable que ces intervalles soient faibles, particulièrement les seuils inférieurs qui définissent pour un risque choisi[6], la sensibilité et la spécificité minimales du modèle. Aussi, comme l'avait introduit précédemment la relation (3.3), en fonction de la précision voulue, on peut déterminer le nombre minimum de patients positifs au test nécessaire à notre étude par :

$$VP + FN \geq \frac{1,96^2 \cdot \hat{S}_e(1-\hat{S}_e)}{\beta^2}\,, \tag{3.9}$$

et le nombre minimum de patients négatifs par :

$$VN + FP \geq \frac{1,96^2 \cdot \hat{S}_p(1-\hat{S}_p)}{\beta^2}\,. \tag{3.10}$$

Par exemple, pour une sensibilité désirée de $\hat{S}_e = 90\%$, on souhaite que cette valeur ne soit pas inférieure au seuil de 80% : définissant la borne inférieure de l'IC à $0,95$, d'où $\beta = 10\%$. Ainsi, avec la relation (3.9), les objectifs fixés sont atteints si le nombre minimum de patients pathologiques de l'échantillon est de 35. Dès lors, en considérant que la prévalence de la maladie dans la population est de 20%, le nombre total de patients participant à l'évaluation prospective du test clinique doit être de 175. Dans ces conditions nous obtenons la précision voulue.

[6] On avait préalablement adopté un risque à 5%.

3.4.2 Courbes de ROC

Nous proposons la courbe de ROC [Metz, 1978; Fawcett, 2005] comme dernière mesure pour évaluer des tests diagnostiques. Cette mesure est fondée sur la théorie statistique de la décision. Elle a été développée initialement pour la détection de signaux associés aux radars durant la seconde guerre mondiale, où les opérateurs radars devaient distinguer les bateaux ennemis des bateaux alliés [Daigle, 2002]. Dans les années 1970, les courbes de ROC sont apparues en médecine afin d'améliorer la prise de décision en imagerie médicale [Hanley and McNeil, 1982]. Depuis quelques années, ces courbes sont présentes dans de nombreux domaines, notamment dans la recherche biomédicale où elles caractérisent les capacités (et les limitations) d'un test à prédire une maladie chez un patient.

Ces courbes analysent les variations de la sensibilité et de la spécificité, afin de visualiser la performance globale du test. Elles sont obtenues en traçant dans un plan la « *sensibilité* » en fonction de « $1 - spécificité$ », en faisant varier le seuil de décision du modèle (voir figure 3.8, page 109). Étant fondée sur les calculs de la sensibilité et de la spécificité, la construction de ces courbes a l'avantage d'être moins sensible au déséquilibre entre les classes, donc de la prévalence de la maladie. Un exemple détaillant la construction de la courbe de ROC est donné en annexe C.

À partir de la courbe de ROC, l'indice permettant d'évaluer numériquement la courbe est l'**aire sous la courbe** de ROC (*AUC*, *Area Under the Curve*). Cet indice peut être analysé comme la probabilité du modèle à prédire correctement la pathologie chez un patient. Deux cas particuliers sont observés sur la figure 3.9, ainsi :

- une aire AUC de $0,5$ indique que le test est proche du hasard et il n'apporte rien au diagnostic : le signe informant de la présence de la maladie est alors indépendant de la pathologie. En d'autres termes, le modèle ne différencie pas les deux classes, cela peut être observé par un recouvrement total des deux densités de probabilité, avec $VP = FP$ (voir figure 3.8) ;

- une aire égale à 1 correspond au meilleur résultat, avec une sensibilité et une spécificité de 100%, dans ce cas les deux classes sont totalement séparées.

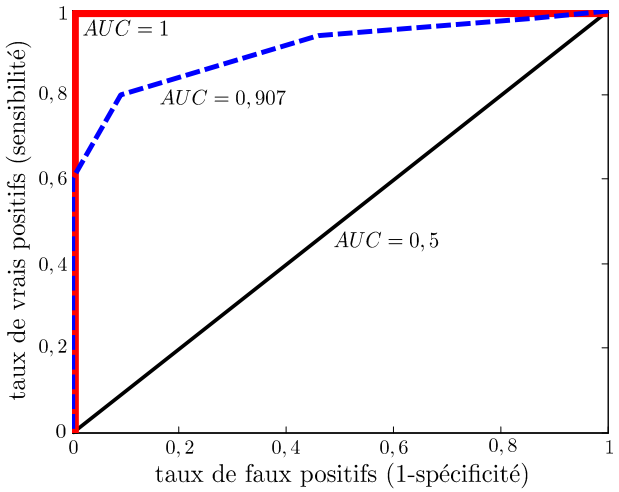


FIG. 3.9 – Exemple de courbes de ROC.

L'analyse des courbes de ROC était initialement prévue pour traiter des problèmes binaires. Cependant, des travaux comme ceux de [Srinivasan, 1999; Landgrebe and Duin, 2007; Landgrebe

and Duin, 2008] ont proposé des extensions aux problèmes multi-classes. Néanmoins, l'utilisation de ces extensions est souvent peu intuitive et leur calcul est généralement coûteux, contraignant la plupart des utilisateurs à employer le taux de classification pour évaluer les performances des problèmes multi-classes.

3.5 Comparaisons de modèles par analyse des courbes de ROC

Nous avons vu qu'il existe de nombreuses approches pour classer les observations et sélectionner les variables, qui génèrent, par conséquent, de multiples modèles. Ainsi, la démarche empirique pour choisir le modèle consiste à tester sur le problème un grand nombre de solutions afin de sélectionner le modèle le plus approprié. La comparaison passe donc par l'estimation des performances de chaque modèle.

Nous avons longuement discuté des méthodes et des indices mesurant la performance d'un modèle, comme la probabilité d'erreur, la sensibilité, la spécificité et l'aire sous la courbe de ROC. Cependant, l'utilisation simultanée de nombreux indices peut nuire à l'interprétation de la performance, où noyée dans les chiffres, l'efficacité du modèle n'apparaît plus forcément très clairement. C'est pourquoi, il est souvent préférable d'utiliser un indice donnant une indication plus globale sur la performance, tel que la courbe de ROC.

La courbe de ROC est un outil d'évaluation et de comparaison de modèles efficace et robuste, avec l'intérêt d'établir visuellement la pertinence d'un modèle. La superposition des courbes permet ainsi d'évaluer rapidement et facilement le meilleur modèle. Dans l'exemple de la figure 3.10, on peut observer que les courbes correspondant aux modèles $\mathcal{M}_1$ et $\mathcal{M}_3$ sont toujours au dessus de celle du modèle $\mathcal{M}_2$. Ces deux modèles seront toujours meilleurs quelle que soit la situation : le modèle $\mathcal{M}_2$ peut alors être éliminé. Les courbes restantes ($\mathcal{M}_1$ et $\mathcal{M}_3$) forment ce que l'on peut appeler l'enveloppe convexe, indiquant la limite supérieure à la résolution du problème par les modèles étudiés. Ainsi, selon les situations et les objectifs, comme privilégier la sensibilité à la spécificité, les modèles $\mathcal{M}_1$ ou $\mathcal{M}_3$ peuvent être utilisés conjointement.

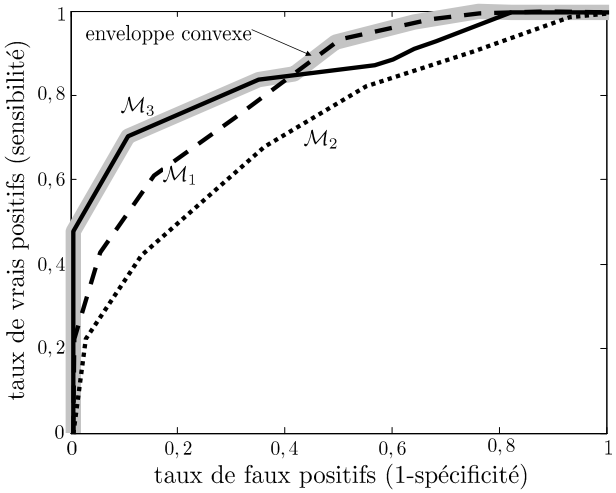


FIG. 3.10 – Comparaison des courbes de ROC.

Dans la suite logique de l'analyse des courbes de ROC, le calcul des aires sous les courbes permet d'évaluer numériquement et plus précisément les modèles. [Hanley and McNeil, 1983; McNeil and Hanley, 1984; Delong *et al.*, 1988] proposent une comparaison statistique des aires sous les courbes provenant de tests effectués sur un même échantillon de patients. En effet, il peut être intéressant de connaître si une modification du modèle permet d'améliorer significativement

la prédiction d'une pathologie. En outre, [Hanley and McNeil, 1982; McNeil and Hanley, 1984] ont proposé une autre comparaison statistique, s'appliquant lorsque l'on cherche à comparer pour une même pathologie la performance de tests effectués sur des échantillons de patients différents ; cette comparaison doit considérer le caractère indépendant et non corrélé des résultats des tests. En effet, la variation dans la distribution des patients entre les études n'est pas contrôlée et peut biaiser l'interprétation de la comparaison.

3.6 Conclusions

Le choix de la méthode d'évaluation s'avère être primordial dans le diagnostic de la maladie chez un patient. L'appréciation de la performance d'un test ou d'un modèle peut se faire par de nombreux indices. Notons qu'il n'y a pas d'indices meilleurs que d'autres, ils mesurent des informations différentes. Cependant, il est peut être préférable d'utiliser un indice permettant de synthétiser l'information de la performance, afin de ne pas noyer le lecteur et soi-même, dans une masse de chiffres et d'indices. C'est pourquoi, les courbes de ROC sont souvent privilégiées pour fournir l'information de la pertinence d'un test. Cette méthode possède de nombreux avantages, comme une indépendance par rapport à la distribution des classes, une visualisation efficace et une estimation globale de la performance. Par sa visualisation expressive, cet indice permet alors aux médecins de comparer efficacement la performance de chaque examen et de chaque outil de diagnostic.

Dans ce chapitre, nous avons pu noter l'importance du nombre d'observations à utiliser pour que l'estimation des performances soit précise et pertinente. Ce paramètre est donc primordial, tout comme cela avait été noté dans la construction des modèle de classification dans le chapitre 1. Cependant, dans le cadre d'études médicales, le nombre de patients n'est pas extensible, ce qui nécessite alors de traiter les données par des méthodes particulières, faisant appel notamment à des techniques de validation croisée et de rééchantillonnage. Aussi, dans le cas de l'étude de pathologie rare, le nombre de patients pathologiques est souvent beaucoup plus faibles que le nombre de patients sains, dès lors, la prévalence de la maladie doit être prise en compte. Dans ce chapitre nous avions noté que les indices de sensibilité et de spécificité ne sont pas touchés par le déséquilibre, comme cela est montré en annexe C. Cependant, [Holt, 2005][7] a démontré que la prévalence pouvait également intervenir dans la lecture de la sensibilité et la spécificité, nous suggérant alors de la considérer comme un paramètre important.

[7] L'article de [Holt, 2005] émet plusieurs critiques sur la pertinence de l'interprétation des résultats publiés par [Bhatikar *et al.*, 2005].

Deuxième partie

Contributions

Chapitre 4

Problématique étudiée : la prédiction de la syncope

4.1 Introduction

La syncope est un terme médical désignant l'évanouissement. [Kapoor, 2000] la caractérise par une perte subite et brève de connaissance et du tonus postural, suivie par un retour spontané à un état de conscience normal. Une apparition isolée d'une syncope ne constitue pas nécessairement un problème, mais peut le devenir lorsque des épisodes d'évanouissement sont répétés.

De nombreuses études ont montré que la syncope pouvait être considérée comme une pathologie fréquente. Les travaux de [Savage *et al.*, 1985] ont étudié l'apparition de la syncope sur un échantillon impressionnant de 5 209 patients. Leurs travaux, provenant de la célèbre étude *The Framingham Study*[1], ont ainsi pu faire apparaître un taux d'apparition de la syncope supérieur à 3%, avec une reproduction récurrente des symptômes dans 30% des cas. Plus récemment, dans les années quatre-vingt-dix, plusieurs études comme celles de [Manolis *et al.*, 1990; Kapoor, 1992; Kapoor, 1995], ont montré que 3% des visites aux salles d'urgence et 6% des hospitalisations étaient directement liées aux symptômes de la syncope ; sachant que la majorité des patients sujets à l'apparition des symptômes ne consulte pas de médecin [Linzer *et al.*, 1997]. [Cottier, 2002] note ainsi que, près d'un tiers des jeunes adultes indiquent avoir eu au moins une perte de connaissance de brève durée.

Les travaux de [Linzer *et al.*, 1997], en rassemblant cinq études en milieu hospitalier entre 1984 et 1990, ont permis de mettre en évidence d'autres observations relatives à la syncope. Ainsi, en considérant 1 002 patients, ils remarquent également que malgré un bilan complet, l'apparition de la syncope est expliquée dans à peine 65% des cas, alors que de nombreux troubles pouvant entraîner l'apparition des symptômes sont connus. Nous les récapitulons à la figure 4.1 et dans le tableau 4.1. Aussi, comme le souligne [Linzer *et al.*, 1997] (voir tableau 4.1), même si la majorité des causes sont bénignes, les syncopes d'origine cardiaque peuvent entraîner de graves traumatismes. En effet, nous verrons par la suite que le taux de mortalité des patients ayant des syncopes récurrentes d'origines cardiaques est très important [Kapoor *et al.*, 1987]. Notons que d'autres travaux de [Getchell *et al.*, 1999], basés sur des études cliniques entre 1992 et 1994, ont analysé dans le détail l'étiologie[2] de la syncope sur un échantillon de 1 516 patients. En plus d'être très fourni, cet échantillon a la particularité de comporter exclusivement des patients hospitalisés

[1] En 1948, un projet de recherche très ambitieux sur la santé a été lancé aux États-Unis, afin d'étudier les facteurs liés au développement des maladies cardiovasculaires. Appelée *The Framingham Study*, cette étude a été conçue comme recherche longitudinale et a pris en compte pas moins de 5 209 hommes et femmes, âgés de 30 à 62 ans.

[2] En médecine, l'étiologie signifie l'étude des causes et des facteurs d'une pathologie.

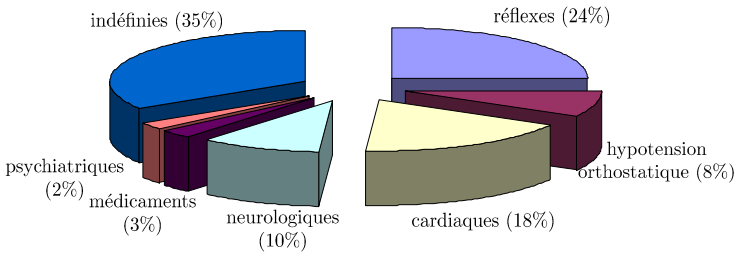


FIG. 4.1 – Causes de la syncope [Linzer *et al.*, 1997].

âgés (73 ans de moyenne et 13, 4 ans d'écart type). [Getchell *et al.*, 1999] ont globalement obtenu une répartition des causes équivalente à celle de [Linzer *et al.*, 1997], avec tout de même un taux de syncope inexpliquée dépassant les 41%.

type ou cause de syncope	**sévérité**	% (écart)
réflexes		
vaso-vagale[3](angoisse, chaleur, douleur, nausée)	bénigne	18 (8-37)
de situation (toux, déglutition)	bénigne	5 (1-8)
autres (sinus carotidien, névralgie)	bénigne	1 (0-4)
hypotension orthostatique[4]	bénigne	8 (4-10)
médicaments	bénigne à sévère	3 (1-7)
psychiatrique	bénigne	2 (1-7)
neurologique (migraine, épilepsie)	modérée	10 (3-32)
cardiaque		
maladie cardiaque organique	sévère	4 (1-8)
arythmies	sévère	14 (4-38)
bradyarythmies[5]	modéré	
tachyarythmies[6]	sévère	
indéfinies	bénigne à sévère	35 (13-41)

TAB. 4.1 – Description des causes de la syncope et de leur degré de sévérité [Linzer *et al.*, 1997].

En répartissant les causes de l'apparition des symptômes de la syncope en trois catégories, causes d'origine cardiaque, non cardiaque et indéfinie, [Kapoor *et al.*, 1987] ont étudié le taux de récidive durant une période de 30 mois, sur un échantillon de 433 patients. Ainsi durant cette période, 146 patients (34%) ont subi des syncopes à répétition. [Kapoor *et al.*, 1987] ont ainsi pu observer que lorsque l'origine de la syncope était cardiaque ou non cardiaque, il y avait respectivement une récidive de 31% et de 36%. Tandis que lorsque la cause de la syncope n'était pas déterminée, la récidive était alors de l'ordre de 43%.

Bien que la syncope est un événement occasionnel dans 70% des cas [Savage *et al.*, 1985], la syncope récurrente inexpliquée représente un problème d'importance. En effet, même si son taux de mortalité est relativement faible, 6% contre 30% et 12%, respectivement lorsque l'origine de la syncope est cardiaque et non cardiaque [Kapoor *et al.*, 1987], la syncope inexpliquée peut

[3]Appelée aussi syncope cardioneurogène, la syncope vaso-vagale apparaît notamment en cas de chute de la pression artérielle et de la fréquence cardiaque, provoquées par des dysfonctionnements du système nerveux qui contrôle ces paramètres cardiovasculaires.

[4]L'hypotension orthostatique est définie par une chute de la pression artérielle lors du passage en position debout et se traduit par une sensation de malaise après un lever brutal ou un alitement prolongé. La majorité des cas est lié au traitement de l'hypotension.

[5]Trouble du rythme cardiaque caractérisé par l'irrégularité et la lenteur des contractions.

[6]Trouble du rythme cardiaque caractérisé par l'irrégularité et la rapidité des contractions.

entraîner une réelle incidence sur la qualité de vie des patients. En effet, elle cause généralement peu de traumatismes, cependant les préoccupations et les angoisses liées aux risques traumatiques et à la crainte de récidive peuvent influer sur la qualité de vie des patients. Aussi, les syncopes inexpliquées sont les plus handicapantes, car l'absence de diagnostic empêche de prescrire un traitement approprié pour isoler l'apparition des symptômes.

4.2 Investigations et démarches diagnostiques

Une perte de connaissance n'est pas forcément provoquée par une syncope, d'autres pathologies peuvent en être la cause, telles que l'hypoglycémie, une crise épileptique, la cataplexie[7], ou encore une crise de panique [Cottier, 2002; Antonini-Revaz *et al.*, 2004]. Il est alors nécessaire d'éliminer ces autres pathologies avant d'orienter le diagnostic d'un malaise vers un type de syncope. Dès lors, en cas de suspicion de syncope, il faut en déterminer la cause. Le diagramme, proposé à la figure 4.2, présente une démarche diagnostique [Antonini-Revaz *et al.*, 2004]. [Cottier, 2002] propose une démarche similaire, en apportant des détails supplémentaires, notamment en fonction de l'âge du patient et de ses antécédents cardiaques.

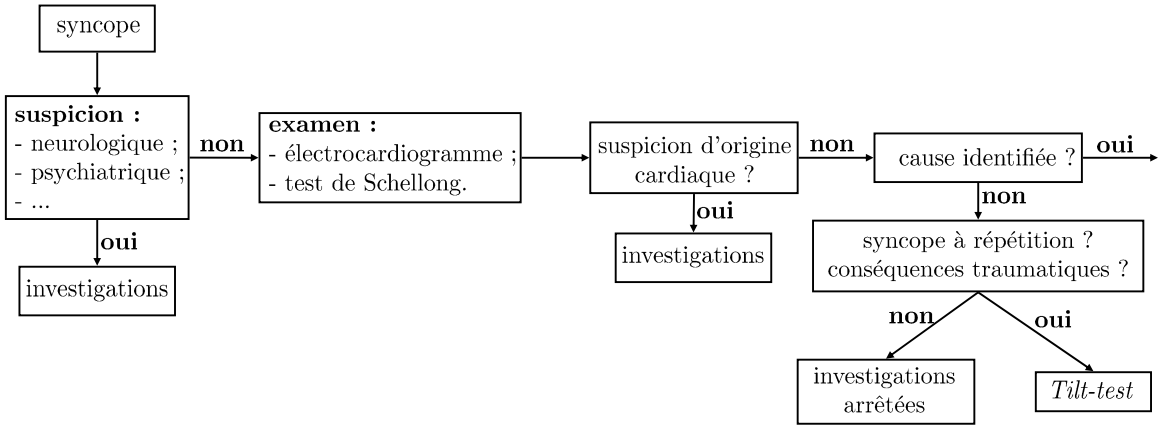


FIG. 4.2 – Démarche diagnostique en cas de suspicion de syncope [Antonini-Revaz *et al.*, 2004].

Le diagnostic peut s'avérer particulièrement difficile étant donné les multiples étiologies identifiées et les récurrences intermittentes et peu fréquentes des symptômes. Aussi, la démarche peut requérir de nombreux examens fastidieux, complets, coûteux et parfois invasifs, afin de trouver les causes responsables de l'apparition de la syncope. On peut notamment évoquer le test de Schellong présent dans le diagramme de la figure 4.2. Cet examen consiste à comparer la pression artérielle et la fréquence cardiaque en position couchée et debout, afin notamment d'écarter la cause d'origine hypotension orthostatique. Depuis quelques années, de nouvelles méthodes et technologies ont permis d'améliorer considérablement l'évaluation et le diagnostic des syncopes. On peut citer notamment, les appareils implantés d'enregistrements de longue durée de l'électrocardiogramme [Seidl *et al.*, 2000; Krahn *et al.*, 2003; Boersma *et al.*, 2004].

Malgré les nombreux examens effectués, il arrive que l'origine de la syncope ne soit pas clairement identifiée. La syncope inexpliquée est donc caractérisée par un diagnostic hésitant entre les différents types de syncope ou encore entre une syncope et les autres types de pertes de connaissances. C'est dans ces cas et lorsque les épisodes sont répétés que le patient peut être amené à réaliser le test de la table d'inclinaison (*Head-Upright Tilt-Test*, HUTT), que nous appellerons plus simplement *tilt-test*.

[7] La cataplexie est une perte soudaine du tonus musculaire, sans perte de la conscience. Elle peut être déclenchée par une émotion.

4.2.1 Test d'inclinaison : *Head-Upright Tilt-Test*

L'examen du *tilt-test* [Benditt *et al.*, 1996; Newby and Grubb, 2006] est privilégié lorsque l'étiologie des syncopes ou plus généralement des malaises, est indéfinie et que ceux-ci ont une allure de type vaso-vagale[8]. De plus, il faut que les épisodes de ces malaises soient répétés.

Le *tilt-test* est une méthode reconnue pour recréer les conditions dans lesquelles le patient ressent les symptômes. Durant toute la période du *tilt-test*, plusieurs paramètres sont analysés, dont la pression artérielle et la fréquence cardiaque qui sont régulièrement enregistrées. Ce test est effectué à jeun, il débute par une période dite de repos, durant environ 10 minutes, où le patient doit rester allongé sur la table d'examen en position horizontale. Cette période stabilise les mesures recueillies, comme la fréquence cardiaque et la pression artérielle. Suite à la phase de stabilisation et sous l'action d'un moteur électrique, la table s'incline à un angle entre 60° et 80° pendant une durée pouvant atteindre 45 minutes (voir figure 4.3). Ainsi, le passage brutal de la position allongée à la position debout peut provoquer une chute de la pression artérielle (hypotension) et un évanouissement [Bellard, 2003; Newby and Grubb, 2006]. Dès lors, le test est considéré comme positif, si l'apparition des symptômes survient avec des modifications de la pression artérielle et de la fréquence cardiaque ; comme l'hypotension et la bradycardie[9], caractéristiques de la syncope vaso-vagale. Ainsi, une chute (par rapport aux valeurs de la phase de repos), de plus de 60% de la pression artérielle et de plus 30% de la fréquence cardiaque, indique l'apparition d'une syncope de type vaso-vagale [Newby and Grubb, 2006]. En identifiant le mécanisme des malaises, ce test permettrait d'apporter des éléments de réponse afin d'adapter le traitement du patient.

patient en position allongée

patient en position inclinée

Note : Les photos illustrant l'examen du *tilt-test* proviennent du site administré par le Dr Damien Tagan à l'adresse suivante : http ://www.hopital-riviera.ch/soins-intensifs/Site_EF/Tilt%20test/Tilt-test.htm .

FIG. 4.3 – Test de la table d'inclinaison.

Le *tilt-test* permet de discriminer les patients présentant des symptômes des patients sans

[8]Une origine vaso-vagale est impliquée dans 10 à 30% des cas où les patients réalisent des syncopes sans cause apparente [Baux *et al.*, 1997].

[9]La bradycardie correspond à un rythme cardiaque trop bas par rapport à la normale.

symptômes apparents, avec un niveau de précision acceptable pour un test médical [Bellard, 2003]. De nombreuses études, comme celles de [Kenny *et al.*, 1986; Strasberg *et al.*, 1989; Fitzpatrick *et al.*, 1991], ont évalué la pertinence de ce test en obtenant un taux de spécificité[10] supérieur à 90%. Aussi, [Perez-Paredes *et al.*, 1999] suggèrent que pour des patients sujets à des syncopes inexpliquées, il est préférable de réaliser le *tilt-test* le plus rapidement après le dernier épisode de la syncope, afin d'améliorer la sensibilité : ils obtiennent dans ces conditions un taux de reproductibilité allant jusqu'à 85%. La sensibilité du *tilt-test* peut être encore améliorée en administrant aux patients des médicaments (injection d'isoprénaline pour accélérer le rythme cardiaque, ou prise de nitroglycérine sublinguale pour diminuer la pression artérielle), afin d'amplifier la provocation des symptômes et donc la reproductibilité du test [Natale *et al.*, 1998; Salamé *et al.*, 2006; Newby and Grubb, 2006].

On pourrait penser, à juste titre, que l'idée de réaliser le *tilt-test*, amenant à reproduire les symptômes, pourrait provoquer des troubles psychologiques chez des patients. Dans ce cas, les résultats des tests pourraient être faussés. Une étude intéressante de [Baux *et al.*, 1997] a cherché à évaluer l'impact des troubles psychiatriques sur les résultats du *tilt-test*. Cette étude a analysé un échantillon de 178 patients souffrant de syncopes inexpliquées. Ces patients ont effectué un entretien psychiatrique avant de réaliser les *tilt-tests*, afin de mesurer la présence ou non de troubles. L'étude a révélé la présence de troubles d'anxiété pour 21% des patients. Ceux-ci avaient la particularité de réaliser fréquemment des syncopes, mais obtenaient des résultats au *tilt-test* majoritairement négatifs. Ainsi, [Baux *et al.*, 1997] ont conclu, en évoquant l'anxiété comme « le seul » facteur prédictif de récidive de syncopes. Ce point est rarement considéré dans les études intégrant le *tilt-test*.

Le *tilt-test* est considéré comme le protocole standard pour le diagnostic de la syncope inexpliquée. Cependant, son principal problème est sa durée. En effet, comme évoqué précédemment, sa durée peut atteindre 55 minutes : 10 minutes en position allongée et 45 minutes en position inclinée. Aussi, la préparation du test et du patient, l'ajout d'agents pharmacologiques, contribuent également à augmenter la durée totale de l'examen et le temps concédé par le personnel médical.

4.2.2 Signaux de mesures : électrocardiogramme et signal d'impédancemétrie thoracique

Comme évoqué précédemment, durant l'examen du *tilt-test* plusieurs signaux sont enregistrés ; il est courant d'acquérir l'électrocardiogramme (ECG) et la pression artérielle (PA). Comme nous l'avons vu, ces signaux sont largement employés pour prédire le résultat du *tilt-test* et donc, l'apparition des symptômes de la syncope. En outre, dans le cadre de la prédiction de la syncope, de récentes études [Bellard *et al.*, 2003; Bellard, 2003; Schang *et al.*, 2003] ont révélé l'intérêt de l'utilisation du signal d'impédancemétrie thoracique [Bonjer *et al.*, 1952; Kubicek *et al.*, 1966; Lababidi *et al.*, 1970].

L'impédancemétrie est basée sur la mesure de l'impédance électrique du thorax [Malmivuo and Plonsey, 1995][11], afin d'analyser durant chaque battement l'hémodynamique cardiaque[12]. Ce signal permet alors d'évaluer le volume d'éjection systolique[13] (VES) durant chaque cycle

[10] La spécificité indique le pourcentage de patients parmi les « non syncopeurs » fidèlement classés comme « non syncopeurs » (*cf.* section 3.4.1).

[11] L'ouvrage est disponible en ligne à l'adresse suivante http ://www.bem.fi/book/index.htm.

[12] L'hémodynamique cardiaque est l'étude de l'écoulement du sang dans les vaisseaux, considérant notamment le volume, le débit et la vitesse.

[13] Le volume systolique indique la quantité de sang éjectée par un ventricule à chaque systole.

cardiaque et donc, de déterminer le **débit cardiaque** (Q) [Newman and Calister, 1999] :

$$\mathrm{Q} = \mathrm{VES} \times \mathrm{FC}\,, \tag{4.1}$$

où FC désigne la fréquence cardiaque.

La technique d'impédancemétrie a l'avantage d'être non invasive, facilement utilisable et permet également de fournir une information instantanée, continue et précise. Cette mesure du débit cardiaque est fondée sur l'analyse des modifications de la résistance transthoracique, causées par les variations du débit sanguin lors de l'entrée et de la sortie du sang dans le thorax. Ainsi, l'application d'un courant alternatif de faible amplitude et de haute fréquence permet alors de mesurer les variations de la résistance thoracique. Comme montré à la figure 4.4, ce courant est envoyé à travers deux électrodes (a et d), l'une placée sur le cou et l'autre sur l'abdomen. Au même moment, une tension est mesurée aux bornes de deux autres électrodes (b et c) placées à proximité des précédentes. Ainsi, en fonction de l'intensité du courant injecté et de la tension recueillie, la loi d'Ohm permet alors d'obtenir une impédance. Le signal obtenu, appelé Z, mesure alors l'impédance thoracique en fonction du débit cardiaque au cours du temps. Le dispositif de mesure est donné pour l'appareil PhysioFlow$^{\mathrm{TM}}$ de Manatec Biomedical[14] [Bour *et al.*, 1994].

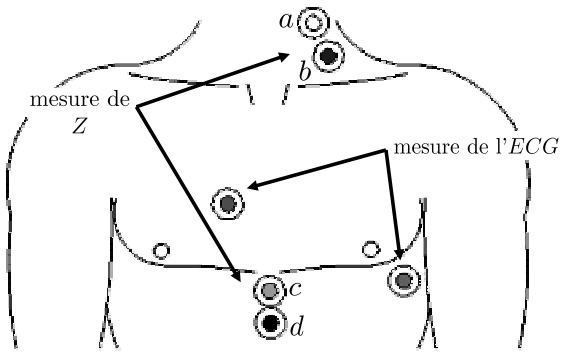


FIG. 4.4 – Placement des électrodes d'acquisition des signaux de l'électrocardiogramme et de l'impédancemétrie thoracique.

Caractérisant l'hémodynamique cardiaque, le signal Z, donné à la figure 4.5, permet d'observer la variation du volume de sang entre les électrodes. Ainsi, par ses caractéristiques (amplitudes et durées), le signal Z reflète alors le volume d'éjection systolique [Bellard, 2003]. Cependant, ce signal possède un faible rapport signal sur bruit, ce qui amène le plus souvent à considérer sa dérivée, notée dZ/dt [Kubicek *et al.*, 1966]. Sur la dérivée de Z, chaque phase de battement du cœur possède une trace électrique particulière, comme sur l'ECG. En effet, l'enregistrement de l'ECG, dont un exemple est donné à la figure 4.5, permet d'observer notamment la contraction des deux oreillettes sur l'onde P et la dépolarisation ventriculaire sur l'onde Q. Aussi, le complexe QRS caractérise la contraction brève des deux ventricules, quant à l'onde T, elle correspond à la repolarisation ventriculaire. Ainsi, à l'image des interprétations réalisées sur l'ECG, [Lababidi *et al.*, 1970] ont extrait sur le signal dZ/dt différents instants de temps particuliers du battement ; ceux-ci ont pu être corrélés aux informations obtenues sur le phonocardiogramme[15] [Nyober *et al.*, 1940]. Sans rentrer dans les détails nous en citons quelques-uns :

[14] http ://www.physioflow.com/

[15] Le phonocardiogramme est la représentation des phénomènes auditifs sur papier des battements cardiaques.

- l'instant A correspond à la contraction des oreillettes ;

- l'instant B correspond à la fermeture de la valve entre l'oreillette droite et le ventricule droit (valve tricuspide) ;

- le début de l'instant t_1 correspond au début de l'éjection ventriculaire gauche dans l'aorte ;

- l'instant dZ_{max}/dt correspond au pic de vitesse de l'éjection ;

- l'instant X correspond à la fermeture de la valve entre le ventricule gauche et l'aorte (valve aortique) ;

- l'instant Y correspond à la fermeture de la valve entre le ventricule droit et l'artère pulmonaire (valve pulmonaire) ;

- l'instant O correspond à l'ouverture de la valve entre l'oreillette gauche et le ventricule gauche (valve mitrale) ;

- l'intervalle de temps entre le début de l'onde Q de l'ECG et l'instant B représente la période de pré-éjection ;

- les intervalles de temps t_1 et t_2 correspondent respectivement à l'accélération positive et négative de l'éjection ventriculaire. L'intervalle de temps entre B et Y représente la période complète d'éjection ventriculaire (TEVG).

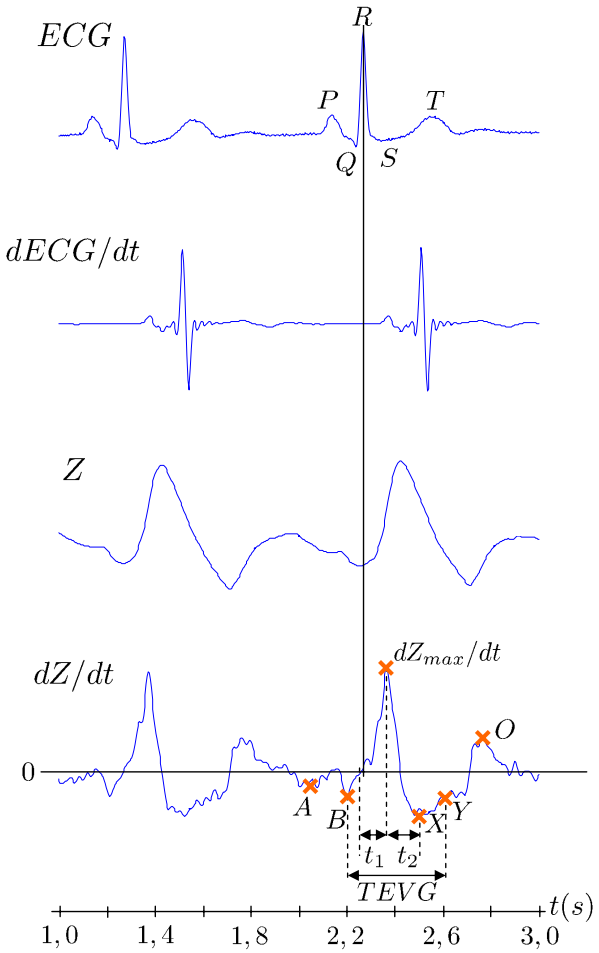


FIG. 4.5 – Caractéristiques extraites sur la dérivée du signal d'impédancemétrie thoracique durant un battement cardiaque.

Rappelons que, la mesure du signal d'impédancemétrie thoracique permet d'obtenir le débit cardiaque, une fois le volume d'éjection systolique déterminé. Ainsi, en considérant le thorax comme un cylindre, [Nyober, 1959] a pu estimer le calcul du VES à partir du signal Z. Les reproches faits sur le signal Z ont amené [Kubicek *et al.*, 1966; Bernstein, 1986] à approcher le calcul du VES en considérant la dérivée dZ/dt ; dans le même temps, [Bernstein, 1986] a proposé de modéliser le thorax par un cône tronqué. Par leur formule, [Kubicek *et al.*, 1966] estiment le VES en utilisant les paramètres dZ_{max}/dt et $TEVG$ directement mesurés sur le signal dZ/dt.

La méthode de mesure du signal d'impédancemétrie thoracique peut se heurter à certaines difficultés, notamment lors de l'acquisition du signal, où celui-ci peut subir des perturbations. En effet, les électrodes étant positionnées sur le cou et le thorax, ces perturbations peuvent être dues à de fortes respirations, des contractions musculaires, ou tout simplement des mouvements des patients. Cependant, des études telles que [Fuller, 1992; Shoemaker *et al.*, 1994] qui, en comparant

la mesure du débit cardiaque obtenue par d'autres techniques, ont permis néanmoins de valider l'utilisation du signal d'impédancemétrie thoracique. Ces autres techniques sont la thermodilution [Secher *et al.*, 1979; Doering *et al.*, 1995], la dilution de colorant [Smith *et al.*, 1970] ou encore la méthode de Fick indirect [Edmunds *et al.*, 1982].

4.3 État de l'art sur la prédiction de la syncope

De nombreuses études ont été citées tout au long de ce chapitre, montrant les préoccupations soulevées par le problème inhérent aux syncopes. Nous présentons dans cette section, quatre études qui se sont attachées à prédire le résultat du *tilt-test* pour des patients sujets à des apparitions récurrentes de syncopes inexpliquées. Aussi, comme nous l'avons évoqué, le principal problème de cet examen est sa durée, c'est pourquoi, chacune de ces études évalue ses performances de prédiction en réduisant la durée d'investigation du *tilt-test*.

Le tableau 4.2 récapitule les résultats des quatre études présentées ci-dessous. On peut y observer deux analyses : l'**analyse rétrospective** et l'**analyse prospective**. L'analyse rétrospective rend compte des résultats obtenus sur le groupe d'apprentissage. Rappelons que ce groupe de patients est utilisé pour déterminer les règles et les paramètres du modèle de classification, afin de discriminer les patients positifs au *tilt-test* des patients négatifs. Quant au groupe prospectif, celui-ci estime la performance et la reproductibilité des paramètres de discrimination, en utilisant un groupe de patients n'ayant pas participé à l'apprentissage des règles. Les groupes rétrospectif et prospectif correspondent respectivement aux groupes d'apprentissage et de test, présentés à la section 3.3.

- [Mallat *et al.*, 1997]

  Cette étude a analysé un échantillon total de 197 patients (98 patients dans l'analyse rétrospective et 99 patients dans l'analyse prospective). Pour cette étude, la durée analysée du *tilt-test* est réduite aux 6 premières minutes après le basculement de la table à 60°. La prédiction du résultat négatif du *tilt-test* est fondée sur une augmentation de la fréquence cardiaque inférieure à 18 bpm (battements par minute) pendant une durée prolongée.

- [Pitzalis *et al.*, 2002]

  Cette étude-ci a analysé un échantillon total de 318 patients (238 patients dans l'analyse rétrospective et 80 patients dans l'analyse prospective). Pour cette étude, la durée analysée du *tilt-test* est réduite aux 15 premières minutes après le basculement de la table à 70°. La prédiction du résultat positif du *tilt-test* est fondée sur la diminution de la valeur de la pression artérielle systolique, en comptabilisant le nombre de fois que cette valeur est passée sous le seuil défini durant la période de repos.

- [Bellard *et al.*, 2003]

  Cette autre étude a analysé un échantillon total de 68 patients. Deux périodes du tilt-test ont été analysées : la période de repos et la période entre la 5-ième et la 10-ième minute après le basculement de la table à 70°. D'autre part, les prédictions sont établies principalement sur des caractéristiques recueillies sur le signal d'impédancemétrie thoracique, dont les paramètres t_1, t_2, dZ_{max}/dt (figure 4.5. Ainsi, durant la période de repos, la prédiction du résultat positif du *tilt-test* est donnée lorsque l'intervalle de temps t_2 est inférieur à $199\,ms$. En considérant la période entre les 5-ième et 10-ième minutes du basculement, la prédiction du test est obtenue lorsque t_2 varie de plus de $40\,ms$ par rapport au seuil défini dans la phase de repos. Une autre analyse a été réalisée en incorporant aux variables

précédentes, des variables hémodynamiques : fréquence cardiaque (FC), pressions systolique (PAS), diastolique (PAD) et différentielle (PD). Ainsi, la prédiction du test positif considère également les seuils de ces nouvelles variables : $FC < 11\,bpm$, $PAS < 2\,$mmHg, $PAD < 7\,$mmHg et $PD < -3\,$mmHg. [Bellard *et al.*, 2003] ont également étudié l'impact du test pharmacologique à la nitroglycérine pour la prédiction du résultat du *tilt-test*. Dans cette analyse, les indices dZ_{max}/dt, t_2, PAS et PD se sont révélés pertinents.

- [Schang *et al.*, 2003]
  Cette étude a analysé un échantillon total de 129 patients (70 patients dans l'analyse rétrospective et 59 patients dans l'analyse prospective). Dans ce cas, la durée analysée du *tilt-test* est réduite à la période de repos. La prédiction est établie principalement sur des caractéristiques recueillies sur le signal d'impédancemétrie thoracique, dont le temps d'éjection ventriculaire. Contrairement aux trois précédentes études, ce travail exploite les réseaux de neurones comme modèle de prédiction.

étude	**période analysée du HUTT**	**analyse**	S_e	S_p	VPP	VPN
[Mallat *et al.*, 1997]	1re à 6-ième minutes	rétrospective	100%	89%	96%	100%
		prospective	96%	87%	75%	98%
[Pitzalis *et al.*, 2002]	1re à 15-ième minutes	rétrospective	93%	58%	28%	98%
		prospective	80%	85%	57%	94%
[Bellard *et al.*, 2003]†	période de repos	rétrospective	68%	63%	63%	68%
		prospective	–	–	–	–
[Bellard *et al.*, 2003]†	5-ième à 10-ième minutes	rétrospective	68%	70%	68%	70%
		prospective	–	–	–	–
[Bellard *et al.*, 2003]‡	5-ième à 10-ième minutes	rétrospective	50%	97%	93%	67%
		prospective	–	–	–	–
[Schang *et al.*, 2003]	période de repos	rétrospective	100%	100%	100%	100%
		prospective	69%	73%	67%	75%

Note : † étude analysant uniquement les variables liées au signal dZ/dt. ‡ étude analysant les variables liées au signal dZ/dt et FC, PAS, PAD et PD.
Rappelons que S_e, S_p, VPP et VPN indiquent respectivement la sensibilité, la spécificité, les valeurs prédictives positive et négative ; leur description est donnée à la section 3.4.1 (page 108).

TAB. 4.2 – Récapitulatif des résultats significatifs de recherches sur la syncope inexpliquée.

Les deux dernières études [Bellard *et al.*, 2003; Schang *et al.*, 2003], ont montré au terme de leurs analyses, l'importance et l'intérêt des variables recueillies sur le signal d'impédancemétrie thoracique, notamment l'intervalle de temps lié à la fin de l'éjection ventriculaire t_2.

4.4 Conclusions

Le risque de chute et de traumatisme inhérent aux syncopes, bien qu'il se soit révélé faible comparé à d'autres pathologies est néanmoins présent. Aussi, pour limiter et prévenir les malaises, le patient doit connaître sa maladie et les circonstances de son apparition. En effet, cette prévention lui permettrait d'éviter certaines conduites ou circonstances à risque [Newby and Grubb, 2006]. Tel est l'enjeu des diagnostics.

On notera enfin les travaux [Brignole *et al.*, 2001; Brignole *et al.*, 2004] qui proposent un document de grande qualité, très complet et synthétique sur la syncope.

Chapitre 5

Études expérimentales pour la prédiction de la syncope

5.1 Introduction

Les travaux présentés dans ce chapitre portent sur la prédiction de la syncope inexpliquée. Rappelons que la syncope inexpliquée concerne approximativement 35% des apparitions de syncopes (*cf.* tableau 4.1, page 118). Comme le précise la démarche diagnostique présentée à la section 4.2 (voir figure 4.2), le diagnostic de ce type de syncope peut requérir l'examen du *tilt-test*. Cet examen reproduit les conditions provoquant la syncope, mais son problème majeur est sa durée. En effet, le protocole de cet examen nécessite une période de repos de 10 minutes afin de stabiliser les signaux à mesurer et une seconde période, où la table sur laquelle le patient est installé bascule, le laissant quasiment à la verticale durant 45 minutes ou jusqu'à la survenue des symptômes. Ainsi, sans apparition de symptômes, l'examen monopolise du personnel médical durant près d'une heure. Dès lors, pour des raisons de coût et de bien-être des patients, il paraît important de pouvoir réduire la durée du test. C'est dans cet objectif que s'inscrivent les études présentées dans ce chapitre, qui tentent de prédire l'apparition des signes liés aux syncopes avant que l'examen n'arrive à son terme, évitant ainsi aux patients de ressentir les symptômes. Dans ce même but, à la section 4.3 ont été exposés plusieurs travaux, dont ceux de [Bellard *et al.*, 2003; Schang *et al.*, 2003] qui parviennent à prédire l'apparition de la syncope pendant la période de repos. Les études de [Mallat *et al.*, 1997; Pitzalis *et al.*, 2002; Bellard *et al.*, 2003] n'y parviennent qu'en utilisant les premières minutes après le basculement. Ainsi, à l'image de ces travaux, nous allons analyser les deux phases du *tilt-test*, la phase de repos et la phase basculée, respectivement aux sections 5.3 et 5.4. D'autre part, comme il a été mentionné à la section 4.3, le signal d'impédancemétrie thoracique semble apporter des informations pertinentes pour prédire la syncope [Bellard *et al.*, 2003; Schang *et al.*, 2003]. C'est ainsi qu'une attention particulière sera portée au signal d'impédancemétrie thoracique à la section 5.5. Au terme de ces différentes analyses, nous apporterons, sous forme de synthèse, un bilan des résultats obtenus.

5.2 Cadres expérimentaux

Les travaux présentés sont basés sur des études réalisées au service de cardiologie du CHU d'Angers. Les patients participant à ces études sont tous sujets à des syncopes récurrentes inexpliquées, et, ont réalisé au moins un épisode de syncope dans les trois derniers mois avant d'accomplir le *tilt-test*. Les résultats négatifs à différents examens, tels que des tests sanguins, un électrocardiogramme à 12 dérivations, une échocardiographie transthoracique ou encore une

échographie carotidienne, ont permis d'inclure dans les études cliniques les patients ne souffrant pas de maladies neurologique, cardiaque et psychiatrique. D'autre part, certains médicaments (tels que les diurétiques, les vasodilatateurs, les bêtabloquants) pouvant interférer avec le test ont été interrompus au moins deux jours avant les études médicales. Tous les patients ont alors réalisé le *tilt-test* sur une table motorisée (FGCK, Couverchel, Draveil, France) dans des conditions similaires : entre 14 heures et 17 heures, dans une chambre à lumière tamisée dont la température est maintenue entre 24° C et 25° C. Comme expliqué précédemment, après une période de repos, la table sur laquelle le patient est installé bascule à un angle de 70° durant 45 minutes. Si des symptômes apparaissent, le sujet retourne en position couchée et le test est arrêté prématurément. Le *tilt-test* est donc considéré **positif** lors de la reproduction de symptômes liées à la syncope, tels qu'une perte de conscience et du tonus postural ou encore des symptômes proches de la syncope tels que des nausées, des étourdissements, une pâleur ou une sensation imminente de syncope. Si aucun symptôme n'apparaît au bout des 45 minutes de basculement, le *tilt-test* est alors considéré comme **négatif**. Les études réalisées au service de cardiologie du CHU d'Angers ont donc conduit à la création de deux échantillons de patients.

L'analyse de la répartition statistique du premier échantillon de 86 patients a conduit à écarter deux patients, estimant la présence de valeurs aberrantes sur ces derniers. À la section 2.2.2, un élément d'une variable (qui suit une loi normale) pouvait se révéler aberrant, si sa valeur était en dehors de l'intervalle $[\mu - 2\sigma; \mu + 2\sigma]$ ou $[\mu - 3\sigma; \mu + 3\sigma]$. Or, la figure 5.1 montre que deux données sont très éloignées de la moyenne, suggérant ainsi d'écarter les deux patients. Le premier échantillon est donc composé de 84 patients (44 hommes et 40 femmes), dont la moyenne d'âge est de 43 ans et l'écart type 15 ans (variant de 18 à 73 ans). À l'issue des *tilt-tests*, 44 patients se sont révélés positifs à l'examen : la prévalence observée dans l'échantillon est donc de 52%. Nous notons cet échantillon $\mathcal{E}_1$ et les détails des variables le composant sont donnés aux tableaux 5.1 et 5.2. Ces tableaux donnent plusieurs informations, telles que la répartition des patients pour chacune des deux classes et les variables possédant des valeurs manquantes. Aussi, sur le tableau 5.2, nous pouvons observer qu'une même variable est mesurée durant les deux phases du *tilt-test* : période de repos et les 10 premières minutes de la période du basculement.

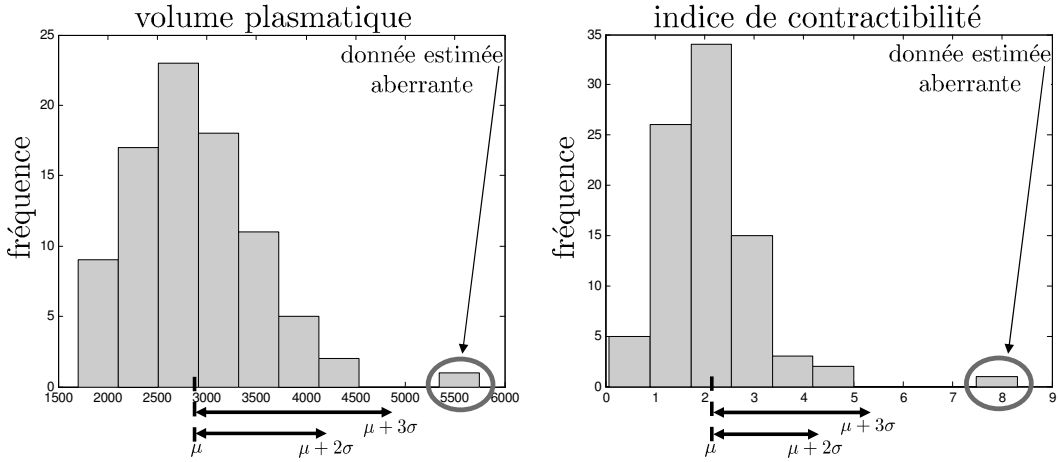


FIG. 5.1 – Illustration des deux données aberrantes supprimées de l'échantillon initial $\mathcal{E}_1$.

Le second échantillon de patients noté $\mathcal{E}_2$, composé initialement de 138 patients, est, pour les mêmes raisons que $\mathcal{E}_1$, réduit à 129 patients (63 hommes et 66 femmes), dont la moyenne d'âge est de 42 ans et l'écart type 14 ans (variant de 18 à 73 ans). À l'issue des *tilt-tests*, 63 patients se sont révélés positifs à l'examen : la prévalence observée dans cet échantillon est donc de 49%. Dans cet échantillon, l'électrocardiogramme (ECG) et le signal d'impédancemétrie thoracique (Z) sont enregistrés en continu durant tout l'examen du *tilt-test*.

Des mesures sur le signal d'impédancemétrie thoracique apparaissent dans les deux échantillons de patients ($\mathcal{E}_1$ et $\mathcal{E}_2$). Le dispositif de mesure utilisé est l'appareil PhysioFlow^{TM} de Manatec Biomedical [Bour *et al.*, 1994]. Le signal Z est obtenu en injectant un courant alternatif de faible ampérage ($1,8$ mA) à haute fréquence (75 KHz) à travers 4 électrodes (Ag/AgCl, 40493E), comme illustré à la figure 4.4 de la page 122. Les signaux Z et ECG sont mesurés avec une période d'échantillonnage de 250 Hz.

variable (unité)	**réponse au *tilt-test***	
	positive	**négative**
	(44 patients)	**(40 patients)**
âge	41 ± 15	45 ± 15
sexe (homme / femme)	22/22	22/18
taille (cm)	$167,5 \pm 8,4$	$167,2 \pm 7,5$
poids (kg)	$69,1 \pm 10,6$	$67,1 \pm 15,7$
eau totale théorique (l)	$36,8 \pm 5,5$	$36,8 \pm 6,9$
surface corporelle (m^2)	$1,80 \pm 0,16$	$1,76 \pm 0,24$
volume plasmatique mesuré (ml)	$2\,921 \pm 694$	$2\,831 \pm 640$
eau totale mesurée (l)	$36,6 \pm 6,6$	$36,6 \pm 7,0$
volume plasmatique théorique / eau totale (sans unité)	$34,2 \pm 6,4$	$34,3 \pm 6,7$
volume plasmatique mesuré / eau totale (sans unité)	$33,6 \pm 6,6$	$33,8 \pm 6,6$
eau totale (l)	$53,2 \pm 7,0$	$55,4 \pm 5,9$
masse grasse (kg)	$18,8 \pm 6,2$	$17,5 \pm 7,4$
pourcentage de masse grasse (%)	$27,1 \pm 7,8$	$25,5 \pm 6,4$
masse maigre (kg)	$50,3 \pm 9,2$	$49,6 \pm 10,4$
pourcentage de masse maigre (%)	$72,9 \pm 7,8$	$74,5 \pm 6,4$
rapport masse maigre / masse grasse (sans unité)	$3,1 \pm 1,4$	$3,2 \pm 1,3$
hématocrite (%)	$42,6 \pm 3,2$	$42,6 \pm 3,6$
hémoglobine (g/100ml)	$13,7 \pm 1,1$	$13,8 \pm 1,5$
osmolarité ($mOsm \cdot kg^{-1}$)	$293,8 \pm 8,3$ *	$293,6 \pm 5,5$ *
variation initiale max. de FC (bpm)	$16,8 \pm 9,9$ *	$20,2 \pm 11,8$ *
évaluation du rebond initial FC (bpm)	$1,1 \pm 0,2$	$1,2 \pm 0,4$
variation initiale de PAS (mmHg)	$-19,6 \pm 15,1$	$-19,1 \pm 17,3$
variation initiale de PAD (mmHg)	$-9,3 \pm 9,9$	$-7.8 \pm 9,2$
pression artérielle minimale (mmHg)	$122,9 \pm 25,7$	$131,0 \pm 25,5$
chute de PAS max (mmHg)	$-13,0 \pm 14,0$	$-7,2 \pm 15,2$
chute de PAD max (mmHg)	$-2,6 \pm 12,7$	$-1.7 \pm 7,2$
FC max (bpm)	$90,4 \pm 13,0$	$91,7 \pm 19,1$
élévation max de FC (bpm)	$23,1 \pm 10,2$	$23,5 \pm 15,4$
delta de PAS (mmHg)	$-16,0 \pm 21,0$	$-11,3 \pm 12,6$
delta de PAD (mmHg)	$-9,2 \pm 9,4$	$-5,3 \pm 6,0$

Note : * indique la présence de valeurs manquantes. Pour les variables continues, nous donnons la moyenne et l'écart type. En présence de valeurs manquantes, les calculs de la moyenne et de l'écart type considèrent uniquement les valeur disponibles.

TAB. 5.1 – Récapitulatif de l'ensemble des variables recueillies pour l'étude de l'apparition de la syncope lors d'un examen du *tilt-test* (partie 1/2).

variable (unité)	**période de repos**		**réponse au *tilt-test* les 10^{re} minutes du basculement**		**écart des deux périodes**	
	positive	négative	positive	négative	positive	négative
fréquence cardiaque (bpm)	$67,1 \pm 9,0$	$68,0 \pm 12,0$	$89,2 \pm 14,0$ *	$86,3 \pm 15,4$ *	$21,6 \pm 9,6$ *	$18,3 \pm 10,9$ *
pression artérielle systolique (mmHg)	$136,4 \pm 22,7$	$138,7 \pm 20,8$	$128,8 \pm 20,9$ *	$136,1 \pm 21,7$ *	$-9,6 \pm 14,4$ *	$-2,6 \pm 16,2$ *
pression artérielle diastolique (mmHg)	$70,7 \pm 11,4$	$77,4 \pm 13,2$	$79,7 \pm 12,2$ *	$84,9 \pm 13,8$ *	$3,5 \pm 9,5$ *	$7,4 \pm 6,5$ *
pression artérielle moyenne (mmHg)	$94,1 \pm 14,1$	$96,9 \pm 14,3$	$93,2 \pm 13,5$ *	$100,5 \pm 14,5$ *	$-0,9 \pm 10,8$ *	$3,6 \pm 8,0$ *
pression artérielle pulsée (mmHg)	$60,7 \pm 17,7$	$61,3 \pm 13,9$	$46,1 \pm 15,4$ *	$51,2 \pm 16,0$ *	$-13,1 \pm 10,4$ *	$-10,0 \pm 13,1$ *
vitesse moyenne artère cérébrale ($m{\cdot}s^{-1}$)	$58,6 \pm 13,2$ *	$59,4 \pm 10,4$ *	$53,1 \pm 23,2$ *	$52,8 \pm 11,1$	$-4,4 \pm 17,8$ *	$-6,6 \pm 4,8$ *
indice de résistance (sans unité)	$1,12 \pm 0,36$ *	$1,13 \pm 0,38$ *	$1,16 \pm 0,48$ *	$1,15 \pm 0,41$	$0,01 \pm 0,27$ *	$0,02 \pm 0,29$ *
résistance vasculaire cérébrale (sans unité)	$1,66 \pm 0,39$ *	$1,71 \pm 0,44$ *	$1,41 \pm 0,47$ *	$1,54 \pm 0,45$ *	$-0,29 \pm 0,27$ *	$-0,18 \pm 0,19$ *
indice de pulsabilité (sans unité)	$0,80 \pm 0,15$ *	$0,78 \pm 0,11$ *	$0,77 \pm 0,14$ *	$0,81 \pm 0,16$ *	$-0,00 \pm 0,11$ *	$-0,03 \pm 0,12$ *
variation du volume du mollet (ml/100ml)	–	–	$3,64 \pm 1,17$ *	$3,49 \pm 1,11$ *	–	–
durée d'accélération positive de l'éjection ventriculaire (ms)	$197,1 \pm 36,9$	$194,0 \pm 34,34$	$294,1 \pm 63,6$ *	$292,9 \pm 55,7$ *	$101,2 \pm 45,6$ *	$97,3 \pm 40,4$ *
maximum de dZ ($\Omega{\cdot}s^{-1}$)	$422,8 \pm 145,6$	$379,3 \pm 170,8$	$384,9 \pm 137,9$ *	$350,0 \pm 118,3$ *	$-34,1 \pm 138,7$ *	$-33,4 \pm 101,4$ *
indice de contractibilité ($m\Omega{\cdot}s^{-2}$)	$2,22 \pm 0,85$	$2,07 \pm 1,29$	$1,32 \pm 0,43$ *	$1,24 \pm 0,48$ *	$-0,93 \pm 0,75$ *	$-0,84 \pm 0,98$ *
durée de la partie négative de l'éjection ventriculaire (ms)	$169,2 \pm 50,1$	$194,1 \pm 96,9$	$275,7,1 \pm 100,6$ *	$261,1 \pm 96,4$ *	$113,7 \pm 95,5$ *	$66,0 \pm 99,4$ *

Note : * indique la présence de valeurs manquantes. Pour les variables continues, nous donnons la moyenne et l'écart type. En présence de valeurs manquantes, les calculs de la moyenne et de l'écart type considèrent uniquement les valeur disponibles.

TAB. 5.2 – Récapitulatif de l'ensemble des variables recueillies pour l'étude de l'apparition de la syncope lors d'un examen du *tilt-test* (partie 2/2).

5.3 Recherche d'indices prédictifs du résultat du *tilt-test* durant la période de repos

5.3.1 Analyse exhaustive des sous-ensembles pertinents de variables initiales

5.3.1.1 Introduction

Dans cette première analyse dont les principaux résultats ont été publiés dans [Feuilloy *et al.*, 2005a; Feuilloy *et al.*, 2005b], nous comparons les différents modèles de classification sur l'échantillon de patients nommé précédemment $\mathcal{E}_1$. Comme nous souhaitons considérer uniquement la période de repos du *tilt-test*, nous devons, au préalable, pré-sélectionner des variables indépendantes de la phase de basculement parmi les 70 variables initiales (présentées aux tableaux 5.1 et 5.2). Une pré-sélection de 15 variables a été réalisée par les médecins du CHU d'Angers. Ces variables leur semblaient particulièrement pertinentes pour prédire durant la phase de repos, l'apparition de la syncope durant le *tilt-test*. Le tableau 5.3 récapitule les 15 variables mesurées pour les 84 patients. Parmi ces 15 variables, nous retrouvons les variables utilisées par les études détaillées à la section 4.3 [Mallat *et al.*, 1997; Pitzalis *et al.*, 2002; Bellard *et al.*, 2003; Schang *et al.*, 2003], telles que la fréquence cardiaque, la pression artérielle systolique, ou encore, des variables liées au signal d'impédancemétrie thoracique.

variables	*symboles*
âge	*âge*
surface corporelle	*BSA* (*body surface area*)
volume plasmatique	*VolPlas*
fréquence cardiaque	*FC*
pression artérielle systolique	*PAS*
pression artérielle diastolique	*PAD*
pression pulsée	*PP*
eau totale	*TBW* (*total body water*)
rapport masse maigre / masse grasse	*LW/FW*
hématocrite	*Ht*
hémoglobine	*Hb*
accélération positive de l'éjection ventriculaire	t_1
partie négative de l'éjection ventriculaire	t_2
maximum de *dZ*	dZ_{max}/dt
indice de contractibilité	C

TAB. 5.3 – Liste des variables pré-sélectionnées par les médecins susceptibles d'être pertinentes pour prédire l'apparition des symptômes de la syncope durant la position couchée du *tilt-test*.

5.3.1.2 Méthodes

Les variables initiales de l'échantillon $\mathcal{E}_1$ sont fournies directement aux modèles. Ainsi, mis à part leur normalisation (*cf.* section 2.2.3), il n'y a pas de pré-traitement particulier. Dès lors, compte tenu du faible nombre de variables (15), pour chaque technique de classification nous évaluons toutes les combinaisons possibles de variables, qui sont de l'ordre de $2^{15} - 1$ (32 767). Ainsi, le processus de sélection utilisé est de type *wrapper* (englobant l'algorithme d'apprentissage dans la sélection, *cf.* section 2.4.2), il est illustré à la figure 5.2.

Les traitements et les expérimentations sont effectués avec MatLab© (The MathWorks Inc., South Natic, MA, USA).

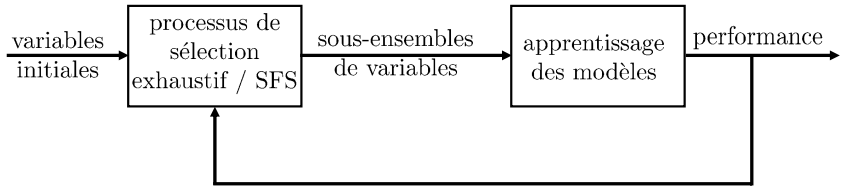


FIG. 5.2 – Processus de sélection des sous-ensembles de variables pertinentes pour prédire le résultat du *tilt-test* en position couchée.

Techniques de classification

Plusieurs méthodes de classification peuvent être appliquées pour séparer les résultats des patients au *tilt-test*. Le résultat étant soit positif, soit négatif, cela nous place dans une tâche de classification binaire. Les techniques utilisées sont des méthodes linéaires et non linéaires, elles sont basées sur des approches génératives et discriminantes. Ainsi, nous nous attachons à comparer les performances des méthodes suivantes : fonctions discriminantes linéaires (FDL) et quadratiques (FDQ), classifieurs de Bayes naïfs (BN), les *support vector machines* (SVM) et les réseaux de neurones. Les détails de ces méthodes sont donnés dans le chapitre 1.

Parmi les approches génératives, nous avons les classifieurs de Bayes naïf, qui comme explicité à la section 1.2.4.2 (page 21), nécessitent le calcul ou l'estimation des densités de probabilité. À la section 1.2.3, nous avons vu que cette estimation peut être paramétrique ou non paramétrique. Rappelons que les estimations paramétriques font des hypothèses sur la forme analytique de la distribution des observations, afin de lier les distributions à des lois connues. Ainsi, nous comparons deux types d'estimations : l'une en faisant l'hypothèse que les variables suivent des lois normales et l'autre en estimant les densités par les k-plus proches voisins (k-ppv) ; nous notons respectivement ces deux approches par $\mathrm{BN}_{\mathrm{gauss}}$ et $\mathrm{BN}_{k\text{-ppv}}$. Dans les expérimentations employant l'estimation non paramétrique par les k-ppv, le paramètre k n'est pas prédéfini, sa valeur est choisie de façon à optimiser les performances en généralisation.

Pour les *support vector machines* (section 1.4.3, page 53), nous avons vu que la transformation des observations de l'espace initial vers un espace de plus grande dimension utilise principalement des fonctions noyaux conventionnelles, telles que des fonctions polynomiales ou encore des fonctions gaussiennes. La mise en œuvre des SVM sous Matlab est réalisée par la *toolbox* « LS-SVMlab1.5 » de [Suykens *et al.*, 2002][1]. Cette *toolbox* a l'avantage d'avoir été expérimentée dans de nombreux travaux comme ceux de [Lukas, 2003; Lu, 2005]. Dans les expérimentations employant les noyaux polynomiaux et gaussiens, les paramètres ne sont pas prédéfinis, leur valeur est choisie de manière à optimiser les performances en généralisation.

Pour les réseaux de neurones, nous avons choisi les perceptrons multicouches (PMC), en imposant pour leur architecture, une seule couche cachée de neurones. Les fonctions d'activation des neurones sont de type sigmoïde (tangente hyperbolique, voir figure 1.19 à la page 41). L'algorithme d'apprentissage utilisé est celui de Levenberg-Marquardt, qui a la particularité de s'adapter à la forme de la surface d'erreur et l'avantage de converger rapidement (*cf.* section 1.4.2.5). D'autre part, compte tenu du nombre relativement faible des entrées (15 au plus), nous avons privilégié une méthode empirique pour déterminer l'architecture idéale des réseaux. Ainsi, pour chaque sous-ensemble de variables d'entrées du modèle, l'architecture du PMC est obtenue en comparant les performances en généralisation du modèle pour plusieurs nombres de neurones de la couche cachée ; dans nos expérimentations ce nombre varie de 2 à 20. D'autre part, pour améliorer la

[1] http ://www.esat.kuleuven.ac.be/sista/lssvmlab/

généralisation des réseaux de neurones, nous avons opté pour la méthode de régularisation de « l'arrêt prématuré ». Comme expliqué à la section 1.4.2.5, cette technique de régularisation évalue périodiquement les performances de généralisation, afin de stopper l'apprentissage avant de voir apparaître un surapprentissage du modèle. Ainsi, dans notre processus d'apprentissage, l'évaluation périodique des performances du réseau est effectuée sur les sous-ensembles de validation ($\mathcal{X}_V$) ; la description de ces sous-ensembles est donnée au paragraphe suivant. Pour finir, compte tenu du caractère aléatoire de l'initialisation des poids des réseaux de neurones qui, par ailleurs, influence considérablement la convergence de l'apprentissage, il est nécessaire de réitérer plusieurs apprentissages pour chaque architecture. Ainsi, pour un même nombre de neurones cachés, 100 apprentissages sont réalisés, permettant ainsi d'estimer précisément les performances de chaque architecture. Dès lors, étant donné la « lourdeur » du processus des PMC pour l'évaluation et la sélection du meilleur réseau de neurones, il paraît déraisonnable de les évaluer sur chacun des 32 767 sous-ensembles de variables. Par conséquent, pour les PMC, la recherche de sous-ensembles de variables pertinentes n'est pas réalisée de manière exhaustive, mais par la méthode de sélection séquentielle SFS (*cf.* section 2.4.3.3, page 88). Pour nos 15 variables initiales, cette méthode a l'avantage de réduire le nombre de combinaisons à évaluer à 120 sous-ensembles de variables.

Évaluation des performances

Pour permettre de mesurer l'impact de la manipulation des données, il est important de mesurer correctement les performances de prédiction du résultat du *tilt-test*. Cependant, le faible nombre d'observations disponibles fait augmenter le risque de biais dans l'estimation des performances. En effet, à la section 3.3.2, nous avons évoqué les difficultés d'estimation des performances de généralisation en présence d'un échantillon composé de peu d'observations. Or, dans notre application, l'échantillon analysé est composé de 84 patients, obligeant par conséquent à utiliser des méthodes d'estimation particulières. La figure 5.3 illustre le processus d'évaluation utilisé. Celui-ci est adapté de [Loughrey and Cunningham, 2005], où l'échantillon initial ($\mathcal{X}$) est divisé en deux groupes. Le premier groupe ($\mathcal{X}_A$) est composé de 48 patients, il est utilisé pour construire les modèles et déterminer leurs caractéristiques. Ce sous-ensemble de patients est lui-même partitionné en 6 sous-ensembles ($K = 6$), afin d'estimer plus précisément les performances de généralisation par validation croisée (*cf.* section 3.3.2.1). Ainsi, nous notons $\mathcal{X}_{A_k}$ et $\mathcal{X}_{V_k}$ ($k = 1, \ldots, K$), respectivement le k-ième sous-ensemble d'apprentissage et de validation. Le second groupe ($\mathcal{X}_T$) compte 36 patients, il est utilisé pour évaluer la reproductibilité des modèles construits. Les patients de ce sous-ensemble ne sont donc ni employés pour la construction, ni pour la validation des modèles ; ils permettent de donner une estimation « aveugle » des performances des modèles. Chaque sous-ensemble (apprentissage, validation et test) est construit aléatoirement, mais en conservant une prévalence équivalente à l'échantillon initial ($51\% \pm 2\%$).

Les modèles et les différentes configurations de leurs entrées sont comparés par les courbes de ROC et sont évalués en considérant la sensibilité (S_e), la spécificité (S_p), les valeurs prédictives (VPP et VPN) et l'indice global de l'aire sous la courbe de ROC (AUC). Les détails de ces indices sont donnés à la section 3.4.1 (page 108). Ainsi, nous noterons AUC_V (moyenne $\pm$ écart type) l'aire moyenne sous la courbe de ROC des K sous-ensembles de validation ($\mathcal{X}_{V_k}$, $k = 1, \ldots, K$) issus de la validation croisée et AUC_T l'aire sous la courbe de ROC pour le sous-ensemble de test.

5.3.1.3 Résultats

Le tableau 5.4 récapitule les résultats des modèles obtenus par les techniques génératives (FDL, FDQ, BN_{gauss} et $\text{BN}_{k\text{-ppv}}$) et les techniques discriminantes (PMC, $\text{SVM}_{\text{lin.}}$, $\text{SVM}_{\text{poly.}}$ et SVM_{RBF}). Dans ce tableau apparaissent également les variables sélectionnées optimisant les

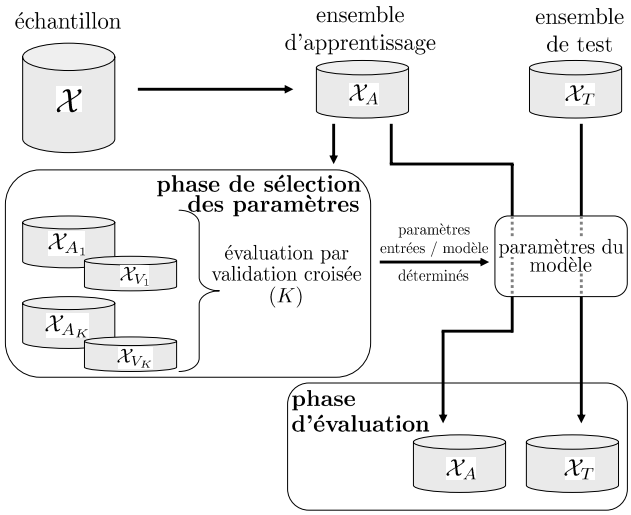


FIG. 5.3 – Processus d'estimation des performances des modèles [Loughrey and Cunningham, 2005].

performances de chacune des techniques de classification. Rappelons que la détermination des sous-ensembles de variables est réalisée de manière exhaustive, excepté pour les PMC, qui utilise une technique de sélection plus rapide (SFS). En effet, avec les PMC les évaluations du cas le plus favorable (1 entrée et 2 neurones) et du plus défavorable (15 entrées et 20 neurones) nécessitent respectivement en moyenne 60 s et 353 s ; sachant que l'évaluation[2] de chaque architecture est répétée 100 fois. Ainsi, par la méthode de sélection SFS, le temps estimé pour évaluer les 120 sous-ensembles est de 58 heures, contre 706 jours, si l'approche exhaustive avait été employée avec les PMC pour déterminer le sous-ensemble optimal parmi 32 767 possibles.

technique de classification	**nombre de variables pertinentes et sous-ensemble optimal de variables**	AUC_V	AUC_T
FDL	$1 : \{\ C\ \}$	$0,662 \pm 0,28$	$0,740$
FDQ	$3 : \{\ \hat{a}ge,\ Ht,\ Hb\ \}$	$0,740 \pm 0,20$	$0,436$
$\mathbf{BN}_{\text{gauss}}$	$7 : \{\ BSA,\ LW/FW,\ Ht,\ Hb,\ C,\ t_2,\ FC\ \}$	$0,726 \pm 0,11$	$0,511$
$\mathbf{BN}_{k\text{-ppv}}$	$2 : \{\ LW/FW\ \ Ht\ \}$	$0,760 \pm 0,13$	$0,641$
PMC	$3 : \{\ FC,\ LW/FW,\ Ht\ \}$	$0,802 \pm 0,15^{\dagger}$	$0,630^{\dagger}$
$\mathbf{SVM}_{\text{lin.}}$	$3 : \{\ \hat{a}ge,\ BSA,\ t_1\ \}$	$0,610 \pm 0,19$	$0,650$
$\mathbf{SVM}_{\text{poly.}}$	$6 : \{\ \hat{a}ge,\ TBW,\ LW/FW,\ Ht,\ Hb,\ FC\ \}$	$0,830 \pm 0,12$	$0,594$
$\mathbf{SVM}_{\text{RBF}}$	$5 : \{\ \hat{a}ge,\ BSA,\ TBW,\ Ht,\ t_1\ \}$	$0,800 \pm 0,14$	$0,607$

Note : Pour les $\mathbf{SVM}_{\text{poly.}}$, le degré du polynôme rendant les performances de validation optimale est de 3.

Pour les PMC, l'architecture optimale est composée de 19 neurones dans la couche cachée, obtenant sur 100 essais les performances moyennes suivantes : $AUC_V = 0,651 \pm 0,09$ et $AUC_T = 0,554 \pm 0,04$. † indique les performances du meilleur réseau parmi les 100 apprentissages réalisés.

TAB. 5.4 – Comparaison des performances des modèles de classification issus d'une sélection exhaustive des variables d'entrée pour prédire le résultat du *tilt-test* en position couchée.

La figure 5.4 compare pour chaque technique de classification, les courbes de ROC des sous-ensembles de validation et de test. Globalement, l'analyse des résultats (tableau 5.4 et figure 5.4) montre de meilleures performances pour les méthodes discriminantes (excepté $\text{SVM}_{\text{lin.}}$). De plus,

[2]La machine utilisée pour réaliser ces tests est un PC de bureau : Pentium IV 3 GHz 1, 5 Go de RAM.

nous devons relativiser les résultats obtenus par les PMC, qui ne reflètent pas réellement les performances optimales : ils les sous-estiment. En effet, pour cette technique, les sous-ensembles de variables sont obtenus par une heuristique qui ne permet pas d'évaluer toutes les combinaisons possibles, contrairement à ce qui est réalisé pour les autres techniques de classification. D'autre part, les différences significatives, observées pour chaque méthode entre les sous-ensembles de validation et de test, montrent une certaine difficulté des modèles construits à généraliser.

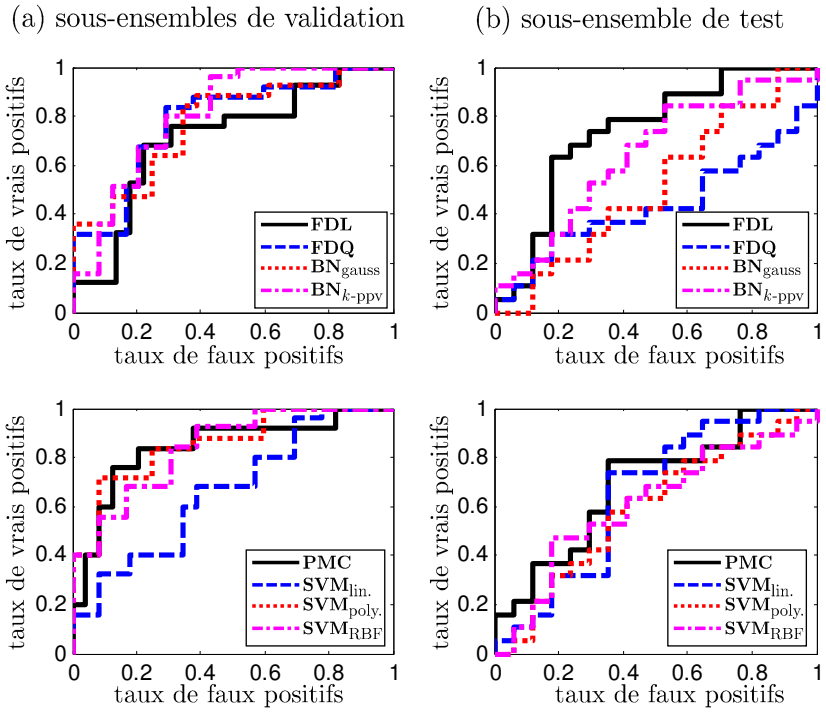


FIG. 5.4 – Comparaison des courbes de ROC des modèles de classification issus d'une sélection exhaustive des variables d'entrée pour prédire le résultat du *tilt-test* en position couchée.

Le tableau 5.5 expose les variables sélectionnées pour chaque technique de classification. Il est évidemment délicat de synthétiser à partir de ce tableau les variables les plus pertinentes comme étant celles les plus souvent sélectionnées. En effet, avec d'autres méthodes de classification, nous aurions pu trouver encore d'autres sous-ensembles de variables. Cependant, la vue globale donnée par le tableau 5.5 permet d'observer une forte concentration de variables sélectionnées liées aux paramètres physiologiques telles que, *âge*, *BSA* et *LW/FW*. De plus, parmi les variables les plus sélectionnées, nous trouvons également le taux d'hématocrite (*Ht*), qui permet de caractériser le sang et donc la dynamique cardiovasculaire (viscosité du sang, expliquant en partie son écoulement [Billat, 2003]). Or, la présence du taux d'hématocrite comme variable pertinente pour la prédiction de la syncope étonne au sens que la différence de valeurs entre les échantillons des patients positifs et négatifs aux *tilt-tests* n'est pas significative, comme le montre la figure 5.5. Cela contribue a montrer qu'une variable prise individuellement peut ne pas être pertinente, mais peut le devenir lorsqu'elle est associée à d'autres variables. Parmi les 15 variables pré-sélectionnées par les médecins, le taux d'hématocrite et la quantité d'hémoglobine semblent être incontournables pour prédire le résultat du *tilt-test*. Il est aussi surprenant de ne pas retrouver les informations liées aux pressions artérielles, et notamment, la pression artérielle systolique qui, comme nous l'avons vu dans le chapitre précédent, peut caractériser la survenue de symptômes inhérents à la syncope [Pitzalis *et al.*, 2002].

variables	**FDL**	**FDQ**	**BN**$_{\text{gauss}}$	**BN**$_{\text{k-ppv}}$	**PMC**	**SVM**$_{\text{lin.}}$	**SVM**$_{\text{poly.}}$	**SVM**$_{\text{RBF}}$
âge (*âge*)		√				√	√	√
surface corporelle (*BSA*)			√			√		√
volume plasmatique (*VolPlas*)								
fréquence cardiaque (*FC*)			√		√		√	
pression artérielle systolique (*PAS*)								
pression artérielle diastolique (*PAD*)								
pression pulsée (*PP*)								
eau totale (*TBW*)							√	√
rapport masse maigre / masse grasse (*LW/FW*)			√	√	√		√	
hématocrite (*Ht*)		√	√	√	√		√	√
hémoglobine (*Hb*)		√	√				√	
accélération positive de l'éjection ventriculaire (t_1)						√		√
partie négative de l'éjection ventriculaire (t_2)			√					
maximum de *dZ* (dZ_{max}/dt)								
indice de contractibilité (*C*)	√		√					

TAB. 5.5 – Récapitulatif des variables sélectionnées par le processus exhaustif pour chaque modèle de classification, afin de prédire le résultat du *tilt-test* en position couchée.

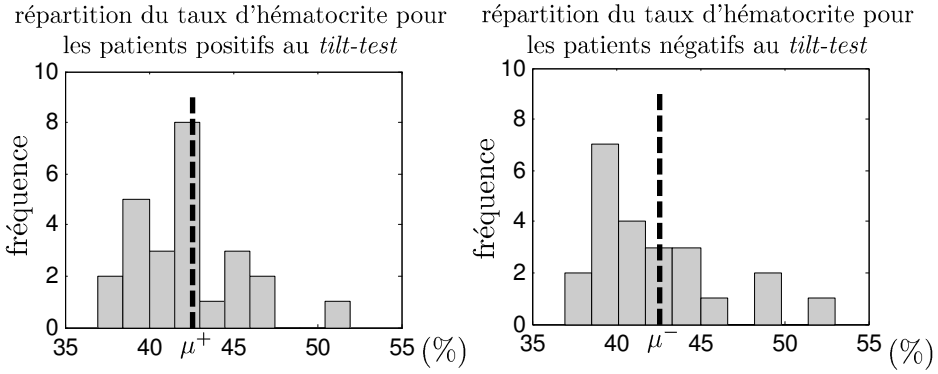


FIG. 5.5 – Comparaison de la distribution de la variable « hématocrite » pour les patients positifs et négatifs au *tilt-test*.

5.3.2 Extraction de caractéristiques pertinentes par combinaison des variables initiales

5.3.2.1 Introduction

Les expérimentations présentées dans cette section, publiées en partie dans [Feuilloy *et al.*, 2005b; Feuilloy *et al.*, 2005c], cherchent à améliorer les résultats de la prédiction du *tilt-test* obtenus précédemment. La démarche utilisée dans cette section ne consiste plus à chercher les sous-ensembles de variables optimisant la séparation des classes, mais des combinaisons des variables initiales obtenues par des techniques de projection linéaires et non linéaires.

Ainsi, nous analysons toujours la phase de repos du *tilt-test*, en considérant les 15 mêmes variables pré-sélectionnées pour les 84 patients issus de l'échantillon $\mathcal{E}_1$ (*cf.* section 5.3.1 et tableau 5.3).

5.3.2.2 Traitement préliminaire

Avant d'évaluer l'impact des méthodes de projection comme traitement des entrées des modèles, nous avons dans un premier temps cherché à éliminer parmi les 15 variables, les variables potentiellement « nuisibles ». Ces variables dites « nuisibles » ont été, à la section 2.1 et par le concours de [Langley, 1996; Gutierrez-Osuna, 2002], associées aux variables redondantes pouvant perturber certaines méthodes statistique. Ainsi, en étudiant la corrélation entre les variables, nous choisissons d'éliminer les variables qui ajoutent une redondance d'information. La figure 5.6 montre l'indice de corrélation linéaire obtenu entre chaque paire de variables. L'analyse de ces corrélations permet d'éliminer cinq variables (*TBW*, *Ht*, *C*, *FC* et *PP*) et donc de conserver les dix suivantes : *âge*, *BSA*, *VolPlas*, *LW/FW*, *Hb*, dZ_{max}/dt, t_1, t_2, *PAS* et *PAD*.

Rappelons qu'à la section 2.4.1, où nous avons repris la démonstration de [Guyon and Elisseeff, 2003], nous avons évoqué qu'une corrélation très élevée entre deux variables ne signifie pas nécessairement une absence de complémentarité entre ces variables. Ainsi, cette remarque amène difficilement à éliminer impunément les variables que nous considérons nuisibles (fortement corrélées à une majorité d'autres variables). Par conséquent, nous allons tout au long de l'étude sur les méthodes de projection, comparer l'ensemble de données initiales (avec les 15 variables) et l'ensemble de données prétraitées (avec les 10 variables).

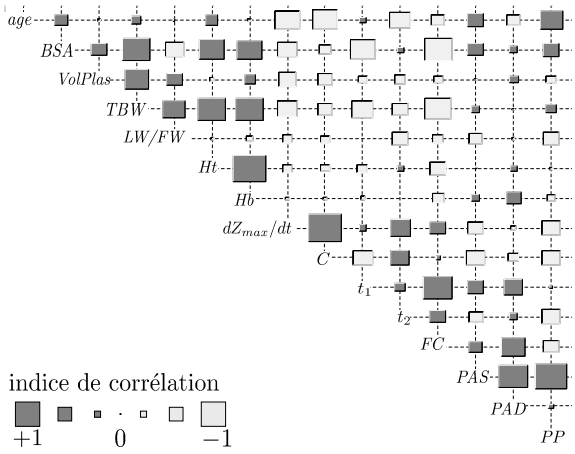


FIG. 5.6 – Indice de corrélation entre chaque paire de variables pré-sélectionnées.

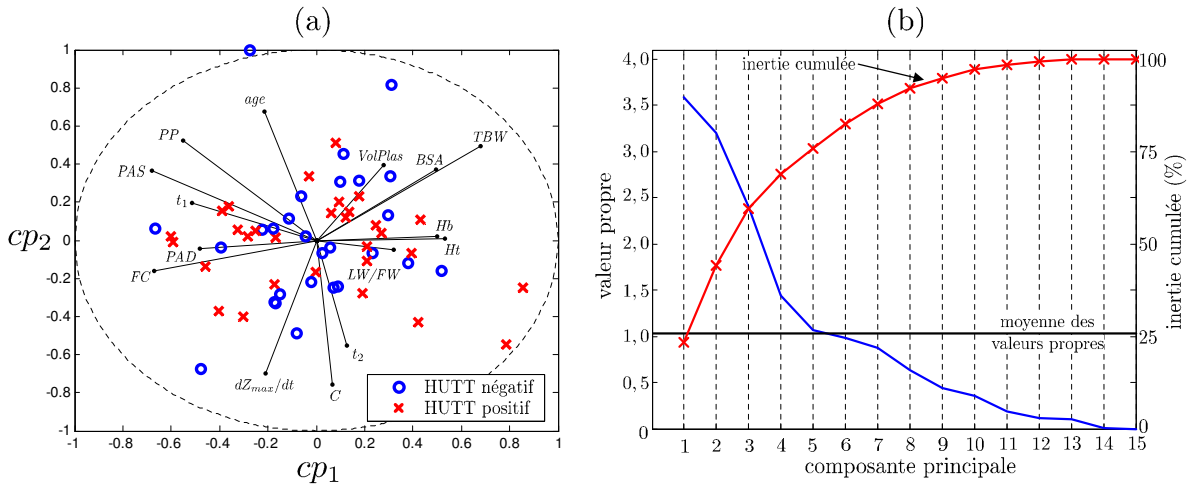
5.3.2.3 Méthodes

Les méthodes de classification et leur évaluation sont identiques à celles utilisées dans l'étude précédente (section 5.3.1.2). Ainsi, les méthodes discriminantes et génératives sont comparées et évaluées par le processus donné à la figure 5.3.

Les techniques de réduction de la dimensionnalité permettent par la diminution des variables d'entrées des modèles de faciliter leur apprentissage (*cf.* section 2.1). Par des techniques de projection, les caractéristiques issues de ces méthodes peuvent être vues comme une compression ou une synthèse des variables initiales. Ces techniques cherchent alors à extraire des composantes conservant le maximum d'informations des données originales. Comme montré à la section 2.3 (page 65), deux approches sont possibles. Une approche linéaire, telle que l'analyse en composantes principales (ACP), qui permet d'obtenir des combinaisons linéaires des variables initiales. L'autre approche utilisée est l'analyse en composantes curvilignes (ACC) qui, contrairement à l'ACP, permet de découvrir des relations non linéaires au sein de l'ensemble des données.

Par un choix judicieux de l'espace de projection, l'ACP réduit la dimension de l'espace d'entrée en conservant le maximum d'informations. La figure 5.7(a) donne une illustration de la représentation des observations dans le repère des deux premiers axes principaux. Nous pouvons y observer un chevauchement important entres les deux classes. Aussi, sur cette même figure apparaît la projection des variables, qui permet d'observer la corrélation entre les variables initiales et les deux premières composantes principales.

Dans la section 2.3.2.1, nous avons indiqué les méthodologies généralement employées pour choisir le nombre de composantes principales (CP) à conserver, en citant notamment la règle de Kaiser [Kaiser, 1961]. Celle-ci préconise de calculer la moyenne des valeurs propres et de conserver les CP associées aux valeurs propres dépassant cette moyenne. Cette méthode est admise uniquement par sa simplicité ; dans la pratique, elle possède une généralisation peu robuste. D'autre part, en observant les valeurs propres à la figure 5.7(b), leur décroissance légèrement irrégulière suggère de conserver les composantes principales avant le premier palier, donc les 4 premières CP. Cette interprétation, connue dans la littérature par le « test de l'éboulis » [Cattell, 1966], note qu'une faible variation entre des valeurs propres entraîne une faible augmentation de la restitution de l'information initiale dans les CP correspondantes à ces valeurs propres. Cependant, cette règle comme la règle de Kaiser, ne se généralise que très difficilement en pratique. C'est ainsi, qu'en présence de 15 variables, il nous a semblés plus judicieux d'évaluer empiriquement le nombre de composantes principales optimales. Dès lors, pour chaque modèle de classification, nous augmentons progressivement le nombre de CP et nous observons l'impact sur leurs performances de généralisation.

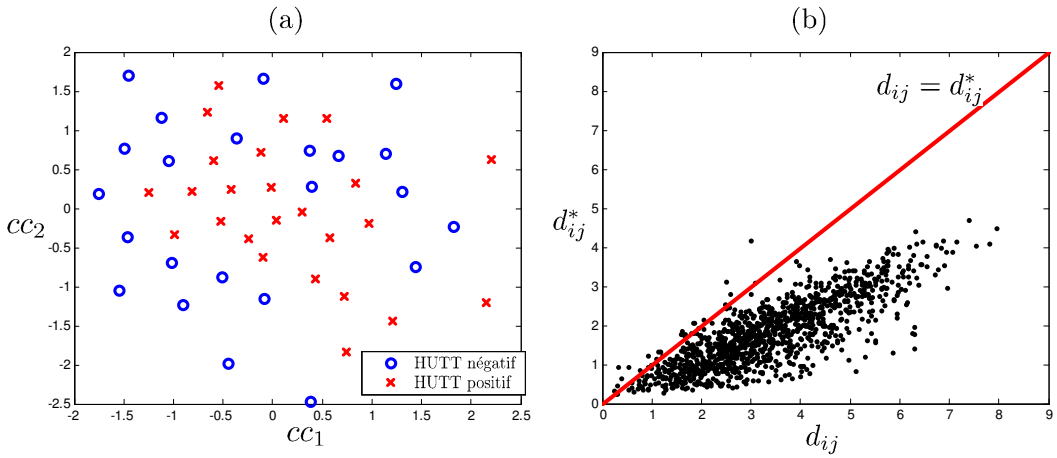


Note : (a) Représentation bidimensionnelle (*biplot*) des variables et des observations dans le plan des deux premiers axes factoriels. (b) Représentation de l'inertie expliquée sur chacune des composantes principales et de l'inertie cumulée sur les CP.

FIG. 5.7 – Représentation du *biplot* et de l'inertie expliquée sur chacune des nouvelles composantes issues de l'analyse en composantes principales.

Pour l'analyse en composantes curvilignes, le processus de sélection du nombre de composantes curvilignes est effectué également de manière empirique : en augmentant progressivement le nombre de composantes curvilignes et en observant les performances des modèles. La figure 5.8 illustre le résultat de la projection par l'ACC sur deux composantes curvilignes (CC_1 et CC_2). La différence observable face à l'ACP porte sur la répartition des classes, où dans le cas de l'ACC, celles-ci sont plus nettement séparables.

L'étape de traitement des entrées par l'ACP et l'ACC est effectuée de plusieurs manières afin de comparer leur efficacité et leur complémentarité. Ainsi, dans un premier temps nous évaluons les deux méthodes de projection indépendamment. Cette première analyse, illustrée par le processus de la figure 5.9(a), considère les sous-ensembles de 15 et de 10 variables. Dans un second temps, nous considérons uniquement les 15 variables et nous associons les deux méthodes de projection afin d'évaluer leur complémentarité. Ainsi, une ACC est effectuée sur les composantes principales



Note : (a) Observations projetées dans les deux premières composantes curvilignes. (b) Représentation « $dy - dx$ » : comparaison des distances de chaque couple d'observations entre l'espace initial (d^*_{ij}) et l'espace réduit d_{ij}.

FIG. 5.8 – Résultats de la projection en deux dimensions de l'analyse en composantes curvilignes.

issues d'une ACP et une ACP est effectuée sur les composantes curvilignes issues d'une ACC, comme le montre la figure 5.9(b).

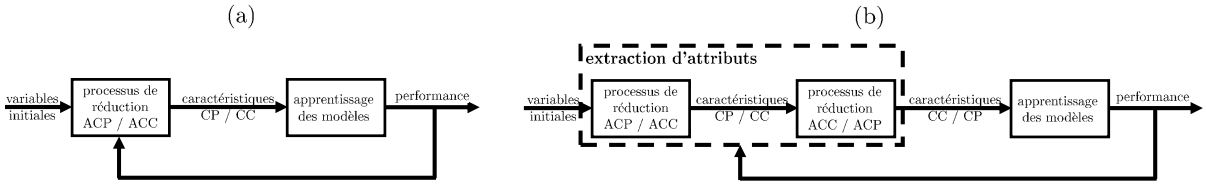


FIG. 5.9 – Processus d'extraction des caractéristiques pertinentes pour prédire le résultat du *tilt-test* en position couchée.

5.3.2.4 Résultats

Les tableaux 5.6 et 5.7 reportent respectivement pour l'ACP et l'ACC, les performances des modèles de classification sur les ensembles de validation (AUC_V) et de test (AUC_T). Dans ces tableaux, nous observons également le nombre de composantes (principales et curvilignes) conservées, donnant ainsi l'aperçu de la réduction de dimension effectuée.

Les résultats donnés aux tableaux 5.6 et 5.7 montrent que le pré-traitement, qui a amené à éliminer cinq variables jugées fortement redondantes, n'améliore pas significativement les performances des modèles. D'autre part, l'ACP, effectuée sur les 15 variables initiales, donne globalement de meilleures performances avec moins de composantes principales que lorsque l'ACP est réalisée sur les 10 variables.

L'analyse des tableaux 5.6 et 5.7 montre globalement de meilleures performances lors de l'utilisation de l'ACC, où la plus grande différence est obtenue par la méthode $\text{SVM}_{\text{lin.}}$. Les aires sous la courbe de ROC du sous-ensemble de validation sont égales à $0,461$ et $0,631$, respectivement pour l'ACP et l'ACC. Cependant, nous pouvons noter que pour obtenir des performances optimales, l'ACP requiert moins de composantes que l'ACC, notamment pour les PMC, où avec uniquement les deux premières composantes principales l'aire moyenne de validation obtenue est de $0,811$.

technique de classification	**nombre de CP conservées**	AUC_V	AUC_T
FDL	3 $(4)^\dagger$	$0,506 \pm 0,20$ $(0,412 \pm 0,24)^\dagger$	$0,635$ $(0,616)^\dagger$
FDQ	3 $(4)^\dagger$	$0,621 \pm 0,19$ $(0,567 \pm 0,16)^\dagger$	$0,619$ $(0,601)^\dagger$
$\mathbf{BN}_{\text{gauss}}$	3 $(4)^\dagger$	$0,631 \pm 0,12$ $(0,609 \pm 0,12)^\dagger$	$0,662$ $(0,625)^\dagger$
$\mathbf{BN}_{k\text{-ppv}}$	3 $(4)^\dagger$	$0,629 \pm 0,21$ $(0,568 \pm 0,20)^\dagger$	$0,517$ $(0,591)^\dagger$
PMC	2 $(6)^\dagger$	$0,811 \pm 0,11$ $(0,724 \pm 0,16)^\dagger$	$0,737$ $(0,755)^\dagger$
$\mathbf{SVM}_{\text{lin.}}$	4 $(2)^\dagger$	$0,464 \pm 0,24$ $(0,412 \pm 0,18)^\dagger$	$0,511$ $(0,523)^\dagger$
$\mathbf{SVM}_{\text{poly.}}$	8 $(5)^\dagger$	$0,692 \pm 0,20$ $(0,640 \pm 0,16)^\dagger$	$0,662$ $(0,523)^\dagger$
$\mathbf{SVM}_{\text{RBF}}$	4 $(5)^\dagger$	$0,692 \pm 0,20$ $(0,619 \pm 0,20)^\dagger$	$0,590$ $(0,635)^\dagger$

Note : $(\cdot)^\dagger$ indique que le processus de réduction de la dimension est réalisé sur l'ensemble de données pré-traitées composé de 10 variables initiales.

Pour les PMC, l'architecture optimale est composée de 13 neurones dans la couche cachée, obtenant sur 100 essais les performances moyennes suivantes : $AUC_V = 0,751 \pm 0,12$ et $AUC_T = 0,684 \pm 0,10$. Les valeurs reportées dans le tableau indiquent les performances du meilleur réseau parmi les 100 apprentissages réalisés.

TAB. 5.6 – Comparaison des performances des modèles de classification issus d'une analyse en composantes principales (ACP) pour prédire le résultat du *tilt-test* en position couchée.

technique de classification	**nombre de CC conservées**	AUC_V	AUC_T
FDL	4 $(2)^\dagger$	$0,589 \pm 0,20$ $(0,724 \pm 0,19)^\dagger$	$0,628$ $(0,421)^\dagger$
FDQ	11 $(5)^\dagger$	$0,693 \pm 0,16$ $(0,662 \pm 0,23)^\dagger$	$0,622$ $(0,715)^\dagger$
$\mathbf{BN}_{\text{gauss}}$	11 $(4)^\dagger$	$0,693 \pm 0,16$ $(0,642 \pm 0,09)^\dagger$	$0,653$ $(0,613)^\dagger$
$\mathbf{BN}_{k\text{-ppv}}$	1 $(5)^\dagger$	$0,631 \pm 0,13$ $(0,654 \pm 0,32)^\dagger$	$0,356$ $(0,594)^\dagger$
PMC	5 $(4)^\dagger$	$0,801 \pm 0,22$ $(0,712 \pm 0,15)^\dagger$	$0,793$ $(0,709)^\dagger$
$\mathbf{SVM}_{\text{lin.}}$	4 $(2)^\dagger$	$0,631 \pm 0,15$ $(0,600 \pm 0,19)^\dagger$	$0,740$ $(0,452)^\dagger$
$\mathbf{SVM}_{\text{poly.}}$	5 $(10)^\dagger$	$0,716 \pm 0,15$ $(0,656 \pm 0,32)^\dagger$	$0,508$ $(0,594)^\dagger$
$\mathbf{SVM}_{\text{RBF}}$	5 $(5)^\dagger$	$0,620 \pm 0,16$ $(0,631 \pm 0,17)^\dagger$	$0,635$ $(0,709)^\dagger$

Note : $(\cdot)^\dagger$ indique que le processus de réduction de la dimension est réalisé sur l'ensemble de données pré-traitées composé de 10 variables initiales.

Pour les PMC, l'architecture optimale est composée de 20 neurones dans la couche cachée, obtenant sur 100 essais les performances moyennes suivantes : $AUC_V = 0,733 \pm 0,17$ et $AUC_T = 0,701 \pm 0,08$. Les valeurs reportées dans le tableau indiquent les performances du meilleur réseau parmi les 100 apprentissages réalisés.

TAB. 5.7 – Comparaison des performances des modèles de classification issus d'une analyse en composantes curvilignes (ACC) pour prédire le résultat du *tilt-test* en position couchée.

Dans le cadre de l'ACP, nous avons introduit, à la section 2.3.2.1, la démarche qui permet d'évaluer la représentation des variables initiales dans les composantes principales. Au chapitre 6, cette démarche sera détaillée plus finement, afin de montrer comment obtenir une représentation graphique de la contribution des variables dans les CP. Un exemple est donné à la figure 5.10, il illustre, dans le cadre de cette étude, la représentation et donc la contribution des 15 variables initiales dans les 6 premières CP.

Ainsi, la figure 5.10 permet d'observer les variables qui ont le plus contribué à la formation des deux premières CP, optimisant ainsi la prédiction par les PMC ($AUC_V = 0,811 \pm 0,11$). Celles-ci sont : *TBW*, *Ht*, *Hb*, *PAS* et *PAD*. Aussi, le tableau 5.6 montre que toutes les méthodes

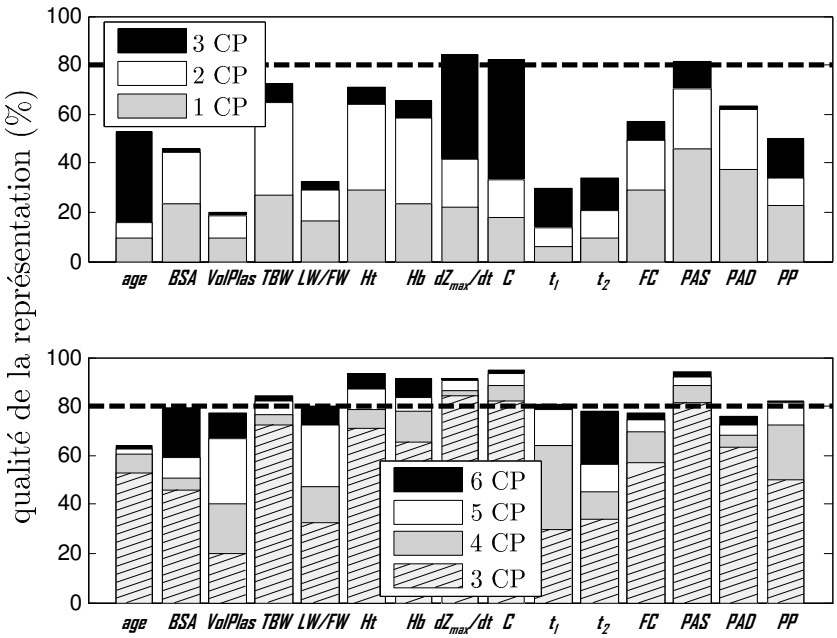


FIG. 5.10 – Qualité de la représentation des variables pré-sélectionnées dans les composantes principales utilisées pour prédire le résultat du *tilt-test* en position couchée.

génératives privilégient l'utilisation des trois premières CP, où avec une contribution supérieure à 80%[3], les variables dZ_{max}/dt, C et PAS sont fortement représentées. En ajoutant la 4-ième CP, SVM_{RBF} optimise ses performances de prédiction ($AUC_V = 0,692 \pm 0,20$), et cela, par une forte contribution supplémentaire de la variable t_1, dont la représentation double à la 4-ième CP.

L'interprétation des composantes issues de l'ACP est relativement facile et intuitive, contrairement à celles issues de l'ACC. Rappelons qu'à la section 2.3.4, nous avons évoqué la difficulté d'interprétation des composantes curvilignes et plus généralement de toutes les composantes issues de processus de réduction non linéaire. Sans outil analytique comparable à l'ACP, les composantes issues de l'ACC ne peuvent, en l'état actuel des choses, être décomposées afin d'extraire la représentation des variables initiales dans les CC. Au chapitre suivant, à la section 6.4, nous donnerons un processus permettant d'adapter la méthodologie de l'ACP à l'ACC afin d'obtenir la contribution des variables aux CC. Une fois la méthode détaillée, nous pourrons enfin identifier, à la section 6.5, la composition des CC qui ont le plus contribué à la prédiction du résultat du *tilt-test*.

Le tableau 5.8 montre les performances des modèles de classification en fonction du processus de réduction utilisant successivement, l'ACP puis l'ACC ou, l'ACC puis l'ACP, nous les notons respectivement ACP → ACC et ACC → ACP. Avec ces processus de réduction, la dimension de l'espace de projection, donc le nombre d'entrées des modèles, diminue. Ce même tableau permet d'observer globalement une réduction de l'espace à 3 ou 4 dimensions. Les résultats obtenus montrent majoritairement de meilleures performances pour la transformation ACC → ACP, avec notamment les PMC qui obtiennent une aire moyenne sous la courbe de ROC de $0,894$.

Ces processus de projection, mêlant l'ACP et l'ACC, ne permettent pas d'obtenir la contribution des variables initiales dans les nouvelles composantes. Ainsi, comme avec l'ACC, nous

[3]Le seuil de 80% est largement employé dans la littérature afin de signifier une bonne représentation d'une variable au sein des composantes principales [Georgin, 2002].

technique de classification	**nombre de composantes ACP → ACC**	AUC_V	AUC_T
FDL	$11 \rightarrow 3^{\dagger}$	$0,633 \pm 0,16$	$0,622$
FDQ	$6 \rightarrow 3^{\dagger}$	$0,663 \pm 0,19$	$0,681$
$\mathbf{BN}_{\text{gauss}}$	$6 \rightarrow 3^{\dagger}$	$0,663 \pm 0,19$	$0,718$
$\mathbf{BN}_{k\text{-ppv}}$	$6 \rightarrow 3^{\dagger}$	$0,674 \pm 0,18$	$0,529$
PMC	$4 \rightarrow 4^{\dagger}$	$0,768 \pm 0,09$	$0,774$
$\mathbf{SVM}_{\text{lin.}}$	$11 \rightarrow 3^{\dagger}$	$0,653 \pm 0,16$	$0,635$
$\mathbf{SVM}_{\text{poly.}}$	$7 \rightarrow 4^{\dagger}$	$0,787 \pm 0,20$	$0,517$
$\mathbf{SVM}_{\text{RBF}}$	$15 \rightarrow 3^{\dagger}$	$0,681 \pm 0,21$	$0,693$

Note : † indique le nombre final de composantes introduites dans les modèles de classification pour leur apprentissage.

Pour les $\mathbf{SVM}_{\text{poly.}}$, le degré du polynôme rendant les performances de validation optimale est de 3.

Pour les PMC, l'architecture optimale est composée de 19 neurones dans la couche cachée, obtenant sur 100 essais les performances moyennes suivantes : $AUC_V = 0,694 \pm 0,15$ et $AUC_T = 0,670 \pm 0,09$. Les valeurs reportées dans le tableau indiquent les performances du meilleur réseau parmi les 100 apprentissages réalisés.

technique de classification	**nombre de composantes ACC → ACP**	AUC_V	AUC_T
FDL	$4 \rightarrow 3^{\dagger}$	$0,693 \pm 0,17$	$0,625$
FDQ	$8 \rightarrow 6^{\dagger}$	$0,693 \pm 0.17$	$0,737$
$\mathbf{BN}_{\text{gauss}}$	$11 \rightarrow 11^{\dagger}$	$0,693 \pm 0,16$	$0,653$
$\mathbf{BN}_{k\text{-ppv}}$	$6 \rightarrow 4^{\dagger}$	$0,684 \pm 0,16$	$0,489$
PMC	$15 \rightarrow 3^{\dagger}$	$0,894 \pm 0,12$	$0,780$
$\mathbf{SVM}_{\text{lin.}}$	$4 \rightarrow 3^{\dagger}$	$0,651 \pm 0,12$	$0,675$
$\mathbf{SVM}_{\text{poly.}}$	$14 \rightarrow 5^{\dagger}$	$0,673 \pm 0,19$	$0,536$
$\mathbf{SVM}_{\text{RBF}}$	$12 \rightarrow 4^{\dagger}$	$0,757 \pm 0,14$	$0,567$

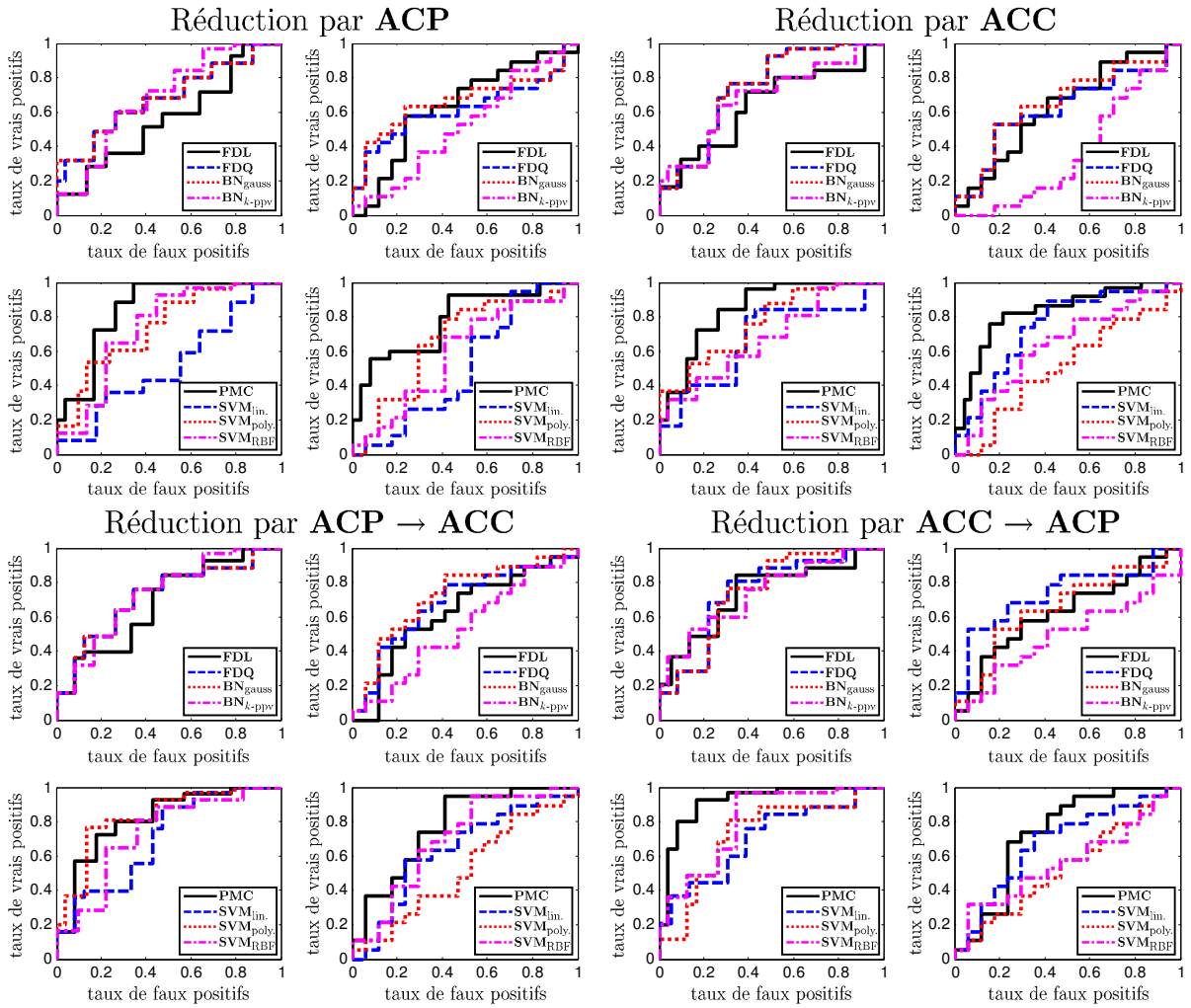
Note : † indique le nombre final de composantes introduites dans les modèles de classification pour leur apprentissage.

Pour les PMC, l'architecture optimale est composée de 14 neurones dans la couche cachée, obtenant sur 100 essais les performances moyennes suivantes : $AUC_V = 0,848 \pm 0,11$ et $AUC_T = 0,714 \pm 0,12$. Les valeurs reportées dans le tableau indiquent les performances du meilleur réseau parmi les 100 apprentissages réalisés.

TAB. 5.8 – Comparaison des performances des modèles de classification issus d'un traitement, associant l'ACP et l'ACC, pour prédire le résultat du *tilt-test* en position couchée.

donnerons, à la section 6.5, des informations sur les composantes les plus pertinentes et notamment celles ayant contribué avec les PMC à obtenir une AUC_V de $0,894$.

La figure 5.11 illustre pour chaque technique de réduction les courbes de ROC des sous-ensembles de validation et de test pour chaque modèle de classification. L'analyse de ces courbes donne plus de clarté à la comparaison des méthodes, nous permettant ainsi d'observer la grande efficacité des approches discriminantes par rapport aux approches génératives. Aussi, il est intéressant de noter que, globalement, l'estimation aveugle des performances des modèles sur l'ensemble de test donne des résultats relativement homogènes avec ceux de l'ensemble de validation ; preuve des bonnes capacités de généralisation des méthodes de traitement et de classification, contrairement à ce qui a pu être perçu dans l'analyse précédente, lors de la sélection exhaustive.



Note : Comme pour la figure 5.4, pour chacune des techniques de projection employées, les courbes de ROC sur les ensembles de validation et de test sont représentées respectivement à gauche et à droite.

FIG. 5.11 – Comparaison des courbes de ROC des modèles de classification issus de processus de projection (linéaire et non linéaire) pour prédire le résultat du *tilt-test* en position couchée.

5.3.3 Discussions et conclusions

La prédiction de la syncope récurrente inexpliquée requiert l'utilisation du *tilt-test* comme procédure de diagnostic qui souffre cependant de sa longue durée (approximativement une heure dans le cas où les symptômes de la syncope n'apparaissent pas). Réduire sa durée est de première importance. Les études présentées précédemment, et publiées dans [Feuilloy *et al.*, 2005b; Feuilloy *et al.*, 2005c; Feuilloy *et al.*, 2005a], proposent une prédiction en considérant uniquement les dix premières minutes de la phase de repos du *tilt-test*. Cette discussion va alors permettre de résumer les remarques.

Le tableau 5.9 et la figure 5.12 récapitulent pour chaque processus de réduction (sélection et projection) les résultats des meilleures méthodes de classification. Nous pouvons ainsi remarquer l'absence des méthodes génératives, qui ont certainement des difficultés de généralisation, dues au faible nombre d'observations disponibles pour l'apprentissage. En effet, ce faible nombre peut entraîner un biais important sur l'estimation des densités de probabilité. Malgré la généralisation de l'utilisation des lois normales dans le milieu médical pour estimer les densités de variables, il apparaît au vu des résultats, que cette hypothèse, dans notre cas, n'est peut être pas raisonnable.

technique de réduction	**technique de classification**	S_e (%)	S_p (%)	VPP (%)	VPN (%)	AUC
sélection exhaustive	$\text{SVM}_{\text{poly.}}$	76 ± 16 $(58)^{\dagger}$	83 ± 13 $(59)^{\dagger}$	83 ± 14 $(61)^{\dagger}$	78 ± 13 $(56)^{\dagger}$	$0,830 \pm 0,12$ $(0,594)^{\dagger}$
projection ACP	PMC	88 ± 14 $(71)^{\dagger}$	74 ± 16 $(68)^{\dagger}$	88 ± 13 $(67)^{\dagger}$	80 ± 11 $(72)^{\dagger}$	$0,811 \pm 0,11$ $(0,737)^{\dagger}$
projection ACC	PMC	82 ± 14 $(76)^{\dagger}$	75 ± 39 $(74)^{\dagger}$	83 ± 23 $(72)^{\dagger}$	73 ± 37 $(78)^{\dagger}$	$0,801 \pm 0,22$ $(0,793)^{\dagger}$
projection ACP $\rightarrow$ ACC	$\text{SVM}_{\text{poly.}}$	80 ± 19 $(84)^{\dagger}$	71 ± 40 $(18)^{\dagger}$	83 ± 19 $(53)^{\dagger}$	70 ± 40 $(50)^{\dagger}$	$0,787 \pm 0,20$ $(0,517)^{\dagger}$
projection ACC $\rightarrow$ ACP	PMC	96 ± 10 $(84)^{\dagger}$	84 ± 20 $(71)^{\dagger}$	87 ± 15 $(80)^{\dagger}$	97 ± 8 $(76)^{\dagger}$	$0,894 \pm 0,12$ $(0,780)^{\dagger}$

Note : † indique que les résultats sont obtenus sur le sous-ensemble de test.

TAB. 5.9 – Récapitulatif des meilleures associations des méthodes de classification et de réduction (sélection et extraction) pour prédire le résultat du *tilt-test* en position couchée.

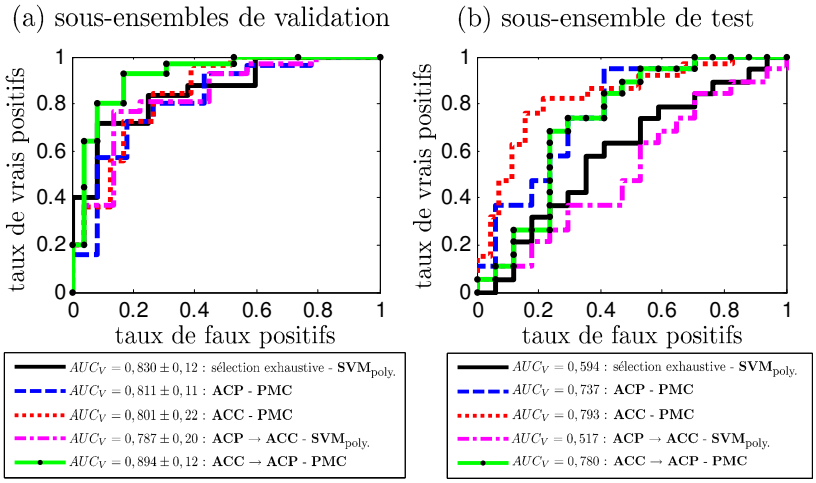


FIG. 5.12 – Comparaison des courbes de ROC des modèles de classification issus de processus de réduction de dimension (sélection et extraction) pour prédire le *tilt-test* en position couchée.

L'utilisation d'un sous-ensemble de test, en plus des sous-ensembles de validation, permet d'estimer sans biais les performances de généralisation des modèles construits. En effet, même si l'utilisation de la validation croisée permet de réduire le biais d'estimation des performances des modèles, l'apprentissage et le choix des modèles sont, malgré tout, affectés par quelques patients particuliers appartenant aux sous-ensembles de validation. Ainsi, l'estimation sans biais, par l'intermédiaire de l'évaluation du sous-ensemble de test, permet d'observer de moins bonnes performances pour les modèles issus d'une sélection de variables. Par cette approche (méthode exhaustive, tableau 5.9), l'écart entre les performances de validation et de test est très important. En effet, la sensibilité/spécificité de test et de validation est respectivement de 58%/59% et 76%/83% (*cf.* tableau 5.9). Nous aurions pu imaginer que, dans un premier temps, le manque de reproductibilité est dû à la non homogénéité des patients des sous-ensembles de validation et de test. Cependant, pour les techniques de projection la disparité des performances entre ces sous-ensembles n'est pas aussi évidente. Dès lors, nous pouvons peut être incriminer les modèles, ou plutôt, les variables issues des recherches exhaustives comme étant défavorables à la généralisation.

Lors des explorations exhaustives des combinaisons de variables, les PMC et SVM (noyaux polynomial et gaussien) ont obtenu les meilleures performances et ont montré la forte influence des variables suivantes : *âge*, *FC*, *TBW*, *LW/FW* et *Ht*. Quant aux modèles basés sur l'ACP, ils ont été influencés par les variables suivantes : *TBW*, *Ht*, *Hb*, dZ_{max}/dt, *C* et *PAS*.

Parmi les 15 variables pré-sélectionnées par les médecins, le taux d'hématocrite et la quantité d'hémoglobine semblent être incontournables pour la prédiction du résultat du *tilt-test*. Or, ces deux variables étant relativement difficiles et coûteuses à obtenir, il serait judicieux de s'en passer. Il reste alors trois autres types de variables fortement représentés : les mesures de fréquence cardiaque, les mesures de la pression artérielle et les mesures du signal d'impédancemétrie thoracique (Z). Notons, que l'eau totale (TBW) et le rapport masse maigre/masse grasse (LW/FW) peuvent être obtenus à partir du signal Z en utilisant des formules appropriées. Ainsi, la présence des variables TBW et LW/FW et des indices extraits directement sur la dérivée de Z (dZ_{max}/dt, t_1 et C) montrent l'importance du signal d'impédancemétrie thoracique, comme l'ont déjà observé [Bellard *et al.*, 2003; Schang *et al.*, 2003; Schang *et al.*, 2006]. D'autre part, la dynamique cardiovasculaire, caractérisée en partie par le taux d'hématocrite qui paraissait si important, pourrait être exprimée par l'intermédiaire de l'indice dZ_{max}/dt qui, comme explicité dans le chapitre précédent, est fortement corrélé au volume d'éjection systolique (VES) et donc au débit sanguin [Charloux *et al.*, 2000]. [Yammanouchi *et al.*, 1996] ont pu observer pour les patients positifs, une diminution rapide du VES corrélée à une augmentation de l'indice dZ_{max}/dt et de contractibilité (C). Ces observations confirment la pertinence des résultats obtenus, permettant d'envisager l'utilisation unique du signal d'impédancemétrie thoracique pour la prédiction de la syncope ; nous analyserons cette situation à la section 5.5.

Dans les expérimentations de cette section, les performances sont obtenues par l'association de variables. Ainsi la pression artérielle et la fréquence cardiaque ont influencé la prédiction au même titre que le signal Z. Rappelons qu'avec uniquement la *PAS*, [Pitzalis *et al.*, 2002] ont prédit, avec une sensibilité de 85%, le résultat du *tilt-test* durant les 15 premières minutes de la phase basculée. D'autre part, [Mallat *et al.*, 1997], en étudiant l'influence de la fréquence cardiaque pendant les 6 premières minutes de la phase basculée du *tilt-test*, ont prédit le résultat du test avec une sensibilité de 96%. Ces études, [Mallat *et al.*, 1997], [Pitzalis *et al.*, 2002] et [Bellard *et al.*, 2003; Schang *et al.*, 2003], ont considéré de manière indépendante chaque groupe de variables ; respectivement la fréquence cardiaque, la pression artérielle et le signal Z. En associant ces variables, notre étude, publiée également [Feuilloy *et al.*, 2005c], a permis d'améliorer la prédiction du résultat du *tilt-test* et cela **durant la phase de repos**, en obtenant une sensibilité de 96% et une spécificité de 84%. Le tableau 5.10 récapitule les résultats obtenus des différentes études citées précédemment ; notons que ce tableau est repris de la section 4.3, où les études ont été détaillées.

Notons enfin que par l'utilisation des méthodes de projection, même si des variables, telles que t_1 ou t_2 (voir figure 5.10), sont faiblement représentées, il n'en demeure pas moins qu'elles ont participé à la création des composantes, et donc, qu'elles ont contribué à obtenir les performances de prédiction. Il est alors délicat d'affirmer la non utilité de certaines variables. Malgré l'efficacité démontrée par les méthodes de projection, l'inconvénient majeur de ces approches est, par conséquent, le manque d'informations sur le rôle joué par les variables originales dans la construction des nouvelles composantes [Illouz and Jardino, 2001; Guérif, 2006]. En d'autres termes, la description des entrées des modèles ne sont pas clairement définies. D'autant plus pour les méthodes de projection non linéaire qui ne disposent pas d'outils analytiques pour extraire la représentation des variables dans les nouvelles composantes, comme le fait l'ACP. Cela influence considérablement le choix pour certains auteurs quant à opter pour des méthodes de sélection plutôt que pour des méthodes de projection, afin de réduire la dimension d'un problème. Rappe-

étude	**période du *tilt-test***	$S_e\,(\%)$	$S_p\,(\%)$	$VPP\,(\%)$	$VPN\,(\%)$
ACC → ACP PMC	période de repos	87 ± 6	94 ± 4	93 ± 4	89 ± 5
		96 ± 10 (84)*	84 ± 20 (71)*	87 ± 15 (80)*	97 ± 8 (76)*
[Schang *et al.*, 2003]	période de repos	100	100	100	100
		69	73	67	75
[Bellard *et al.*, 2003]	période de repos	68	63	63	68
		–	–	–	–
[Bellard *et al.*, 2003]†	5 à 10-ième min du basculement	68	70	68	70
		–	–	–	–
[Bellard *et al.*, 2003]‡	5 à 10-ième min du basculement	50	97	93	67
		–	–	–	–
[Pitzalis *et al.*, 2002]	1re à 15-ième min du basculement	93	58	28	98
		80	85	57	94
[Mallat *et al.*, 1997]	1re à 6-ième min du basculement	100	89	96	100
		96	87	75	98

Note : † étude analysant uniquement les variables liées au signal dZ/dt. ‡ étude analysant les variables liées au signal dZ/dt et FC, PAS, PAD et PD.
Les lignes en blanc correspondent aux résultats obtenus lors d'une analyse rétrospective et les lignes grisées correspondent aux résultats obtenus lors d'une analyse prospective. * Indique que les résultats sont obtenus sur le sous-ensemble de test ; ce sous-ensemble n'est pas utilisé pour l'apprentissage et la sélection du modèle, il donne les performances sans aucune forme de biais.

TAB. 5.10 – Comparaison des résultats de prédiction de la réponse du *tilt-test* en position couchée, avec les principales études analysant la syncope inexpliquée.

lons qu'au chapitre 6, nous proposerons une adaptation du processus d'extraction d'information de l'ACP aux processus de projection non linéaire, afin d'améliorer l'interprétation des méthodes non linéaires en les rendant encore plus efficaces.

5.4 Recherche d'indices prédictifs du résultat du *tilt-test* durant les deux périodes de l'examen : couchée et basculée

5.4.1 Introduction

Dans cette section, nous allons pour prédire le résultat du *tilt-test* considérer les deux phases suivantes : la période couchée et les 10 premières minutes du basculement. Pour cela, nous utilisons l'échantillon $\mathcal{E}_1$ (composé de 84 patients) et ses 70 variables disponibles (voir tableaux 5.1 et 5.2). Cependant, comme noté dans ces tableaux, des valeurs manquantes apparaissent et parfois en quantité importante pour certains patients. Ainsi, nous avons fait le choix de ne pas chercher à les remplacer, mais simplement d'éliminer les patients pour lesquels les variables sont manquantes. À l'issue de ce pré-traitement, nous obtenons 58 patients pour lesquels les 70 variables ont été mesurées. Les 58 patients (31 hommes et 27 femmes), de ce nouvel échantillon, ont une moyenne d'âge de 40 ans et un écart type de 13 ans (variant de 18 à 73 ans). Dans cet échantillon, 26 patients se sont révélés positifs au *tilt-test* : la prévalence observée de cet échantillon est donc de 45%.

Les travaux présentés dans cette section, et publiés dans [Feuilloy *et al.*, 2006a], ont comme objectif de chercher les variables et les sous-ensembles de variables pertinentes pour la prédiction du résultat du *tilt-test*. Afin de faciliter la lecture des résultats, nous allons au travers du tableau 5.11, attribuer un indice pour chacune des 70 variables.

variable	indice
âge	1
sexe	2
taille	3
poids	4
eau totale théorique	5
surface corporelle	6
volume plasmatique mesurée	7
eau totale mesurée	36
volume plasmatique théorique/eau totale	37
volume plasmatique mesuré/eau totale	38
eau totale	39
masse grasse	40
pourcentage de masse grasse	41
masse maigre	42
pourcentage de masse maigre	43
masse maigre/masse grasse	44
hématocrite	45
hémoglobine	46
osmolarité	47
variation initiale max. de FC.	48
évaluation du rebond initial de FC.	49
variation initiale de PAS	50
variation initiale de PAD	51
pression artérielle minimale	52
chute de PAS maximale	53
chute de PAD maximale	54
fréquence cardiaque maximale	55
élévation maximale de FC.	56
delta de PAS maximal	57
delta de PAD maximal	58

variable	indice au repos	indice au TILT	indice des deltas
fréquence cardiaque (FC)	8	17	27
pression artérielle systolique (PAS)	9	18	28
pression artérielle diastolique (PAD)	10	19	29
pression artérielle moyenne (PM)	11	20	30
pression artérielle pulsée	12	21	31
vitesse moyenne artère cérébrale	13	22	32
index de résistance	14	23	33
résistance vasculaire cérébrale	15	24	34
index de pulsabilité	16	25	35
variation du volume du mollet		26	
accélération positive de l'éjection ventriculaire (t_1)	59	63	67
maximum de dZ ($\frac{dZ_{max}}{dt}$)	60	64	68
indice de contractibilité (C)	61	65	69
partie négative de l'éjection ventriculaire (t_2)	62	66	70

Note : Indices au repos : variables recueillies pendant la phase de repos. Indices au TILT : attributs recueillis durant les 10 premières minutes du basculement et indices des deltas : rapport entre les attributs des deux phases.

TAB. 5.11 – Récapitulatif des variables et de leur indice utilisées pour prédire l'apparition des symptômes de la syncope durant les deux positions du *tilt-test*.

Dans cette étude, deux facteurs sont considérés : le coût calculatoire et les performances des sous-ensembles de variables sélectionnées.

5.4.2 Méthodes

5.4.2.1 Techniques de sélection de variables

Comme dans l'étude présentée à la section 5.3.1, nous cherchons les sous-ensembles de variables capables d'optimiser la séparation des classes et donc de prédire au mieux le résultat du *tilt-test*. Précédemment, nous avons considéré 15 variables (donc 32 767 combinaisons possibles), rendant possible la recherche exhaustive. Dans cette analyse, les 70 variables initiales sont utilisées, le nombre de combinaisons croît donc considérablement, en étant égal à $2^{70} \approx 1,18 \cdot 10^{21}$. Dès lors, une recherche exhaustive n'est plus envisageable. Ainsi, pour faire face à cette explosion combinatoire, d'autres types de recherches doivent être utilisés. Dans cette étude, nous allons comparer l'efficacité de plusieurs méthodes de sélection de variables (*cf.* section 2.4). Parmi les approches de type *filter*, détaillées à la section 2.4.2, l'algorithme RELIEF et la mesure basée sur le critère de Fisher (FDR) sont évalués et comparés aux autres algorithmes de la catégorie *wrapper*. Rappelons que contrairement aux approches *filters*, les algorithmes de type *wrapper* utilisent un algorithme d'apprentissage durant les étapes de sélection des sous-ensembles de variables.

Aux sections 2.4.3.3 et 2.4.3.4, nous avons présenté les principales méthodes appartenant à la catégorie *wrapper*, basées respectivement sur des approches heuristiques et non déterministes. Aussi, dans cette analyse, différentes méthodes de sélection sont comparées, notamment les méthodes de sélection séquentielle ascendante et descendante (SFS et SBS), ainsi que leurs dérivées LRS, SFFS et SFBS. Rappelons que ces méthodes dérivées sont apparues afin d'améliorer le processus de sélection des variables, en permettant d'éliminer une variable sélectionnée ou de sélectionner une variable éliminée. Ces approches sont ainsi connues pour leur capacité à effectuer des retours en arrière (*backtracking*). Contrairement à SFFS et SFBS qui réalisent les retours en arrière de manière autonome, l'utilisation de LRS nécessite de définir deux paramètres : L et R qui représentent respectivement le nombre de variables à sélectionner et le nombre de variables à éliminer. Après plusieurs essais, nous avons opté pour deux cas : $L = 3$ et $R = 2$, ainsi que $L = 2$ et $R = 3$. Face à ces méthodes heuristiques, des méthodes non déterministes sont également utilisées telles que, l'algorithme RGSS et les algorithmes génétiques (AG). Ces deux autres méthodes offrent une plus grande capacité d'exploration de l'espace des combinaisons de variables, par l'utilisation de procédés aléatoires.

L'utilisation des algorithmes génétiques nécessite de définir plusieurs paramètres, comme la taille de la population (ici de 80 individus) ou encore, la probabilité de mutation (ici de $0,05$). Dans nos expérimentations, la sélection par tournoi est utilisée pour choisir les 40 parents pour la reproduction. La fonction d'adaptation qui évalue chaque individu (donc chaque sous-ensemble de variables) est donnée par la relation suivante :

$$J = AUC_V + 0,01 \times n\_bits\_à\_0\,, \tag{5.1}$$

où AUC_V est la moyenne de AUC sur les échantillons de validation et $n\_bits\_à\_0$ est le nombre de variables ignorées. L'arrêt de l'algorithme est effectué lorsque le nombre de 500 générations est atteint.

Le but commun à toutes ces méthodes de sélection est de trouver un sous-ensemble de variables pertinentes en agissant sur un compromis entre l'exploration de l'espace des combinaisons (nombre de sous-ensembles de variables à évaluer) et le coût calculatoire. En effet, l'augmentation de la complexité du processus de sélection peut être corrélée avec le nombre de combinaisons évaluées. Dès lors, on peut raisonnablement estimer que plus le nombre de combinaisons évaluées augmente, meilleur est le sous-ensemble trouvé. Ainsi, en excluant l'emploi d'un algorithme d'apprentissage, les méthodes de type *filter* parviennent rapidement à trier les variables par pertinence.

Les méthodes heuristiques extraient des sous-ensembles avec une rapidité relative, sans explorer complètement l'ensemble des combinaisons possibles. Nous avons vu également que l'amélioration des méthodes heuristiques est apparue grâce à la notion de retours en arrière avec les méthodes telles que, LRS, SFFS et SFBS. Cependant comme relaté à la section 2.4.3.3, ces trois dernières méthodes ont l'inconvénient d'être soit paramétriques (LRS), soit d'être confrontées à des problèmes de convergence (SFFS et SFBS). Le problème principal des méthodes séquentielles réside dans le fait que la sélection ou l'élimination des variables est totalement dépendante des variables déjà sélectionnées ou déjà éliminées. Cela a pour effet de réduire l'exploration de l'espace des combinaisons des variables et d'attirer les méthodes vers des minimums locaux ; les heuristiques apportant des retours en arrière n'endiguent pas nécessairement ces phénomènes. Ce problème est en partie réduit par les méthodes non déterministes, par l'intégration de procédés aléatoires dans le processus de sélection. Gourmands en calcul, les algorithmes génétiques permettent néanmoins de réduire le coût et d'améliorer le compromis exploration de l'espace des combinaisons/coût calculatoire, en agissant notamment sur le paramètre de la taille de la population.

Afin de faciliter la lecture, nous dénommerons « méthodes classiques », l'ensemble des méthodes de sélection précédemment décrites.

5.4.2.2 Sélection séquentielle de variables avec retours en arrière par les algorithmes génétiques

Afin d'améliorer encore le compromis entre l'exploration de l'espace des variables et le coût calculatoire, nous proposons dans cette étude de combiner les algorithmes génétiques avec des méthodes classiques de sélection séquentielle, connues pour leur rapidité : la sélection naïve basée sur les critères de pertinence des variables et la sélection séquentielle ascendante (SFS). À l'image des méthodes LRS, SFFS et SFBS, ces combinaisons de méthodes effectuent des retours en arrière durant le processus de recherche des sous-ensembles de variables. Cependant, contrairement aux méthodes classiques, les retours en arrière ne sont plus réalisés de manière séquentielle, mais par des algorithmes génétiques, permettant notamment de converger sans se soucier des choix de paramètres (L et R pour LRS) et à la présence de minimums locaux (SFFS et SFBS).

Pour décrire notre processus de sélection, nous allons prendre l'exemple de SFS. À chaque itération, cette approche ajoute au sous-ensemble de variables, la variable optimisant les performances de classification en validation. Ainsi, de manière aléatoire et temporaire, nous arrêtons la progression de SFS afin d'optimiser le sous-ensemble courant, par les AG. Une fois cette optimisation réalisée, le processus SFS continue, en partant du sous-ensemble de variables réduit par les AG. Avec les retours en arrière, ce processus peut ne pas s'arrêter, c'est pourquoi nous choisissons de le stopper après un certain nombre d'itérations. Dans nos expérimentations, le processus s'arrête après avoir effectué 200 itérations, sans considérer celles des AG.

Dans notre analyse, la probabilité d'arrêt de l'algorithme séquentiel pour le retour en arrière est de $0,1$. L'optimisation par les AG est faite sur 20 générations, la population est composée de 20 individus et la probabilité de mutation est de $0,05$. Avec ces paramètres, la convergence et l'exécution des AG sont rapides. Ces méthodes peuvent être vues comme des processus de sélection séquentiels avec optimisations locales. L'annexe D propose des explications complémentaires et détaillées.

Comme il sera montré, ces méthodes ont l'avantage d'obtenir de bonnes performances en sélectionnant peu de variables, tout en minimisant le nombre de combinaisons à évaluer. Par opposition aux « méthodes classiques », cette nouvelle classe de méthodes sera appelée par la

suite « méthodes combinées ». Ainsi, les nouvelles méthodes, fondées sur les approches de sélection séquentielle naïve (SFS_{Fisher} et SFS_{RELIEF}) et ascendantes (SFS), seront notées respectivement $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$.

5.4.2.3 Techniques de classification et d'évaluation

Face à ce problème combinatoire conséquent, nous avons pu, dans un premier temps, réduire le « coût » de recherche des sous-ensembles de variables pertinentes par l'emploi de méthodes de sélection adaptées. Aussi, afin de rendre exploitables ces méthodes de sélection, nous avons choisi d'utiliser un algorithme d'apprentissage relativement rapide. Lors des études précédentes, nous avons pu observer que les techniques de classification fondées sur des approches génératives sont les plus rapides et que le classifieur de Bayes naïf (nommé BN_{gauss}) possède le meilleur compromis rapidité/efficacité. Cependant, comme nous l'avons remarqué à la section 5.3.3, ces méthodes nécessitent de posséder un nombre important d'observations afin d'estimer, au plus juste, les différentes densités de probabilité. Précédemment, avec l'échantillon composé de 84 patients, ces problèmes avaient déjà été soulevés. Or, dans cette nouvelle étude, nous avons uniquement 58 patients, cela suggère que nous pouvons être confrontés aux mêmes problèmes d'estimation. C'est ainsi que nous avons fait le choix d'utiliser le classifieur des k-plus proches voisins (k-ppv). Cette méthode de classification a été peu abordée dans ce manuscrit, elle est cependant très efficace lorsque le nombre d'observations n'est pas trop élevé. Cette approche, évoquée succinctement à la section 1.2.3.2 dans le cadre de l'estimation des densités, est définie par [Loosli *et al.*, 2006] comme un classifieur universel de référence défini à partir d'une règle, qui stipulent que chaque observation de test prend l'étiquette de la classe dominante parmi ses k-ppv. Ainsi, dans la phase de sélection des sous-ensembles de variables, l'algorithme d'apprentissage nécessaire à la mise en place des méthodes de type *wrapper* est donc basé sur les k-ppv. Aussi, nous utiliserons les perceptrons multicouches (PMC) afin d'évaluer les performances et la pertinence des sous-ensembles de variables sélectionnées sur l'échantillon de test. Cela permettra de comparer, plus objectivement, ces nouvelles études avec celles qui ont utilisées les méthodes de projection (*cf.* section 5.3). Les techniques d'utilisation et d'évaluation des PMC sont identiques à celles utilisées dans les analyses précédentes, décrites à la section 5.3.1.2.

Dans cette section, nous ne comparons pas les méthodes de classification, mais les méthodes de sélection de variables. Les performances de ces dernières sont toutefois évaluées par le même processus utilisé précédemment et illustré à la figure 5.3 (page 134). Dès lors, l'échantillon de 58 patients est divisé en deux groupes (apprentissage et test), chacun d'eux étant composé de 29 patients. De même, la validation croisée mesurant les performances moyennes de validation (AUC_V) est réalisée sur 6 sous-ensembles de patients ($K = 6$).

5.4.3 Expérimentations et résultats

5.4.3.1 Résultats de la sélection de variables par les méthodes « classiques »

Le tableau 5.12 fait état de l'ensemble des résultats. Les deux méthodes *filters*, basées sur les critères de Fisher et RELIEF, donnent les coefficients de pertinence pour chaque variable. Calculés sans considérer de dépendance entre les variables, ces coefficients permettent de trier les variables par ordre de pertinence. En combinant l'information des critères avec un processus de sélection séquentielle, nous obtenons ce que nous avons appelé à la section 2.4.3.1 une approche de sélection « naïve ». Cette approche consiste alors à ajouter progressivement une nouvelle variable en conservant l'ordre de pertinence induit par les critères. Notés SFS_{Fisher} et SFS_{RELIEF}, ces processus de sélection séquentielle sont extrêmement rapides, mais restent bien moins performants que les processus de sélection SFS et SBS. Les sous-ensembles de variables

méthode de sélection	**variables sélectionnées**	**nombre de variables**	AUC_V
	évaluation sans processus de sélection		
pas de sélection	$\{1, \ldots, 70\}$	70	0,550
Aléatoire (20 *essais*)	–	$30,10 \pm 21,08$	$0,425 \pm 0,095$
Aléatoire (*optimal*)	–	56	0,600
	sélection de variables par des processus séquentiels naïfs		
$\mathbf{SFS}_{Fisher}$	$\left\{ \begin{array}{c} 13,14,19,20,22,28, \\ 29,30,34,49,70 \end{array} \right\}$	11	0,700
$\mathbf{SFS}_{RELIEF}$	–	65	0,650
	sélection de variables par des processus séquentiels		
SFS	$\left\{ \begin{array}{c} 2,3,5,13,25,28,29,30, \\ 33,35,36,45,46,49,62 \end{array} \right\}$	15	0,900
SBS	$\left\{ \begin{array}{c} 25,26,28,37,42,44,45, \\ 46,49,53,55,56,58,59, \\ 62,63,67,68,69,70 \end{array} \right\}$	20	0,850
	sélection de variables par des processus séquentiels avec retours en arrière		
$\mathbf{LRS}_{(L=3,\, R=2)}$	$\{3,4,28,29,33,35,46,49,62\}$	9	0,900
$\mathbf{LRS}_{(L=2,\, R=3)}$	$\{16,26,36,39,46,49\}$	6	1
SFFS	$\{2,3,5,28,29,33,49,62\}$	8	0,900
SFBS	$\left\{ \begin{array}{c} 1,2,25,33,35,38,41,42, \\ 43,45,46,49,51,52,56,62, \\ 63,64,65,67,68,69,70 \end{array} \right\}$	23	0,900
	sélection de variables par des processus non déterministes		
RGSS (20 *essais*)	-	$12,55 \pm 2,24$	$0,825 \pm 0,044$
RGSS (*optimal*)	$\left\{ \begin{array}{c} 16,20,21,26,29,35, \\ 45,49,51,62,67,69 \end{array} \right\}$	12	0,900
AG (20 *essais*)	-	$14,75 \pm 3,45$	$0,975 \pm 0,038$
AG (*optimal*)	$\{25,34,35,41,49,51,54,62,69\}$	9	1

TAB. 5.12 – Comparaison des performances de classifieurs (k-ppv) issus de méthodes de sélection dites « classiques » pour prédire le résultat du *tilt-test* en position couchée et basculée.

pour ces quatre méthodes et pour les processus utilisant des retours en arrière (LRS, SFFS et SFBS) sont obtenus pour des performances optimales en validation (AUC_V). Rappelons que pour éviter un temps d'exécution démesuré, les méthodes de sélection sont associées au classifieur des k-plus proches voisins. Les recherches des sous-ensembles de variables optimales par les méthodes non déterministes (RGSS et AG) sont obtenues en répétant 20 fois le processus avec différentes initialisations. Pour la méthode RGSS, le processus est réitéré avec un sous-ensemble de départ déterminé aléatoirement. Quant aux algorithmes génétiques, le processus est répété avec différentes initialisations de la population initiale. Ainsi, parmi les 20 sous-ensembles obtenus par RGSS et AG, les sous-ensembles que nous considérons optimaux (meilleur AUC_V parmi les 20 essais) sont notés « RGSS (*optimal*) » et « AG (*optimal*) ». Notons que les résultats moyens (sur les 20 essais) apparaissent également dans le tableau 5.12.

Afin de vérifier l'impact des méthodes de sélection utilisées, nous comparons leurs résultats à ceux de deux autres approches n'utilisant pas de processus de sélection (tableau 5.12). La première conserve les 70 variables initiales et la seconde sélectionne aléatoirement les variables et leur nombre (20 sélections sont réalisées par ce processus). Ces méthodes sont appelées respectivement « pas de sélection » et « aléatoire ». Comme nous pouvons le remarquer, avec une $AUC_V = 0,550$, la sélection des 70 variables n'est pas pertinente. Ce faible résultat est à relativiser, car contrairement à d'autres méthodes de classification (notamment les réseaux de neurones), la méthode des k-ppv ne permet pas de pondérer l'implication des variables non pertinentes (nuisibles à la séparation des classes) dans la classification.

En observant le tableau 5.12, nous pouvons globalement observer que les performances s'améliorent avec la complexité des processus de sélection ; la complexité peut s'interpréter en termes d'exploration de l'espace des combinaisons de variables (nombre de sous-ensembles évalués). Ainsi, dès lors que les méthodes explorent davantage l'espace, les performances apparaissent supérieures. Aussi, quelle que soit la complexité de la méthode de sélection, le nombre de variables sélectionnées reste relativement important ; notons néanmoins un avantage pour les méthodes effectuant des retours en arrière.

La figure 5.13 récapitule les performances de prédiction obtenues par les k-ppv en validation, en comparant les courbes de ROC. Ainsi, comme nous pouvons le voir, les méthodes sont comparées suivant quatre catégories : sans sélection (approches appelées « pas de sélection » et « aléatoire »), sélection séquentielle (SFS_{Fisher}, SFS_{RELIEF}, SFS et SBS), sélection séquentielle avec retours en arrière (LRS, SFFS et SFBS) et sélection non déterministe (RGSS et AG).

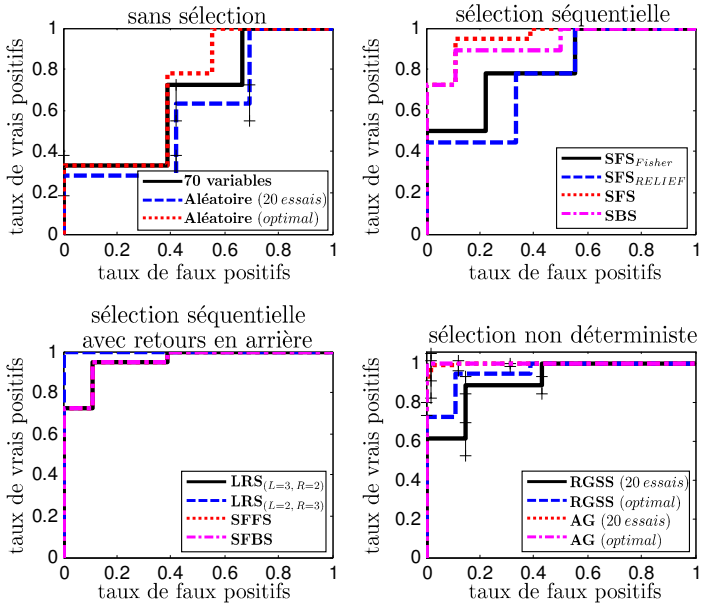


FIG. 5.13 – Comparaison des courbes de ROC de classifieurs (k-ppv) issus de méthodes de sélection dites « classiques » pour prédire le résultat du *tilt-test* en position couchée et basculée.

5.4.3.2 Résultats de la sélection de variables par les « méthodes combinées »

Dans cette seconde phase d'expérimentation, trois des méthodes classiques de sélection (Fisher, RELIEF et SFS) sont combinées avec les algorithmes génétiques afin de réaliser dans des processus de sélection, des retours en arrière stochastiques et non plus séquentiels.

Les performances sont montrées dans le tableau 5.13. Comme ces méthodes sont basées sur des processus aléatoires (par l'intermédiaire des AG et le choix de l'instant où les retours en arrière doivent être effectués), nous répétons leurs exécutions 20 fois afin de réduire le biais de l'estimation de leurs performances. Comme précédemment, les résultats montrent les performances optimales et moyennes sur les 20 essais. La figure 5.14 compare pour chacune des méthodes combinées, les courbes de ROC moyennes suivant les 20 essais et les courbes de ROC optimales pour les performances de validation.

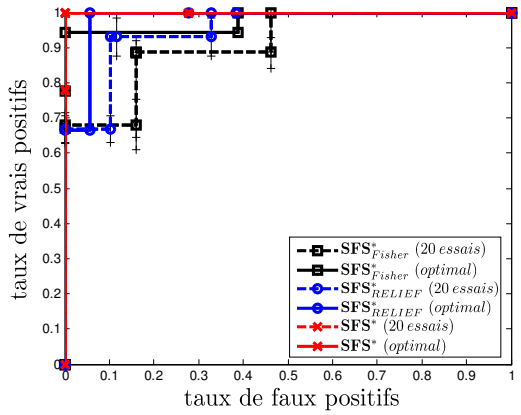


FIG. 5.14 – Comparaison des courbes de ROC de classifieurs (k-ppv) issus de méthodes de sélection combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée.

méthode de sélection	**variables sélectionnées**	**nombre de variables**	AUC_V
$\mathbf{SFS}^*_{Fisher}$ $(20\,essais)$	-	$4,50 \pm 1,01$	$0,830 \pm 0,047$
$\mathbf{SFS}^*_{Fisher}$ $(optimal)$	$\{18, 20, 49, 62\}$	4	$0,950$
$\mathbf{SFS}^*_{RELIEF}$ $(20\,essais)$	-	$4,77 \pm 1,32$	$0,880 \pm 0,057$
$\mathbf{SFS}^*_{RELIEF}$ $(optimal)$	$\{30, 49, 62, 68\}$	4	$0,950$
$\mathbf{SFS}^*$ $(20\,essais)$	-	$6,31 \pm 1,85$	1 ± 0
$\mathbf{SFS}^*$ $(optimal)$	$\{17, 30, 49, 62, 68\}$	5	1

TAB. 5.13 – Comparaison des performances de classifieurs (k-ppv) issus de méthodes de sélection combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée.

En comparant les résultats des méthodes classiques (tableau 5.12) avec ceux des méthodes combinées (tableau 5.13), nous pouvons observer, à la figure 5.15, que ces dernières obtiennent de meilleures performances. Ces améliorations de performances sont obtenues tout en réduisant le nombre de variables sélectionnées. En effet, le nombre moyen de variables sélectionnées pour les méthodes combinées $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$ atteint respectivement $4,77$ et $6,31$, contre 15 et 16 pour les méthodes séquentielles SFS_{RELIEF} et SFS.

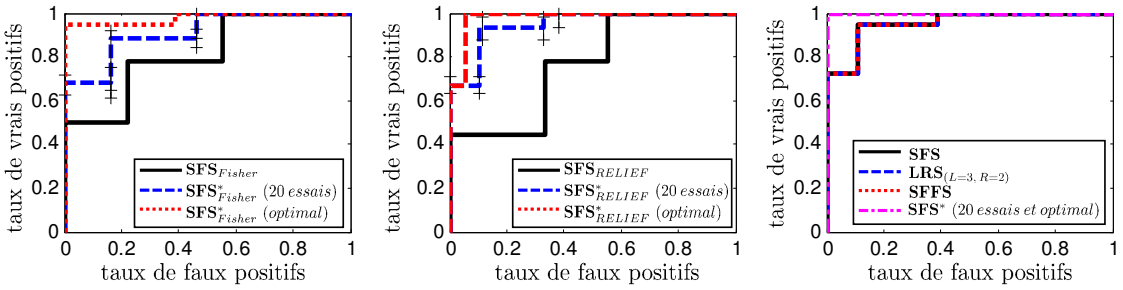


FIG. 5.15 – Comparaison des courbes de ROC de classifieurs (k-ppv) issus de méthodes de sélection « classiques » et combinées aux AG pour prédire le *tilt-test* en position couchée et basculée.

Comparaison des courbes de ROC de classifieurs (k-ppv) issus de méthodes de sélection combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée.

5.4.3.3 Performances de la prédiction

Comme observé dans les études précédentes, les perceptrons multicouches sont très efficaces. Cependant compte tenu de la lourdeur de leur processus d'apprentissage et du choix de l'architecture, nous ne les avons pas appliqués dans les processus de sélection. Toutefois, nous les employons sur les meilleurs sous-ensembles de variables obtenus, en l'occurrence, par les méthodes de sélection combinées. Les performances des PMC sont estimées en répétant 100 apprentissages avec différentes initialisations des poids, et cela, pour différentes architectures. Dès lors, nous cherchons, comme précédemment, à trouver la valeur des poids et le nombre idéal de neurones dans la couche cachée qui maximisent les performances de validation. Le tableau 5.14 relate les performances obtenues par les PMC sur les sous-ensembles optimaux des méthodes combinées ($\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$). Ce tableau indique pour chacun des trois sous-ensembles, l'AUC moyenne de test (AUC_T) pour les 100 apprentissages effectués sur l'architecture optimale et la meilleure AUC ($\widehat{AUC_T}$), obtenue par le modèle optimisant les performances de validation. Rappelons que le sous-ensemble de test ne participe ni à l'apprentissage, ni à la sélection du modèle (entrée et architecture du modèle).

Parmi les trois cas évalués, le meilleur sous-ensemble de variables semble être obtenu par la méthode $\mathbf{SFS}^*_{RELIEF}$, même si les performances des trois méthodes sont très proches. Les quatre variables composant le sous-ensemble de $\mathbf{SFS}^*_{RELIEF}$ sont : pression artérielle moyenne (30), évaluation du rebond initial de FC (49), t_2 (62) et dZ_{max}/dt (68). Le modèle, obtenu par ces entrées et par le réseau de neurones possédant 9 neurones dans sa couche cachée, permet d'obtenir une AUC sur un groupe de patients parfaitement inconnus de $0,971$. Avec une sensibilité de 100%, une spécificité de 94% et des valeurs prédictives positive et négative de 92% et de 100%, ce modèle permet de prédire efficacement le résultat du *tilt-test*.

Notons également qu'un nombre significatif de variables sélectionnées est commun aux trois sous-ensembles : évaluation du rebond initial de FC (49) et t_2 (62). Cela peut justifier le fait que les résultats des apprentissages des PMC sont très proches (voir tableau 5.14) et peut également révéler les bonnes capacités de convergence des méthodes effectuant des retours en arrière par les AG.

méthode de sélection	**variables sélectionnées**	**PMC**	
		AUC_T	$\widehat{AUC_T}$
$\mathbf{SFS}^*_{Fisher}$ (*optimal*)	$\{18, 20, 49, 62\}$	$0,795 \pm 0,037$	$0,941$
$\mathbf{SFS}^*_{RELIEF}$ (*optimal*)	$\{30, 49, 62, 68\}$	$0,768 \pm 0,078$	$0,971$
$\mathbf{SFS}^*$ (*optimal*)	$\{17, 30, 49, 62, 68\}$	$0,753 \pm 0,038$	$0,902$

TAB. 5.14 – Comparaison des performances (en test) de classifieurs (PMC) issus de méthodes de sélection combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée.

D'autre part, comme le montre le tableau 5.15, le modèle, fondé sur $\mathbf{SFS}^*_{RELIEF}$ et un PMC, peut être comparé très favorablement aux autres études. En effet, les résultats obtenus sont largement supérieurs et sont bien plus pertinents, si l'on considère que pour estimer leurs performances, [Mallat *et al.*, 1997; Pitzalis *et al.*, 2002] ont utilisé leur sous-ensemble de validation et [Bellard *et al.*, 2003] ont employé leur sous-ensemble d'apprentissage.

étude	**période du *tilt-test***	$S_e\,(\%)$	$S_p\,(\%)$	$VPP\,(\%)$	$VPN\,(\%)$
$\mathbf{SFS}^{*}_{RELIEF}$ (*optimal*) **PMC**	1re à 10-ième min du basculement	100	94	92	100
[Bellard *et al.*, 2003]†	5 à 10-ième min du basculement	68*	70*	68*	70*
[Bellard *et al.*, 2003]‡	5 à 10-ième min du basculement	50*	97*	93*	67*
[Pitzalis *et al.*, 2002]	1re à 15-ième min du basculement	80	85	57	94
[Mallat *et al.*, 1997]	1re à 6-ième min du basculement	96	87	75	98

Note : † étude analysant uniquement les variables liées au signal dZ/dt. ‡ étude analysant les variables liées au signal dZ/dt et FC, PAS, PAD et PD.

* Indique que les résultats sont obtenus sur le sous-ensemble d'apprentissage.

TAB. 5.15 – Comparaison des résultats de prédiction de la réponse du *tilt-test* en positions couchée et basculée, avec les principales études analysant la syncope inexpliquée.

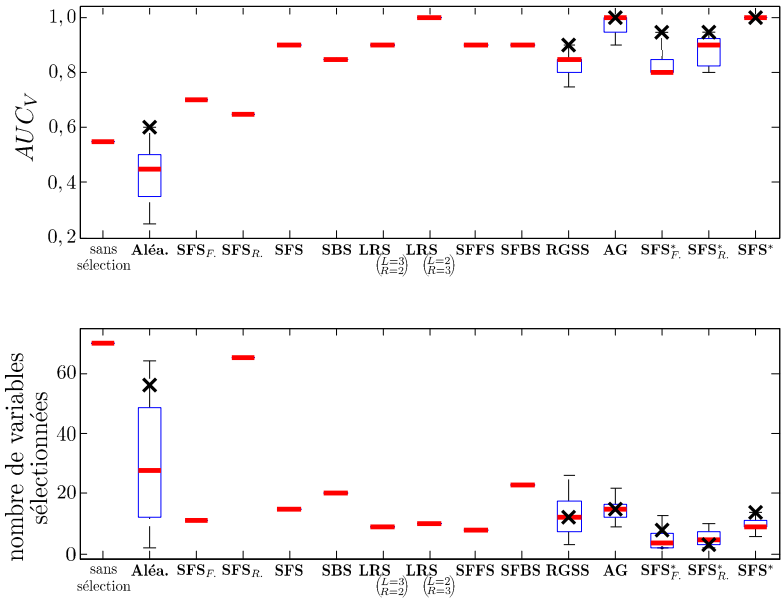
5.4.4 Discussions et conclusions

Dans cette section, nous venons d'analyser les différentes variables acquises lors des périodes couchée et basculée du *tilt-test*, pour des patients sujets à l'apparition récurrente de syncopes inexpliquées. Les dix premières minutes de la période basculée ont été étudiées, sachant que l'examen réclame initialement une durée de 45 minutes. Cette analyse a donc permis de prévoir le résultat du *tilt-test* bien avant que ce dernier n'arrive à son terme.

L'étude réalisée a permis de comparer différentes méthodes de sélection de variables, principalement des approches heuristiques (SFS_{Fisher}, SFS_{RELIEF}, SFS, SBS, LRS, SFFS et SFBS) et non déterministes (RGSS et AG). Face à ces méthodes, nous avons proposé une nouvelle approche fondée sur ces deux types de techniques afin d'améliorer les performances de sélection et de classification [Feuilloy *et al.*, 2006a]. Rappelons que ce nouveau processus de sélection est proche de la méthode SFFS, dont les retours en arrière sont, dorénavant, effectués aléatoirement par les algorithmes génétiques. Cette approche est une combinaison de SFS et des AG, et fait donc coïncider, d'une certaine manière, une recherche locale (SFS) avec une recherche globale (AG).

Les tableaux 5.12 et 5.13 ont permis de relever l'avantage des méthodes combinées face aux méthodes dites « classiques ». En effet, si nous considérons simultanément les performances de classification et le nombre de variables sélectionnées, les méthodes combinées surpassent largement les autres méthodes. La figure 5.16 récapitule, pour chaque méthode, les performances obtenues (AUC_V) et le nombre de variables sélectionnées. Malgré les bonnes performances des AG, ces derniers ne permettent pas de réduire le nombre de variables autant que les méthodes combinées. Il est à noter que l'exécution des AG a été réalisée sans information sur le problème. Dès lors, comme évoqué à la section 2.4.3.4, dans ces conditions, les réponses données par les AG ne sont pas forcément optimales. Malgré ces remarques, les AG restent la meilleure méthode, parmi toutes les méthodes classiques évaluées. Cependant, comme le montre la figure 5.17, les AG nécessitent pour cela d'évaluer un nombre de combinaisons très important (40 000). Pour arriver à un sous-ensemble optimal, les méthodes de sélection séquentielle (SFS et SBS) en évaluent 2 485 sous-ensembles, contre 12 002 et 26 806 pour les méthodes effectuant des retours en arrière, respectivement LRS et SFFS/SFBS. Dès lors, nous apercevons aisément que les bonnes

performances des AG sont obtenues grâce à une recherche extrêmement lourde. Cette même figure permet d'observer également les qualités de recherche des méthodes combinées. En effet, ces dernières obtiennent les sous-ensembles optimaux après avoir évalué en moyenne 3 612 et 8 894 combinaisons, respectivement pour $\mathbf{SFS}^*_{Fisher}$/$\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$. Pour des performances supérieures aux AG, les méthodes combinées ont évalué 5 à 10 fois moins de sous-ensembles ; le gain de temps apporté est considérable.

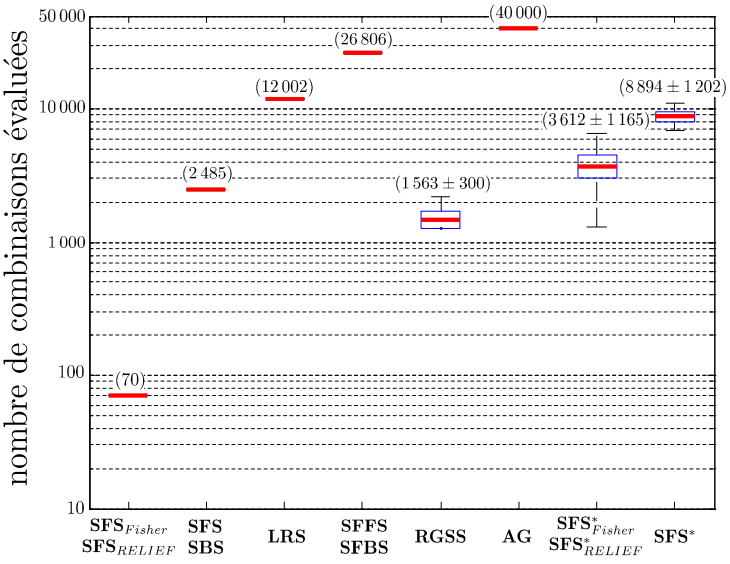


Note : Pour les méthodes de sélection faisant intervenir un processus aléatoire, la distribution des résultats est donnée par une boîte à moustaches (*boxplot*), indiquant la médiane (trait rouge), les premier et troisième quartiles (respectivement minimum et maximum de la boîte bleue), ainsi que les valeurs extrêmes (traits noirs). D'autre part, les croix noires indiquent les performances optimales (AUC_V et le nombre de variables sélectionnées).

FIG. 5.16 – Performances et nombre de variables sélectionnées pour les méthodes de sélection « classiques » et combinées aux AG pour prédire le *tilt-test* en position couchée et basculée.

L'observation des variables sélectionnées par les méthodes combinées montre une prédominance pour celles provenant du signal d'impédancemétrie thoracique. D'autre part, en considérant l'ensemble des méthodes de sélection de variables, nous retrouvons, comme dans l'étude précédente, des variables liées à la pression artérielle et à la fréquence cardiaque. Il est intéressant de remarquer que le sous-ensemble de variables qui semble le plus pertinent, obtenu par la méthode $\mathbf{SFS}^*_{RELIEF}$, est composé de variables appartenant à ces trois catégories. En effet, en sélectionnant quatre variables parmi les 70 initiales, ce sous-ensemble associé à un perceptron multicouches donne, selon notre analyse, les meilleures performances pour la prédiction du résultat du *tilt-test*. Ainsi, avec la pression artérielle moyenne, l'évaluation du rebond initial de la FC, l'intervalle de temps t_2, l'amplitude dZ_{max}/dt, et un PMC composé de 9 neurones dans la couche cachée, le modèle obtenu permet de prédire la syncope avec une sensibilité de 100% et une spécificité de 94%. Dès lors, avec uniquement quatre variables d'entrées, le modèle est rapide, peu complexe et parcimonieux ; comme préconisé dans la littérature (*cf.* section 1.4.2.5, page 48). L'autre avantage de la configuration obtenue réside dans la simplification de l'acquisition des variables, où quatre variables doivent être enregistrées et traitées au lieu des 70 initiales.

Sur le signal dZ, les variables les plus souvent sélectionnées sont t_2 et dZ_{max}/dt. Comme vu dans l'étude précédente (*cf.* section 5.3.3), des études telles que [Bellard *et al.*, 2003; Schang *et al.*, 2003; Schang *et al.*, 2006] ont établi la pertinence de ce signal. [Bellard *et al.*, 2003] ont



Note : Le nombre de variables initiales est de 70. La méthode LRS concerne les deux cas **LRS**$_{(L=3,\,R=2)}$ et **LRS**$_{(L=3,\,R=2)}$.
Pour les méthodes de sélection faisant intervenir un processus aléatoire, la distribution des résultats est donnée par une boîte à moustaches (*box-plot*), indiquant la médiane (trait rouge), les premier et troisième quartiles (respectivement minimum et maximum de la boîte bleue), ainsi que les valeurs extrêmes (traits noirs). D'autre part, la moyenne et l'écart type apparaissent numériquement sur le graphique.

FIG. 5.17 – Nombre de combinaisons de variables évaluées par les méthodes de sélection « classiques » et combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée.

ainsi observé que t_2 diminue significativement pour les patients positifs au *tilt-test* ; cette observation se vérifie dans nos données comme le montre le tableau 5.2 à la page 130 (en moyenne, la variable t_2 est égale à $169,2$ ms et à $194,1\,ms$, respectivement pour les patients positifs et négatifs au *tilt-test*). Dès lors, en établissant un seuil à 199 ms, [Bellard *et al.*, 2003] ont pu prédire la réponse positive de l'examen avec une sensibilité de 68% et une spécificité de 63% (sur un groupe rétrospectif de 68 patients). D'autre part, ces mêmes études ont pu montrer la corrélation entre le volume d'éjection systolique (donc le débit sanguin), le temps d'éjection ventriculaire (TEV$\approx t_1 + t_2$, *cf.* section 4.2.2) et le pic maximum sur dZ (dZ_{max}/dt). L'utilité du TEV est justifiée par le fait que des contractions excessives des ventricules pourraient provoquer une syncope (*i.e.* Bezold-Jarish reflex[4] [Fenton *et al.*, 2000]). Ainsi, en utilisant le TEV, [Schang *et al.*, 2006] ont prédit la réponse positive du *tilt-test* avec une sensibilité de 85% et une spécificité de 39% sur un groupe prospectif de 59 patients.

Par l'intermédiaire de [Mallat *et al.*, 1997], nous avons préalablement montré l'utilité de la fréquence cardiaque pour la prédiction de la syncope. Dans cette analyse, l'évaluation du rebond initial de la FC apparaît plus pertinent que la FC, même si ces deux variables sont fortement dépendantes. Durant l'examen du *tilt-test*, le rebond initial de FC est obtenu lors du passage de la position allongée à la position debout. Après le basculement, l'enclenchement des mécanismes adaptatifs (du maintien de l'équilibre) est réalisé par la transition liquidienne vers l'abdomen et les jambes, entraînant une baisse du remplissage cardiaque (appelé retour veineux ou précharge). Cela entraîne une baisse du volume éjection systolique, donc une baisse de la pression artérielle,

[4]Le réflexe de Bezold-Jarisch est un réflexe dont le point de départ se situe au niveau du ventricule gauche. Il empreinte les voies vagales et entraîne une bradycardie, une hypotension et une vasodilatation (agrandissement du calibre des vaisseaux) périphérique.

laissant ainsi apparaître une tachycardie[5] compensatrice suivie d'une bradycardie[6]. La plus profonde modification hémodynamique sur l'accession à la position debout se produit dans les 30 premières secondes [Hainsworth and Mark, 1993]. Le changement de posture, résultant du *tilt-test*, peut induire une diminution de la pression artérielle systolique qui, pour [Hainsworth and Mark, 1993], peut également être considérée comme le stress orthostatique lié au changement de posture. Le rebond initial de la FC se mesure alors entre le maximum de la FC de la tachycardie compensatrice et la bradycardie qui la suit.

Dans ces derniers commentaires, nous avons vu une nouvelle fois les relations entre la pression artérielle et la fréquence cardiaque, caractéristiques de la dynamique cardiaque. C'est ainsi que contrairement à [Mallat *et al.*, 1997] et [Pitzalis *et al.*, 2002] qui ont analysé séparément la FC et la pression artérielle, il nous est apparu intéressant d'analyser conjointement toutes ces variables. Notre modèle basé sur un PMC, utilisant la pression artérielle moyenne, le rebond initial de la FC, t_2 et dZ_{max}/dt, a permis de prédire le résultat positif du *tilt-test* à la 10-ième minute de basculement avec une sensibilité de 100% et une spécificité de 94% (VPP de 92% et VPN de 100%). Rappelons que [Mallat *et al.*, 1997] ont obtenu une sensibilité et spécificité respectivement de 96% et 87% (VPP de 75% et VPN de 98%) et [Pitzalis *et al.*, 2002] ont quant à eux obtenu une sensibilité et spécificité respectivement de 80% et 85% (VPP de 55% et VPN de 94%). Ces résultats ont été obtenus lors de la prédiction du résultat positif du *tilt-test* à la 6-ième minute du basculement pour [Mallat *et al.*, 1997] et à la 15-ième minute du basculement pour [Pitzalis *et al.*, 2002] ; le tableau 5.15 récapitule l'ensemble de ces résultats.

5.5 Évaluation de la pertinence du signal d'impédancemétrie thoracique dans la prédiction de la syncope

5.5.1 Introduction

Comme nous avons pu le voir dans les études présentées aux sections 5.3 et 5.4, certaines caractéristiques liées au signal d'impédancemétrie thoracique se sont révélées pertinentes, notamment celles extraites sur la dérivée de ce signal (dZ). Nos résultats ont ainsi concordé avec les observations faites par [Bellard *et al.*, 2003; Schang *et al.*, 2003] et exposées à la section 4.3 qui, en utilisant les caractéristiques rappelées à la figure 5.18, ont pu prédire l'apparition des symptômes de la syncope lors de l'examen du *tilt-test*. Les caractéristiques principalement extraites du signal dZ sont jusqu'à présent dZ_{max}/dt, l'indice de contractibilité C et les intervalles de temps t_1 et t_2 liés au temps d'éjection ventriculaire (TEV). Rappelons que les signaux ont été enregistrés sur l'appareil PhysioFlow^{TM} de Manatec (*cf.* section 4.2.2).

L'objectif de cette section est d'analyser plus spécifiquement et plus largement le signal d'impédancemétrie thoracique et cela indépendamment des autres variables. Cette démarche va alors permettre d'évaluer réellement l'impact de ce signal dans la prédiction de la syncope lors de l'examen du *tilt-test*.

Dans un premier temps, nous allons détailler les étapes à réaliser avant d'exploiter qualitativement le signal d'impédancemétrie, notamment, en présentant le prétraitement et la méthodologie employée pour extraire chaque complexe représentatif d'un battement cardiaque. En effet, rappelons que les différentes caractéristiques t_1, t_2, C et dZ_{max}/dt sont habituellement mesurées sur plusieurs complexes et que les variables exploitées dépendent de leur moyenne. Ensuite, nous présenterons certains processus proposés par la littérature qui permettent de sélectionner les complexes

[5]La tachycardie correspond à un rythme cardiaque plus rapide que la normale.
[6]La bradycardie correspond à un rythme cardiaque trop bas par rapport à la normale.

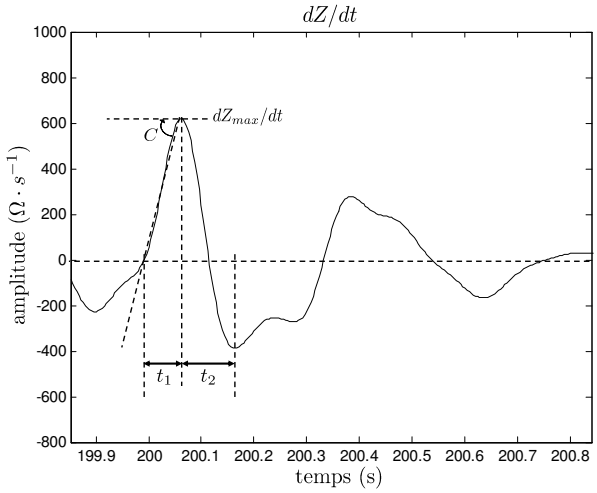


FIG. 5.18 – Récapitulatif des caractéristiques extraites sur le signal d'impédancemétrie thoracique (Z) pour prédire le résultat du *tilt-test* [Schang *et al.*, 2003].

sur lesquels les caractéristiques sont extraites. Ces méthodes de sélection seront également confrontées à d'autres processus que nous avons développés durant ces travaux [Feuilloy *et al.*, 2006b; Feuilloy *et al.*, 2006c], dans le but d'améliorer la sélection des complexes et, par conséquent, la pertinence des caractéristiques extraites. Enfin, tout au long de cette section, réservée à l'analyse du signal d'impédancemétrie thoracique, nous évaluerons également la pertinence de nouvelles caractéristiques pour la prédiction de la syncope, à partir des signaux Z et dZ, et cela, dans les domaines temporel [Feuilloy *et al.*, 2006b; Schang *et al.*, 2007] et fréquentiel [Feuilloy *et al.*, 2006c].

5.5.2 Prétraitement et extraction des complexes dZ

Avant d'extraire et d'analyser les caractéristiques issues du signal d'impédancemétrie thoracique, le signal est filtré afin d'éliminer notamment les nuisances liées à la fréquence d'alimentation du moteur faisant basculer la table du *tilt-test* (50 Hz).

Le filtre passe-bas (fréquence de coupure à 30 Hz ou à 40 Hz selon les études) est basé sur un système à réponse impulsionnelle finie. Ce filtre d'ordre 128 est associé à une fenêtre de Hamming [Kunt, 1981; Oppenheim and Schafer, 1989]. Cette fenêtre de pondération joue un rôle primordial dans l'analyse spectrale qui suivra, en contrôlant l'influence (largeur et amplitude) des lobes secondaires des estimateurs spectraux [Cottet, 1997].

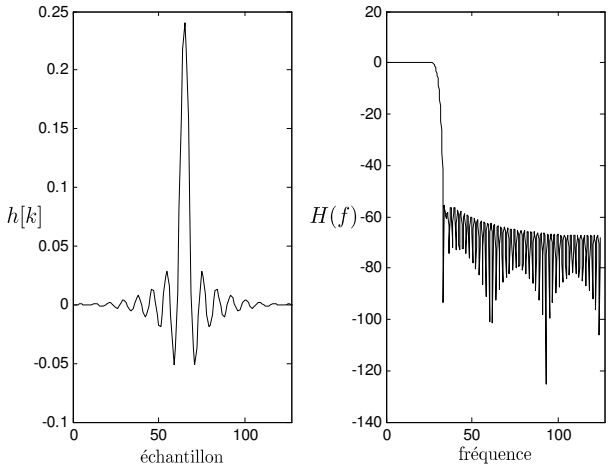


FIG. 5.19 – Réponses impulsionnelle et fréquentielle du filtre utilisé pour réduire les nuisances lors de l'analyse des signaux ECG et Z.

Les réponses impulsionnelle et fréquentielle du filtre sont données à la figure 5.19, l'opération de filtrage est effectuée à l'aide de la relation :

$$y(n) = \sum_{k=0}^{K-1} h(k)\, x(n-k)\,, \tag{5.2}$$

où n indique le n-ième élément du signal discret, h est la réponse impulsionnelle, et x et y désignent respectivement le signal original et le signal filtré.

D'autre part, à la section 4.2.2, nous avons évoqué que le signal Z possède un faible rapport signal sur bruit (RSB) amenant ainsi [Kubicek *et al.*, 1966] à suggérer l'utilisation de sa dérivée dZ. Cette dérivée est calculée par dérivée centrale [Friesen *et al.*, 1990], comme le montre la relation suivante :

$$dZ(n) = Z(n+1) - Z(n-1)\,. \tag{5.3}$$

Rappelons que l'acquisition du signal Z est obtenue pour une fréquence d'échantillonnage de 250 Hz. Ainsi, si nous considérons uniquement la période de repos de 10 minutes, le signal Z posséderait 150 000 points ; par conséquent, n serait compris entre 2 et 149 999. Notons que dans le cadre de la détection du complexe QRS, [Friesen *et al.*, 1990] a utilisé ce calcul pour déterminer la dérivée de l'ECG :

$$dECG(n) = ECG(n+1) - ECG(n-1)\,. \tag{5.4}$$

La figure 5.20 illustre les signaux ECG et Z, ainsi que leur dérivée respective $dECG$ et dZ.

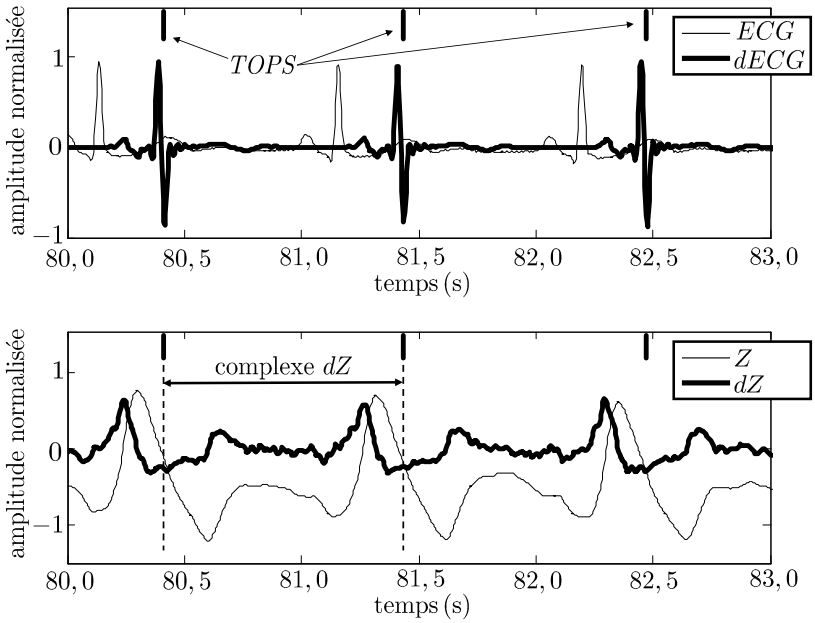


FIG. 5.20 – Extraction d'un complexe sur les signaux ECG et Z et leur dérivée $dECG$ et dZ.

L'évolution du signal dZ, et donc de ses caractéristiques, est perceptible en comparant le signal entre chaque battement cardiaque. Dès lors, il est nécessaire d'extraire chaque partie du signal représentative d'un battement, chacune de ces parties a été appelée « complexe dZ » (voir figure 5.20).

Proposée par [Friesen *et al.*, 1990], la méthode d'extraction employée fixe un seuil sur la dérivée de l'ECG à $0,3 \cdot \max(dECG)$ afin de déterminer ce que nous avons appelé les $TOPS$, comme le montre la figure 5.20. Ainsi, entre deux $TOPS$, nous avons un complexe dZ sur lequel les caractéristiques vont être déterminées.

5.5.3 Sélection des complexes par minimisation de l'erreur quadratique moyenne totale

Avant de sélectionner des complexes, les signaux sont filtrés et les complexes sont séparés par l'algorithme basé sur la détection des battements cardiaques (*cf.* section 5.5.2).

Pour l'analyse du signal dZ et de ses paramètres, [Bellard *et al.*, 2003; Bellard, 2003] ont réalisé le processus illustré par la figure 5.21, dont les étapes sont les suivantes :

(i) sélectionner une période de 5 minutes ;

(ii) extraire, normaliser et ré-échantillonner les complexes dZ de cet intervalle, afin que chaque complexe possède le même nombre de points ;

(iii) choisir visuellement un « beau » complexe dZ comme modèle de référence ;

(iv) calculer l'erreur quadratique moyenne (EQM), entre le modèle et chaque complexe dZ ; déduire l'erreur quadratique moyenne totale (EQMT) et σ_{EQM} qui correspondent respectivement à la moyenne et à l'écart type des EQM ;

(v) exclure les complexes dont l'EQM est hors de l'intervalle :

$$[EQMT - \sigma_{EQM}\,;\, EQMT + \sigma_{EQM}]\ ;$$

(vi) recalculer la moyenne des complexes restants afin de fournir le complexe moyen, sur lequel les caractéristiques de dZ vont être déterminées.

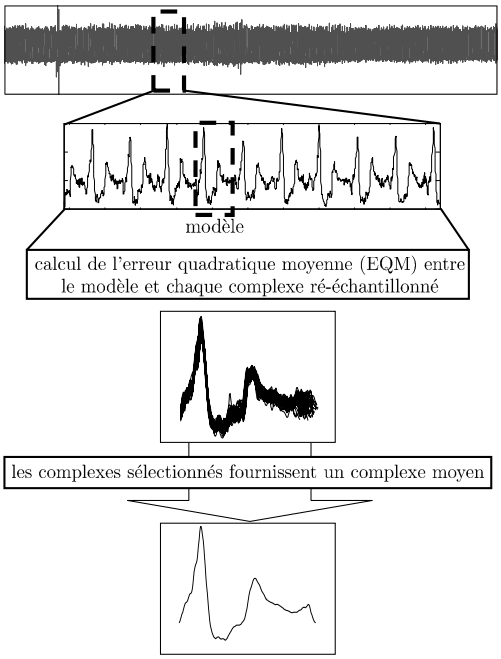


FIG. 5.21 – Extraction et sélection de complexes dZ par minimisation de l'erreur quadratique moyenne totale [Bellard *et al.*, 2003; Bellard, 2003].

Une fois les complexes sélectionnés, l'algorithme proposé par [Bellard *et al.*, 2003; Bellard, 2003] détecte sur le complexe moyen dZ/dt les variables suivantes (voir figure 5.18) :

- l'intervalle de temps t_1 (en ms) ;
- l'intervalle de temps t_2 (en ms) ;
- l'amplitude maximale du complexe dZ_{max}/dt (en $\Omega \cdot \mathrm{s}^{-1}$) ;
- l'indice de contractibilité C (en $\mathrm{m}\Omega \cdot \mathrm{s}^{-2}$), rappelons qu'il est obtenu par $(dZ_{max}/dt)/t_1$;
- l'indice de résistance vasculaire noté RVI (en $\mathrm{mmHg} \cdot \Omega^{-1} \cdot \mathrm{s}^2$), obtenu par :

$$\frac{\text{pression artérielle moyenne}}{dZ_{max}/dt \cdot \mathrm{FC}}\,, \tag{5.5}$$

FC désigne la fréquence cardiaque.

Le tableau 5.16 récapitule les résultats obtenus par [Bellard *et al.*, 2003] lors de l'analyse de ces cinq caractéristiques. Ils ont ainsi pu définir, qu'une valeur de $t_2 < 199\,\text{ms}$ permet de prédire le résultat positif du *tilt-test* durant la phase de repos avec une sensibilité de 68% et une spécificité de 63%. Cette étude rétrospective considère une population de 68 patients.

variable	**résultat du *tilt-test***	**période de repos**	**5 à 10-ième min. du basculement**	**syncope ou fin du *tilt-test***
t_1	négatif	200 ± 5	$292 \pm 9^{***}$	$324 \pm 9^{***}$
	positif	200 ± 6	$299 \pm 11^{***}$	$324 \pm 10^{***}$
t_2	négatif	233 ± 14	235 ± 14	$272 \pm 15^{*}$
	positif	$183 \pm 10^{\dagger\dagger}$	$280 \pm 18^{***}$	$271 \pm 13^{***}$
dZ_{max}/dt	négatif	542 ± 37	501 ± 29	496 ± 33
	positif	485 ± 26	432 ± 31	$405 \pm 24^{\dagger}$
C	négatif	$2,76 \pm 0,20$	$1,76 \pm 0,11^{***}$	$1,56 \pm 0,12^{***}$
	positif	$2,52 \pm 0,15$	$1,50 \pm 0,13^{***}$	$1,30 \pm 0,10^{***}$
RVI	négatif	$0,184 \pm 0,019$	$0,172 \pm 0,015$	$0,176 \pm 0,022$
	positif	$0,192 \pm 0,012$	$0,186 \pm 0,017$	$0,178 \pm 0,014$

Note : Les informations données dans le tableau expriment la moyenne ± l'écart type. Les différences statistiques des résultats appartenant aux échantillons « période de repos » et « 5 à 10-ième minute du basculement » sont indiquées par * ($p < 0,05$) et *** ($p < 0,001$). De même, les différences statistiques des résultats appartenant aux échantillons des patients ayant eu une « réponse négative » et une « réponse positive » au *tilt-test* sont données par † ($p < 0,05$) et †† ($p < 0,001$).

TAB. 5.16 – Évolution des variables issues du signal d'impédancemétrie thoracique dZ durant les positions couchée et basculée du *tilt-test* [Bellard *et al.*, 2003].

5.5.4 Sélection des complexes par optimisation manuelle du rapport signal sur bruit et évaluation de nouvelles caractéristiques prédictives

Certaines caractéristiques utilisées par [Bellard *et al.*, 2003] (t_1, t_2, C et dZ_{max}/dt) ont montré leur efficacité, comme nous avons pu l'observer dans nos analyses présentées aux sections 5.3 et 5.4. Dans cette section, nous proposons d'une part, d'améliorer le processus de sélection des complexes mis en place par [Bellard *et al.*, 2003] et d'autre part, d'analyser de nouvelles caractéristiques issues du signal d'impédancemétrie thoracique. La détermination et l'analyse qualitative des nouvelles variables seront décrites après la procédure de sélection des complexes.

Pour l'analyse des signaux Z et dZ et de leurs paramètres, nous avons décrit le processus de sélection des complexes et d'extraction des caractéristiques en prenant l'exemple du calcul du paramètre $LVET$ (intervalle de temps entre a et b de la figure 5.22). Ce processus décrit dans [Schang *et al.*, 2007] suit les étapes suivantes :

(i) sélectionner manuellement (donc visuellement) un intervalle de temps durant lequel le signal Z possède la plus grande quantité de complexes ayant le plus fort rapport signal sur bruit ;

(ii) extraire dans cet intervalle de temps, 12 complexes dZ consécutifs ;

(iii) normaliser et ré-échantillonner les complexes dZ pour réduire l'influence du changement des intervalles entre les battements. Pour une séquence de 12 battements, les 12 complexes dZ sont normalisés et ré-échantillonnés afin d'avoir le même nombre de points. Ainsi, si la durée d'un battement est convenue à 100 ms, et que la valeur du $LVET_{norm}$ est de 25 ms, alors son intervalle de temps occupe 25% d'un battement ;

(iv) calculer les valeurs des 12 $LVET_{norm}$ de dZ ;

(v) calculer $\overline{LVET_{norm}}$, comme la valeur moyenne des 12 $LVET_{norm}$, en :
- rejetant les trois plus petites valeurs ;
- rejetant les trois plus grandes valeurs ;
- calculant la moyenne sur les six valeurs restantes.

(vi) extraire du signal dZ les 12 battements suivants ;

(vii) SI la fin de la période de repos n'a pas été atteinte ALORS revenir à (iii) ;

(viii) calculer la valeur moyenne ($\overline{\overline{LVET_{norm}}}$) sur l'ensemble des $\overline{LVET_{norm}}$.

Notons que, comme pour le processus de [Bellard *et al.*, 2003], les signaux ont été préalablement filtrés et les complexes extraits par l'algorithme basé sur la détection des battements cardiaques (*cf.* section 5.5.2). Comme nous pouvons le remarquer, la procédure de sélection des complexes est proche de la méthodologie précédente proposée par [Bellard *et al.*, 2003].

Comme évoqué dans l'introduction de cette section, nous profitons de l'amélioration de la sélection des complexes pour évaluer de nouvelles caractéristiques liées aux signaux Z et dZ. Ainsi, conformément à la figure 5.22, nous allons désormais considérer les caractéristiques suivantes : $\overline{\overline{Slope_{1\,norm}}}$, $\overline{\overline{Slope_{2\,norm}}}$, $\overline{\overline{Area_{1\,norm}}}$, $\overline{\overline{Area_{2\,norm}}}$, $\overline{\overline{Z_{max} - Z_{min}}}$ et $\overline{\overline{dZ_{max}/dt - dZ_{min}/dt}}$. Les estimations des caractéristiques liées aux pentes ($Slope_{1\,norm}$ et $Slope_{2\,norm}$) et aux aires ($Aire_{1\,norm}$ et $Aire_{2\,norm}$) sur Z et dZ sont décrites en annexe A.2.

L'évaluation de ces nouvelles caractéristiques est réalisée sur l'échantillon de patients nommé $\mathcal{E}_2$ décrit à la section 5.2. Rappelons que cet échantillon, qui était initialement composé de 138 patients, a vu son nombre réduit à 129. Dans l'analyse qui suit, un autre patient a été exclu de l'étude en raison d'une mauvaise qualité du signal d'impédancemétrie enregistré (due certainement à des mouvements prolongés ou excessifs du patient). Ainsi, pour les 128 patients restants, 65 (51%, 40 ± 15 ans, 35 femmes, 30 hommes) ont eu une réponse négative au *tilt-test* et 63 (49%, 45 ± 15 ans, 31 femmes, 32 hommes) ont eu une réponse positive au *tilt-test*.

Comme dans les études précédentes (présentées aux sections 5.3.1, 5.3.2 et 5.4), le processus d'évaluation est identique à celui illustré à la figure 5.3 de la page 134. Ainsi, parmi les 128 patients, le groupe d'apprentissage réalisant l'étude rétrospective est composé de 64 patients (32 ont eu une réponse négative au *tilt-test*) et le groupe de test servant à l'analyse prospective est composé de 64 patients (33 ont eu une réponse négative au *tilt-test*).

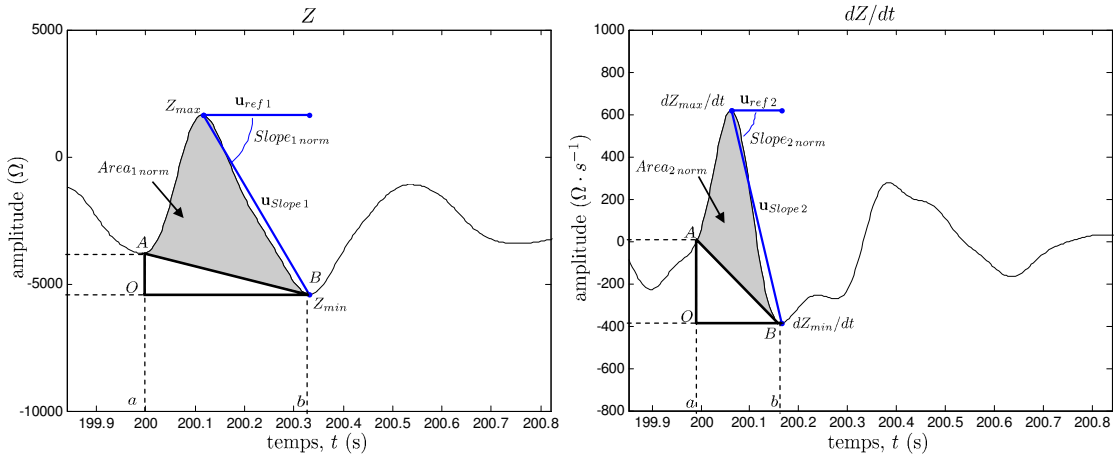


FIG. 5.22 – Nouvelles caractéristiques extraites du signal d'impédancemétrie thoracique (Z et dZ) dans le domaine temporel.

Associés aux nouvelles caractéristiques ($\overline{\overline{Slope_{1\,norm}}}$, $\overline{\overline{Slope_{2\,norm}}}$, $\overline{\overline{Area_{1\,norm}}}$ et $\overline{\overline{Area_{2\,norm}}}$, $\overline{\overline{Z_{max}-Z_{min}}}$ et $\overline{\overline{dZ_{max}/dt-dZ_{min}/dt}}$), le temps d'éjection ventriculaire gauche ($\overline{\overline{LVET_{norm}}}$), le sexe et l'âge des patients sont également considérés dans cette étude. Afin d'évaluer la pertinence de ces caractéristiques pour prédire le résultat du *tilt-test* en position couchée, nous utilisons un modèle fondé sur les *support vector machines*. Le tableau 5.17 compare les différentes performances obtenues en fonction du type de noyau choisi (linéaire, polynomial, sigmoïde et gaussien). Notons, qu'en considérant neuf caractéristiques, l'apprentissage par les SVM permet d'essayer toutes les combinaisons possibles comme cela a pu être réalisé à la section 5.3.1. Ainsi, les résultats du tableau 5.17 sont donnés pour chaque meilleure combinaison de variables telle que :

- le sous-ensemble $\left\{\overline{\overline{LVET_{norm}}}, \overline{\overline{Slope_{2\,norm}}}, \overline{\overline{Area_{2\,norm}}}\right\}$ optimise la prédiction du résultat du *tilt-test* lors de l'utilisation du noyau linéaire avec les SVM ;
- le sous-ensemble $\left\{\overline{\overline{Slope_{1\,norm}}}, \overline{\overline{Slope_{2\,norm}}}, \overline{\overline{Area_{2\,norm}}}\right\}$ optimise la prédiction du résultat du *tilt-test* lors de l'utilisation du noyau polynomial avec les SVM ;
- le sous-ensemble $\left\{\overline{\overline{LVET_{norm}}}, \overline{\overline{Slope_{1\,norm}}}, \overline{\overline{Slope_{2\,norm}}}, \overline{\overline{Area_{1\,norm}}}, \overline{\overline{Area_{2\,norm}}}\right\}$ optimise la prédiction du résultat du *tilt-test* lors de l'utilisation du noyau sigmoïde avec les SVM ;
- le sous-ensemble $\left\{\overline{\overline{LVET_{norm}}}, \overline{\overline{Slope_{1\,norm}}}, \overline{\overline{Slope_{2\,norm}}}, \overline{\overline{Area_{2\,norm}}}\right\}$ optimise la prédiction du résultat du *tilt-test* lors de l'utilisation du noyau gaussien avec les SVM.

Notons que chacun des sous-ensembles considère également les variables sexe et âge du patient.

noyau des SVM	**mesure de performance**			
	S_e (%)	S_p (%)	VPP (%)	VPN (%)
linéaire	90	15	50	62
polynomiale	87	75	77	86
sigmoïde	67	70	67	68
gaussien	94	79	74	93

TAB. 5.17 – Comparaison des performances de classifieurs (SVM), associés aux nouvelles caractéristiques extraites du signal Z et dZ, pour prédire le résultat du *tilt-test* en position couchée.

Les résultats du tableau 5.17, nous amènent à penser que le sous-ensemble de variables

$$\left\{ \hat{a}ge, sexe, \overline{\overline{LVET_{norm}}}, \overline{\overline{Slope_{1\,norm}}}, \overline{\overline{Slope_{2\,norm}}}, \overline{\overline{Area_{2\,norm}}} \right\}$$

combiné à un modèle basé sur les SVM (noyau gaussien) donne les meilleures performances de prédiction, avec une sensibilité de 94% et une spécificité de 79%. Ce travail, détaillé dans [Schang *et al.*, 2007], avait préalablement été engagé lors d'une autre étude [Schang *et al.*, 2006], dans laquelle d'autres modèles étaient utilisés, basés en l'occurrence sur des fonctions discriminantes linéaires (seuils) et des réseaux de neurones. Les principaux résultats de l'étude de [Schang *et al.*, 2006] sont exposés dans le tableau 5.18. Comme nous pouvons l'observer, les mêmes variables associées à un réseau de neurones ont permis aux auteurs de prédire, dans la position couchée, le résultat du *tilt-test* avec une sensibilité de 88% et une spécificité de 64% sur un groupe prospectif de 59 patients. Dans ce même tableau nous pouvons observer également la pertinence de chaque variable pour la prédiction en adoptant simplement un seuil de décision. Ce seuil, obtenu sur un groupe rétrospectif de 70 patients, a permis de révéler que la variable $\overline{\overline{LVET_{norm}}}$ est la plus robuste, donc certainement la plus reproductible.

seuil	**mesure de performance**			
	S_e (%)	S_p (%)	VPP (%)	VPN (%)
$\overline{\overline{LVET_{norm}}} < 23,9\%$	85 (73)	39 (73)	52 (75)	76 (71)
$\overline{\overline{Slope_{1\,norm}}} > 85,6\,^\circ$	27 (68)	42 (67)	27 (69)	42 (65)
$\overline{\overline{Slope_{2\,norm}}} > 84,8\,^\circ$	81 (86)	24 (45)	46 (64)	62 (75)
$\overline{\overline{Area_{2\,norm}}} > 2,3 \cdot 10^4$	46 (54)	48 (48)	41 (54)	53 (48)
PMC	88 (100/86)	64 (100/64)	66 (100/64)	88 (100/86)

Note : $(\cdots)$ indique que les résultats sont obtenus sur le groupe rétrospectif de patients ; groupe sur lequel les seuils ont été établis. Pour les PMC, $(\cdots/\cdots)$ précise d'autre part, que les résultats sont obtenus respectivement sur l'échantillon d'apprentissage et l'échantillon de validation.

TAB. 5.18 – Évaluation de la pertinence des nouvelles caractéristiques extraites du signal Z et dZ, pour prédire le résultat du *tilt-test* en position couchée [Schang *et al.*, 2006].

5.5.5 Amélioration du processus de sélection des complexes par optimisation automatique du rapport signal sur bruit

Le principal problème des deux processus de sélection de complexes décrits précédemment, réside dans le fait que chacun d'eux nécessite l'action d'un « opérateur » ; ils ne sont donc pas parfaitement automatisés, rendant leur utilisation et leur implémentation difficile. Afin, d'améliorer la sélection des complexes, nous proposons un processus de sélection automatique [Feuilloy *et al.*, 2006b; Feuilloy *et al.*, 2006c]. Ce dernier est développé autour de trois paramètres, N, T et L. Le signal échantillonné de N points est considéré « périodique », où chaque période (approximativement de T points[7]) représente un complexe dZ, comme le montre la figure 5.23. Ainsi, la méthode extrait une partie du signal composée de L points, dans laquelle le nombre de complexes peut être estimé par $K = L/T$. Les paramètres de ce processus doivent respecter la condition suivante : $2\,T < L < N - T$.

À l'image des procédures précédentes de [Bellard *et al.*, 2003] et de [Schang *et al.*, 2007], notre procédure de sélection de complexes cherche dans le signal dZ une partie, appelée aussi fenêtre, où la variation du signal est minimale. Bien évidemment, nous pouvons, en définissant la partie

[7]La variable T peut être estimée en connaissant la fréquence d'échantillonnage (f_e) et la fréquence cardiaque (FC) tel que $T = \frac{f_e \cdot 60}{\mathrm{FC}}$.

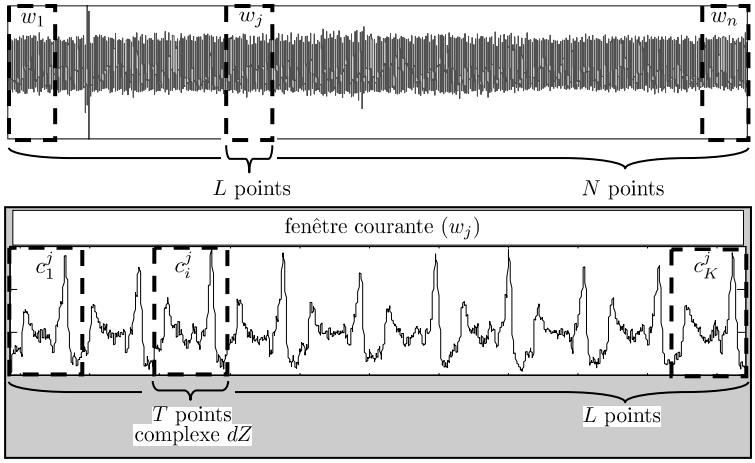


FIG. 5.23 – Illustration des paramètres liés aux méthodes de sélection automatique d'une fenêtre de complexes dZ.

du signal de N points, pré-sélectionner une période du *tilt-test* (phase couchée ou phase basculée).

Le signal dZ est considéré comme un signal périodique bruité. Ainsi, la mesure de l'évolution de la variabilité du signal est obtenue en calculant le rapport signal sur bruit (RSB). Le rapport RSB_{c_i} définit le RSB d'un complexe i :

$$RSB_{c_i} = 10 \cdot \log_{10} \left(\frac{P_{c_i}}{P_{bruit}} \right) , \tag{5.6}$$

où P_{c_i} désigne la puissance moyenne du complexe c_i durant la période de temps T telle que

$$P_{c_i} = \frac{1}{T} \int_{t}^{t+T} [c_i(t)]^2 \, dt \, . \tag{5.7}$$

Le bruit d'un complexe c_i est évalué en le comparant à un complexe défini comme le modèle (noté $c_{modèle}$). Dans nos travaux, deux approches ont été considérées pour déterminer P_{bruit} permettant ainsi d'obtenir le RSB d'un complexe dZ issu d'une fenêtre w_j :

- dans la première approche (voir figure 5.24), le modèle du complexe dZ est défini par la moyenne de tous les complexes issus de l'intervalle L. Au préalable, chaque complexe de L est extrait pour être interpolé et normalisé en un même nombre de points T. Ainsi, le modèle, noté $c_{modèle}$, est calculé par la moyenne de tous les complexes prétraités et le RSB du complexe c_i de la fenêtre w_j est donné par

$$RSB_{c_i^j} = 10 \cdot \log_{10} \left(\frac{P_{c_i^j}}{P_{c_{modèle}^j} - P_{c_i^j}} \right) . \tag{5.8}$$

  Le RSB sur l'ensemble de la fenêtre w_j est donc obtenu par

$$RSB_{w_j} = \frac{1}{L/T} \cdot \sum_{i=1}^{K} RSB_{c_i^j} \, . \tag{5.9}$$

  Cette méthode est appelée sélection automatique par optimisation globale (OG) du RSB ;

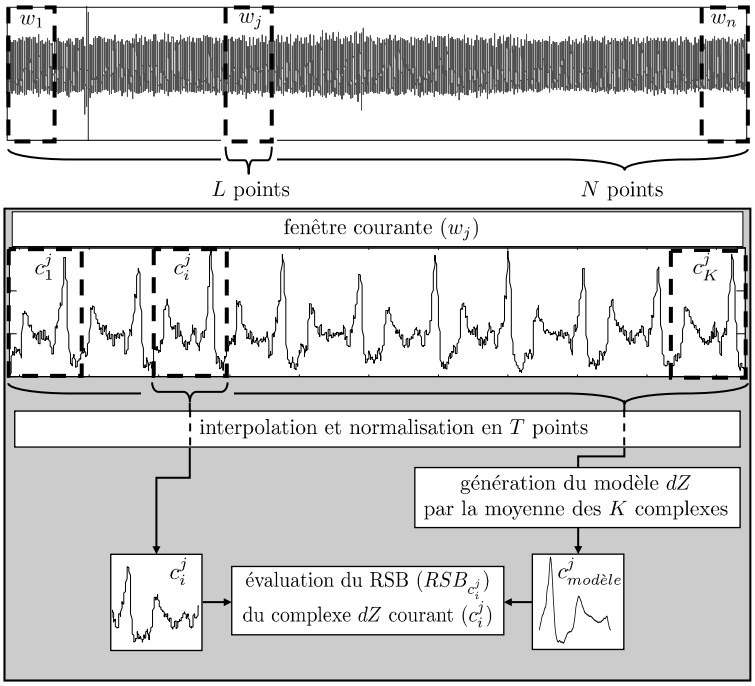


FIG. 5.24 – Sélection automatique d'une fenêtre de complexes dZ par optimisation globale (OG) du rapport signal sur bruit.

- dans la seconde approche (voir figure 5.25), le modèle est défini par le complexe précédant le complexe à évaluer c_i : $c_{modèle} \equiv c_{i-1}$. Dans ce cas, le RSB d'un complexe c_i est déterminé en utilisant deux complexes successifs. Comme précédemment, les deux complexes extraits de L doivent être interpolés et normalisés en un même nombre de points T. Le RSB du complexe c_i de la fenêtre w_j est donné par

$$RSB_{c_i^j} = 10 \cdot \log_{10} \left(\frac{P_{c_i^j}}{P_{c_{i-1}^j} - P_{c_i^j}} \right) . \tag{5.10}$$

  Le RSB sur l'ensemble de la fenêtre w_j est donc obtenu par

$$RSB_{w_j} = \frac{1}{L/T - 1} \cdot \sum_{i=2}^{K} RSB_{c_i^j} . \tag{5.11}$$

  Cette méthode est appelée sélection automatique par optimisation locale (OL) du RSB.

L'évaluation de la fenêtre w_{j+1} est réalisée en éliminant le complexe c_1^j de la fenêtre w_j et en ajoutant le complexe c_{K+1}^j, qui devient c_K^{j+1}.

L'étape finale consiste à choisir parmi toutes les fenêtres évaluées la fenêtre $w_{opt.}$ ayant le plus fort RSB tel que :

$$w_{opt.} = \underset{\forall w_j}{\operatorname{argmax}} \, RSB_{w_j} , \tag{5.12}$$

révélant ainsi la fenêtre possédant une faible variabilité et une périodicité optimale. Cette propriété aura une grande importance lorsqu'à la section 5.5.5.2 nous travaillerons dans le domaine fréquentiel.

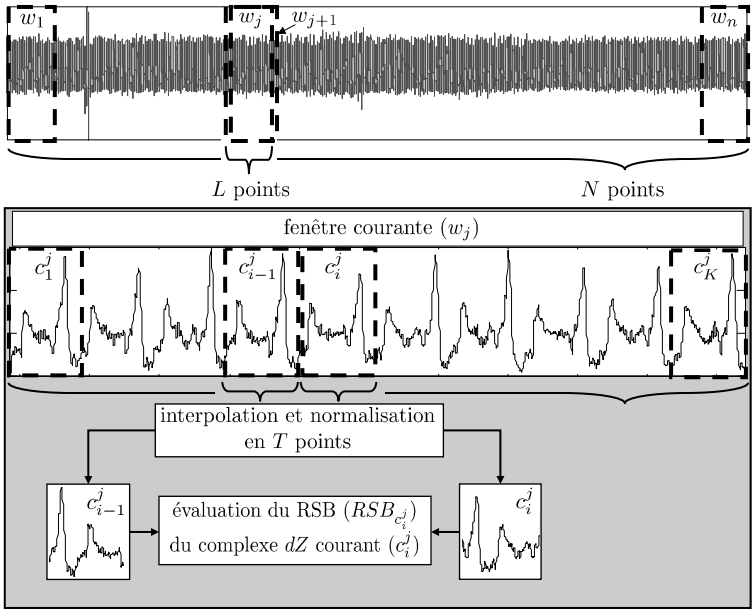


FIG. 5.25 – Sélection automatique d'une fenêtre de complexes dZ par optimisation locale (OL) du rapport signal sur bruit.

Ces deux méthodes de sélection sont comparées dans une analyse expérimentale (*cf.* section 5.5.5.1). Toutefois, nous pouvons déjà observer que le processus nommé OL apparaît plus rapide que son homologue OG. En effet, l'évaluation du RSB de la fenêtre w_{j+1} nécessite le calcul supplémentaire d'un seul RSB d'un complexe, si les RSB des complexes de la fenêtre w_j sont connus, contrairement à la méthode OG, qui doit recalculer le modèle dZ de la nouvelle fenêtre en considérant le nouveau complexe.

5.5.5.1 Évaluation de la pertinence de l'extraction des caractéristiques liées au signal dZ en fonction des processus automatiques de sélection des complexes

Dans cette section, nous allons démontrer l'efficacité des processus de sélection des complexes, en évaluant la capacité des caractéristiques liées au signal dZ à prédire le résultat du *tilt-test*.

Pour réaliser ces expérimentations publiées dans [Feuilloy *et al.*, 2006b], nous considérons une nouvelle fois l'échantillon appelé $\mathcal{E}_2$ décrit à la section 5.2. Rappelons que cet échantillon, qui était initialement composé de 138 patients, a vu son nombre réduit à 128 (*cf.* section 5.5.4). Ainsi, pour les 128 patients restants, 65 (51%, 40±15 ans, 35 femmes, 30 hommes) ont eu une réponse négative au *tilt-test* et 63 (49%, 45±15 ans, 31 femmes, 32 hommes) ont eu une réponse positive au *tilt-test*.

Les processus de sélection étant fondés sur une estimation du RSB, il nous est apparu judicieux de comparer l'évolution de la pertinence des caractéristiques en fonction des niveaux de RSB relevés sur les fenêtres et les complexes. En outre, afin d'améliorer l'interprétation des résultats, nous évaluons chaque caractéristique individuellement. Cette étude ne cherche donc pas à optimiser scrupuleusement la prédiction du résultat du *tilt-test*, mais elle s'attache à analyser l'amélioration de la pertinence des caractéristiques en fonction des complexes sélectionnés. C'est ainsi que nous avons fait le choix d'utiliser une mesure particulière pour évaluer l'évolution de la pertinence des caractéristiques ; celle-ci peut être définie comme la qualité de discrimination. Dès lors, contrairement aux analyses précédentes, nous pouvons considérer l'ensemble des patients disponibles pour évaluer la discrimination des classes. Cette discrimination peut être optimisée

suivant la règle de Bayes (*cf.* section 1.2.2.1, page 16) en minimisant le chevauchement des classes comme le montre la figure 5.26.

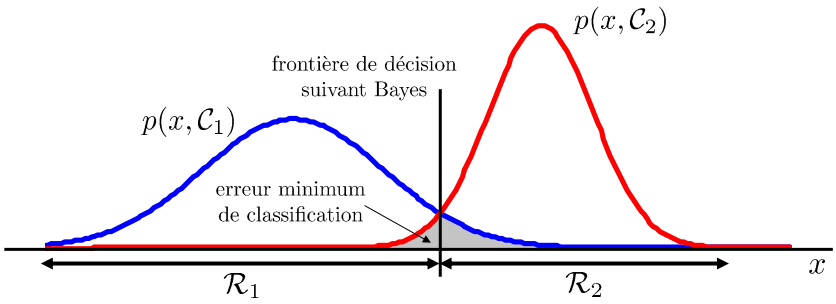


FIG. 5.26 – Illustration de la probabilité d'erreur obtenue par la règle de Bayes.

La pertinence d'une variable est alors considérée en fonction du recouvrement entre les classes, et notre mesure de performance est tout simplement le rapport entre la surface du recouvrement et les surfaces des densités de probabilité $p(x, C_1)$[8] et $p(x, C_2)$[8]. Les détails de cette mesure, notée P_{err} sont donnés à la section A.3. Notons que [Duda and Hart, 1973] l'interprètent comme une probabilité d'erreur telle que

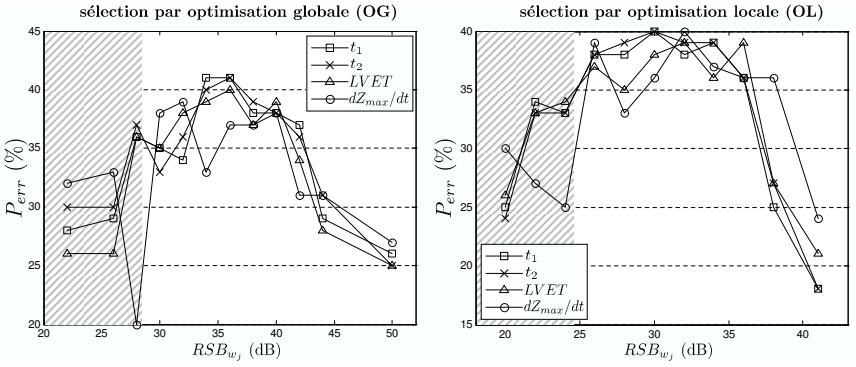
$$\begin{aligned} P_{err} &= P(x \in R_2, C_1) + P(x \in R_1, C_2) \\ &= \int_{R_2} p(x|C_1)P(C_1)\,dx + \int_{R_1} p(x|C_2)P(C_2)\,dx\,. \end{aligned} \tag{5.13}$$

Si les deux densités se chevauchent complètement, alors la probabilité du modèle à réaliser une erreur de classification est de 50 % ; la mesure indique par ailleurs le taux de recouvrement des classes, qui peut s'interpréter comme le risque d'erreur.

Une fois la mesure de performance établie, nous pouvons désormais évaluer l'influence de la sélection des complexes pour prédire le résultat du *tilt-test* lors de la phase couchée ; la durée de cette phase définit l'intervalle de temps de N points. La figure 5.27 relate la pertinence des caractéristiques issues de dZ (t_1, t_2, $LVET$ et dZ_{max}/dt), en fonction du niveau du RSB mesuré sur une fenêtre donnée et nommée précédemment RSB_{w_j}). Cette fenêtre est sélectionnée par les méthodes basées sur l'optimisation globale et locale ; les équations sont données respectivement par les relations (5.9) et (5.11). La taille (L points) de la fenêtre w_j est d'une minute : pour une fréquence cardiaque de 60 bpm, 60 complexes seraient alors sélectionnés. Les caractéristiques sont alors mesurées sur chaque complexe de la fenêtre, afin d'en obtenir une valeur moyenne. Dès lors, la mesure de la probabilité de l'erreur de classification (P_{err}) est évaluée en considérant les valeurs moyennes de tous les patients. Les résultats donnés sur cette figure permettent d'observer globalement une amélioration des performances lorsque le RSB de la fenêtre sélectionnée augmente, notamment pour la méthode de sélection automatique par optimisation locale, où l'on obtient au final un recouvrement des classes autour de 20 %.

L'obtention des caractéristiques sur l'ensemble de la fenêtre n'est pas forcément optimal. En effet, si la taille de la fenêtre est grande, des complexes potentiellement incorrects ou aberrants présents dans la fenêtre peuvent biaiser les calculs des caractéristiques. Ainsi, nous réalisons une seconde phase d'expérimentation dans laquelle les caractéristiques sont extraites sur un **seul** complexe de la fenêtre. Ces caractéristiques ne sont donc plus considérées comme les moyennes des

[8] Rappelons qu'à la section 1.2.2.2, nous avons défini $p(x, C_k) = p(x|C_k)\,P(C_k)$.



Note : Les zones hachurées indiquent que la mesure du RSB est effectuée sur un échantillon possédant un déséquilibre entre les classes ou un nombre limité de patients : la précision des résultats donnés dans cette zone est donc approximative.

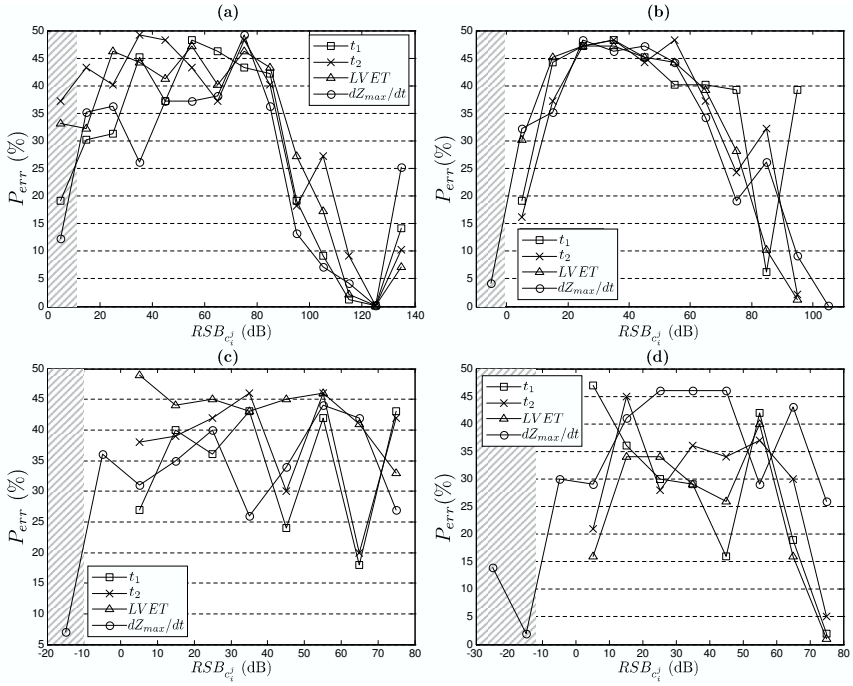
FIG. 5.27 – Influence du RSB de la fenêtre (RSB_{w_j}) sur la pertinence des caractéristiques issues du signal dZ, lors de la sélection de fenêtre par optimisation globale et locale.

caractéristiques des complexes de la fenêtre. La figure 5.28 montre l'évolution de la probabilité de l'erreur de classification en fonction du niveau du RSB mesuré sur le complexe, où les caractéristiques sont extraites ; ce rapport signal sur bruit est noté $RSB_{c_i^j}$ pour les méthodes d'optimisation globale (5.8) et locale (5.10). Notons que dans ce cadre expérimental, les méthodes d'optimisation globale et locale déterminent chacune la fenêtre w_j qui maximise le RSB pour chaque patient. Plus précisément, la fenêtre w_j, de L points consécutifs, est extraite parmi les N points originaux (voir figure 5.23). C'est donc sur ces deux fenêtres que sont choisis les complexes c_i^j, sur lesquels les caractéristiques sont extraites.

L'évolution de la probabilité d'erreur, visible sur la figure 5.28, confirme qu'il est préférable de travailler sur une fenêtre possédant le plus important RSB, quelle que soit la méthode de sélection de la fenêtre utilisée. D'autre part, cette même figure montre qu'il est également important d'extraire les caractéristiques sur des complexes possédant le plus grand RSB.

L'extraction des caractéristiques par la moyenne des complexes (voir figure 5.27), ou par le complexe ayant le plus grand RSB de la fenêtre (voir figure 5.28), ressemble respectivement aux approches présentées par [Bellard *et al.*, 2003] et [Schang *et al.*, 2007]. L'analyse des figures 5.27 et 5.28, nous montre que la pertinence des caractéristiques extraites est fortement dépendante de leur mode de calcul. Ainsi, l'extraction des caractéristiques doit-elle se faire en considérant les complexes ayant les plus grands RSB de la fenêtre ou en déterminant la valeur moyenne des caractéristiques sur tous les complexes de la fenêtre ? D'autre part, ces méthodes de sélection de complexes et d'extraction de caractéristiques ont-elles réellement un impact sur les performances obtenues ? Pour répondre à ces questions, nous proposons une dernière analyse dans laquelle nous comparons nos processus à des sélections aléatoires de fenêtres et de complexes. La distribution des résultats obtenus par la sélection ou l'extraction aléatoire est donnée à la figure 5.29. L'« extraction aléatoire » signifie que les caractéristiques issues du signal dZ sont mesurées à partir de complexes choisis aléatoirement.

Les résultats des combinaisons associant chaque processus de sélection et chaque méthode d'extraction sont donnés dans le tableau 5.19. Les distributions illustrées par la figure 5.29 montrent que l'utilisation d'une méthode de sélection (par optimisation globale ou locale) influence sensiblement la probabilité d'erreur. Cependant, ce n'est que par l'association d'un processus de sélection et d'une méthode d'extraction des caractéristiques que la probabilité d'erreur baisse significati-



Note : Les zones hachurées indiquent que la mesure du RSB est effectuée sur un échantillon possédant un déséquilibre entre les classes ou un nombre limité de patients : la précision des résultats donnés dans cette zone est donc approximative.

(a) sélection par optimisation globale et utilisation du plus grand $RSB_{c_i^j}$. (b) sélection par optimisation locale et utilisation du plus grand $RSB_{c_i^j}$. (c) sélection par optimisation globale et utilisation du plus petit $RSB_{c_i^j}$. (d) sélection par optimisation locale et utilisation du plus petit $RSB_{c_i^j}$.

FIG. 5.28 – Influence du RSB du complexe ($RSB_{c_i^j}$) sur la pertinence des caractéristiques issues du signal dZ, lors de la sélection de fenêtre par optimisation globale et locale.

vement, comme le montre le tableau 5.19. Pour finir, même si une méthodologie ne se détache pas réellement des autres, le processus de sélection par optimisation locale du RSB semble être le plus performant, comme cela a pu être observé dans les analyses précédentes.

5.5.5.2 Extraction de caractéristiques prédictives dans le domaine fréquentiel

Dans ce manuscrit, nous avons jusqu'à présent extrait des variables sur les signaux Z et dZ dans le domaine temporel. Or, certaines études ont montré que le domaine fréquentiel pouvait apporter de précieuses informations ; citons notamment les travaux de [Pagani *et al.*, 1986; Malliani, 1999; Ebden *et al.*, 2004]. Ces auteurs ont étudié l'évolution dans le domaine fréquentiel de caractéristiques extraites à partir de l'ECG, dans le cadre de la prédiction de la syncope pour des patients réalisant l'examen du *tilt-test*.

Dans notre application, le passage dans le domaine fréquentiel n'est pas aussi intuitif qu'il pourrait sembler. En effet, les signaux, provenant du signal d'impédancemétrie thoracique, présentent, comme une grande majorité des signaux extraits de phénomènes réels, un aspect aléatoire[9]. Dès lors, les techniques habituellement utilisées pour obtenir la transformée de Fourier discrète de signaux déterministes, telles que l'algorithme très populaire de la Transformée de Fourier Rapide, ne peuvent être appliquées directement à des signaux aléatoires et non stationnaires [Kunt, 1981], car le contenu spectral de ces signaux évolue en fonction du temps. Cependant,

[9]L'aspect aléatoire ne permet pas de prévoir la forme des signaux ni de les décrire de manière analytique.

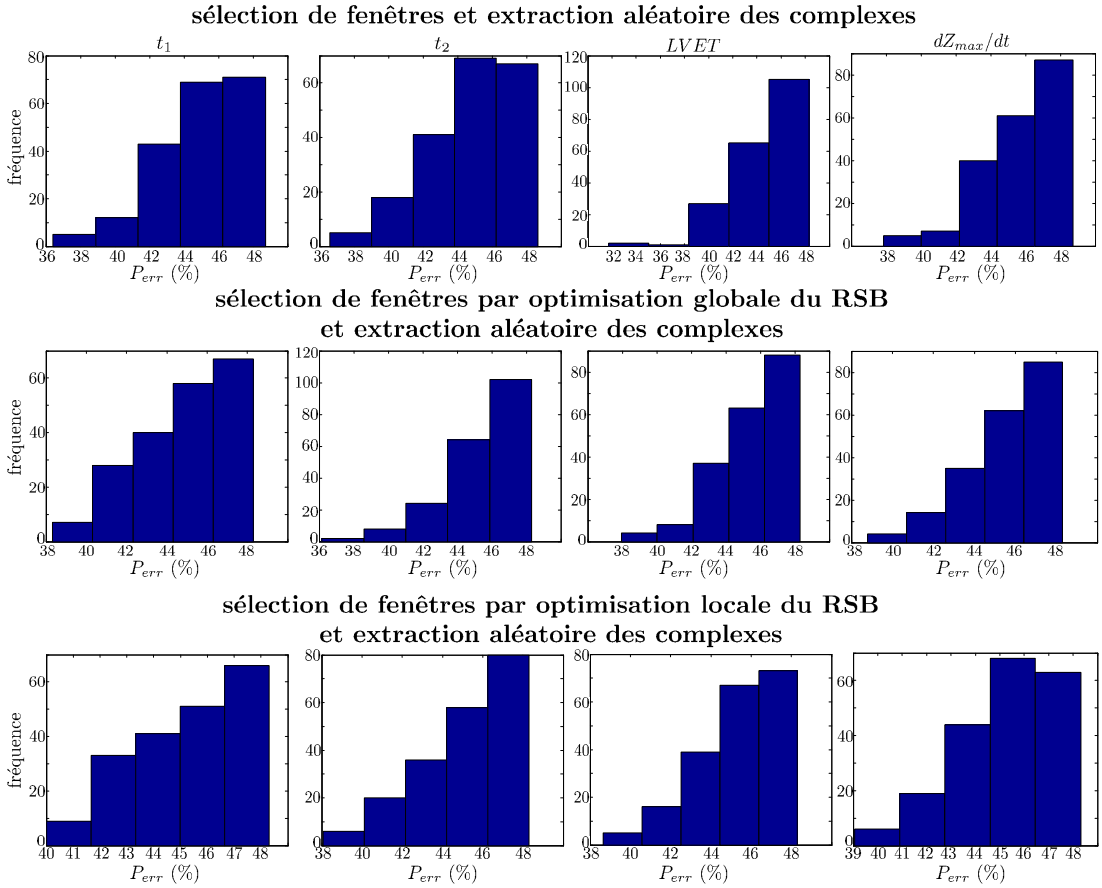


FIG. 5.29 – Influence de la sélection de fenêtres (aléatoire, optimisation globale et locale) sur la pertinence (P_{err}) des caractéristiques issues du signal dZ, lors d'extraction aléatoire de complexes.

l'allure générale des signaux (ECG et Z) étant connue, l'aspect aléatoire n'est donc que relatif et peut surtout être attribué à du bruit provenant de l'acquisition. En effet, comme évoqué à la section 4.2.2, de nombreux facteurs, tels que des fortes respirations, des contractions musculaires, des mouvements ou encore le moteur faisant basculer la table d'examen, peuvent perturber l'acquisition des signaux en introduisant un bruit non négligeable qui contribue, par conséquent, à accentuer l'aspect aléatoire des signaux. La présence de ce bruit nous incite alors à extraire la portion du signal présentant le plus fort rapport signal sur bruit, ce qui permettrait de réduire significativement l'aspect aléatoire des signaux en conservant la portion du signal la plus stationnaire (caractérisée par la plus grande périodicité).

C'est dans cette optique qu'apparaît l'intérêt de nos processus de sélection automatique d'une fenêtre de complexes, notamment par optimisation locale (*cf.* section 5.5.5), qui privilégie une fenêtre où l'allure de chaque complexe varie le moins possible. Ainsi, la portion du signal extrait peut être caractérisée comme un **signal stationnaire**. Par cette propriété, [Kunt, 1981] introduit alors la notion de densité spectrale de puissance (DSP, en anglais PSD pour *Power Spectral Density*) afin de déduire la description fréquentielle des signaux aléatoires. Par ailleurs, rappelons qu'à la section 5.5.5.1, cette méthode de sélection de complexes s'est révélée être la plus performante, nous encourageant d'autant plus à l'utiliser afin d'extraire la partie du signal dZ à analyser.

Dans cette section, nous considérons une nouvelle fois l'échantillon appelé $\mathcal{E}_2$ décrit à la section 5.2. Rappelons qu'une fois le prétraitement effectué (*cf.* section 5.5.5.1), le nombre de patients restant est de 128.

extraction des caractéristiques	**probabilité de l'erreur de classification (P_{err})** t_1	t_2	$LVET$	dZ_{max}/dt
sélection aléatoire de la fenêtre contenant les complexes				
par choix aléatoire des complexes	$45 \pm 3^*$	$45 \pm 2^*$	$45 \pm 3^*$	$46 \pm 2^*$
par la moyenne des complexes	46 ± 2	45 ± 2	44 ± 3	44 ± 2
par les complexes optimisant le RSB	44 ± 3	44 ± 3	44 ± 3	46 ± 2
sélection de la fenêtre par la méthode d'optimisation globale				
par choix aléatoire des complexes	$45 \pm 3^*$	$46 \pm 2^*$	$46 \pm 2^*$	$46 \pm 2^*$
par la moyenne des complexes	30	41	40	41
par les complexes optimisant le RSB	48	48	47	42
sélection de la fenêtre par la méthode d'optimisation locale				
par choix aléatoire des complexes	$45 \pm 3^*$	$45 \pm 2^*$	$45 \pm 2^*$	$45 \pm 2^*$
par la moyenne des complexes	30	40	29	37
par les complexes optimisant le RSB	42	46	48	40

Note : * indique que les résultats sont liés aux distributions représentées à la figure 5.29.

TAB. 5.19 – Évaluation de la pertinence (P_{err}) des caractéristiques issues du signal dZ en fonction du processus de sélection de fenêtres (aléatoire, optimisation globale et locale) et de la méthode d'extraction des caractéristiques (choix aléatoire des complexes, moyenne des complexes et complexes optimisant le RSB).

L'approche proposée ici, cherche de nouvelles caractéristiques en travaillant dans le domaine fréquentiel, par le biais du calcul de la densité spectrale de puissance. La DSP de dZ (5.14), notée $\widehat{S}$, est estimée par la méthode de Welch [Welch, 1967], qui est une version améliorée du périodogramme. La méthode de Welch permet de réduire la variance de l'estimation, par le calcul de la moyenne de plusieurs DSP, chacune réalisée sur une portion du signal original. Ces portions, appelées plus couramment segments, peuvent également se chevaucher. D'autre part, l'utilisation d'une fenêtre (en l'occurrence Hamming) sur chaque segment modifie également le biais : nous noterons $\widetilde{x}(n) = x(n) \cdot w_h(n)$, le résultat de l'utilisation de la fenêtre w_h sur un segment x du signal dZ. Ces différents paramètres amènent à donner la relation de $\widehat{S}$ suivante :

$$\hat{S}(n) = \frac{1}{M} \sum_{m=0}^{M-1} \left(\frac{1}{K} \left| \sum_{k=0}^{K-1} \tilde{x}(mK/2 + k) e^{-2j\pi \frac{kn}{K}} \right|^2 \right) . \tag{5.14}$$

Dans notre étude, nous estimons la DSP sur une durée d'une minute du signal dZ, enregistrée pendant la période de repos du *tilt-test* (équivalent à un échantillon de 15 340 points). En considérant un segment K de 2 048 points (8, 2 secondes) et un recouvrement de 50%, le nombre de segments M est alors égal à 15. D'autre part, avec une fréquence d'échantillonnage de 250 Hz et en considérant un segment de 2 048 points, la résolution spectrale est alors de 0, 122 Hz.

Les amplitudes des fréquences obtenues lors du calcul de la DSP déterminent alors les nouvelles caractéristiques, qui doivent permettre de prédire le résultat du *tilt-test*. Cependant comme le montre la figure 5.30, toutes les fréquences ne peuvent pas contribuer à discriminer les deux réponses du *tilt-test*, et d'autre part, l'utilisation complète de la DSP, comme entrée du modèle, entraînerait un problème majeur de dimensionnalité (*cf.* section 2.1). Nous avons donc fait le choix d'extraire un certain nombre de fréquences qui pourraient être caractéristiques de l'apparition des symptômes de la syncope.

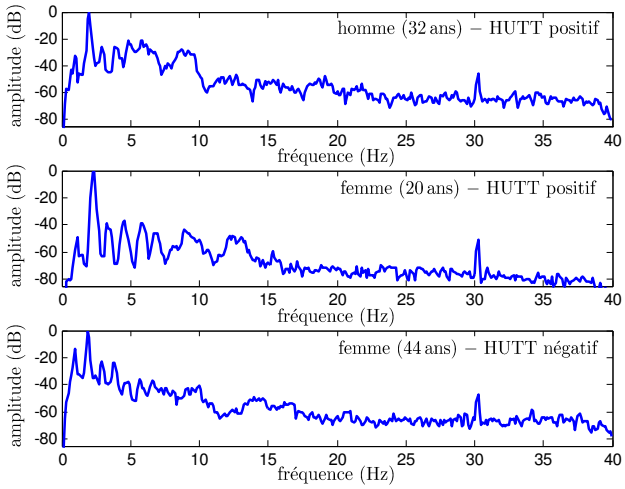


FIG. 5.30 – Illustrations de la densité spectrale de puissance (DSP) obtenue par la méthode de Welch.

La procédure mise en place est quelque peu empirique, due à un manque d'information dans la littérature quant au traitement fréquentiel du signal d'impédancemétrie thoracique. Ainsi, nous avons, dans un premier temps, relevé consciencieusement pour chaque patient de l'échantillon d'apprentissage, toutes les fréquences apparentes[10] sur la DSP estimée. Puis, dans un second temps, nous avons conservé un ensemble de fréquences à $\pm$ 0, 122 Hz, communes à chaque classe de patients ; celles-ci sont représentées à la figure 5.31, où l'identification des 40 indices est donnée dans le tableau 5.20.

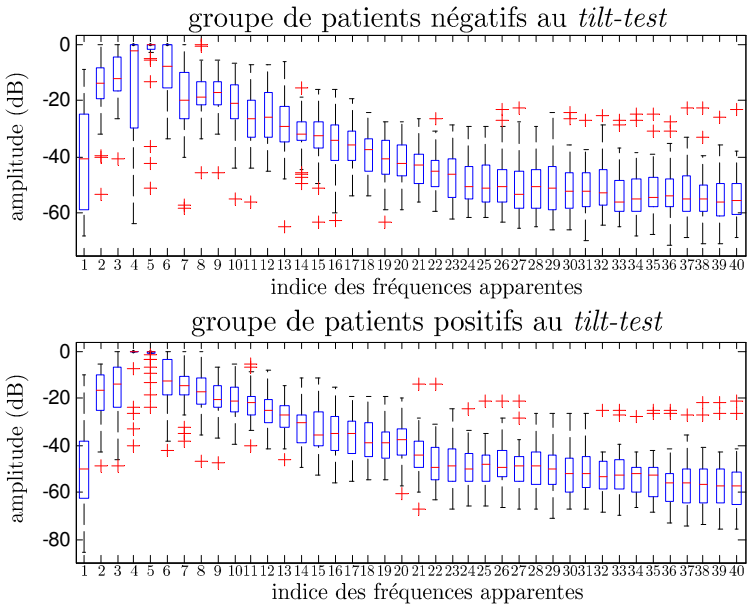
indice	1	2	3	4	5	6	7	8	9	10
fréquence (Hz)	1, 05	1, 17	2, 23	2, 34	4, 34	4, 69	5, 74	6, 77	7, 76	8, 79
indice	11	12	13	14	15	16	17	18	19	20
fréquence (Hz)	10, 81	11, 62	12, 75	13, 80	14, 71	15, 66	16, 70	17, 58	18, 72	19, 69
indice	21	22	23	24	25	26	27	28	29	30
fréquence (Hz)	20, 78	21, 74	22, 69	23, 79	24, 84	25, 76	26, 25	27, 54	28, 79	29, 68
indice	31	32	33	34	35	36	37	38	39	40
fréquence (Hz)	30, 24	31, 82	32, 93	33, 77	34, 64	35, 86	36, 69	37, 64	39, 26	39, 87

Note : Chacune des fréquences données correspond à la fréquence moyenne relevée sur l'ensemble des patients à $\pm$0, 122 Hz.

TAB. 5.20 – Identification des indices correspondant aux 40 fréquences pré-sélectionnées sur la densité spectrale de puissance.

Les 40 amplitudes correspondantes à ces 40 fréquences sont considérées comme nos nouvelles variables d'entrées, dont quelques mesures statistiques sont fournies à la figure 5.32 (les amplitudes dont les fréquences étaient manquantes sur certains patients n'ont pas été remplacées).

[10]Une fréquence est considérée apparente lorsque celle-ci correspond à un maximum local de la DSP, autrement dit à un pic.



Note : Les distributions des amplitudes de chaque fréquence sont données par des boîtes à moustaches (*boxplot*), indiquant la médiane (trait rouge), les premier ($Q1$) et troisième ($Q3$) quartiles (respectivement minimum et maximum de la boîte bleue), ainsi que les valeurs extrêmes (traits noirs). Les croix rouges indiquent la présence de valeurs aberrantes selon le seuil $[Q1 - 1,5(Q3 - Q1)\,,\, Q3 + 1,5(Q3 - Q1)]$, donné à la section 2.2.2 (page 61).

FIG. 5.31 – Analyse de la distribution des amplitudes de chaque fréquence pré-sélectionnée sur la DSP pour chaque groupe de patients (positif et négatif) de l'échantillon d'apprentissage.

D'autre part, pour qu'une variable puisse être considérée en tant que telle, il est nécessaire qu'elle n'apparaisse que si le résultat d'un patient au *tilt-test* est négatif ou positif. Ainsi, aux termes de notre analyse nous avons pré-sélectionné 14 variables comme étant les plus robustes. L'indice de Fisher (*cf.* section 2.4.2.2) nous a permis de les trier par ordre de pertinence. Représentées à la figure 5.33, les fréquences à $\pm 0,122$ Hz sur lesquelles les amplitudes sont mesurées, sont par ordre décroissant de pertinence $1,17$ Hz, $1,05$ Hz, $4,69$ Hz, $17,58$ Hz, $39,26$ Hz, $5,85$ Hz, $26,25$ Hz, $2,34$ Hz, $4,34$ Hz, $5,74$ Hz, $27,54$ Hz, $30,24$ Hz, $32,93$ Hz et $2,23$ Hz.

Après la pré-sélection des 14 variables, il reste à définir les variables les plus adaptées pour prédire le résultat du *tilt-test*. Pour atteindre cet objectif, nous avons réalisé une démarche similaire à celle utilisée à la section 5.4.2.1, en adoptant une recherche du sous-ensemble optimal de variables par une sélection naïve. Rappelons qu'à chaque itération, cette méthode de sélection ajoute une nouvelle variable en suivant l'ordre induit par un critère ; dans notre étude, nous avons utilisé le critère de Fisher. D'autre part, avec un nombre limité de 14 variables, la méthode de sélection naïve produit uniquement 14 combinaisons possibles. Dès lors, l'évaluation de chaque sous-ensemble peut être réalisée par des réseaux de neurones.

Comme dans les analyses précédentes (*cf.* section 5.3.1.2), pour chaque sous-ensemble de variables, nous recherchons la meilleure architecture du réseau en imposant au préalable une seule couche cachée, dont le nombre de neurones varie entre 2 et 20. Les fonctions d'activation des neurones sont encore de type sigmoïde (tangente hyperbolique) et l'algorithme d'apprentissage, basé sur Levenberg-Marquardt, est associé à la méthode de régularisation de « l'arrêt prématuré », afin d'améliorer les performances de généralisation. Chaque combinaison de variables est évaluée, pour chaque architecture, au cours de 100 apprentissages avec des initialisations différentes des poids.

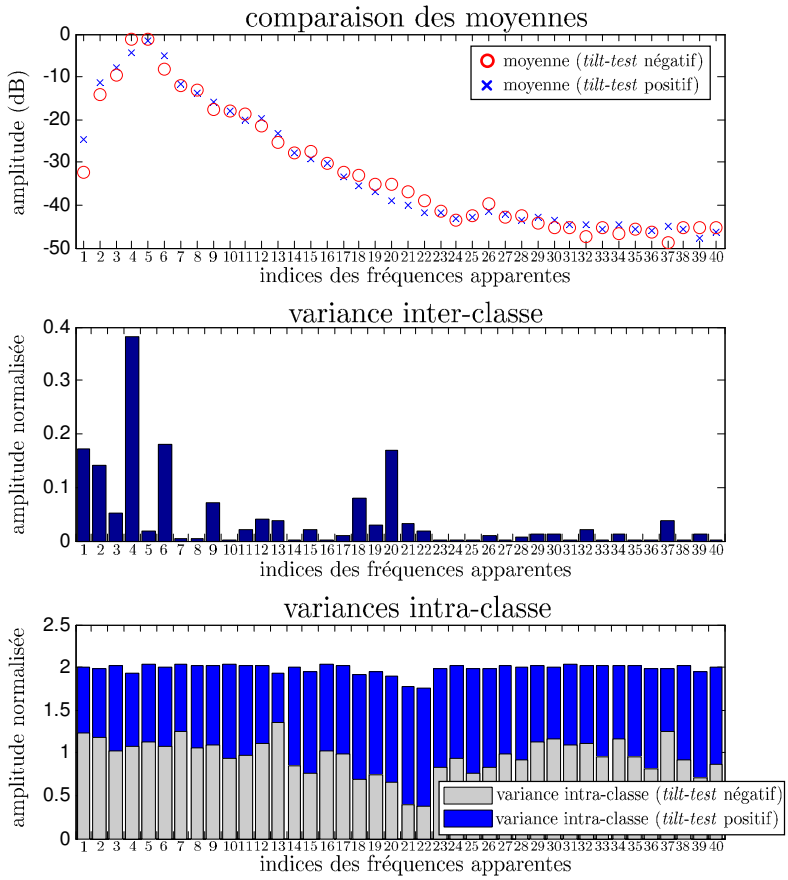


FIG. 5.32 – Analyse de mesures statistiques des amplitudes de chaque fréquence pré-sélectionnée sur la DSP pour chaque groupe de patients (positif et négatif) de l'échantillon d'apprentissage.

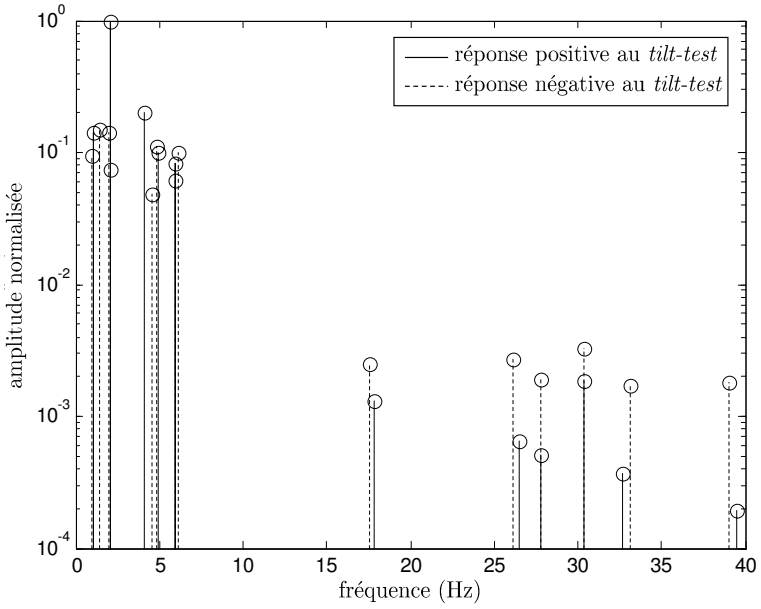
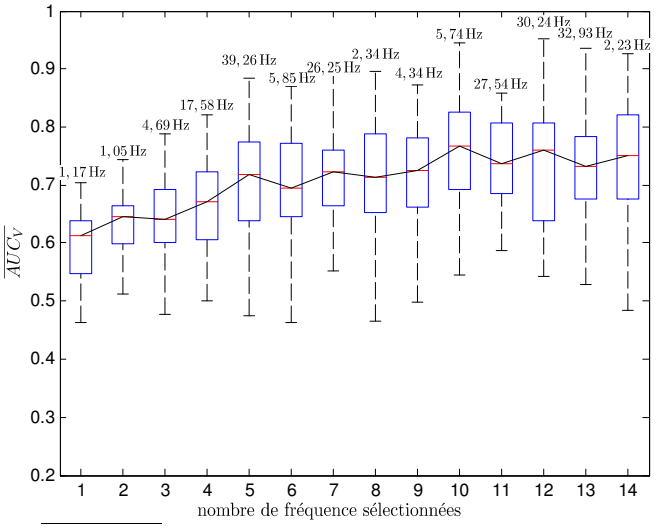


FIG. 5.33 – Illustration des 14 fréquences sélectionnées et de leur amplitude pour un patient de chaque classe (positif et négatif au *tilt-test*).

Le processus d'évaluation par validation croisée est identique à celui utilisé dans les études précédentes et décrit à la section 5.3.1.2 (page 131). Ainsi, l'échantillon initial de 128 patients est divisé en deux groupes : le premier groupe ($\mathcal{X}_A$), composé de 69 patients, est utilisé pour construire les modèles et déterminer leurs caractéristiques. Ce sous-ensemble de patients est lui-même partitionné en 7 sous-ensembles afin d'estimer plus précisément les performances de généralisation (la validation) par validation croisée. Le second groupe ($\mathcal{X}_T$), composé de 59 patients, est utilisé pour évaluer la reproductibilité des modèles construits. Rappelons que les patients de ce sous-ensemble ne sont ni employés pour la construction, ni pour la validation des modèles ; ils permettent de donner une estimation sans biais des performances des modèles.

La figure 5.34 montre les performances moyennes de validation ($\overline{AUC_V}$), obtenues par les réseaux de neurones dont le nombre de neurones de la couche cachée est optimisé. Ces résultats sont donnés en augmentant progressivement le nombre d'entrées du modèle : par exemple, si le nombre de fréquences sélectionnées est 4, alors les variables d'entrées du modèles sont liées aux amplitudes des fréquences 1, 17 Hz, 1, 05 Hz, 4, 69 Hz et 17, 58 Hz. Ainsi, comme le montre la figure 5.34, en utilisant les dix premières fréquences, la prédiction du résultat du *tilt-test* est optimale, avec une aire moyenne sous la courbe de ROC de 0, 766. Notons que cette performance moyenne est obtenue avec des réseaux composés de 9 neurones dans la couche cachée.



Note : La distribution des résultats est donnée par des boîtes à moustaches (*boxplot*), indiquant la médiane (trait rouge), les premier ($Q1$) et troisième ($Q3$) quartiles (respectivement minimum et maximum de la boîte bleue), ainsi que les valeurs extrêmes (traits noirs).

FIG. 5.34 – Évolution de l'aire moyenne sous la courbe de ROC (en validation) des classifieurs PMC issus de la sélection naïve de fréquence (SFS_{Fisher}).

Les amplitudes de ces dix fréquences sont extraites de la DSP estimée, sur chacun des patients de l'échantillon de test (noté $\mathcal{X}_T$). Sur les 100 modèles, générés lors de l'apprentissage avec 9 neurones dans la couche cachée, l'estimation moyenne de la généralisation atteint une valeur similaire à celle de validation, avec une aire moyenne sous la courbe de ROC de 0, 794 (0, 728 et 0, 856 pour respectivement le premier et le troisième quartile). Aussi, avec le meilleur modèle (optimisant la validation), l'AUC obtenue sur le sous-ensemble de test atteint 0, 967, avec une sensibilité de 100% et une spécificité de 97%.

5.5.6 Discussions et conclusions

Dans cette section, réservée exclusivement à l'analyse du signal d'impédancemétrie thoracique, nous avons, dans le cadre de la prédiction de la syncope, étudié la pertinence des caractéristiques extraites dans les domaines temporel et fréquentiel en fonction, notamment, de la partie du signal sélectionnée. Le tableau 5.21 récapitule les résultats les plus importants obtenus et présentés dans cette section. Ces résultats révèlent le caractère informatif du signal d'impédancemétrie pour la prédiction du résultat du *tilt-test* sur des patients sujets à l'apparition de syncopes inexpliquées. En effet, lors de l'utilisation de modèles non linéaires [Feuilloy *et al.*, 2006c; Schang *et al.*, 2007] (par des PMC et des SVM), nous avons pu obtenir une sensibilité à la prédiction de la syncope

dans la phase de repos, respectivement de 100% et de 94% (avec une spécificité de 97% et de 79%). Notons cependant que [Bellard *et al.*, 2003] se sont plus attachés à déterminer des règles simples et compréhensibles plutôt que d'optimiser la performance de prédiction du résultat du *tilt-test*. Ils ont ainsi cherché pour chaque variable le seuil idéal : c'est par cette approche, qu'ils ont pu prédire l'apparition de la syncope avec une sensibilité de 68% et une spécificité de 63%, lorsque $t_2 < 199\,\text{ms}$. La qualité de la comparaison entre ces études est, malgré tout, renforcée par le fait que les échantillons de patients de ces quatre études proviennent de mêmes études cliniques réalisées au CHU d'Angers.

étude du signal d'impédancemétrie	**période du *tilt-test***	**nombre de patients**	$S_e\,(\%)$	$S_p\,(\%)$	$VPP\,(\%)$	$VPN\,(\%)$
dans le domaine fréquentiel [Feuilloy *et al.*, 2006c]	période de repos	59	100	97	96	100
dans le domaine temporel [Schang *et al.*, 2007]	période de repos	64	94	79	74	93
dans le domaine temporel [Schang *et al.*, 2006]	période de repos	59	88	64	66	88
dans le domaine temporel [Bellard *et al.*, 2003]	période de repos	68	68*	63*	63*	68*
dans le domaine temporel [Bellard *et al.*, 2003]	5 à 10-ième min du basculement	68	68*	70*	68*	70*

Note : * Indique que les résultats sont obtenus sur le sous-ensemble d'apprentissage.

TAB. 5.21 – Comparaison des résultats obtenus lors de l'analyse exclusive du signal d'impédancemétrie thoracique, dans le cadre de la prédiction du résultat du *tilt-test* en position couchée, avec les principales études analysant la syncope inexpliquée.

Les travaux développés par [Feuilloy *et al.*, 2006c] ont montré que par la seule utilisation du signal d'impédancemétrie thoracique, la prédiction de l'apparition des symptômes lors de la phase couchée du *tilt-test* pouvait être réalisée avec une sensibilité de 100% et une spécificité de 97%. Ainsi, en comparant ces résultats avec les meilleures performances obtenues quand la phase basculée était utilisée (*cf.* section 5.4, où la sensibilité et la spécificité étaient respectivement de 100% et de 94%), nous pouvons remarquer que le signal d'impédancemétrie thoracique peut contenir suffisamment d'informations pour suggérer sa seule utilisation pour la prédiction de la syncope. De ce fait, le processus de prédiction serait simplifié, en évitant l'acquisition de nombreuses autres variables.

Les résultats, exposés à la section 5.5.5.1 lors de la comparaison des procédures de sélection des complexes, ont montré l'importance et l'influence de cette phase sur la qualité des caractéristiques extraites sur le signal d'impédancemétrie thoracique. En effet, en comparant nos méthodes de sélection à des sélections aléatoires, la qualité de la discrimination des caractéristiques a pu être considérablement améliorée. D'autre part, l'aspect automatique du processus facilite désormais la possibilité d'intégrer, dans des appareils d'acquisition, une phase de traitement et d'analyses des variables liées au signal d'impédancemétrie thoracique.

5.6 Discussions

Les travaux, développés dans ce chapitre, ont permis d'explorer la problématique liée à l'apparition récurrente de syncopes inexpliquées. La thématique concernait plus précisément la prédiction du résultat du *tilt-test* pour des patients sujets à ce type de syncope. La description de l'examen du *tilt-test* avait révélé son principal défaut : sa durée. En effet, rappelons que celle-ci peut atteindre une heure, si le patient n'éprouve pas de symptômes. D'autre part, la préparation du test et du patient contribue à augmenter la durée totale de l'examen et le temps concédé par le personnel médical. Face à ces enjeux économiques et bien évidemment au bien-être des patients, la réduction de la durée de cet examen est donc de première importance.

Le *tilt-test* s'exécutant en deux phases (couchée et basculée), nous nous sommes attachés à explorer la phase de repos [Feuilloy *et al.*, 2005b; Feuilloy *et al.*, 2005c; Feuilloy *et al.*, 2005a; Feuilloy *et al.*, 2006b; Feuilloy *et al.*, 2006c; Schang *et al.*, 2007] et les premières minutes de la phase basculée [Feuilloy *et al.*, 2006a]. C'est ainsi que nous avons pu révéler l'importance et l'influence de certaines variables. D'autre part, les travaux engagés ont permis de mettre en œuvre un grand nombre de méthodes évoquées dans l'état de l'art de ce manuscrit et d'attirer l'attention sur les difficultés qui sont liées à leur utilisation dans des applications réelles. Ces méthodes concernent des algorithmes d'apprentissage artificiel et des approches pour la sélection de variables et l'extraction de caractéristiques. Le tableau 5.22 récapitule l'ensemble des études et des résultats pertinents évoqués dans ce chapitre. La procédure mise en place pour l'évaluation des modèles obtenus a permis d'estimer « sans biais » leur performance de généralisation. Rappelons que [Bellard *et al.*, 2003] ont utilisé l'échantillon d'apprentissage afin d'évaluer la performance de prédiction, ou encore [Mallat *et al.*, 1997; Pitzalis *et al.*, 2002] qui eux, ont utilisé l'échantillon de validation. De là, il n'est pas impossible que leur performance soit sur-estimée et que la prédiction ne soit plus aussi sensible en présence de nouveaux patients. Cette remarque conforte le bien-fondé des comparaisons que nous pouvons faire entre nos travaux et ces études. Nous pouvons néanmoins, relativiser les performances obtenues par [Mallat *et al.*, 1997; Pitzalis *et al.*, 2002; Bellard *et al.*, 2003] qui, par l'analyse d'une seule variable, ont utilisé des modèles de décision linéaires basés sur des seuils ; il paraît maintenant évident que ce type de modèles ne permettait pas d'obtenir des performances significatives.

L'avantage d'utiliser des modèles très parcimonieux permet d'établir des règles de décision très simples, facilitant la compréhension de la prédiction donnée. Dès lors, avec l'augmentation du nombre d'entrées dans le modèle, les règles se compliquent de manière à intégrer simultanément l'information contenue par chacune des variables. Par conséquent, les modèles deviennent en quelque sorte des « boîtes noires », les rendant alors très peu accessibles. Il en est de même lors de l'utilisation de processus de réduction de la dimensionnalité par des méthodes de projection qui, comme nous avons pu le voir, combinent les variables initiales afin d'obtenir en plus faible nombre des composantes synthétisant l'information originale. Comme nous avons déjà eu l'occasion de le dire, ces composantes sont peu exploitables, notamment celles provenant de processus de projections non linéaires. Cela est dommageable au vu de nos résultats ; en effet, rappelons que l'utilisation de l'analyse en composantes curvilignes à la section 5.3.2, a permis d'obtenir des performances de prédiction très intéressantes. Pour pallier ce problème d'interprétation des composantes provenant de projection non linéaire, nous développons dans le chapitre suivant une approche permettant d'interpréter la nature des composantes non linéaires.

L'analyse spécifique du signal d'impédancemétrie thoracique a permis de montrer sa pertinence dans la prédiction du résultat du *tilt-test* [Feuilloy *et al.*, 2006b; Feuilloy *et al.*, 2006c; Schang *et al.*, 2007]. En effet, la précision des résultats obtenus lors de son utilisation a permis de le comparer très

favorablement avec des indices plus classiques tels que la fréquence cardiaque ou encore la pression artérielle. D'autre part, ce signal apporte par ces caractéristiques des informations importantes sur la dynamique cardiaque suggérant, comme il a été dit à la section 5.3.3, de l'utiliser afin de limiter l'emploi de variables difficilement accessibles, comme le taux d'hématocrite par exemple. En effet, à l'image de l'électrocardiogramme, le signal d'impédancemétrie peut être enregistré en continu, sans nécessiter « d'intervention humaine », en laissant le soin aux « machines » de surveiller l'évolution des courbes. Cette sorte d'automatisation a été rendue possible par l'élaboration de processus d'analyse automatique (*cf.* section 5.5.5).

étude	méthode	période du *tilt-test*	nombre de patients	prévalence de la maladie (%)	S_e (%)	S_p (%)	VPP (%)	VPN (%)
[Feuilloy *et al.*, 2006c] (*cf.* section 5.5.5.2)	signal dZ **PMC**	période de repos	59	53	100	97	96	100
[Schang *et al.*, 2007] (*cf.* section 5.5.4)	signaux Z et dZ $\mathbf{SVM_{RBF}}$	période de repos	64	53	94	79	74	93
[Feuilloy *et al.*, 2005b] [Feuilloy *et al.*, 2005c] (*cf.* section 5.3.2)	mesures physiologiques et signal dZ **PMC** **ACC → ACP**	période de repos	36	52	84	71	80	76
[Feuilloy *et al.*, 2006a] (*cf.* section 5.4)	mesures physiologiques et signal dZ **PMC** $\mathbf{SFS^*_{RELIEF}}$ (*optimal*)	1re à 10-ième min du basculement	29	45	100	94	92	100
[Schang *et al.*, 2006] (*cf.* section 5.5.4)	signaux Z et dZ **PMC**	période de repos	59	53	88	64	66	88
[Schang *et al.*, 2003] (*cf.* section 4.3)	signal dZ **PMC**	période de repos	59	53	69	73	67	75
[Bellard *et al.*, 2003] (*cf.* sections 4.3 et 5.5.3)	signal dZ **seuil**	période de repos	68	56	68*	63*	63*	68*
[Bellard *et al.*, 2003] (*cf.* sections 4.3 et 5.5.3)	signal dZ **seuil**	5 à 10-ième min du basculement	68	56	68*	70*	68*	70*
[Bellard *et al.*, 2003] (*cf.* section 4.3)	mesures physiologiques et signal dZ **seuil**	5 à 10-ième min du basculement	68	56	50*	97*	93*	67*
[Pitzalis *et al.*, 2002] (*cf.* section 4.3)	pression artérielle **seuil**	1re à 15-ième min du basculement	80	30	80	85	57	94
[Mallat *et al.*, 1997] (*cf.* section 4.3)	fréquence cardiaque **seuil**	1re à 6-ième min du basculement	99	28	96	87	75	98

Note : * Indique que les résultats sont obtenus sur le sous-ensemble d'apprentissage.

TAB. 5.22 – Récapitulatif des études analysant l'apparition des symptômes de la syncope lors de l'examen du *tilt-test*.

Chapitre 6

Nouvelle approche pour l'extraction d'informations et l'interprétation des méthodes de projection non linéaire

6.1 Introduction

La caractéristique commune à toutes les méthodes de projection est la propriété des nouvelles composantes : elles sont obtenues par la transformation et la combinaison (linéaire ou non linéaire) des variables initiales. Par conséquent, les variables initiales sont représentées dans les nouvelles composantes proportionnellement à leur influence dans la construction des composantes. Dans le cadre de l'analyse en composantes principales (ACP), nous avons identifié cela par la qualité de la représentation des variables dans les composantes principales (*cf.* section 2.3.2.1, page 67). Ce point est fondamental, comme le souligne justement [Saporta, 2006] qui évoque que la méthode la plus naturelle pour trouver et donner une signification à une composante principale, c'est de la relier aux variables initiales.

On pourrait légitimement généraliser la remarque de [Saporta, 2006] aux composantes issues de méthodes de réduction non linéaire. Cependant, contrairement à l'ACP et en l'état actuel des choses, il n'existe pas de moyens analytiques qui permettent d'extraire la représentation des variables dans des composantes issues de processus de réduction non linéaire. Cette difficulté avait déjà été abordée en conclusion de la section 2.3.4 et à la discussion de la section 5.3.2. Cela correspond, pour [Illouz and Jardino, 2001; Guérif, 2006], à l'inconvénient majeur dans l'utilisation et l'interprétation de la plupart des méthodes de réduction non linéaire.

Cet handicap rend les projections non linéaires peu exploitables dans de grandes dimensions. En effet, même si la plupart de ces méthodes peuvent projeter des données dans n'importe quelle dimension, elles sont, par la force des choses, utilisées principalement pour des projections en deux ou trois dimensions, afin d'obtenir une représentation graphique et donc, faciliter leur interprétation. Cette utilisation contraignante dans de très faibles dimensions reste sans effet pour obtenir la relation entre les variables originales et les nouvelles composantes. Ainsi, les composantes non linéaires sont exploitées sans chercher à les relier aux variables initiales ; cela, malgré l'importance de ces relations qui sont fondamentales pour transmettre la description physique des variables initiales aux nouvelles composantes [Saporta, 2006].

D'une manière plus générale, la réduction de la dimension induit forcément une perte d'information, donc plus la réduction est importante, plus l'information perdue peut l'être également. Dès lors, l'utilisation quasiment exclusive des méthodes de réduction non linéaire en deux ou trois dimensions peut limiter fortement leur efficacité et leur capacité à synthétiser l'information contenue initialement dans les données.

La méthodologie proposée dans ce chapitre, publiée dans [Feuilloy *et al.*, 2007], résout cette difficulté, en permettant l'interprétation des composantes résultantes de processus de réduction non linéaire en les reliant aux variables initiales. La section 6.2 donnera les fondements de la méthodologie, qui est une adaptation du processus utilisé par l'ACP, décrit lui, à la section 6.3. Suite aux développements de notre méthode à la section 6.4, nous donnerons dans la même section une validation expérimentale sur des ensembles de données synthétiques. Enfin, à la section 6.5, nous adapterons le procédé d'extraction d'informations à notre problème de prédiction de l'apparition des symptômes de la syncope lors de l'examen du *tilt-test*. Rappelons qu'à la section 5.3.2, nous avons utilisé un processus de projection non linéaire (analyse en composantes curvilignes) qui a permis d'obtenir des performances intéressantes (*cf.* résultats de la section 5.3.2.4). Cependant, lors de ces travaux, publiés également dans [Feuilloy *et al.*, 2005c], nous n'avions pas pu interpréter profondément les composantes non linéaires utilisées pour la prédiction, en les liant aux variables initiales ; contrairement à ce qui a été fait pour l'analyse en composantes principales. Ainsi, à la section 6.5, nous compléterons l'analyse de la projection non linéaire employée pour la prédiction de la syncope, en exprimant précisément quelles variables ont le plus contribué à la formation des composantes curvilignes utilisées pour la prédiction.

6.2 Fondements

Notre approche d'extraction d'informations sur des composantes non linéaires est fondée sur une reproductibilité et une adaptation de la technique utilisée pour l'interprétation des composantes principales. Rappelons que dans la description de l'ACP, cette approche avait été brièvement abordée à la section 2.3.2.1 ; nous la détaillerons à la section 6.3 afin de faciliter le parallèle avec le processus développé pour les méthodes de projection non linéaire.

Ainsi, la qualité de la représentation d'une variable dans une composante principale est dépendante des valeurs et des vecteurs propres obtenus lors de l'ACP. Or, comme l'a fait remarquer [Dreyfus *et al.*, 2002], l'analyse en composantes curvilignes (ACC) est une méthode de réduction non linéaire (*cf.* section 2.3.3.2), qui peut être vue comme une extension non linéaire de l'ACP ; ils évoquent alors une ACP « par parties ». Ce principe a également été évoqué par [Bishop, 2006], où il suggérait de réaliser plusieurs ACP dans l'espace original afin d'améliorer la projection. Cette information a été pour nous fondamentale, car elle nous donnait la possibilité de lier « analytiquement » le processus d'extraction de l'information de l'ACP pour l'ACC.

Avant de poursuivre, rappelons que la matrice des observations $\mathbf{X}$ contient n observations $\mathbf{x}_1, ..., \mathbf{x}_n$, chacune décrite par p variables. Le vecteur $\mathbf{x}_i = (x_{i1}, ..., x_{ip})^T$ donne alors les p éléments de l'observation i, et l'élément (i, j) de $\mathbf{X}$ (noté x_{ij}) correspond à la j-ème variable de la i-ème observation.

6.3 Extraction de la contribution des variables dans le cas d'une analyse en composantes principales

6.3.1 Description de la méthode

Le détail de l'extraction d'informations est donné sur un exemple simple illustré à la figure 6.1(a). Les données sont caractérisées par deux variables x_1 et x_2, sur lesquelles une ACP est réalisée qui, comme le montre cette même figure, permet d'obtenir la direction principale et les deux vecteurs propres :

$$\mathbf{u}_1 \begin{pmatrix} u_{cp_1}^{x_1} \\ u_{cp_1}^{x_2} \end{pmatrix} \quad \text{et} \quad \mathbf{u}_2 \begin{pmatrix} u_{cp_2}^{x_1} \\ u_{cp_2}^{x_2} \end{pmatrix} . \tag{6.1}$$

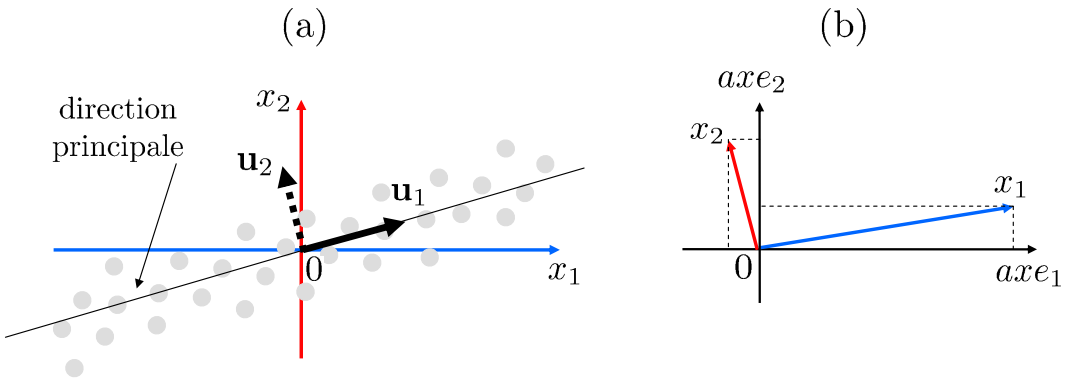
Ainsi, à partir de ces vecteurs propres, ou de la matrice des vecteurs propres associée :

$$\mathbf{U} = \begin{bmatrix} u_{cp_1}^{x_1} & u_{cp_2}^{x_1} \\ u_{cp_1}^{x_2} & u_{cp_2}^{x_2} \end{bmatrix} , \tag{6.2}$$

nous pouvons désormais obtenir les coordonnées factorielles des deux « points − variables[1] » dans le repère des axes principaux, comme le montre la figure 6.1(b). Les coordonnées des variables sont alors données par les relations suivantes :

$$\text{coordonnées de } x_1 \begin{pmatrix} u_{cp_1}^{x_1}\sqrt{\lambda_{cp_1}} \\ u_{cp_2}^{x_1}\sqrt{\lambda_{cp_2}} \end{pmatrix} \text{ et coordonnées de } x_2 \begin{pmatrix} u_{cp_1}^{x_2}\sqrt{\lambda_{cp_1}} \\ u_{cp_2}^{x_2}\sqrt{\lambda_{cp_2}} \end{pmatrix} , \tag{6.3}$$

où λ_{cp_1} et λ_{cp_2}, donnent respectivement les valeurs propres des deux premières composantes principales (CP).



Note : (a) Représentation des données et des vecteurs propres dans l'espace initial des deux variables (x_1 et x_2). (b) Représentation de la projection des variables dans le repère des axes factoriels : coordonnées des « points − variables » projetés sur les axes.

FIG. 6.1 – Illustration du résultats d'une analyse en composantes principales.

La projection des « points − variables » sur les axes nous permettent alors d'obtenir l'information relative à la contribution des variables initiales dans la création de chaque CP. En

[1]Les coordonnées « points − variables » permettent de représenter les variables dans l'espace factoriel.

d'autres termes, nous connaissons désormais la contribution relative Q_j^i de la variable i dans la composante principale j, parmi les q composantes principales :

$$Q_j^i = \frac{\left(\sqrt{\lambda_j}\, u_j^i\right)^2}{\sum_{l=1}^{p} \left(\sqrt{\lambda_l}\, u_l^i\right)^2}\,; \tag{6.4}$$

dans notre exemple, $i, j = 1, 2$, le nombre de variables p est de 2 et $q = p$.

La figure 6.2(b) montre pour notre exemple, la relation entre les variables initiales et les composantes principales. Ainsi, deux lectures sont possibles, où d'une part, nous pouvons déterminer le lien d'une variable dans les CP, et d'autre part, nous pouvons obtenir la proportion des variables dans une CP.

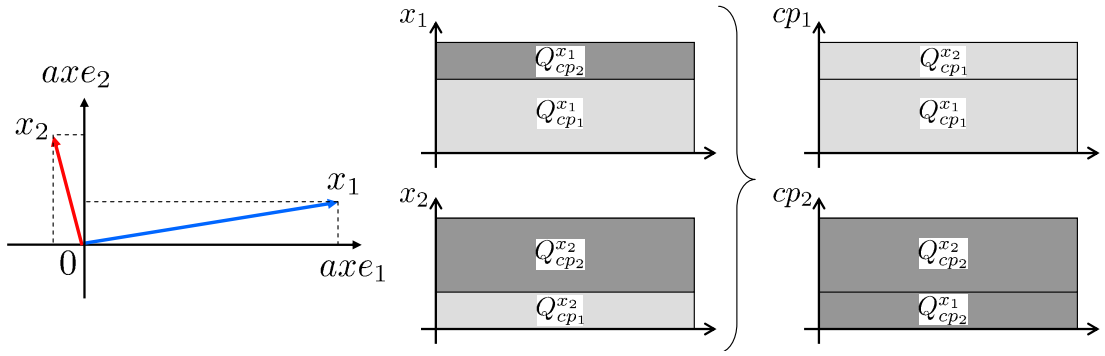


FIG. 6.2 – Interprétation de la qualité de la représentation des variables initiales dans les composantes principales sur un exemple en deux dimensions.

Suite à cet exemple introductif, nous allons généraliser le processus au cas de p variables projetées sur q axes principaux (q CP), avec $q \leq p$.

Dans l'exemple précédent, nous avons évoqué la projection des variables sur les axes principaux, où les coordonnées factorielles des p « points − variables » sur le j-ème axe sont obtenues par $\mathbf{u}_j\sqrt{\lambda_j}$. À partir de là, nous pouvons isoler chaque variable et obtenir ainsi la projection de la i-ème variable sur la j-ème composante principale par $u_j^i\sqrt{\lambda_j}$, où λ_j correspond à la valeur propre de la j-ème CP et u_j^i est le i-ème élément (variable) du vecteur propre de cette même CP. Nous pouvons regrouper l'ensemble des vecteurs propres dans la matrice suivante :

$$\mathbf{U} = \begin{bmatrix} u_1^1 & \cdots & u_j^1 & \cdots & u_q^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_1^i & \cdots & u_j^i & \cdots & u_q^i \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u_1^p & \cdots & u_j^p & \cdots & u_q^p \end{bmatrix} \quad , \quad q \leq p\,. \tag{6.5}$$

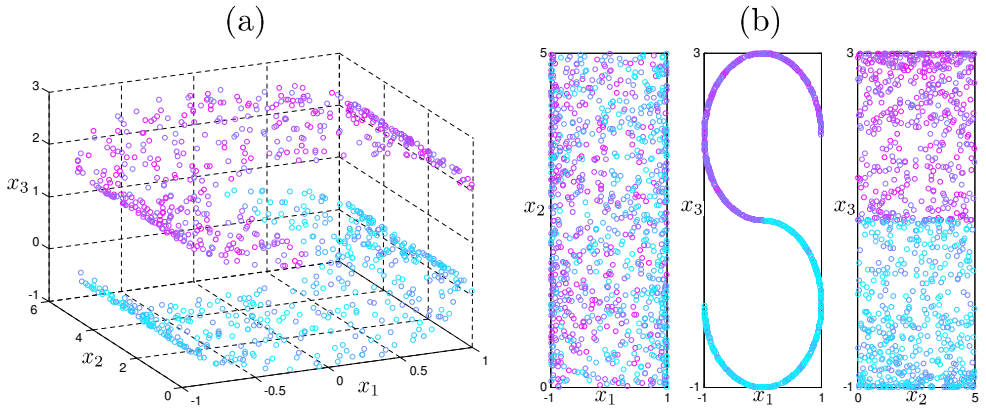
La qualité de la représentation d'une variable dans une composante (6.4), peut bien évidemment être obtenue pour les q premières composantes principales, comme le montre la relation suivante :

$$Q_j^i = \frac{\sum_{j=1}^{q} \left(\sqrt{\lambda_j}\, u_j^i\right)^2}{\sum_{l=1}^{p} \left(\sqrt{\lambda_l}\, u_l^i\right)^2} \tag{6.6}$$

Les ouvrages de [Saporta, 2006; Lebart *et al.*, 2006] permettront au lecteur d'obtenir des détails supplémentaires.

6.3.2 Validation expérimentale

Nous illustrons dans cette section l'interprétation de la qualité de la représentation des variables dans les composantes principales. L'exemple, donné à la figure 6.3, représente une forme de « S » obtenue à partir de trois variables x_1, x_2 et x_3. Cet ensemble de données est souvent utilisé pour valider et illustrer les performances de méthodes de projection non linéaire, à l'image de l'ensemble de données représentant un « petit suisse » utilisé à la section 2.3.3.4.



Note : (a) Représentation des données dans l'espace initial des trois variables (x_1, x_2 et x_3). (b) Représentation des données dans l'espace initial pour chaque paire de variables (x_1 avec x_2, x_1 avec x_3 et x_2 avec x_3).

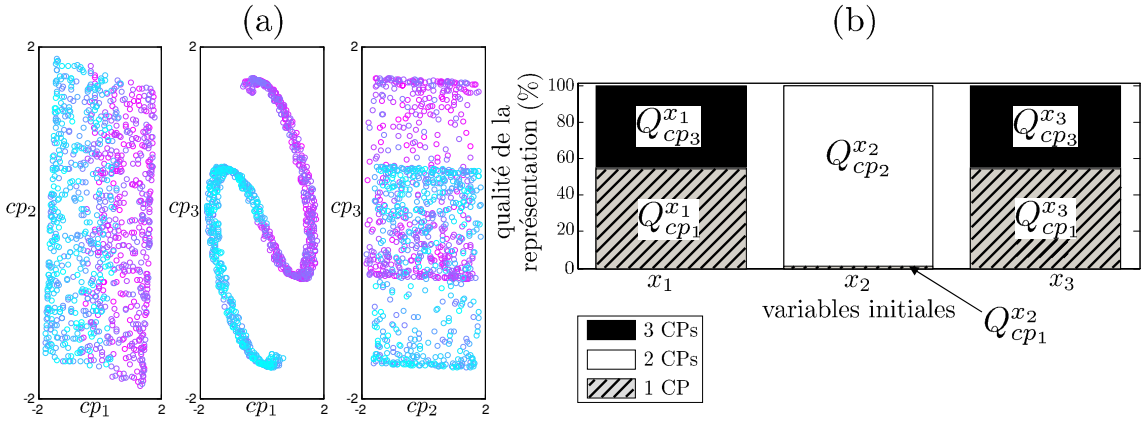
FIG. 6.3 – Ensemble de données, représentant une forme de « S », utilisé pour valider l'estimation de la représentation des variables dans les nouvelles composantes.

À partir de ces données, une analyse en composantes principales est réalisée, les résultats de la projection sont donnés à la figure 6.4. Nous y visualisons ainsi, la projection des observations dans l'espace réduit à deux dimensions (pour chaque paire de composantes principales) et également, la qualité de la représentation des variables dans ces mêmes composantes. Dès lors, dans le repère cp_1 associé à cp_3 de la figure 6.4(a), nous pouvons observer une forme de « S » proche de celle obtenue avec l'association des variables initiales x_1 et x_3 (figure 6.3b). De cette comparaison, nous pourrions déduire que les variables x_1 et x_3 ont dû fortement contribuer à la construction des composantes cp_1 et cp_3. Effectivement, nous pouvons facilement le vérifier en analysant les diagrammes de la figure 6.4(b), représentatifs des contributions des variables. En effet, les variables x_1 et x_3 sont quasiment entièrement représentées dans les cp_1 et cp_3, tandis que x_2 est uniquement représentée dans la cp_2.

6.4 Extraction de la contribution des variables dans le cas d'une réduction de dimension non linéaire

6.4.1 Introduction

La présentation de la méthode et l'exemple expérimental sur la contribution des variables, dans le cadre de l'ACP, ont permis de montrer que cette technique est facilement exploitable. Cependant, la détermination de l'indice Q nécessite de considérer les **valeurs propres** et les **vecteurs propres**, qui sont induits par le processus de l'ACP. Ainsi, en l'état actuel des choses, cette



Note : (a) Représentation des données projetées dans l'espace réduit par l'ACP pour chaque paire de composantes principales (cp_1 avec cp_2, cp_1 avec cp_3 et cp_2 avec cp_3). (b) Qualité (Q) de la représentation des trois variables initiales dans les trois composantes principales.

FIG. 6.4 – Résultats de l'analyse en composantes principales sur les données représentant une forme de « S ».

technique est difficilement adaptable aux méthodes de réduction non linéaire, qui ne permettent pas d'obtenir des éléments équivalents aux valeurs et vecteurs propres. L'approche, proposée dans cette section, permet d'estimer des grandeurs similaires aux valeurs et vecteurs propres, que nous appellerons respectivement « **pseudo-valeurs propres** » et « **vecteurs propres locaux** ». Ces nouveaux éléments permettront alors de généraliser pour l'analyse en composantes curvilignes, le processus qui détermine la qualité de la représentation des variables dans les nouvelles composantes.

6.4.2 Evaluation et vérification de la projection par estimation des « pseudo-valeurs propres »

Avant d'aborder la généralisation des valeurs propres par l'estimation des « pseudo-valeurs propres », nous allons revenir sur un point abordé à la section 2.3.3.3. Ce point fait référence à l'évaluation du résultat de la projection obtenue par des méthodes de réduction non linéaire, où nous avons introduit une représentation (« $dy - dx$ ») proposée par [Demartines, 1992]. Celle-ci permet de vérifier la préservation locale de la topologie et de valider globalement la projection. Pour l'ACC, [Demartines and Hérault, 1997] ont proposé une adaptation de cette représentation, afin de comparer les distances entre les observations situées dans l'espace d'origine (d^*_{ij}) et celles projetées dans l'espace réduit (d_{ij}). Cela permet alors d'interpréter la projection et sa distorsion comme suit :

- si $d_{ij} \approx d^*_{ij}$ alors la projection peut être considérée comme linéaire ;
- si $d_{ij} > d^*_{ij}$ alors les observations dans l'espace réduit sont étirées ;
- si $d_{ij} < d^*_{ij}$ alors les observations dans l'espace réduit sont regroupées.

La conservation de la topologie locale est vérifiée lorsque l'écart des distances des observations entre les deux espaces est faible et reste sous un certain seuil ; au dessus de ce seuil, les distances séparant les observations ne sont plus considérées comme locales (*cf.* section 2.3.3.2, page 73).

L'indice dr donne une information numérique sur la distorsion de la projection, en calculant la corrélation entre les distances de l'espace original et celles de l'espace réduit, tel que :

$$dr = \frac{\sum_i \sum_{j \neq i} (d^*_{ij} - \bar{d}^*)(d_{ij} - \bar{d})}{\sqrt{\sum_i \sum_{j \neq i} (d^*_{ij} - \bar{d}^*)^2} \sqrt{\sum_i \sum_{j \neq i} (d_{ij} - \bar{d})^2}}, \tag{6.7}$$

avec $\bar{d}^*$ et $\bar{d}$, leur moyenne respective. Ainsi, si le taux de distorsion dr se rapproche de 0, alors la distorsion de la projection devient importante. Cet indice est très important. Il est apparu très pertinent lorsque nous avons pu observer sa forte corrélation avec les valeurs propres associées aux CP [Feuilloy *et al.*, 2005c]. En effet, des simulations ont montré qu'à l'issue d'une ACP, une similitude apparaissait entre l'évolution de l'indice dr et celle des valeurs propres. Par conséquent, l'indice dr, nous permet d'estimer le taux d'inertie expliqué sur les composantes non linéaires, à l'image de ce qui se fait sur l'ACP[2]. Ainsi, le taux d'inertie expliquée sur les q composantes non linéaires peut être obtenu par :

$$I_q = \frac{\sum_{j=1}^{q} (dr_j)^{-1}}{\sum_{l=1}^{p} (dr_l)^{-1}} \ , \ q \leq p \, . \tag{6.8}$$

Cette généralisation du calcul du taux d'inertie de l'ACP, aux méthodes de réduction non linéaire, permet d'obtenir une information globale sur le pourcentage d'information des données initiales transmises aux composantes non linéaires. Cette information permet alors d'obtenir les « pseudo-valeurs propres », qui sont une estimation de l'information restituée sur chacune des q composantes, telle que : $\lambda_j \approx 1/dr_j$, $j = 1, \ldots, q$.

6.4.3 Estimation des « vecteurs propres locaux »

Dans l'ACP, les valeurs propres donnent globalement l'information des données initiales restituées dans les nouvelles composantes principales. C'est par les vecteurs propres que nous avons la possibilité d'obtenir l'information suivant chaque variable. À l'image de l'estimation des valeurs propres par les pseudo-valeurs propres, il est donc nécessaire d'estimer les vecteurs propres dans le cadre d'une projection non linéaire, afin de déterminer la qualité de la représentation des variables dans les composantes non linéaires.

Contrairement aux valeurs propres, l'estimation des « vecteurs propres » est moins intuitive. Avec l'ACP, la direction principale des observations dans l'espace initial donne le premier axe principal pour lequel est associé le vecteur propre de la première composante principale. Ainsi, comme montré précédemment, la relation entre les variables initiales et la première CP est donc donnée par cet **axe de projection** qui, étant linéaire, donne par conséquent une projection uniforme sur la totalité de l'espace des variables. Aussi, une projection non linéaire induit une transformation non uniforme et irrégulière dans l'espace des variables. Dès lors, cette constatation implique d'avoir, en fonction de l'emplacement dans l'espace des variables, une contribution différente des variables dans la création des composantes curvilignes.

L'idée proposée est de trouver non plus un axe de projection faisant référence à la transformation des données, mais une **courbe de projection** sur laquelle plusieurs vecteurs propres sont extrapolés afin d'obtenir localement les coefficients de la transformation des données : ces nouveaux vecteurs sont appelés « vecteurs propres locaux ». Cette courbe de projection fait donc

[2]Rappelons que le taux d'inertie dans le cadre de l'ACP, correspond au pourcentage de la représentation des données initiales sur les q premières composantes principales et s'obtient par : $\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{l=1}^{p} \lambda_l}$, $q \leq p$.

le lien entre l'espace initial et l'espace réduit, elle peut être considérée comme le support des « vecteurs propres locaux » en différents points. La figure 6.5 illustre cette courbe de projection, en la comparant à l'axe de projection obtenu par une ACP.

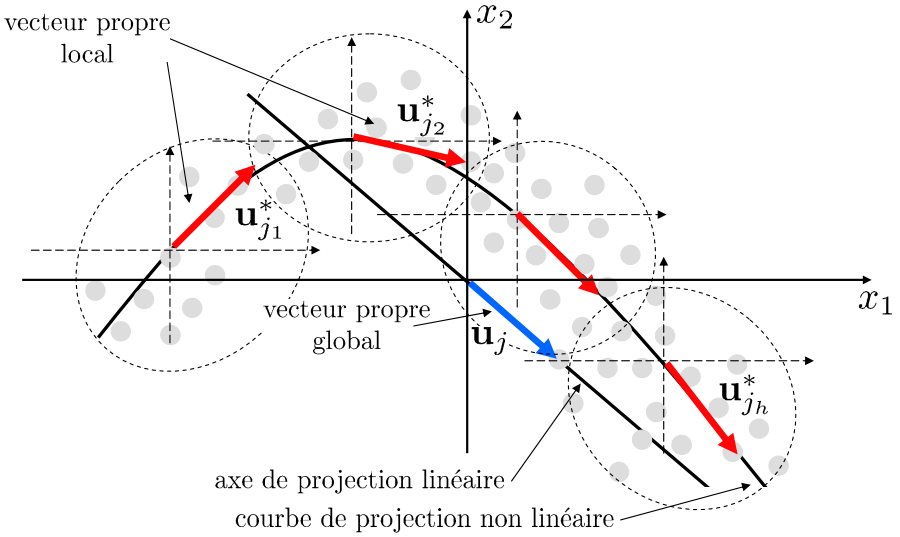


FIG. 6.5 – Illustration de l'extrapolation des « vecteurs propres locaux » ($\mathbf{u}^*_{jl}$, $l = 1, \cdots, h$) sur la courbe de projection.

Notons que sur la figure 6.5, nous avons nommé le vecteur propre obtenu par l'ACP « vecteur propre global », par opposition aux « vecteurs propres locaux » extrapolées sur la courbe de projection. Ainsi, on peut observer la discrétisation, ou plutôt, le partitionnement de l'espace en h *clusters*, sur lesquels, les « vecteurs propres locaux » sont estimés. Ainsi, nous notons $\mathbf{u}_j$, le « vecteur propre global » associé à la j-ème composante issue de l'ACP et $\mathbf{u}^*_{jl}$ le vecteur propre local de la l-ième partition de l'espace associé à la j-ème composante curviligne (CC). De cette notation, nous pouvons pour plus de clarté, comparer l'écriture d'un vecteur propre associé à la j-ème composante principale $\begin{bmatrix} u^1_j \\ \vdots \\ u^i_j \\ \vdots \\ u^p_j \end{bmatrix}$ et à la j-ème composante curviligne $\begin{bmatrix} u^1_{j_1} & \cdots & u^1_{j_l} & \cdots & u^1_{j_h} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u^i_{j_1} & \cdots & u^i_{j_l} & \cdots & u^i_{j_h} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ u^p_{j_1} & \cdots & u^p_{j_l} & \cdots & u^p_{j_h} \end{bmatrix}$.

Nous venons de voir comment estimer les vecteurs propres locaux, cependant, il reste encore à déterminer leur support que nous avons appelé la courbe de projection.

Refaisons un dernier parallèle avec l'ACP, où la projection d'une observation $\mathbf{x}_i$ sur le j-ème axe factoriel est donnée par $y_j = \mathbf{x}_i^T \mathbf{u}_j$. Ainsi, les relations linéaires, liant les observations de l'espace initial à l'espace réduit, sont données par les vecteurs propres $\mathbf{u}_j$. Par conséquent, pour l'ACC, la projection d'une observation $\mathbf{x}_i$ sur la j-ème composante curviligne est alors obtenue par une transformation non linéaire des variables initiales (x_i, avec $i = 1, \cdots, p$), telle que $y_j = \Phi(\mathbf{x}_i)$. La transformation Φ permet alors de passer de l'espace initial à l'espace réduit, et peut être approchée par une fonction non linéaire. Les paramètres de cette fonction, faisant état de la transformation des données, peuvent être déterminés par plusieurs techniques telles que, les réseaux de neurones ou encore les fonctions polynomiales. Cependant, [Saporta, 2006] concède que les fonctions polynomiales ont l'inconvénient d'être trop rigides et qu'elles considèrent les données dans leur globalité. Il suggère alors l'utilisation de transformations polynomiales par morceaux, appelées **fonctions splines**[3].

[3]Cette remarque apparaît dans [Saporta, 2006] dans les présentations des extensions non linéaires de l'ACP.

La figure 6.6 résume l'adaptation de la méthode d'extraction de la qualité de la représentation des variables dans les nouvelles composantes, en comparant le cas d'une ACP à celui d'une ACC. Ainsi, cette figure illustre dans le cas non linéaire, les variations de la qualité de la représentation des variables.

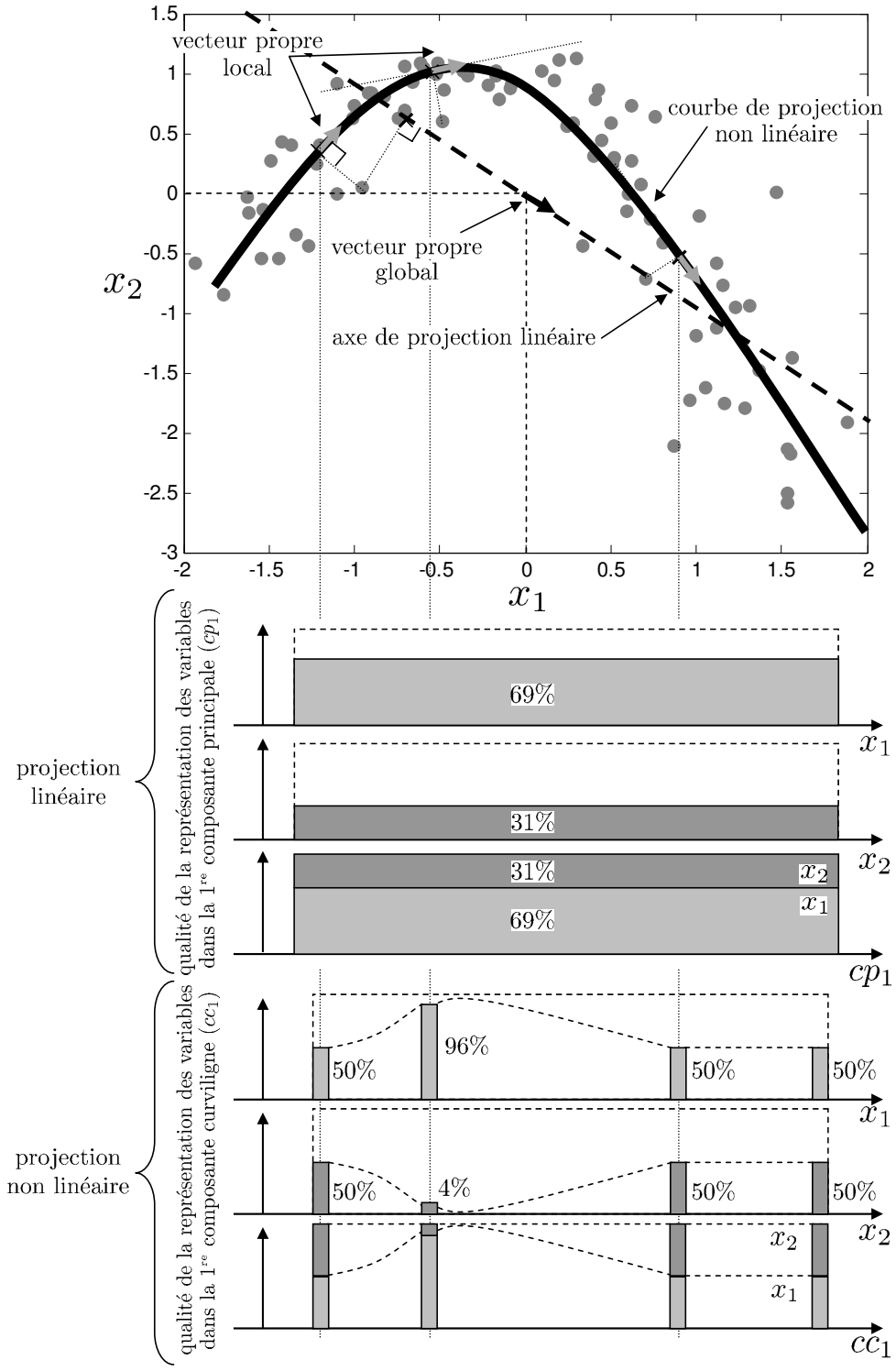


FIG. 6.6 – Comparaison du résultat de l'extraction de la qualité de la représentation des variables dans les composantes principales et dans les composantes curvilignes.

6.4.4 Extension de la procédure d'estimation de la qualité de la représentation aux différentes méthodes de projections

Nous avons détaillé le processus d'estimation de la qualité de la représentation des variables dans les composantes non linéaires pour le cas de l'analyse en composantes curvilignes. Ce choix a été porté par le fait que l'ACC est, d'après son auteur, une ACP par parties. Cependant, comme nous l'avons noté précédemment (introduction de la section 2.3.3.2), la plupart des méthodes de projection non linéaire sont fondées sur une même notion, qui est la préservation locale de la topologie ou de la structure des données. Ainsi, cela nous a amené à penser que le processus d'estimation développé pour l'ACC peut s'adapter à tous les types de projection non linéaire.

6.4.5 Validation expérimentale

Comme dans le cadre de l'ACP, nous proposons une validation expérimentale, afin de vérifier l'efficacité du processus d'estimation de la représentation des variables dans les nouvelles composantes issues des méthodes de projection non linéaire. Dans cette validation expérimentale, les données sont celles utilisées pour valider l'interprétation du processus d'extraction de l'ACP à la section 6.3.2. On retrouve ainsi un ensemble de données représentant une forme de « S » en trois dimensions (*cf.* figure 6.3, page 187).

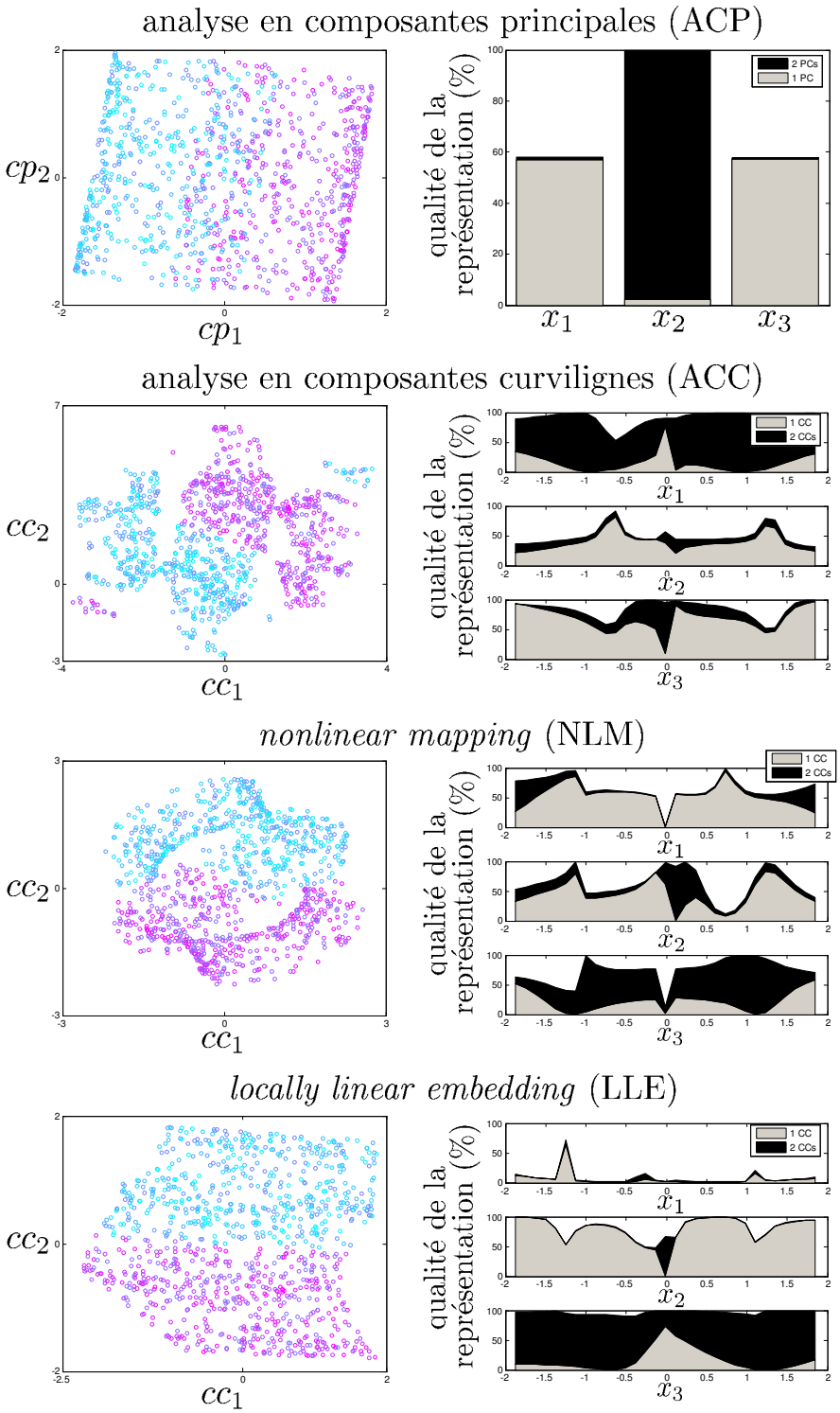
La figure 6.7 donne le résultat des projections pour quatre méthodes (ACP, ACC, NLM et LLE), ainsi que les diagrammes résultant du processus d'extraction de la représentation des variables dans les nouvelles composantes linéaires et non linéaires. Pour cet exemple, où les données forment un « S », nous pouvons remarquer la qualité de la méthode LLE pour déplier le « S ». Rappelons que les détails de cette méthode de réduction non linéaire sont donnés à la section 2.3.3.1 ; dans cet analyse, nous avons choisi de prendre $k = 12$, représentant les k plus proches voisins à considérer pour la reconstruction des poids.

L'estimation de la qualité de la représentation des variables dans les composantes non linéaires issues de LLE montre que la variable x_1 n'a quasiment pas participé à la construction des deux premières composantes non linéaires (cc_1 et cc_2). D'autre part, nous pouvons grâce au diagramme, observer que la variable initiale x_2 est fortement représentée dans cc_1, tout comme x_3 dans cc_2. Ainsi, nous pouvons imaginer que les représentations x_3 en fonction de x_2 et cc_2 en fonction de cc_1 devraient être très similaires. Or, c'est ce que nous pouvons observer en comparant les figures 6.3(b) et 6.7, où les données bleues du haut du « S » sont séparées des données magentas du bas.

Notons que pour l'ACP, la qualité de la représentation à été réalisée par le même processus d'extraction utilisé pour les méthodes de projection non linéaire. Aussi pour l'ACP, une fonction linéaire est utilisée comme support des « vecteurs propres locaux », contre une fonction polynomiale de second degré pour les méthodes de projection non linéaire.

6.4.6 Renforcement de la pertinence de la lecture des estimations des contributions des variables dans les composantes non linéaires

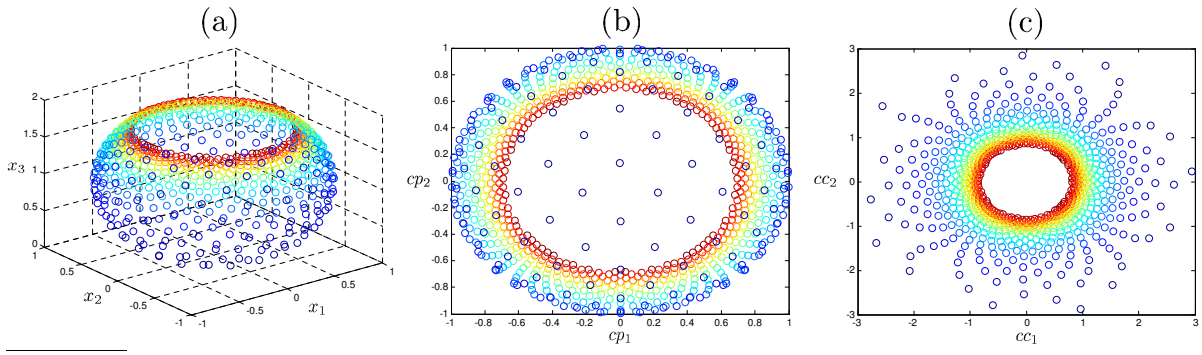
Dans le cadre d'une transformation non linéaire, nous avons vu que la projection peut être différente en chaque point de l'espace d'une variable. Aussi, il paraît évident qu'une région de l'espace sous-représentée par les données n'a pas pu contribuer à la projection, ou pas de manière importante. Dès lors, afin de renforcer la pertinence de la lecture des diagrammes représentatifs de la qualité de la représentation des variables, nous pouvons alors pondérer dans ces diagrammes, chaque région de l'espace des variables par la distribution des données associées.



Note : (à gauche) Représentation des données projetées dans l'espace réduit par ACP, ACC, NLM et LLE suivant les deux premières nouvelles composantes. (à droite) Qualité de la représentation des trois variables initiales dans les deux nouvelles composantes.

FIG. 6.7 – Comparaison des projections par des méthodes de réduction linéaire (ACP) et non linéaire (ACC, LLE et NLM) sur les données représentant une forme de « S ».

Pour illustrer ce point, nous proposons l'exemple donné à la figure 6.8(a), représentant une sphère en trois dimensions.



Note : (a) Représentation des données dans l'espace initial des 3 variables (x_1, x_2 et x_3). (b) Représentation des données projetées dans l'espace réduit par l'ACP (cp_1 et cp_2). (c) Représentation des données projetées dans l'espace réduit par l'ACC (cc_1 et cc_2).

FIG. 6.8 – Ensemble de données, représentant une sphère en trois dimensions, dans l'espace initial et dans les espaces réduits par l'analyse en composantes principales et curvilignes.

Comme le montre la figure 6.8, la projection par l'ACC est plus efficace que la projection par l'ACP. Dans ce dernier cas, un recouvrement de données empêche une analyse efficace de la représentation donnée par les deux premières composantes principales. En revanche, l'ACC, par un dépliage efficace des données, permet de donner un bon aperçu de la structure initiale des données, comme le montre la figure 6.8(c).

L'analyse des deux premières composantes curvilignes (cc_1 et cc_2) donnée à la figure 6.9, permet de lier les variables originales aux composantes. Cette figure montre alors la représentation des variables dans les deux premières composantes curvilignes, où nous pouvons observer une forte représentation des variables x_2 et x_1, respectivement dans les composantes cc_1 et cc_2. Le graphique montrant les contributions globales sur cc_2, révèle notamment une légère participation de la variable x_3 dans la formation de cette composante curviligne. Notons cependant que cette participation est significative dans l'intervalle $[-3\,;\,-1,5]$ de cette même composante.

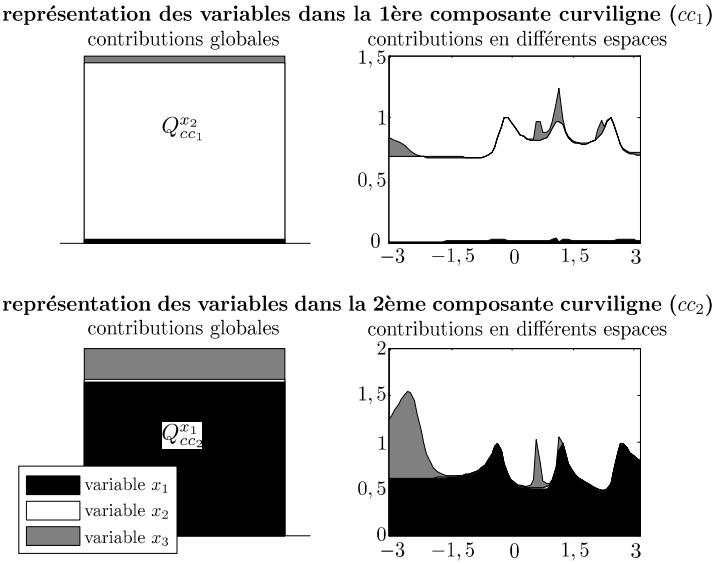
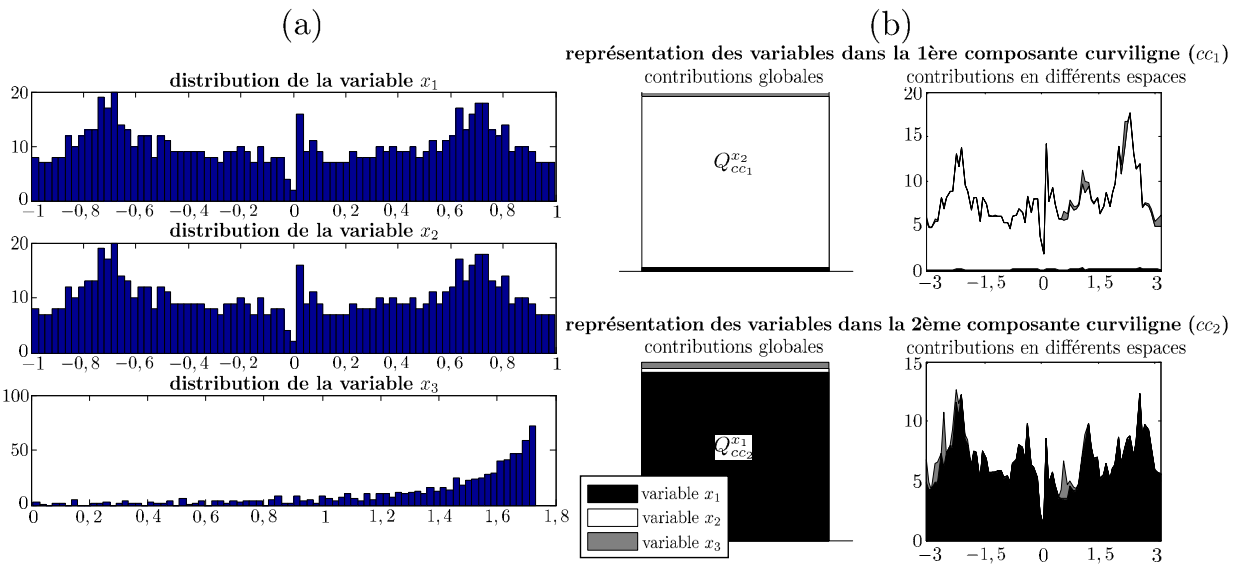


FIG. 6.9 – Qualité de la représentation des variables initiales dans les composantes curvilignes pour l'exemple de la sphère en trois dimensions.

La figure 6.10(a) montre la distribution des variables, où nous remarquons que la variable x_3 est sous-représentée dans l'intervalle $[0\,;\,1,2]$. Ainsi, dans cette région la variable n'a pas pu contribuer fortement à la projection, malgré ce qu'indiquent les diagrammes illustrant la qualité de la contribution des variables dans les composantes de la figure 6.9. Dès lors, pour améliorer la précision des diagrammes, nous avons considéré la distribution des variables dans l'estimation de

chacune des contributions, comme le montre la figure 6.10(b). C'est ainsi que nous remarquons la forte diminution de la variable x_3 dans la contribution à la création de la composante cc_2.



Note : (a) Représentation de la distribution des variables initiales. (b) Qualité de la représentation des variables initiales dans les composantes curvilignes en considérant la distribution des variables initiales.

FIG. 6.10 – Qualité de la représentation des variables initiales, pondérées par leur distribution, dans les composantes curvilignes pour l'exemple de la sphère en trois dimensions.

L'interprétation faite sans considérer la distribution de la variable peut faire penser à la manipulation de données contenant des valeurs aberrantes. L'augmentation du nombre de *clusters* lors du partitionnement de l'espace des variables permettrait de réduire ce problème. Cependant, dans ce cas, la projection dans chaque *cluster* pourrait être biaisée par le manque de données. Ainsi, un compromis pourrait permettre de s'affranchir de l'utilisation de la distribution pour estimer la qualité de la représentation. Aussi, nous pouvons également intégrer l'information de la distribution de chaque variable dans le calcul des pseudo-valeurs propres. En effet, dans ce cas, nous aurions, à l'image de vecteurs propres locaux, différentes valeurs des pseudo-valeurs propres en fonction de l'emplacement dans l'espace initial.

Rappelons que les diagrammes ne donnent qu'une estimation de la participation des variables dans la création des composantes, la précision n'est donc que relative. Cela permet néanmoins, comme nous l'avons montré dans les validations expérimentales, de donner un aperçu du contenu des composantes.

6.5 Application expérimentale à la prédiction de la syncope

6.5.1 Introduction

Dans cette section, nous allons employer la méthodologie développée dans ce chapitre, afin de renforcer l'interprétation des résultats obtenus lors de l'utilisation de l'analyse en composantes curvilignes pour la prédiction du résultat du *tilt-test* durant la période de repos. En effet, à la section 5.3.2, l'ACC a permis d'obtenir des performances très intéressantes et supérieures à l'ACP, comme le montre le tableau 5.9 (page 144) récapitulant les résultats obtenus par les méthodes de projection. Aussi, dans ces travaux, publiés également [Feuilloy *et al.*, 2005c], nous n'avons pas pu identifier réellement quelles étaient parmi les variables initiales, celles qui ont le plus contribué à la formation des composantes curvilignes, et donc, les variables les plus pertinentes

pour prédire la syncope. Dès lors afin de compléter l'interprétation de l'ACC, nous allons, en utilisant la méthode présentée dans ce chapitre, estimer l'information « contenue » dans chaque composante curviligne, en la liant aux variables initiales.

6.5.2 Rappel du contexte

À la section 5.3, nous avons étudié la phase couchée du *tilt-test*, dans le cadre de la prédiction de la syncope. Pour cette étude, rappelons que nous disposions de 84 patients issus de l'échantillon nommé $\mathcal{E}_1$ (*cf.* section 5.2), sur lesquels 15 variables pré-sélectionnées par les médecins avaient été relevées. Le tableau 6.1 redonne la description des 15 variables.

variables	*symboles*
âge	*âge*
surface corporelle	*BSA* (*body surface area*)
volume plasmatique	*VolPlas*
fréquence cardiaque	*FC*
pression artérielle systolique	*PAS*
pression artérielle diastolique	*PAD*
pression pulsée	*PP*
eau totale	*TBW* (*total body water*)
rapport masse maigre / masse grasse	*LW/FW*
hématocrite	*Ht*
hémoglobine	*Hb*
accélération positive de l'éjection ventriculaire	t_1
partie négative de l'éjection ventriculaire	t_2
maximum de *dZ*	dZ_{max}/dt
indice de contractibilité	*C*

TAB. 6.1 – Liste des variables pré-sélectionnées par les médecins susceptibles d'être pertinentes pour prédire l'apparition des symptômes de la syncope durant la position couchée du *tilt-test*.

La résolution de ce problème avait dans un premier été réalisée par une recherche exhaustive de sous-ensembles de variables (*cf.* section 5.3.1). La lecture du tableau 5.4 de la page 134, avait permis d'observer des résultats intéressants par l'utilisation de méthodes génératives, avec notamment les perceptrons multicouches (PMC) et les *support vector machines* (SVM), comme le rappelle le tableau 6.2, extrait du tableau 5.4.

technique de classification	**nombre de variables pertinentes et sous-ensemble optimal de variables**	AUC_V	AUC_T
PMC	3 : { *FC*, *LW/FW*, *Ht* }	$0,802 \pm 0,15^{\dagger}$	$0,630^{\dagger}$
$\mathbf{SVM}_{\text{poly.}}$	6 : { *âge*, *TBW*, *LW/FW*, *Ht*, *Hb*, *FC* }	$0,830 \pm 0,12$	$0,594$
$\mathbf{SVM}_{\text{RBF}}$	5 : { *âge*, *BSA*, *TBW*, *Ht*, t_1 }	$0,800 \pm 0,14$	$0,607$

Note : Pour les $\mathbf{SVM}_{\text{poly.}}$, le degré du polynôme rendant les performances de validation optimales est de 3. Pour les PMC, l'architecture optimale est composée de 19 neurones dans la couche cachée, obtenant sur 100 essais les performances moyennes suivantes : $AUC_V = 0,651 \pm 0,09$ et $AUC_T = 0,554 \pm 0,04$. † indique les performances du meilleur réseau parmi les 100 apprentissages réalisés.

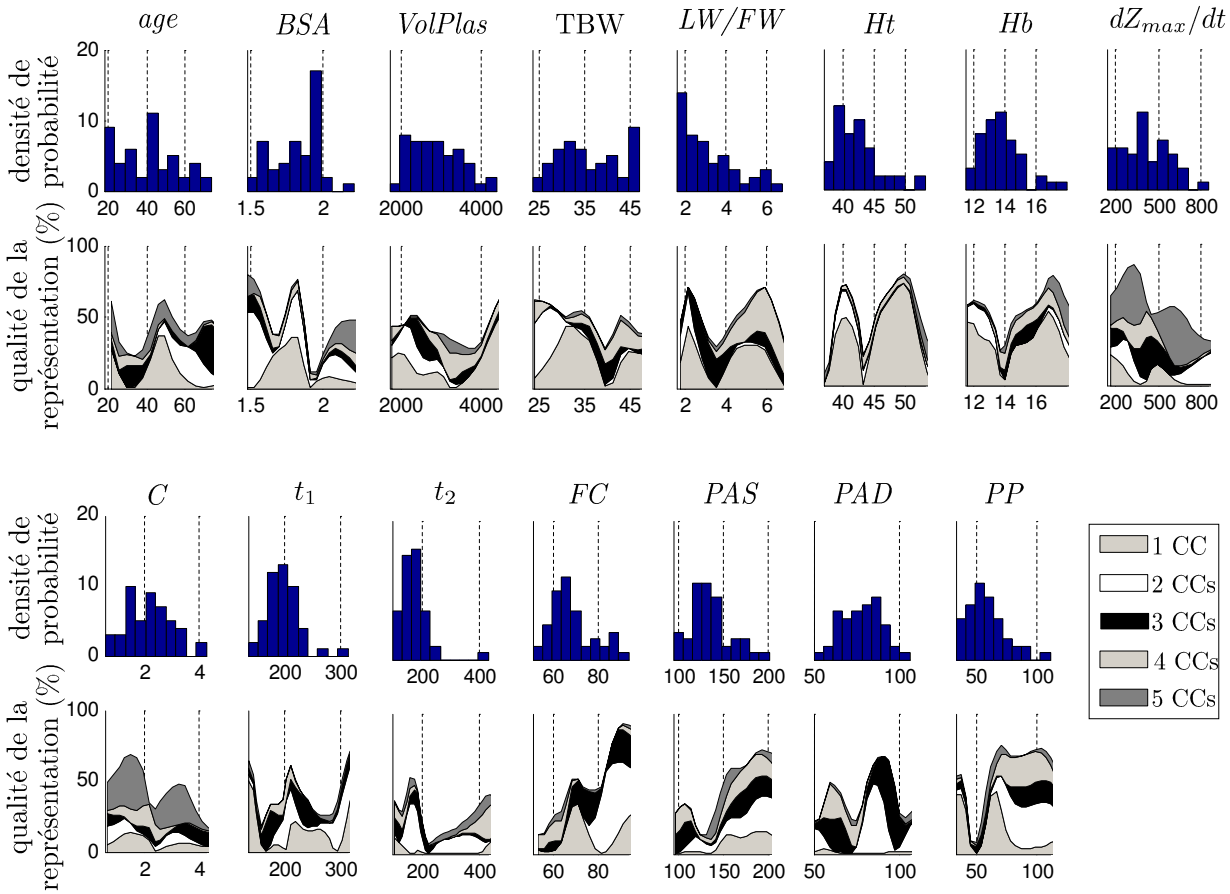
TAB. 6.2 – Comparaison des performances des modèles de classification issus d'une sélection exhaustive des variables pour prédire le *tilt-test* en position couchée, extrait du tableau 5.4.

Ces résultats ont été améliorés en recherchant des combinaisons linéaires et non linéaires des variables originales, respectivement par l'ACP et l'ACC. Ainsi, associé à un PMC, les deux premières composantes résultantes de l'ACP permettaient d'obtenir une aire sous la courbe de ROC de $0,811 \pm 0,11$ et $0,737$, respectivement sur les ensembles de validation et de test (tableau 5.6, page 140). D'autre part, le procédé détaillé à la section 6.4 de ce chapitre, a permis d'extraire les variables qui ont le plus participé à la création de ces deux CP, en l'occurrence : *TBW*, *Ht*, *Hb*, *PAS*, *PAD*, dZ_{max}/dt et *C*. Dans le même temps, avec cinq composantes curvilignes, l'ACC a obtenu une aire sur l'échantillon de test de $0,793$ (en validation $0,801 \pm 0,22$).

Ces résultats ont montré l'efficacité et la pertinence de cette dernière méthode pour notre problématique, dont il nous reste à obtenir la nature des composantes curvilignes.

6.5.3 Extraction de la contribution des composantes curvilignes liées à la prédiction de la syncope

La figure 6.11 montre la qualité de la contribution des 15 variables initiales pour la construction des cinq composantes curvilignes. La transformation non linéaire caractéristique de la projection des données de l'espace original à l'espace réduit est ici, estimée par une fonction polynomiale de second degré. Rappelons que cette fonction permet de positionner les supports des « vecteurs propres locaux », liant ainsi les composantes curvilignes aux variables initiales.



Note : (en haut) Représentation de la distribution des variables initiales ; (en bas) Qualité de la représentation des variables initiales dans les composantes curvilignes.

FIG. 6.11 – Qualité de la représentation des variables pré-sélectionnées dans les composantes curvilignes utilisées pour prédire le résultat du *tilt-test* en position couchée.

La lecture et l'interprétation de la figure n'est pas intuitive, elle est d'autant plus difficile qu'il n'y a pas de structures ou de « formes » qui se dégagent dans la projection, contrairement aux données synthétiques utilisées pour la validation de la méthode. Cela a comme effet, de voir toutes les variables contribuer de manière équivalente à la formation des CC. Dès lors, il n'y a pas réellement de variables qui se démarquent profondément des autres.

Néanmoins, comme évoqué à la section 6.4.6, la pertinence de la lecture de la figure 6.11 peut être considérablement renforcée en considérant la distribution des variables. En effet, dans le cas où une forte qualité de la représentation est liée à une faible distribution de la variable, alors dans ce cas, la représentation de la variable dans la CC n'est pas pertinente, comme lorsqu'une faible qualité de la représentation est associée à une forte distribution de la variable. Pour évaluer cela, il suffit d'évaluer la corrélation entre la qualité de la représentation de la variable et sa distribution dans son espace.

Dès lors, en considérant ce nouveau paramètre, l'analyse de la figure 6.11 montre que dans les cinq premières composantes curvilignes, les variables les plus représentées sont : *Ht*, dZ_{max}/dt, *C* et *PP*. Il est intéressant de remarquer que ces quatre variables sont liées à des degrés divers aux variables extraites par l'ACP. En effet, *Ht*, dZ_{max}/dt et *C* sont apparues pertinentes lors de l'utilisation de l'ACP, tout comme la pression pulsée PP, qui est liée aux variables *PAS* et *PAD* apparues également par l'ACP. Cette observation est d'autant plus remarquable que les résultats obtenus sont très proches, comme le montre le tableau 6.3, extrait du tableau 5.9 de la page 144.

technique de réduction	**technique de classification**	$S_e\,(\%)$	$S_p\,(\%)$	$VPP\,(\%)$	$VPN\,(\%)$	AUC
projection ACP	PMC	88 ± 14 $(71)^{\dagger}$	74 ± 16 $(68)^{\dagger}$	88 ± 13 $(67)^{\dagger}$	80 ± 11 $(72)^{\dagger}$	$0,811 \pm 0,11$ $(0,737)^{\dagger}$
projection ACC	PMC	82 ± 14 $(76)^{\dagger}$	75 ± 39 $(74)^{\dagger}$	83 ± 23 $(72)^{\dagger}$	73 ± 37 $(78)^{\dagger}$	$0,801 \pm 0,22$ $(0,793)^{\dagger}$

Note : † indique que les résultats sont obtenus sur le sous-ensemble de test.

TAB. 6.3 – Récapitulatif des meilleures associations des méthodes de classification et de réduction (sélection et extraction) pour prédire le résultat du *tilt-test* en position couchée, extrait du tableau 5.9.

6.6 Discussions et conclusions

Dans ce dernier chapitre, nous avons proposé une méthodologie qui lie les variables initiales aux composantes issues de projections non linéaires [Feuilloy *et al.*, 2007], à l'image de ce qui est réalisé dans l'interprétation des composantes principales. Notre méthode permet alors de répondre aux difficultés rencontrées et citées notamment par [Illouz and Jardino, 2001; Guérif, 2006] lors de l'utilisation de ces méthodes de projection, en établissant le rôle et la présence des variables originales dans les nouvelles composantes formant l'espace réduit.

Les fondements de notre méthode reposent sur une analogie avec l'analyse en composantes principales, qui est renforcée par le fait que les méthodes de projection non linéaire sont-elles même des extensions de l'ACP [Saporta, 2006]. Cette analogie a donc permis de définir un cadre analytique pour l'extraction d'informations des composantes non linéaires, mais de fait, nous rencontrons les mêmes problèmes d'interprétation que pour l'ACP. En effet, si la contribution des variables est uniforme, l'extraction de ces contributions ne nous permet pas d'obtenir des infor-

mations très pertinentes. C'est ce que nous avons rencontré dans notre étude sur la prédiction de la syncope, car comme l'ACP, la projection par l'ACC n'a pas permis de révéler catégoriquement des variables particulières. Contrairement à ce qui a été observé lors de nos validations expérimentales sur l'ensemble de données synthétiques ; aussi cet ensemble est utilisé dans la littérature principalement pour étudier des méthodes de projection.

De part sa simplicité, notre méthode peut être utilisée non seulement comme élément d'interprétation, comme le suggère notre description, mais aussi comme critère de projection. Plus précisément, nous pouvons imaginer « diriger » ou orienter la projection des données en spécifiant, quelles variables doivent être les plus représentées. Dans cette démarche, nous pouvons aussi bien spécifier quelles parties de l'espace des variables originales doivent participer activement à la projection. Évidemment, cette orientation n'est possible qu'avec une connaissance approfondie du problème et des variables initiales.

Conclusion Générale

Grâce aux avancées des outils informatiques, des soins médicaux et plus largement du domaine biomédical, les recherches dans les systèmes intelligents médicaux et dans l'apprentissage artificiel se sont fortement développées en s'intéressant à des challenges communs. Les contributions produites par ces recherches ont ainsi permis d'améliorer substantiellement la prise en charge des soins des patients. Les travaux présentés dans ce manuscrit de thèse s'inscrivent dans ce contexte, où nous avons étudié des algorithmes d'apprentissage artificiel dans le cadre de la prédiction de la syncope chez l'homme.

Dans cette dernière partie, nous récapitulons les points clés détaillés dans ce mémoire, ainsi que les principales remarques évoquées lors des travaux réalisés sur notre problématique. Pour conclure, nous donnons quelques directions pour de futures recherches.

Résumé

Dans ce manuscrit de thèse, nous avons donné un aperçu des techniques habituellement utilisées en apprentissage artificiel et nous les avons appliquées pour prédire la syncope chez l'homme. Plus précisément, nos études ont consisté à déterminer pour des patients sujets à l'apparition récurrente de syncopes inexpliquées, le résultat du test de la table d'inclinaison (*tilt-test*) qui, rappelons-le, est utilisé pour recréer les conditions dans lesquelles les patients ressentent les symptômes. Les méthodes et leurs utilisations ont été présentées afin de réaliser un système complet de discrimination à partir d'un ensemble de patients, en suivant notamment les étapes liées aux processus de reconnaissance de formes. Ces étapes ont été détaillées avec soin, à l'image du chapitre 1 qui décrit de nombreuses méthodes de discrimination pouvant être employées pour séparer les patients positifs au *tilt-test* des patients négatifs. Le chapitre 2 était consacré aux méthodes de prétraitement et de réduction de la dimensionnalité des données. Ce chapitre donnait un aperçu des techniques utilisées pour extraire et obtenir l'information pertinente liée à notre problématique. L'association des méthodes des deux premiers chapitres permettait de réaliser des modèles de discrimination très variés, qu'il fallait par conséquent évaluer et comparer. Les techniques et les méthodologies du chapitre 3 donnaient des indications afin de sélectionner le modèle le plus approprié pour résoudre notre problème. La présentation des éléments de ces trois premiers chapitres s'est voulue assez générale, de manière à s'ouvrir à d'autres applications. Pour autant, la spécificité du contexte médical, lié à la prédiction de la syncope, n'a pas été altérée et a été largement commentée dans ces chapitres. Comme préalable à la présentation de nos contributions, le chapitre 4 présentait un bref état de l'art du domaine de la prédiction de la syncope lors de l'examen du *tilt-test*.

Les difficultés de ce travail n'ont pas seulement été liées au domaine de l'apprentissage artificiel, mais également au traitement de données provenant d'un environnement médical. En effet, les difficultés s'accentuent avec le nombre de données disponibles, qui est bien souvent insuffisant pour appliquer sereinement les algorithmes d'apprentissage artificiel. Dans de telles configurations, les résultats obtenus peuvent sur-estimer ou sous-estimer les performances réelles des modèles, biaisant ainsi leur interprétation. Il est alors nécessaire de mettre en œuvre des méthodes particulières pour évaluer et estimer les performances des modèles. L'évaluation des performances est primordiale et nécessite une attention particulière. Cette tâche est certainement aussi importante que l'élaboration des modèles de discrimination. Les techniques de discrimination posent

d'autres difficultés, dont certaines sont une nouvelle fois liées au type de données analysées. En effet, certaines techniques nécessitent de faire des hypothèses sur les données qui, provenant de phénomènes biologiques, sont très sensibles et bruitées. C'est pour prévenir ces difficultés, que les approches discriminantes sont souvent privilégiées afin de traiter la tâche de discrimination dans le domaine médical. Dès lors, nous avons vu que les principales méthodes utilisées sont les réseaux de neurones et les *support vector machines* (SVM), qui sont tous les deux capables d'obtenir des modèles linéaires et non linéaires. Les réseaux de neurones ont l'avantage d'être utilisés depuis de très nombreuses années, donnant par conséquent un recul non négligeable sur leur utilisation. Malgré ce *background*, leur manipulation est peu intuitive et nécessite d'employer fréquemment des heuristiques pour améliorer les performances. Cela est certainement lié à leurs origines qui s'appuyaient en partie sur des intuitions, sur lesquelles un cadre théorique a été « greffé ». Les *support vector machines* se sont développés sur la base des critiques faites à l'encontre des réseaux de neurones, apportant ainsi un cadre théorique plus clair. Aussi, l'engouement très actif de la communauté scientifique pour cette approche a permis d'offrir de nombreuses contributions, amenant les *support vector machines* à résoudre, plus aisément que les réseaux de neurones, des challenges actuels tels que, le traitement de gros ensembles de données de très grandes dimensions.

Un élément important des méthodes de discrimination concerne leur complexité qui doit être réduite au maximum afin d'obtenir des modèles les plus parcimonieux possibles. Cet élément est nécessaire pour réaliser un apprentissage performant et une interprétation claire et efficace du modèle. Ainsi, les méthodes de discrimination voient leurs performances s'améliorer considérablement, une fois associées à des techniques de sélection et d'extraction de variables. Cette phase est essentielle et a été largement traitée dans notre travail. Les techniques disponibles sont extrêmement diversifiées, leur utilisation est de nouveau dépendante de la nature des données. Globalement, ces techniques permettent de réduire la dimension qui caractérise l'espace des variables (entrée du modèle) et dans le cadre de notre travail, elles permettent également d'aider à comprendre le mécanisme d'apparition des symptômes de la maladie et ont un impact économique sur l'acquisition des données.

L'élément majeur qui ressort de ce résumé concerne l'utilisation des techniques d'apprentissage artificiel qui, malgré de nombreuses recherches, ne peuvent s'appliquer qu'en considérant scrupuleusement la nature et la structure des données ; l'aspect générique de ces techniques évoqué dans la littérature n'apparaît plus aussi formellement dans la pratique.

Principales contributions

Nos travaux se sont appuyés sur des études menées au service de cardiologie du CHU d'Angers, impliquant des groupes de patients sujets à l'apparition récurrente de syncopes inexpliquées. Afin de déterminer la cause d'apparition des syncopes, ces patients sont amenés à réaliser l'examen du *tilt-test*. Reconnu pour recréer les conditions d'apparition des symptômes, cet examen a l'inconvénient de monopoliser du personnel médical durant près d'une heure. Les travaux présentés dans cette thèse ont eu comme objectif de prédire l'apparition des signes de la syncope avant que l'examen n'arrive à son terme, afin de réduire le coût de cet examen et d'éviter aux patients de ressentir les symptômes. Les travaux engagés sur notre problématique peuvent se répartir en trois parties.

La première partie concerne l'analyse de la phase de repos du *tilt-test*, durant laquelle nous avons comparé de nombreuses techniques de discrimination telles que, les réseaux de neurones, les *support vector machines*, le classifieur de Bayes naïf ou encore, les fonctions discriminantes. Les études, menées sur un ensemble de variables pré-sélectionnées par les médecins, ont per-

mis d'observer de meilleures performances pour les modèles basés sur les réseaux de neurones (notamment les perceptrons multicouches, PMC) et les *support vector machines*. D'autre part, nous avons pu observer également une prépondérance de la participation de certaines variables dans la construction des modèles telles que, l'âge, la surface corporelle ou encore, le taux d'hématocrite. Bien qu'intéressants, ces résultats [Feuilloy *et al.*, 2005b; Feuilloy *et al.*, 2005a] ne surpassent que de peu les performances d'autres études similaires parues dans la littérature. Dès lors, afin d'améliorer nos performances, nous avons adopté une autre démarche qui, consiste à chercher par des méthodes de projections, des combinaisons linéaires et non linéaires dans les variables initiales ; l'utilisation de l'analyse en composantes principales (ACP) et de l'analyse en composantes curvilignes (ACC) a permis de synthétiser au sein de nouvelles composantes l'information contenue dans les variables initiales. Les résultats obtenus [Feuilloy *et al.*, 2005b; Feuilloy *et al.*, 2005c] ont montré l'impact de ces méthodes de projection sur notre problématique, en améliorant considérablement les performances, et cela, en dépassant maintenant les résultats mentionnés dans la littérature. En effet, la sensibilité moyenne du modèle de discrimination qui donne le résultat du *tilt-test*, atteint désormais 84 % (avec une spécificité de 71 %) sur un ensemble de patients totalement inconnus (ensemble de test), contre 58 % (avec une spécificité de 59 %) lors de l'analyse précédente. Notons que ces deux résultats sont obtenus respectivement par des modèles basés sur des PMC et sur des SVM. L'ACP permet d'estimer les variables qui ont le plus contribué à la construction des composantes principales (CP). Par ce procédé propre à l'ACP, nous avons pu observer parmi les variables initiales, que celles liées au signal d'impédancemétrie thoracique sont fortement représentées dans les CP conservées.

La seconde partie concerne l'analyse des dix premières minutes de la phase de basculement du *tilt-test*, durant laquelle nous avons comparé cette fois-ci de nombreuses techniques de sélection de variables. L'objectif était de découvrir les sous-ensembles de variables capables d'optimiser la séparation des classes et donc de prédire au mieux le résultat du *tilt-test*. Les études menées au CHU d'Angers nous ont permis de traiter un ensemble de patients caractérisés par 70 variables, entraînant de fait un nombre de combinaisons considérable ($1,18 \cdot 10^{21}$) et rendant impossible une recherche exhaustive du sous-ensemble de variables pertinentes. Dès lors, pour faire face à cette explosion combinatoire, de nombreuses techniques proposées par la littérature ont été comparées telles que l'algorithme RELIEF, la mesure du critère de Fisher, les techniques de recherches séquentielles et les approches non déterministes. La comparaison de ces méthodes a révélé une dépendance élevée entre la complexité des méthodes (en termes du nombre de combinaisons évaluées et donc en termes de coût calculatoire) et les performances obtenues, nous amenant à observer la supériorité des algorithmes génétiques (AG). Cependant, les techniques de sélection séquentielle ont vu certainement leurs performances entachées par des difficultés d'exploration de l'espace des combinaisons. Par ailleurs, en dépit de leurs performances plus faibles, il est apparu que l'exploration et les performances de ces méthodes pouvaient être améliorées. La méthode la plus populaire, l'algorithme de sélection ascendante séquentielle (SFS), a déjà connu des déclinaisons par l'intégration de processus de retours en arrière. Ces retours en arrière effectués de manière séquentielle, améliorent l'exploration des combinaisons, en évitant à l'algorithme SFS de s'isoler trop rapidement dans des minimums locaux. C'est dans ce cadre que nous avons développé une nouvelle approche pour la sélection de variables, fondée sur une combinaison entre l'algorithme de recherche séquentielle SFS et les algorithmes génétiques. Cette nouvelle approche permet à SFS d'effectuer durant son exploration, des retours en arrière stochastiques et non plus séquentiels. Ainsi, les algorithmes génétiques interviennent de manière aléatoire sur un sous-ensemble de variables préalablement sélectionné par le processus SFS, uniquement pour réaliser les retours en arrière et optimiser le sous-ensemble de variables. Cette approche fait donc coïncider une recherche locale (SFS) avec une recherche globale (AG). Les observations faites une fois ces nouvelles méthodes comparées aux précédentes ont permis de constater leur grande capacité à

sélectionner des sous-ensembles de variables optimisant la qualité de discrimination, surpassant les performances des algorithmes génétiques [Feuilloy *et al.*, 2006a]. D'autre part, cette combinaison de méthodes accroît leurs performances en réduisant considérablement le nombre de variables sélectionnées, tout en minimisant le nombre de combinaisons évaluées et par conséquent le coût calculatoire. Ainsi, avec uniquement quatre variables sélectionnées par notre nouvelle approche, liées notamment au signal d'impédancemétrie thoracique et à la pression artérielle, la sensibilité du fondé sur un PMC atteint désormais 100 % (avec une spécificité de 94 %). Aussi, cette nouvelle approche de sélection de variables a montré sa reproductibilité lors de tests sur des ensembles de données provenant de l'*UCI repository of machine learning databases*.

L'observation générale des travaux présentés précédemment nous a amené à nous intéresser plus particulièrement au signal d'impédancemétrie thoracique, établissant par conséquent notre troisième partie. Ainsi, l'analyse spécifique du signal d'impédancemétrie thoracique a démontré son impact et sa pertinence dans la prédiction du résultat du *tilt-test*. En effet, les performances obtenues lors de son utilisation a permis de le comparer très favorablement avec des mesures plus classiques, telles que la fréquence cardiaque ou encore la pression artérielle. D'autre part, ce signal apporte par ces caractéristiques des informations importantes sur l'hémodynamique cardiaque, suggérant son utilisation à la place de variables difficilement accessibles, comme le taux d'hématocrite qui s'est révélé être pertinent dans notre problématique. En effet, à l'image de l'électrocardiogramme, le signal d'impédancemétrie peut être enregistré en continu sans nécessiter « d'intervention humaine », en laissant le soin aux « machines » de surveiller l'évolution des courbes. Cette surveillance peut être en l'occurrence améliorée par les procédés de prétraitement du signal d'impédancemétrie thoracique que nous avons développés dans nos travaux. En effet, nous avons cherché à améliorer l'utilisation de ce signal en perfectionnant le processus d'extraction des caractéristiques, de manière à optimiser leur pouvoir discriminant. Les techniques développées [Feuilloy *et al.*, 2006b; Feuilloy *et al.*, 2006c] sont fondées sur des méthodes issues du traitement du signal et ont pour objectif de trouver sur le signal temporel, la région dans laquelle les caractéristiques extraites seront les plus pertinentes. Ce travail de préparation des données peut être considéré comme une étape de prétraitement à la sélection et à la discrimination. Aussi, nous avons cherché à évaluer d'une part, l'efficacité de notre procédé de prétraitement et d'autre part, la pertinence de nouvelles caractéristiques de ce signal, telles que des caractéristiques extraites dans le domaine fréquentiel. Les résultats obtenus [Feuilloy *et al.*, 2006c; Schang *et al.*, 2007] ont permis d'améliorer considérablement le bénéfice qu'apporte l'utilisation de ce signal dans la prédiction du résultat du *tilt-test* en période de repos, en fournissant une sensibilité de 100 % (avec une spécificité de 97 %).

La considération exclusive de la phase de repos a montré que nous pouvons nous passer du basculement de la table, au détriment d'une erreur de classement des patients obtenant un résultat négatif, de l'ordre de 3 %. On peut faire la même observation, en considérant les dix premières minutes de la période basculée, où cette fois l'erreur du modèle pour la classification des patients négatifs atteint 6 %. Dès lors, au vu de ces résultats, nous devrions pouvoir nous passer de l'examen complet du *tilt-test*, en réduisant à 10 ou 20 minutes la période de l'examen, sachant qu'elle est initialement d'une heure.

Au cours de nos recherches, nous avons pu constater la difficulté pour interpréter les méthodes de projection non linéaire et plus particulièrement les composantes résultantes de ces processus. Or, l'utilisation d'une de ces méthodes, en l'occurrence l'ACC, a permis d'aboutir à un modèle performant et capable de prédire l'apparition des symptômes de la syncope durant la phase de repos. Aussi, contrairement à l'ACP, aucune technique ne permettait de lier les variables initiales aux composantes non linéaires résultantes de l'ACC. Ce point est fondamental, car lier une

composante aux variables initiales permet de trouver et de donner une signification à cette composante. C'est dans ce contexte que nous avons développé une méthodologie qui permet d'extraire la quantité de la représentation des variables initiales dans les composantes non linéaires [Feuilloy *et al.*, 2007]. Notre approche d'extraction d'informations est fondée sur une reproductibilité et une adaptation de la technique utilisée pour l'interprétation des composantes principales. Le fondement de notre approche a été motivé par le fait que l'auteur de l'ACC voit sa méthode comme une extension non linéaire de l'ACP, en la considérant comme une ACP par parties. D'autre part, le fait que la plupart des méthodes de projection non linéaire sont fondées sur une même notion, qui est la préservation locale de la topologie ou de la structure des données, nous a amené à penser que le processus d'estimation de la représentation des variables dans les composantes développé pour l'ACC peut s'adapter à d'autres techniques de projection non linéaire. Cela s'est vérifié lors d'une validation expérimentale sur des données synthétiques. Par le développement de ce processus d'extraction d'informations, il est apparu une nouvelle fois que les variables liées au signal d'impédancemétrie étaient fortement représentées dans les composantes curvilignes utilisées par le modèle de discrimination. Ce procédé permet sans conteste d'enrichir les méthodes de projection non linéaire, en permettant d'accroître la connaissance sur le résultat de la projection. D'autre part, l'habitude a été prise d'utiliser ces méthodes non linéaires pour réduire l'espace en deux ou trois dimensions, en ignorant des dimensions plus élevées, afin certainement de rendre l'interprétation possible. Notre méthodologie devrait donc permettre de travailler dans des dimensions plus importantes, tout en disposant d'une certaine connaissance sur le résultat de la projection.

Perspectives de recherche

Les différentes études et expérimentations réalisées dans ces recherches ont permis d'explorer de nombreuses pistes pour prédire l'apparition des symptômes de la syncope lors de l'examen du *tilt-test*. Néanmoins, plusieurs points mériteraient d'être explorés ou encore améliorés.

L'un d'eux serait d'apporter une information supplémentaire aux modèles de discrimination, de manière à obtenir une connaissance quant à la certitude de la conclusion donnée par le modèle sur le résultat du *tilt-test*. Cette certitude est implicite pour les modèles probabilistes et à l'image des travaux de [Lu, 2005], elle pourrait être implémentée sur les techniques discriminantes afin d'obtenir une probabilité d'appartenance [Suykens *et al.*, 2002; Lu, 2005; Bishop, 2006].

L'amélioration des performances pourrait être envisagée par la combinaison de techniques de discrimination [Haykin, 1999]. En conservant les caractéristiques et l'efficacité de classement de chacune des méthodes combinées, cette technique augmente son pouvoir discriminant et par conséquent, voit ses performances s'améliorer avec l'avantage de réduire la complexité de chacun des modèles mis en œuvre.

L'un des axes de recherche qui nous semble intéressant concerne l'interprétation des modèles. En effet, l'utilisation des méthodes discriminantes telles que, les réseaux de neurones et les *support vector machines*, ne permettent pas de comprendre précisément les règles de classement des patients. Cette interprétation est difficile à mettre en œuvre, compte tenu notamment de la complexité des modèles issus des réseaux de neurones et des SVM. En effet, ces derniers sont connus pour être des « boîtes noires », et cela malgré la volonté d'obtenir des modèles les plus parcimonieux possibles. Cependant, l'intérêt de pouvoir ouvrir les modèles et de les comprendre a poussé la communauté scientifique à engager des recherches, à l'image des travaux de [Andrews *et al.*, 1995; Zhou and Jiang, 2003; Diederich, 2008].

Les travaux menés durant cette thèse ont permis d'apporter une contribution importante sur l'interprétation des méthodes de projection non linéaire. Aussi, cette nouvelle méthode a été utilisée afin d'obtenir les relations entre les variables initiales et les composantes résultantes de la projection. Il a été envisagé une autre utilisation, notamment comme support dans les processus de projection non linéaire, afin de guider ou de diriger la projection des données en spécifiant quelles variables doivent être les plus représentées. Cette utilisation mériterait d'être employée, notamment dans notre problématique, afin d'influencer la réduction de la dimension pour privilégier les variables les moins coûteuses à acquérir.

Annexes

Annexe A

Compléments mathématiques

Dans ce manuscrit, certains détails mathématiques n'ont pas été développés afin de conserver une fluidité dans la lecture. Parmi les points occultés, nous retrouvons dans cette annexe les détails de l'algorithme de rétropropagation (*cf.* section 1.4.2.5), de l'extraction des nouvelles caractéristiques sur le signal d'impédancemétrie thoracique (*cf.* section 5.5.4) et de l'estimation de la probabilité d'erreur (*cf.* section 5.5.5.1).

A.1 Algorithme de rétropropagation

La démonstration de l'algorithme de rétropropagation est réalisée sur un perceptron multicouches possédant une couche cachée de neurones. Les notations utilisées sont données à la figure A.1, sur laquelle est également illustrée la propagation des observations d'entrée dans le réseau et la rétropropagation de l'erreur à travers ses couches.

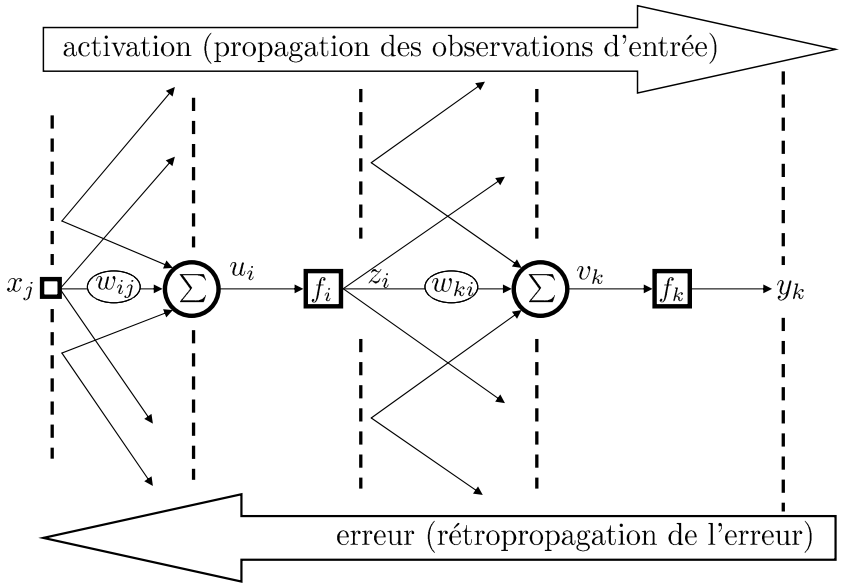


FIG. A.1 – Illustration des notations utilisées pour la démonstration de l'algorithme de rétropropagation.

Une observation d'entrée est propagée dans le réseau à travers ses couches, jusqu'aux neurones de sortie. L'erreur de chaque neurone de sortie est calculée par :

$$E = \frac{1}{2}(t_k - y_k)^2 \,, \tag{A.1}$$

amenant à obtenir l'erreur globale :

$$E_g = \frac{1}{2}\sum_k (t_k - y_k)^2 \,. \tag{A.2}$$

Fondée sur la règle *Delta* (*cf.* section 1.4.2.4), la rétropropagation ajuste les poids de la même manière, par :

$$\Delta w_{ki} = -\rho \frac{\partial E}{\partial w_{ki}} \,, \tag{A.3}$$

où ρ définit le pas d'apprentissage. La dérivée $\frac{\partial E}{\partial w_{ki}}$ doit donc être évaluée, celle-ci se décompose par la formulation suivante :

$$\frac{\partial E}{\partial w_{ki}} = \frac{\partial E_k}{\partial y_k} \cdot \frac{\partial y_k}{\partial v_k} \cdot \frac{\partial v_k}{\partial w_{ki}} \,. \tag{A.4}$$

Le calcul de $\frac{\partial E}{\partial w_{ki}}$ passe alors par la détermination de chacune des dérivées partielles. Sachant que $E = \frac{1}{2}(t_k - y_k)^2$ définit l'erreur du neurone de sortie k, on obtient alors :

$$\frac{\partial E}{\partial y_k} = -(t_k - y_k) \,. \tag{A.5}$$

De même, comme la sortie $y_k = f_k(v_k)$, alors :

$$\frac{\partial y_k}{\partial v_k} = f'_k(v_k) \,. \tag{A.6}$$

Aussi, l'activation du neurone de sortie étant $v_k = \sum w_{ki} z_i + w_{k0}$ permet de déduire que :

$$\frac{\partial v_k}{\partial w_{ki}} = z_i \,. \tag{A.7}$$

Enfin, l'adaptation des poids des neurones de sortie $\Delta w_{ki} = -\rho \frac{\partial E}{\partial w_{ki}}$ peut alors s'écrire :

$$\Delta w_{ki} = -\rho \, (t_k - y_k) \cdot f'_k(v_k) \cdot z_i \tag{A.8}$$

Posons, $\delta_k = \frac{\partial E}{\partial v_k}$, que nous appellerons signal d'erreur.

L'adaptation des neurones de la couche cachée ne peut pas suivre le même processus. Comme nous l'avons dit précédemment, la sortie désirée des neurones cachés est inconnue. Pour contourner ce problème, on considère que l'erreur des neurones de sortie est liée proportionnellement à l'activation des neurones cachés. Ainsi, on peut exprimer l'erreur des neurones de sortie en fonction de leur entrée, donc en fonction des potentiels des neurones cachés. L'erreur du neurone k peut être formulée en fonction des M neurones de la couche cachée : $E = E(u_0, u_1, \ldots, u_M)$.

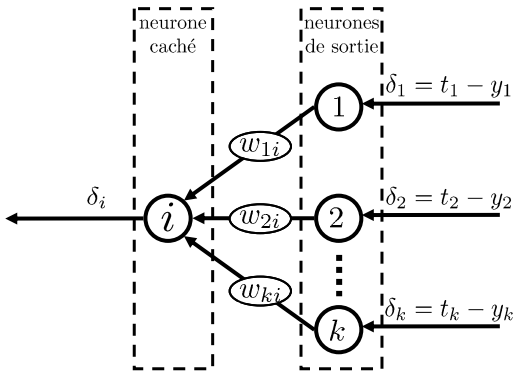


FIG. A.2 – Illustration de la rétropropagation de l'erreur, des k neurones de sortie vers le neurone i de la couche cachée.

Ainsi, par identification aux neurones de sortie (A.5), l'erreur du neurone caché i est :

$$\frac{\partial E}{\partial u_i} = \delta_i = \frac{\partial E}{\partial z_i} \cdot \frac{\partial z_i}{\partial u_i} \,. \tag{A.9}$$

Comme précédemment, il faut déterminer chacune des dérivées partielles. Aussi, par la rétropropagation, l'erreur d'un neurone caché dépend de tous les neurones de sortie, donc de l'erreur globale E_g (A.2). Dès lors, la dérivée $\frac{\partial E}{\partial z_i}$ est égale à :

$$\begin{aligned} \tfrac{\partial E}{\partial z_i} &= \tfrac{\partial}{\partial z_i}\left(\tfrac{1}{2}\sum_k (t_k - y_k)^2\right) , \\ &= \tfrac{1}{2}\sum_k \tfrac{\partial}{\partial z_i}(t_k - y_k)^2 \,, \\ &= -\sum_k (t_k - y_k) \cdot \tfrac{\partial y_k}{\partial z_i} \,, \end{aligned} \tag{A.10}$$

avec $\frac{\partial y_k}{\partial z_i} = \frac{\partial y_k}{\partial v_k} \cdot \frac{\partial v_k}{\partial z_i} = f'_k(v_k) w_{ki}$, d'où :

$$\frac{\partial E}{\partial z_i} = -\sum_k (t_k - y_k) \cdot f'_k(v_k) w_{ki} \,. \tag{A.11}$$

Rappelons que l'activation $v_k = \sum w_{ki} z_i + w_{k0}$ et $z_i = f_j(u_i)$, alors :

$$\frac{\partial z_i}{\partial u_i} = f'_i(u_i) \,. \tag{A.12}$$

Le signal d'erreur des neurones cachés est donc :

$$\begin{aligned} \delta_i = \tfrac{\partial E}{\partial u_i} &= f'_i(u_i) \cdot \sum_k -(t_k - y_k) \cdot f'_k(v_k)\, w_{ki} \,, \\ &= f'_i(u_i) \cdot \sum_k \delta_k w_{ki} \,. \end{aligned} \tag{A.13}$$

Ainsi, $\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial u_i} \cdot \frac{\partial u_i}{\partial w_{ij}} = \delta_i \, x_j$ et l'adaptation des poids des neurones de la couche cachée est donnée par :

$$\Delta w_{ij} = \rho \, \delta_i \cdot x_j \quad \text{avec, } \delta_i = f'_i(u_i) \cdot \sum_k \delta_k w_{ki} \,. \tag{A.14}$$

Rappelons que l'adaptation des poids de toutes les couches ne s'effectue qu'une fois la rétropropagation terminée. La démonstration ne précise pas le choix des fonctions d'activation, nous l'avons présentée dans le cas général. Prenons l'exemple d'une fonction d'activation de type logistique :

$$f_k(v_k) = \frac{1}{1 + e^{-v_k}} \,, \tag{A.15}$$

sa dérivée est donnée par :

$$f'_k(v_k) = \frac{e^{-v_k}}{(1 + e^{-v_k})^2} = f_k(v_k)[1 - f_k(v_k)] \,. \tag{A.16}$$

Comme $y_k = f_k(v_k)$, alors, $f'_k(y_k) = y_k(1 - y_k)$. L'adaptation des neurones de la couche de sortie devient alors :

$$\Delta w_{kj} = \rho \, (t_k - y_k) \cdot y_k(1 - y_k) \cdot z_j \,. \tag{A.17}$$

Par identification, l'adaptation des neurones de la couche cachée est :

$$\Delta w_{ij} = \rho \, z_i(1 - z_i) \cdot \sum_k (t_k - y_k) \cdot y_k(1 - y_k) \cdot z_i \cdot w_{ki} \cdot x_j \,. \tag{A.18}$$

A.2 Estimation des pentes et des aires sur le signal d'impédancemétrie thoracique et sa dérivée

Dans cette section, nous proposons de détailler les étapes des calculs des nouvelles caractéristiques extraites sur le signal d'impédancemétrie thoracique. Illustrées de nouveau à la figure A.3, ces caractéristiques représentent l'aire ($Area_{norm}$) et la pente ($Slope_{norm}$) sur le signal Z et sa dérivée dZ. Ces paramètres ont été utilisés lors de la prédiction du résultat du *tilt-test* (*cf.* section 5.5.4).

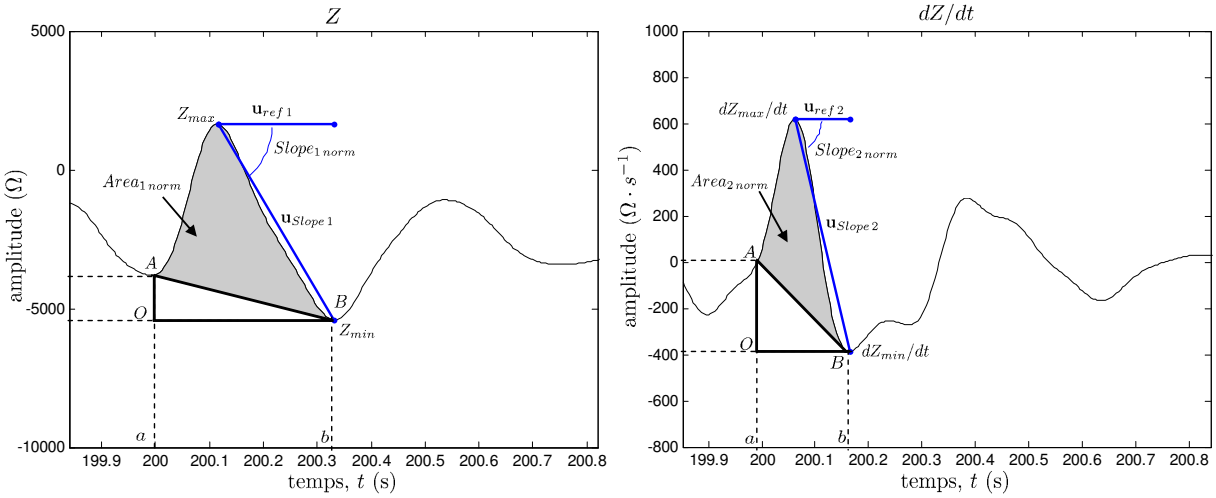


FIG. A.3 – Nouvelles caractéristiques extraites du signal d'impédancemétrie thoracique dans le domaine temporel.

A.2.1 Détermination des pentes ($Slope_{norm}$)

Le calcul des pentes est déterminé par l'angle obtenu entre deux vecteurs, ceux-ci sont notés $\mathbf{u}_{ref}$ et $\mathbf{u}_{Slope}$ sur la figure A.3. Afin de s'assurer que les normes des deux vecteurs soient égales, nous les normalisons par $\mathbf{u}_{Slope} = \frac{\mathbf{u}_{Slope}}{\|\mathbf{u}_{Slope}\|}$ et $\mathbf{u}_{ref} = \frac{\mathbf{u}_{ref}}{\|\mathbf{u}_{ref}\|}$, dès lors $\|\mathbf{u}_{ref}\| = \|\mathbf{u}_{Slope}\| = 1$. Le cosinus de l'angle représentatif de la pente, est déterminé par le produit scalaire $\mathbf{u}_{ref} \cdot \mathbf{u}_{Slope}$, qui lui même est obtenu par :

$$\mathbf{u}_{ref} \cdot \mathbf{u}_{Slope} = \frac{1}{2}\left(\|\mathbf{u}_{ref} + \mathbf{u}_{Slope}\|^2 - \|\mathbf{u}_{ref}\|^2 - \|\mathbf{u}_{Slope}\|^2\right) . \tag{A.19}$$

A.2.2 Détermination des aires ($Area_{norm}$)

Le calcul des aires est obtenu par la règle de Simpson [Burden and Faires, 2001], qui est une technique d'intégration numérique. Comme la règle du trapèze, l'intégrale à déterminer est « cassée » en sous-intégrales. Cependant, contrairement à la technique trapézoidale qui approxime chaque sous-intervalle par des trapèzes, la règle de Simpson utilise des formes quadratiques (A.20) pour améliorer la précision de l'approximation, telle que :

$$\int_a^b f(t)\,\mathrm{d}t \approx \frac{b-a}{6}\left(f(a) + 4f\left(\frac{b+a}{2}\right) + f(b)\right) . \tag{A.20}$$

Pour calculer les aires sur l'intervalle $[a\,;\,b]$ (voir figure A.3), il est nécessaire que les signaux soient définis positifs sur cet intervalle. Ainsi, dans l'approximation des aires $\int_a^b f(t)\,\mathrm{d}t$ ($f(t)$ représente Z ou dZ/dt), le changement d'origine amène à considérer pour chacun des signaux, la nouvelle origine au niveau du point O. L'intégrale $\int_a^b f(t)\,\mathrm{d}t$ est divisée en M sous-intervalles : $[a + 2ih\,;\,a + 2(i+1)h]$, de même largeur $2h = \frac{b-a}{M}$, où i varie entre 0 et $M-1$.

L'aire sous le signal dans l'intervalle $[a\,;\,b]$ est alors exprimée comme la somme des aires des sous-intervalles :

$$\int_a^b f(t)\,\mathrm{d}t = \sum_{i=0}^{M-1} \int_{a+2ih}^{a+2(i+1)h} f(t)\,\mathrm{d}t\,. \tag{A.21}$$

Ainsi, en remplaçant chaque sous-intégrale par l'approximation donnée en (A.21), nous obtenons l'approximation de notre intégrale par :

$$\int_a^b f(t)\,\mathrm{d}t \approx \frac{h}{3} \sum_{i=0}^{M-1} \left(f(a+2ih) + 4f(a+(2i+1)h) + f(a+2(i+1)h)\right)\,. \tag{A.22}$$

Enfin, les aires $Area_{1\,norm}$ et $Area_{2\,norm}$ sont obtenues en enlevant de l'intégrale calculée, l'aire du triangle (OAB), comme le montre la figure A.3.

A.3 Estimation de la probabilité d'erreur de classification

À la section 5.5.5.1, nous avons abordé une mesure singulière pour estimer la pertinence d'une variable. Cette mesure est fondée sur une estimation du recouvrement entre les classes de la variable (voir figure A.4), afin d'évaluer sa capacité à discriminer les classes.

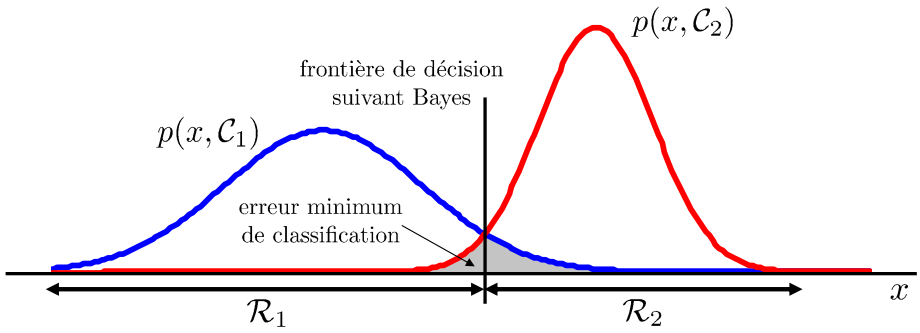


FIG. A.4 – Illustration de l'erreur de classification obtenue par la règle de Bayes.

L'estimation du recouvrement repose sur le théorème de Bayes, défini préalablement à la section 1.2.2.1 (page 16). Le modèle de classification obtenu par ce théorème produit en chaque point de l'espace une règle qui permet d'assigner chaque nouvelle observation à l'une des K classes connues. L'espace des variables est par conséquent divisé en plusieurs régions, entre lesquelles apparaissent les frontières de décision. La règle de Bayes (définition 1, page 16) cherche les frontières qui réduisent le recouvrement entre les classes, et donc qui minimisent la probabilité d'erreur de classification, comme le montre la figure A.4. Dans un problème à deux classes, [Duda and Hart, 1973] estiment cette probabilité d'erreur simplement par :

$$\begin{aligned} P_{err} &= P(x \in \mathcal{R}_2, \mathcal{C}_1) + P(x \in \mathcal{R}_1, \mathcal{C}_2)\,, \\ &= P(x \in \mathcal{R}_2 | \mathcal{C}_1) P(\mathcal{C}_1) + P(x \in \mathcal{R}_1 | \mathcal{C}_2) P(\mathcal{C}_2)\,, \\ &= \int_{\mathcal{R}_2} p(x|\mathcal{C}_1) P(\mathcal{C}_1) \mathrm{d}x + \int_{\mathcal{R}_1} p(x|\mathcal{C}_2) P(\mathcal{C}_2)\,\mathrm{d}x\,, \end{aligned} \tag{A.23}$$

où $P(x \in \mathcal{R}_2, \mathcal{C}_1)$ donne la probabilité que l'observation x est assignée à la classe $\mathcal{C}_1$, sachant qu'elle appartient réellement à la classe $\mathcal{C}_2$.

Si les formes analytiques des densités de probabilité $p(x|\mathcal{C}_k)$[1] sont connues, alors nous pouvons déterminer précisément le recouvrement entre les classes. Dans le cas contraire, l'estimation de ces densités peut amener à déterminer la probabilité d'erreur, par des approximations d'intégrales, ou encore par des simulations de Monte-Carlo [Kay, 1993].

Prenons le cas d'un problème à deux classes caractérisé par une variable x, où les densités de probabilité $p(x|\mathcal{C}_k)$ suivent des lois normales. Les paramètres (moyenne, écart type) de ces densités sont notés par $(\mu_1\,,\,\sigma_1)$ et $(\mu_2\,,\,\sigma_2)$.

Pour obtenir la probabilité d'erreur, nous utilisons par exemple, l'approximation d'intégrales par la règle de Simpson (*cf.* section A.2.2). Cette méthode approxime chaque sous-intervalle d'une fonction $f(x)$ en une forme quadratique. L'intégrale $\int_a^b f(x)\,\mathrm{d}x$, dans notre cas $\int_a^b p(x, C_k)\,\mathrm{d}x$, est divisée en M sous-intervalles $[a{+}2ih,\, a{+}2(i{+}1)h]$ de taille égale $2h = \frac{b-a}{M}$. L'aire sous les densités dans l'intervalle $[a\,;\,b]$ est exprimée comme une somme des aires des sous-intervalles, comme le montre l'expression suivante :

$$\int_a^b p(x, \mathcal{C}_k)\,\mathrm{d}x = \sum_{i=0}^{M-1} \int_{a+2ih}^{a+2(i+1)h} p(x, \mathcal{C}_k)\,\mathrm{d}x\,. \tag{A.24}$$

La difficulté réside alors dans la détermination des limites (a et b) de l'intervalle du calcul de l'intégrale : les limites extérieures des densités de probabilité et les points d'intersection (frontières de décision). En faisant l'hypothèse que les densités de probabilité suivent des lois normales, les difficultés du problème se réduisent. Dans ce cas, nous pouvons considérer que la répartition de chaque classe k est comprise dans l'intervalle $[\mu_k - 3\,\sigma_k\,;\,\mu_k + 3\,\sigma_k]$. Ainsi, comme le montre la figure A.5, l'intégrale représentative du recouvrement des classes est calculée entre $[\mu_{min} - 3\,\sigma_{max}\,;\,\mu_{max} + 3\,\sigma_{max}]$, où μ_{min} et μ_{max} sont respectivement la moyenne la plus forte et la plus faible (parmi les μ_k) et σ_{max} correspond à l'écart type le plus grand (parmi les σ_k).

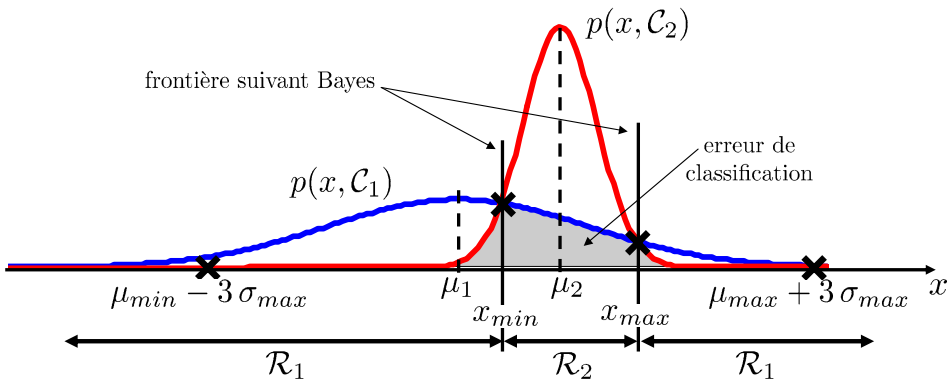


FIG. A.5 – Représentation des paramètres utilisés dans l'estimation de la probabilité de l'erreur de classification.

Les intersections des densités (notées x_{min} et x_{max} à la figure A.5) signifient qu'en ces points, les densités de probabilité sont égales : $p(x|\mathcal{C}_1) = p(x|\mathcal{C}_2)$. Ainsi, une fois les points d'intersection trouvés, la probabilité d'erreur est estimée dans les trois intervalles : $[\mu_{min} - 3\,\sigma,\, x_{min}]$, $[x_{min},\, x_{max}]$ et $[x_{max},\, \mu_{max} + 3\,\sigma]$.

[1] Rappelons que $p(x, \mathcal{C}_k) = p(x|\mathcal{C}_k)\,P(C_k)$, comme évoqué à la section 1.2.2.2.

Pour chaque intervalle (chaque région), les densités de probabilité de chaque classe sont calculées et nous retenons la plus faible ; celle représentative de l'erreur, et donc du recouvrement. Ainsi, en ajoutant les densités retenues pour chaque intervalle, nous obtenons en considérant un facteur de normalisation (A.25), la probabilité d'erreur de classification P_{err} (A.26).

$$\text{facteur de normalisation} = \sum_{k} \int_{\mu_k - 3\,\sigma_k}^{\mu_k + 3\,\sigma_k} p(x, \mathcal{C}_k)\,\mathrm{d}x \tag{A.25}$$

$$P_{err} \approx \frac{\int_{\mathcal{R}_2} p(x, \mathcal{C}_1)\,\mathrm{d}x + \int_{\mathcal{R}_1} p(x, \mathcal{C}_2)\,\mathrm{d}x}{\text{facteur de normalisation}} \tag{A.26}$$

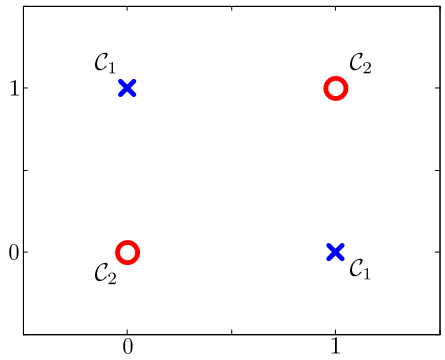
Cette probabilité d'erreur indique alors la pertinence de la variable x telle que, plus cette probabilité est faible, plus les classes sont séparées et plus la variable est pertinente.

Annexe B

Exemples illustratifs de l'apprentissage du OU-exclusif

L'illustration des méthodes de classification non linéaire est souvent réalisée sur un exemple classique : le problème « OU-exclusif » (*XOR*). En effet, il est facile de voir en consultant la figure B.1 que les méthodes de séparation linéaire ne peuvent pas résoudre ce problème. Cette annexe propose donc des exemples numériques pour résoudre ce problème, par les principales méthodes de classification vues dans ce manuscrit. Les démonstrations présentées pour les réseaux de neurones et les SVM peuvent être complétées par les ouvrages de [Abdi, 1994; Haykin, 1999; Rennard, 2006; Theodoridis and Koutroumbas, 2006].

Les quatre observations sont représentées par deux variables x_1 et x_2, composantes des vecteurs d'entrée $\mathbf{x}_i$.



x_1	x_2	*XOR*	classe
0	0	0	$\mathcal{C}_2$
0	1	1	$\mathcal{C}_1$
1	0	1	$\mathcal{C}_1$
1	1	0	$\mathcal{C}_2$

FIG. B.1 – Illustration et table de vérité du problème *XOR*.

B.1 Résolution du problème *XOR* par des réseaux de neurones

Le perceptron élémentaire, constitué d'une unique couche de sortie, ne possède pas les capacités pour séparer des observations non linéairement séparables. Seule, cette architecture ne peut donc pas résoudre le problème *XOR*.

Cependant, [Abdi, 1994] montre qu'en recodant judicieusement la relation *XOR*, celle-ci peut être apprise par un simple perceptron. L'ajout d'une nouvelle entrée x_3 au problème, générée par la multiplication des deux entrées connues ($x_3 = x_1 \times x_2$), permet de résoudre le problème. Le tableau ci-contre reprend la table de vérité du problème *XOR*, en considérant le recodage des entrées.

x_1	x_2	x_3	*XOR*	classe
0	0	**0**	0	$\mathcal{C}_2$
0	1	**0**	1	$\mathcal{C}_1$
1	0	**0**	1	$\mathcal{C}_1$
1	1	**1**	0	$\mathcal{C}_2$

La transformation du problème amène à considérer trois entrées ($p = 3$) et donc l'architecture proposée à la figure B.2, où la sortie du réseau $y(\mathbf{x})$, s'exprime par :

$$y(\mathbf{x}) = f\left(\sum_{j=1}^{p} w_j\, x_j + w_0\right) . \qquad \text{(B.1)}$$

L'utilisation d'une fonction d'activation f de type seuil, telle que :

$$f(a) = \begin{cases} 0 & \text{si } a < 0 \\ 1 & \text{si } a \geq 0\,, \end{cases} \qquad \text{(B.2)}$$

et en considérant les poids $w_1 = 1$, $w_2 = 1$, $w_3 = -2$ et le biais $w_0 = -0,5$, le problème *XOR* peut alors être résolu.

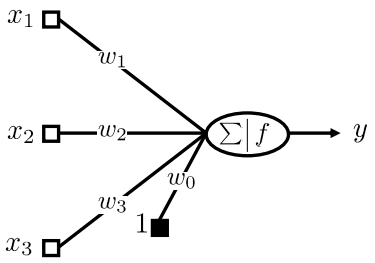


FIG. B.2 – Architecture du perceptron pour résoudre le problème *XOR*.

L'expression de la sortie du réseau est donc :

$$y(\mathbf{x}) = f(x_1 + x_2 - 2x_3 - 0,5)\,. \qquad \text{(B.3)}$$

Cette approche montre qu'un problème non linéairement séparable peut, grâce à un recodage approprié, être transformé en un problème linéairement séparable [Abdi, 1994]. Dans notre cas, la troisième variable x_3 construite à partir des deux autres, a permis de résoudre ce problème, grâce à la bonne connaissance de ce dernier. En pratique, il est rare qu'une connaissance des variables soit autant approfondie, ce qui rend cette approche difficilement généralisable. C'est pourquoi, il est souvent préférable d'utiliser des méthodes plus adaptées pour traiter des problèmes non linéairement séparables. Prenons le cas d'un perceptron multicouches, à une couche cachée de neurones (voir figure B.3), où comme précédemment, la fonction d'activation de tous les neurones est de type seuil (B.2).

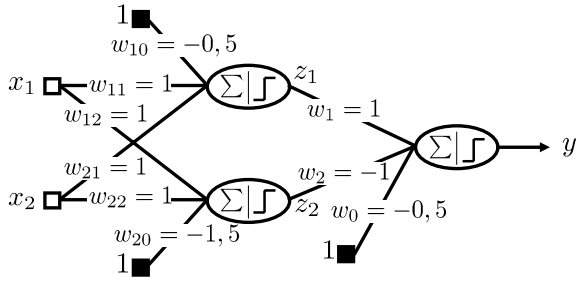


FIG. B.3 – Architecture du perceptron multicouches pour résoudre le problème *XOR*.

La première phase du réseau transforme le problème, afin de le rendre linéairement séparable. Cette phase est le résultat des sorties des neurones cachés $z_i = f(a_i(\mathbf{x}))$, avec $i = 1, 2$. En (B.4) et (B.5), nous retrouvons leur forme analytique liant chaque sortie des neurones cachés aux deux entrées ($p = 2$).

$$z_1(\mathbf{x}) = f\left(\sum_{j=1}^{p} w_{1j}^{(1)}\, x_j + w_{10}^{(1)}\right) = f(x_1 + x_2 - 0,5) \quad \text{(B.4)}$$

$$z_2(\mathbf{x}) = f\left(\sum_{j=1}^{p} w_{2j}^{(1)}\, x_j + w_{20}^{(1)}\right) = f(x_1 + x_2 - 1,5) \quad \text{(B.5)}$$

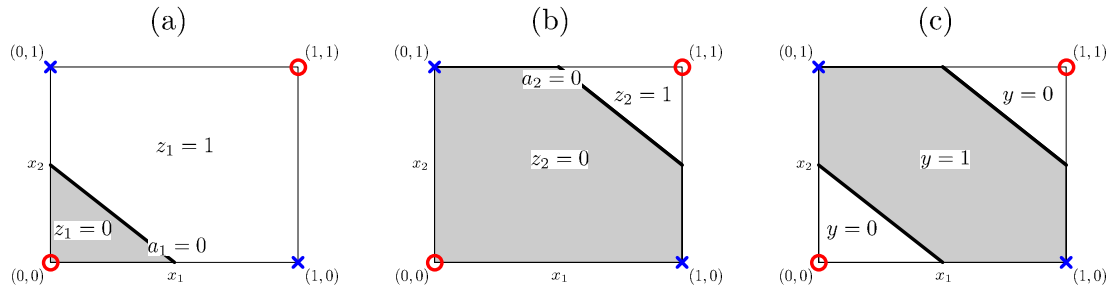
La valeur du neurone de sortie, correspondant au résultat de la fonction *XOR*, est donnée par :

$$\begin{aligned} y(\mathbf{x}) &= f\left(\sum_{i=1}^{M} w_i^{(2)}\, f\left(\sum_{j=1}^{p} w_{ij}^{(1)}\, x_j + w_{i0}^{(1)}\right) + w_0^{(2)}\right), \\ &= f\left(f\left(x_1 + x_2 - 0,5\right) + f\left(x_1 + x_2 - 1,5\right)\right), \\ &= f\left(z_1 - z_2 - 0,5\right), \end{aligned} \quad \text{(B.6)}$$

où conformément à la relation générale (1.57) donnée à la page 43, M indique le nombre de neurones dans la couche cachée, dans notre exemple $M = 2$.

À partir des relations précédentes, nous pouvons déduire naturellement les hyperplans discriminants :

- $a_1(\mathbf{x}) = x_1 + x_2 - 0,5 = 0$;
- $a_2(\mathbf{x}) = x_1 + x_2 - 1,5 = 0$;
- $a(\mathbf{x}) = z_1 - z_2 - 0,5 = 0$.



Note : (a) Frontière de décision pour le neurone dont la sortie est notée z_1. (b) Frontière de décision pour le neurone dont la sortie est notée z_2. (c) Frontière de décision pour le neurone dont la sortie est notée y.

FIG. B.4 – Décomposition des frontières de décision associées à chaque neurone du perceptron multicouches (voir figure B.3) pour résoudre le problème *XOR*.

Le tableau B.1 récapitule les valeurs pouvant être prises par les sorties de ces deux neurones, en fonction des différentes observations d'entrée.

entrées		1re phase				2-ième phase	
x_1	x_2	z_1	a_1	z_2	a_2	y	a
0	0	0	$(-0,5)^\dagger$	0	$(-1,5)^\dagger$	0	$(-0,5)^\dagger$
0	1	1	$(0,5)^\dagger$	0	$(-0,5)^\dagger$	1	$(0,5)^\dagger$
1	0	1	$(0,5)^\dagger$	0	$(-0,5)^\dagger$	1	$(0,5)^\dagger$
1	1	1	$(1,5)^\dagger$	1	$(0,5)^\dagger$	0	$(-0,5)^\dagger$

Note : † donne la valeur de la sortie du neurone sans considérer la fonction d'activation.

TAB. B.1 – Récapitulatif des différents potentiels associés à chaque neurone du perceptron multicouches (voir figure B.3), en fonction des observations d'entrée du problème *XOR*.

L'exemple que nous venons de voir décompose et transforme le problème, afin de le rendre linéairement séparable. Une autre solution, proposée par [Abdi, 1994], est de reprendre la résolution par le perceptron (voir figure B.2) et générer automatiquement la variable construite x_3 par un neurone dans la couche cachée. La figure B.5 illustre cette adaptation, où les poids du neurone de sortie sont équivalents à ceux du perceptron de la figure B.2.

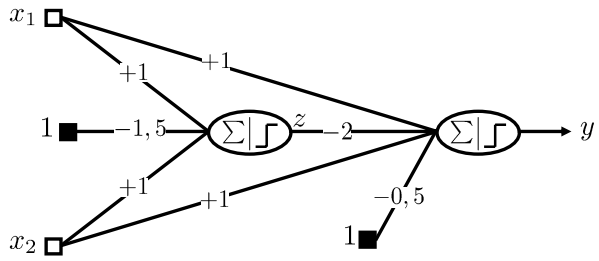


FIG. B.5 – Architecture du perceptron multicouches fondé sur le perceptron (voir figure B.2) pour résoudre le problème *XOR*.

Cette nouvelle architecture permet d'obtenir la relation du neurone de sortie :

$$y(\mathbf{x}) = f\left(x_1 + x_2 - 2z - 0,5\right) , \tag{B.7}$$

avec $z = f\left(x_1 + x_2 - 1,5\right)$. Le tableau B.2 récapitule les valeurs pouvant être prises par les sorties de ces neurones, en fonction des différentes observations d'entrée.

x_1	x_2	z		y	
0	0	0	$(-1,5)^\dagger$	0	$(-0,5)^\dagger$
0	1	0	$(-0,5)^\dagger$	1	$(0,5)^\dagger$
1	0	0	$(-0,5)^\dagger$	1	$(0,5)^\dagger$
1	1	1	$(0,5)^\dagger$	0	$(-0,5)^\dagger$

Note : † donne la valeur de la sortie du neurone sans considérer la fonction d'activation.

TAB. B.2 – Récapitulatif des différents potentiels associés à chaque neurone du perceptron multicouches (voir figure B.5), en fonction des observations d'entrée du problème *XOR*.

B.2 Résolution du problème *XOR* par un réseau RBF

Comme évoqué à la section 1.4.2.6, les réseaux RBF permettent de déformer l'espace afin de rendre linéairement séparable un problème qui ne l'est pas initialement. Ainsi, [Haykin, 1999; Rennard, 2006] proposent une résolution subtile du problème *XOR* en utilisant un réseau RBF avec deux neurones cachés, comme celui donné à la figure B.6.

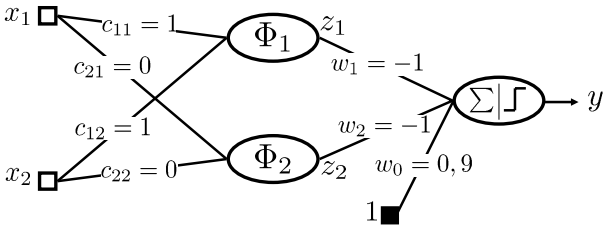


FIG. B.6 – Architecture du réseau RBF pour résoudre le problème *XOR*.

L'activation des deux neurones de la couche cachée est obtenue par les fonctions noyaux suivantes :

$$\Phi_i(\mathbf{x}) = \exp\left(-||\mathbf{x} - \mathbf{c}_i||^2\right) \ , \ i = 1, 2\,, \tag{B.8}$$

avec les centres $\mathbf{c}_1 = (1,1)^T$ et $\mathbf{c}_2 = (0,0)^T$. Les deux fonctions Φ_1 et Φ_2 transforment alors l'espace d'entrée en un nouvel espace, dans lequel le problème *XOR* devient désormais linéairement séparable, comme le montre la figure B.7.

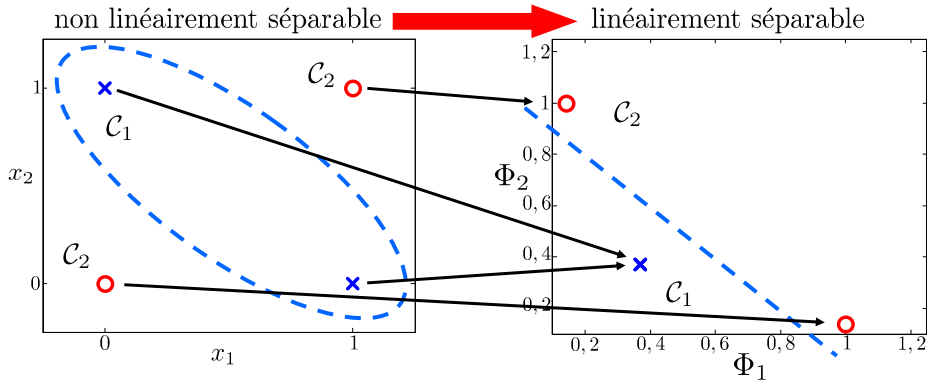


FIG. B.7 – Transformation de l'espace d'entrée par le réseaux RBF pour rendre le problème de discrimination linéaire et résoudre le problème *XOR*.

La sortie du réseau $y(\mathbf{x})$ est donnée par la relation :

$$y(\mathbf{x}) = f\left(\sum_i \left(w_i \exp\left(-||\mathbf{x} - \mathbf{c}_i||^2\right)\right) + w_0\right)\,, \tag{B.9}$$

où la fonction d'activation du neurone de sortie est une nouvelle fois la fonction seuil (B.2). Le tableau B.3 récapitule les valeurs pouvant être prises par les sorties des neurones, en fonction des différentes observations d'entrée.

x_1	x_2	$z_1 = \Phi_1(\mathbf{x})$	$z_2 = \Phi_2(\mathbf{x})$	y	
0	0	0,14	1	0	$(-0,24)^\dagger$
0	1	0,37	0,37	1	$(0,16)^\dagger$
1	0	0,37	0,37	1	$(0,16)^\dagger$
1	1	1	0,14	0	$(-0,24)^\dagger$

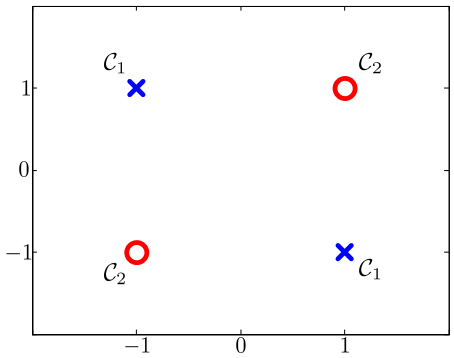
Note : † donne la valeur de la sortie du neurone sans considérer la fonction d'activation.

TAB. B.3 – Récapitulatif des différents potentiels associés à chaque neurone du réseau RBF (voir figure B.6), en fonction des observations d'entrée du problème *XOR*.

B.3 Résolution du problème *XOR* par les SVM

Dans un premier temps, nous pouvons normaliser les observations (voir figure B.8), afin d'en faciliter leur exploitation[1].

[1]La normalisation va apporter dans ce cas particulier, une simplification des calculs en évitant le calcul du bias, dû à la symétrie parfaite des observations.



x_1	x_2	t	classe
-1	-1	-1	$\mathcal{C}_2$
-1	1	$+1$	$\mathcal{C}_1$
1	-1	$+1$	$\mathcal{C}_1$
1	1	-1	$\mathcal{C}_2$

FIG. B.8 – Normalisation du problème *XOR* pour les SVM (illustration et table de vérité.

La transformation des observations du problème est réalisée par un noyau polynomial d'ordre 2 : $K(\mathbf{x},\mathbf{x}') = (1+\mathbf{x}^T\mathbf{x}')^2$. Les nouvelles caractéristiques provenant de la transformation Φ sont obtenues par :

$$\begin{aligned}\mathbf{K}(\mathbf{x},\mathbf{x}') &= (1+\mathbf{x}^T\mathbf{x}')^2 = (1+x_1x'_1+x_2x'_2)^2\,,\\ &= 1+x_1^2x_1'^2+x_2^2x_2'^2+2x_1x_1'x_2x_2'+2x_1x_1'+2x_2x_2'\,,\\ &= \left(1,x_1^2,x_2^2,\sqrt{2}x_1x_2,\sqrt{2}x_1\sqrt{2}x_2\right)^T\left(1,x_1'^2,x_2'^2,\sqrt{2}x_1'x_2',\sqrt{2}x_1'\sqrt{2}x_2'\right)\,,\\ &= \Phi(\mathbf{x})^T\Phi(\mathbf{x}')\,.\end{aligned} \tag{B.10}$$

Ainsi, la projection Φ, correspondant à $(1,x_1^2,x_2^2,\sqrt{2}x_1x_2,\sqrt{2}x_1\sqrt{2}x_2)$, indique la transformation dans un espace de 6 dimensions, les observations initialement représentées en deux dimensions. Nous avons vu que l'optimalité de l'hyperplan par les SVM, donc l'obtention des paramètres de l'hyperplan, est obtenue en introduisant les multiplicateurs de Lagrange dans la formulation du problème. Ainsi, en reprenant la formulation duale (1.76), redonnée ci-dessous (B.11), il suffit de trouver les multiplicateurs de Lagrange α_i, tels que :

$$\begin{cases}\max\limits_{\alpha}\left\{\sum\limits_{i=1}^{n}\alpha_i-\frac{1}{2}\sum\limits_{i,j=1}^{n}\alpha_i\alpha_jt_it_jK(\mathbf{x}_i,\mathbf{x}_j)\right\}\,,\\ \alpha_i\geq 0, i=1,\ldots,n\,,\\ \sum\limits_{i=1}^{n}\alpha_it_i=0\,.\end{cases} \tag{B.11}$$

Ce qui conduit à la fonction suivante :

$$\begin{aligned}\mathcal{L}(\alpha) = \quad &\alpha_1+\alpha_2+\alpha_3+\alpha_4-\tfrac{1}{2}(9\alpha_1^2-2\alpha_1\alpha_2-2\alpha_1\alpha_3+2\alpha_1\alpha_4\\ &+9\alpha_2^2+2\alpha_2\alpha_3-2\alpha_2\alpha_4+9\alpha_3^2-2\alpha_3\alpha_4+9\alpha_4^2)\,,\end{aligned} \tag{B.12}$$

avec les contraintes suivantes : $\alpha_i\geq 0$ $(i=1,\ldots,4)$ et $-\alpha_1+\alpha_2+\alpha_3-\alpha_4=0$. L'optimisation de $\mathcal{L}(\alpha)$ conduit alors au système d'équations suivant :

$$\begin{cases}9\alpha_1-\alpha_2-\alpha_3+\alpha_4 &= 1\,,\\ -\alpha_1+9\alpha_2+\alpha_3-\alpha_4 &= 1\,,\\ -\alpha_1+\alpha_2+9\alpha_3-\alpha_4 &= 1\,,\\ \alpha_1-\alpha_2-\alpha_3+9\alpha_4 &= 1\,.\end{cases} \tag{B.13}$$

La résolution de ce système d'équation, maximisant l'expression (B.11), permet d'obtenir les valeurs optimales des multiplicateurs de Lagrange : dans notre problème, $\alpha_i = \frac{1}{8}$, pour $i = 1, \ldots, 4$. De plus, comme tous les $\alpha_i > 0$, alors les quatre observations sont des vecteurs de support. Sachant que $\mathcal{L}(\alpha) = \frac{1}{4}$, nous permet de déduire que la marge $||\mathbf{w}||$ est égale à $\frac{1}{\sqrt{2}}$.

Aussi, la relation (1.77) de l'hyperplan dans le cadre des SVM, donnée à la page 56, permet de déduire le vecteur des paramètres :

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i t_i \Phi(\mathbf{x_i}) \,, \tag{B.14}$$

qui prend comme valeur :

$$\mathbf{w} = \frac{1}{8}\left[-\begin{pmatrix} 1 \\ 1 \\ 1 \\ \sqrt{2} \\ -\sqrt{2} \\ -\sqrt{2} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ -\sqrt{2} \\ -\sqrt{2} \\ \sqrt{2} \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 1 \\ -\sqrt{2} \\ \sqrt{2} \\ -\sqrt{2} \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \\ 1 \\ \sqrt{2} \\ \sqrt{2} \\ \sqrt{2} \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -\sqrt{2}/2 \\ 0 \\ 0 \end{pmatrix} . \tag{B.15}$$

Le vecteur des paramètres se simplifie alors par : $\mathbf{w}^T = \begin{pmatrix} 0 & 0 & 0 & -\sqrt{2}/2 & 0 & 0 \end{pmatrix}$. Le premier élément de ce vecteur indique que le biais est nul. Enfin, la fonction discriminante $y(\mathbf{x})$ est de la forme :

$$y(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x}) = \begin{pmatrix} 0 & 0 & 0 & -\sqrt{2}/2 & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \end{pmatrix} , \tag{B.16}$$

et se simplifie par : $y(\mathbf{x}) = -x_1x_2$. Ainsi, comme le montre la figure B.9, si :

$$\begin{cases} y(\mathbf{x}) > 0 \quad \text{alors} \quad \mathbf{x} \in \mathcal{C}_1 \,, \\ y(\mathbf{x}) < 0 \quad \text{alors} \quad \mathbf{x} \in \mathcal{C}_2 \,. \end{cases} \tag{B.17}$$

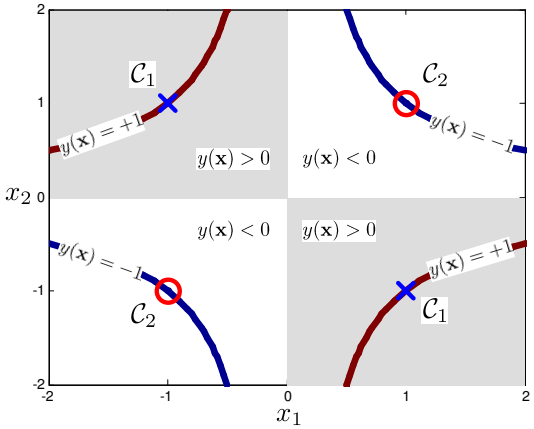


FIG. B.9 – Frontières de décision obtenues par les SVM pour résoudre le problème *XOR*.

Annexe C

Illustration de la pertinence des indices mesurant la performance d'un modèle

Dans ce manuscrit, nous avons insisté fréquemment sur l'importance de l'évaluation des modèles, et notamment sur l'influence de l'indice de mesure utilisé. Dans le chapitre 3, plusieurs indices ont été décrits, où il est apparu que certains étaient fortement influencés par le déséquilibre entre les classes, tels que le taux de classification ou d'erreur de classification. Cette annexe démontre cette dépendance, en comparant les principaux indices de mesure au travers de deux exemples. Aussi, dans le chapitre 3, les courbes de ROC ont été présentées comme une mesure de performance relativement efficace ; c'est pourquoi, cette annexe propose également une illustration de la méthodologie pour construire la courbe de ROC.

C.1 Influence du déséquilibre entre les classes : prévalence de la maladie

Un déséquilibre dans la représentation des classes d'un ensemble d'observations peut biaiser fortement l'interprétation des résultats, si les indices et les méthodes d'évaluation n'ont pas été judicieusement choisis. Il n'est pas rare de trouver ce type de configuration dans certains problèmes. En effet, l'étude de pathologies rares implique, par le fait, un petit échantillon de patients atteints par ce type de pathologies. Dès lors, comparer cet échantillon à des patients « sains » ou atteints d'autres pathologies plus courantes entraîne un déséquilibre entre le nombre de patients de chaque groupe. Pour illustrer ces propos, nous proposons au tableau C.1 un exemple comparant deux études, où les répartitions des classes sont différentes : équilibrées (prévalence de la maladie de 50%) et déséquilibrées (prévalence de la maladie de 10%).

A partir des deux matrices de confusion du tableau C.1, plusieurs indices sont récapitulés dans ce même tableau. Ainsi, la comparaison de ces indices nous montre que pour un même taux de classification de 90%, nous pouvons observer une grande disparité dans la valeur de la sensibilité, de la spécificité et des valeurs prédictives positive et négative. En effet, sans l'observation de la sensibilité et de la spécificité, le test, dans le cas de l'échantillon déséquilibré, aurait paru aussi pertinent que l'autre (échantillon équilibré), or leur performance diverge totalement. Cela confirme le biais que peut apporter l'utilisation unique du taux de classification sur l'interprétation des résultats.

classes équilibrées

	positive	négative
positive	47	7
négative	3	43
total	50	50

prévalence de la maladie : 50%

classes déséquilibrées

	positive	négative
positive	2	2
négative	8	88
total	10	90

prévalence de la maladie : 10%

synthèse des résultats et calculs des indices

	classes équilibrées	classes déséquilibrées
taux de classification	90%	90%
sensibilité (S_e)	94%	20%
spécificité (S_p)	86%	98%
valeur prédictive positive (VPP)	87%	50%
valeur prédictive négative VPN	93%	92%
rapport de vraisemblance positif (RV^+)	6, 71	10
rapport de vraisemblance négatif (RV^-)	0, 07	0, 82

TAB. C.1 – Évaluation de l'influence du déséquilibre entre les classes (prévalence de la maladie) sur les indices de performances usuels.

L'autre exemple présenté dans le tableau C.2 provient de [Huguier and Flahault, 2003]. Il montre l'influence et la dépendance de la prévalence de la maladie aux valeurs prédictives. Dans cet exemple, les deux études comparées possèdent un taux de classification, une sensibilité et une spécificité égales à 90%.

La figure C.1 montre l'évolution des valeurs prédictives positive et négative en fonction de la prévalence de la maladie dans l'échantillon, donc en fonction du déséquilibre des classes. Pour cette simulation, la sensibilité et la spécificité sont égales à 90%. Ainsi, si la prévalence augmente on observe alors une augmentation de VPP et une diminution de VPN, montrant par conséquent la dépendance des valeurs prédictives à la prévalence. De ce fait, les indices VPP et VPN peuvent être utilisés, si la prévalence de la maladie dans l'échantillon correspond à la prévalence réelle de la maladie dans la population.

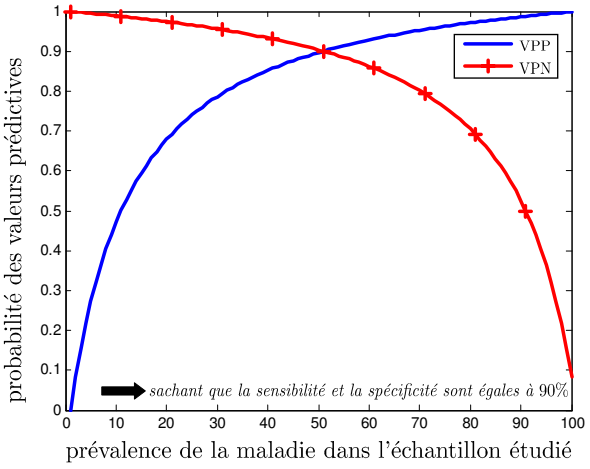


FIG. C.1 – Évolution des valeurs prédictives positive et négative en fonction de la prévalence de la maladie dans l'échantillon.

exemple 1

	positive	négative
positive	45	10
négative	5	90
total	50	100

prévalence de la maladie : 33%

exemple 2

	positive	négative
positive	18	20
négative	2	180
total	20	200

prévalence de la maladie : 11%

synthèse des résultats et calculs des indices

	exemple 1	exemple 2
taux de classification	90%	90%
sensibilité (S_e)	90%	90%
spécificité (S_p)	90%	90%
valeur prédictive positive (VPP)	87%	47%
valeur prédictive négative (VPN)	93%	99%
rapport de vraisemblance positif (RV^+)	9	9
rapport de vraisemblance négatif (RV^-)	$0,11$	$0,11$

TAB. C.2 – Évaluation de l'influence du déséquilibre entre les classes (prévalence de la maladie) sur les valeurs prédictives, indépendamment des indices usuels.

C.2 Construction de la courbe de ROC

Nous proposons dans cette section un exemple de construction de la courbe de ROC. L'exemple reprend la prédiction de l'état du patient en observant sa température corporelle (T °C), à l'image de ce qui a été donné à la section 3.4.1. Nous considérons donc deux classes, patients malades et sains, contenant respectivement 50 et 100 patients. Le modèle de classification est basé sur une approche linéaire qui considère uniquement la variable température.

Chaque point de la courbe de ROC est obtenu en déterminant la matrice de confusion, une fois un seuil de décision établi. Ce seuil, en l'occurrence une valeur de température, permet au modèle de distinguer les patients sains des patients malades. Dès lors, la matrice de confusion permet d'obtenir les quatre variables VP, FN, VN et FP, avec lesquelles nous calculons la sensibilité et la spécificité du modèle pour le seuil de classification choisi. Cette opération est réitérée pour différentes valeurs de seuils, comme montré à la figure C.2 où trois cas sont détaillés.

Notons que le seuil idéal qui maximise la prédiction est donné pour une température de $37,5\,°C$, avec lequel, le meilleur compromis entre la sensibilité et la spécificité est obtenu. Par ailleurs, comme le montre la figure C.2, ce seuil amène à minimiser la probabilité d'erreur.

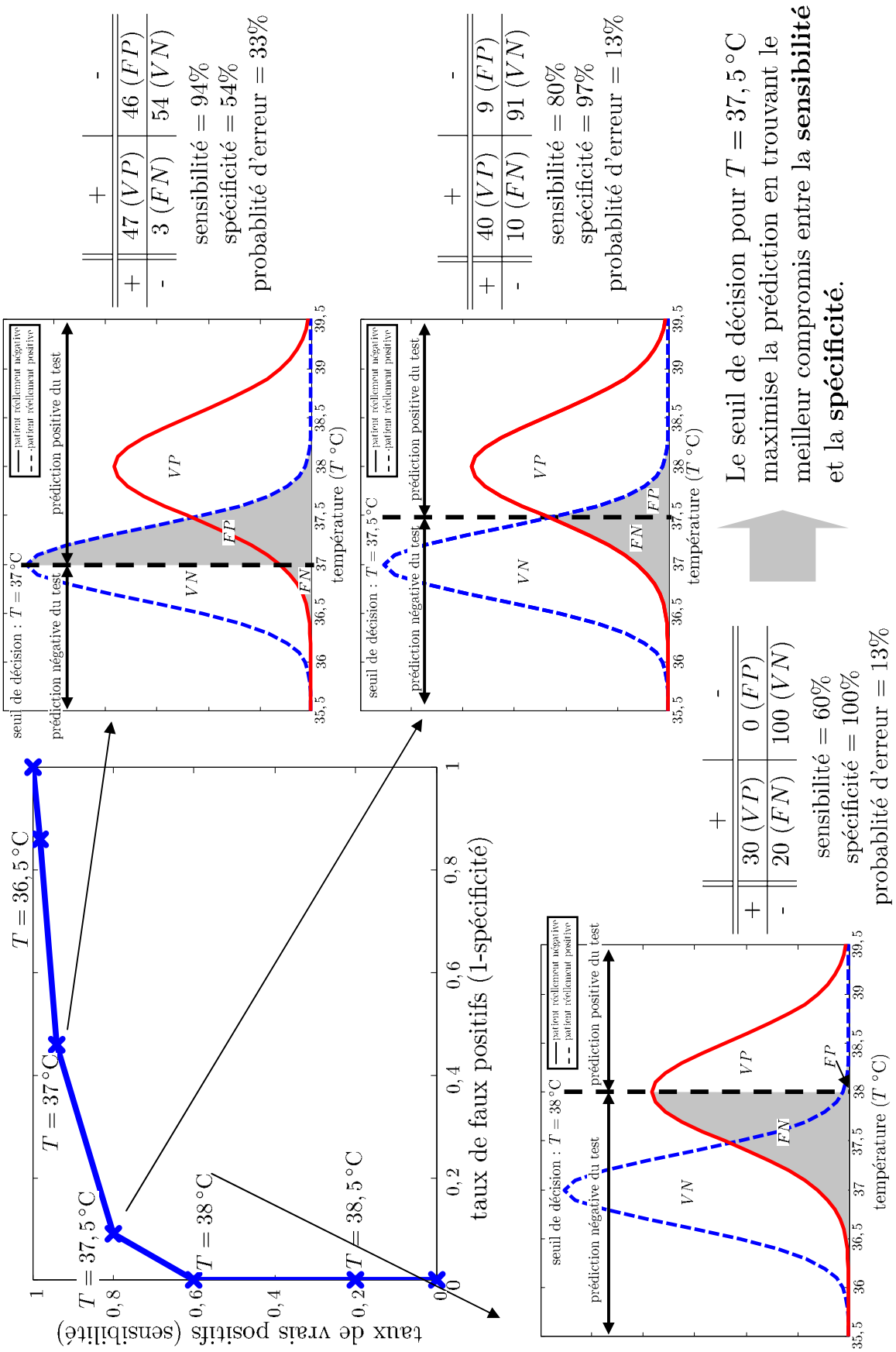


FIG. C.2 – Exemple de construction de la courbe de ROC.

Annexe D

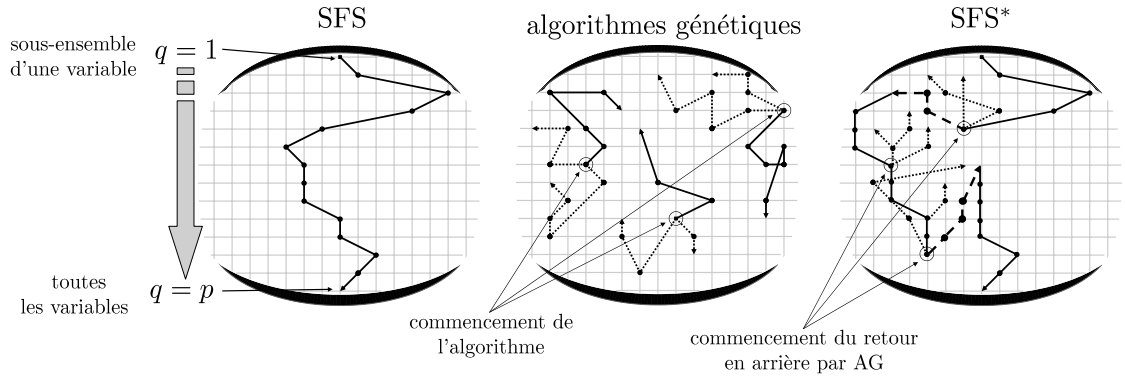
Méthodes de sélection de variables combinant sélection séquentielle et algorithmes génétiques

Dans le cadre de la prédiction de la syncope lors de la phase basculée du *tilt-test* (*cf.* section 5.4), nous avons présenté une nouvelle approche pour la sélection de variables. Les méthodes liées à cette nouvelle approche, que nous avons appelé « méthodes combinées », sont fondées sur une combinaison d'un processus de sélection séquentielle (SFS) et des algorithmes génétiques (AG). Cette approche avait permis d'obtenir de très bonnes performances pour la classification de patients sujets à l'apparition de la syncope, tout en réduisant le nombre de variables sélectionnées. Ces bons résultats ont été très favorablement comparés à ceux des méthodes plus classiques (SFS_{Fisher}, SFS_{RELIEF}, SFS, SBS, LRS, SFFS, SFBS, RGSS et AG), avec l'intérêt d'avoir nécessité l'évaluation de moins de combinaisons de variables (*cf.* section 5.4.4, page 155).

Dans cette annexe, nous proposons d'évaluer notre approche pour la sélection de variables sur d'autres ensembles de données. Le principal challenge est d'extraire le plus petit nombre de variables pertinentes optimisant la classification.

D.1 Rappel de l'approche pour la sélection de variables

Notre méthode de sélection de variables, évoquée à la section 5.4.2.2 et détaillée ci-dessous, a été développée dans un but d'améliorer l'exploration de l'espace des combinaisons de variables, en réduisant le coût calculatoire. Cette approche repose simplement sur une combinaison d'un processus de sélection séquentielle et des algorithmes génétiques. Cette combinaison permet d'associer la rapidité du processus SFS et l'efficacité des AG. En effet, à chaque itération, la méthode SFS ajoute au sous-ensemble, la variable optimisant la classification ; cette méthode voit ainsi son processus converger assez rapidement, en évaluant $p \cdot (p+1)/2$ combinaisons, où p désigne le nombre de variables initiales. Cependant, par ce processus, la recherche apparaît très locale, ne permettant pas d'évaluer l'ajout de plusieurs variables simultanément. Le résultat final de ce processus de sélection est alors dépendant des premières itérations et de la première variable sélectionnée. L'efficacité des AG est, quant à elle, obtenue par une recherche globale qui, contrairement à SFS et à l'ensemble des méthodes séquentielles, évalue simultanément l'association de plusieurs variables. Cependant, les résultats des AG peuvent être biaisés par l'initialisation de l'algorithme, contrairement à la méthode séquentielle. Ainsi, la combinaison de SFS et des AG permet d'associer simultanément les recherches locale et globale dans le nouveau processus de sélection. La figure D.1 illustre cette combinaison en donnant un exemple de progression dans l'espace des variables.



Note : Sur les trois graphiques, chaque intersection de la q-ième ligne correspond une combinaison de q variables sélectionnées.

FIG. D.1 – Comparaison de la progression dans l'espace des variables de la méthode de sélection de variables combinant la sélection séquentielle ascendante et les algorithmes génétiques.

Fondée sur la méthode SFS, notre approche utilise des AG pour optimiser à des moments aléatoires, les sous-ensembles de variables obtenus durant le processus de sélection SFS. En d'autres termes, nous avons ajouté des facultés de retours en arrière « efficaces » au processus de sélection séquentielle le plus classique. L'algorithme D.1 montre le déroulement du nouveau processus de sélection basé sur SFS, nommé **SFS***. Notons qu'une version de cet algorithme est réalisée en utilisant le processus naïf de sélection séquentielle. Certaines méthodes séquentielles fondées elles aussi sur SFS, possèdent des capacités de retours en arrière, comme SFFS, SFBS et LRS (*cf.* section 2.4.3.3). Cependant, pour ces méthodes, les retours effectués de manière séquentielle,

Algorithme D.1 : Pseudo-code de l'algorithme de sélection de variables **SFS***, fondé sur une combinaison des algorithmes séquentiels et évolutionnaires.

Données :
$\langle FS \rangle = \{x_1, \ldots, x_p\}$: ensemble des variables initiales
i_{max} : nombre maximum d'itérations
Résultat : $\langle SS_{optimal} \rangle$: sous-ensemble de variables sélectionnées

début
  $\langle SS_0 \rangle \leftarrow \emptyset$;
  $i \leftarrow 0$;
  tant que $i < i_{max}$ **faire**
    $i \leftarrow i + 1$;
    $x_+ \leftarrow \underset{x_j \in \{\langle FS \rangle - \langle SS_{i-1} \rangle\}}{\text{argmax}} \{J(\langle SS_{i-1} \rangle \cup \{x_j\})\}$;
    $\langle SS_i \rangle \leftarrow \langle SS_{i-1} \rangle \cup \{x_+\}$;
    si valeur_aléatoire() $\leq 0,1$ **alors**
      /* valeur_aléatoire() renvoie une valeur aléatoire dans l'intervalle $[0\,;1]$ suivant une loi uniforme */
      $\langle SS_i \rangle \leftarrow$ Algorithme_Génétique($\langle SS_i \rangle$) ;
      /* Algorithme_Génétique() renvoie un sous-ensemble de variables optimisé par les algorithmes génétiques */
    fin
  fin
fin
$\langle SS_{optimal} \rangle \leftarrow \underset{\forall \langle SS_j \rangle, j=1,\ldots,i_{max}}{\text{argmax}} \{J(\langle SS_j \rangle)\}$;

nuisent à la souplesse et à l'efficacité du processus de sélection. En effet, les retours en arrière de ces méthodes, sélectionne une seule variable à la fois, parmi celles préalablement sélectionnées, et l'élimine du sous-ensemble de variables ; nous retrouvons par ce processus une notion d'optimisation locale. Le retour en arrière, réalisé par les AG, améliore le processus de sélection, en permettant d'optimiser le sous-ensemble de manière globale. C'est-à-dire, qu'il ne considère plus uniquement une seule variable à éliminer, mais une combinaison de plusieurs variables.

D.2 Cadres expérimental et méthodologique

Les ensembles de données présentés dans cette annexe proviennent de *UCI Machine Learning Repository* [Newman *et al.*, 1998], dans lesquels les variables continues sont uniquement conservées[1]. Aussi, comme le montre le tableau D.1 qui décrit ces ensembles, excepté l'ensemble *Arrhythmia*, les six autres ne considèrent que deux classes.

dénomination	**nombre d'observations**	**nombre de variables**	**nombre de classes**
Sonar	208	60	2
Ionosphere	351	33	2
Spectf	267	44	2
Wpbc	198	33	2
Wdbc	569	30	2
Musk Clean1	476	166	2
Arrhythmia	452	206	16

TAB. D.1 – Description des ensembles de données de *UCI* utilisés, pour évaluer les performances des méthodes de sélection de variables combinées aux AG.

Ces ensembles de données n'ont pas subi de prétraitement particulier (détection et traitement de valeurs aberrantes) avant la phase de sélection des sous-ensembles de variables. Les méthodes employées pour réaliser la sélection sont les mêmes que celles utilisées dans la section 5.4, nous pouvons les regrouper en six catégories, dont la dernière fait référence à nos nouvelles méthodes :

- sans processus de sélection : « pas de sélection » et « aléatoire » ;
- sélection de variables par des processus séquentiels naïfs : SFS_{Fisher} et SFS_{RELIEF} ;
- sélection de variables par des processus séquentiels : SFS et SBS ;
- sélection de variables par des processus séquentiels avec retours en arrière : LRS, SFFS et SFBS ;
- sélection de variables par des processus non déterministes : RGSS et AG ;
- sélection de variables par des processus séquentiels avec retours en arrière par algorithmes génétiques : $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$.

[1] L'ensemble de données *Arrhythmia* est le seul concerné par ce pré-traitement, où 73 variables des 279 initiales, ont été éliminées.

Ces méthodes ont été détaillées aux sections 2.4 et 5.4.2.2, comme leurs modalités d'utilisation expérimentale, décrites à la section 5.4.2.1. Rappelons néanmoins que la méthode LRS est utilisée avec $L = 3$, $R = 2$ et $L = 2$, $R = 3$; de manière à réaliser respectivement une sélection ascendante et descendante. Pour les algorithmes génétiques, la taille de la population est toujours de 80 individus, la probabilité de mutation est de $0,05$ et la sélection des individus pour la reproduction est réalisée par tournoi. La fonction d'adaptation est donnée par la relation suivante :

$$J = taux\_de\_classification + 0,01 \times n\_bits\_à\_0\,, \tag{D.1}$$

où $taux\_de\_classification$ définit la moyenne du taux de classification sur les échantillons de validation et $n\_bits\_à\_0$ est le nombre de variables ignorées. L'algorithme est arrêté au bout de 500 générations.

Compte tenu du grand nombre de variables pour certains ensembles de données (*Musk Clean1* et *Arrhythmia*, possédant respectivement 166 et 206 variables), les méthodes combinées vont s'exécuter sur 100 itérations ; précédemment sur la problématique de la syncope, le nombre d'itérations était de 200. Aussi, la probabilité d'arrêt de l'algorithme séquentiel pour l'exécution des retours en arrière est de $0,1$. La population des AG est composée de 20 individus, la probabilité de mutation est de $0,05$ et l'optimisation des AG est faite sur 20 générations.

Ces processus de sélection doivent être associés à un outil de classification. Dans cette analyse, nous utilisons deux approches utilisées précédemment : le classifieur de Bayes naïf (BN_{gauss}) et les k-plus proches voisins (k-ppv), ces algorithmes sont décrits respectivement aux sections 5.3.1.2 et 5.4.2.3. Le choix de ces deux méthodes repose simplement sur le fait qu'elles s'exécutent relativement vite et nécessitent de définir peu de paramètres. Les performances sont évaluées par le processus utilisé tout au long de ce manuscrit, illustré à la figure 5.3 (page 134). Dès lors, l'échantillon de chaque ensemble de données est divisé en deux groupes (apprentissage/validation et test). La validation croisée mesure les performances moyennes de classification sur le groupe validation, elle est réalisée sur 10 sous-ensembles d'observations ($K = 10$).

Chacun des processus de sélection faisant appel à des procédés aléatoires (« aléatoire », RGSS, AG, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), est réalisé dix fois pour chaque ensemble de données et pour chaque classifieur, réduisant ainsi le biais de l'estimation.

D.3 Résultats expérimentaux

Les résultats obtenus pour les ensembles de données sont récapitulés dans les tableaux D.2 à D.8, respectivement pour les ensembles *Sonar*, *Ionosphere*, *Spectf*, *Wpbc*, *Wdbc*, *Musk Clean 1* et *Arrhythmia*.

Comme le montre le tableau D.8 associé aux résultats de l'ensemble *Arrhythmia*, le classifieur de Bayes Naïf n'a pas été évalué. En effet, certaines classes de cet ensemble de données étant sous-représentées, l'interprétation des résultats de ce classifieur aurait été biaisée. Rappelons que cette méthode est fondée sur l'estimation des densités de probabilité, nécessitant de ce fait, un nombre conséquent d'observations.

méthode de sélection	BNgauss			k-ppv		
	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)
évaluation sans processus de sélection						
pas de sélection	60	72,3	65,4	60	77,6	72,1
Aléatoire	32,1±12,5 (30)	68,5±2,8 (70,0)	65,3±4,2 (70,2)	32,1±16,5 (27)	78,5±2,1 (74,8)	75,7±2,9 (79,8)
sélection de variables par des processus séquentiels naïfs						
$\mathbf{SFS}_{Fisher}$	2	85,0	83,4	20	89,2	85,7
$\mathbf{SFS}_{RELIEF}$	33	78,1	32,6	6	89,4	95,4
sélection de variables par des processus séquentiels						
SFS	36	84,4	71,1	20	94,6	77,9
SBS	13	86,8	66,3	21	90,8	79,8
sélection de variables par des processus séquentiels avec retours en arrière						
$\mathbf{LRS}_{(L=3,\ R=2)}$	38	86,9	66,4	8	94,3	69,2
$\mathbf{LRS}_{(L=2,\ R=3)}$	12	87,9	76,0	6	93,9	74,0
SFFS	28	89,9	74,0	20	93,9	77,9
SFBS	14	91,1	77,9	15	92,9	75,0
sélection de variables par des processus non déterministes						
RGSS	22,4±12,3 (12)	83,0±2,7 (85,9)	69,2±1,3 (70,2)	20,4±9,2 (14)	90,2±1,5 (91,6)	75,9±3,0 (78,8)
AG	21,3±2,8 (17)	88,6±0,7 (88,9)	70,8±3,1 (76,0)	24,0±1,9 (25)	93,8±1,1 (93,6)	78,1±3,2 (84,6)
sélection de variables par des processus séquentiels avec retours en arrière par des AG						
$\mathbf{SFS}^*_{Fisher}$	11,3±5,9 (12)	83,1±0,8 (83,8)	70,8±3,3 (75,0)	4,9±3,8 (10)	81,0±6,7 (86,9)	73,5±4,1 (80,8)
$\mathbf{SFS}^*_{RELIEF}$	8,1±4,5 (11)	76,4±4,9 (86,1)	64,3±5,1 (70,2)	21,1±18,9 (11)	81,8±3,3 (85,9)	73,1±2,4 (77,9)
$\mathbf{SFS}^*$	14,5±2,7 (13)	88,9±2,2 (91,1)	74,5±1,2 (76,0)	15,8±4,3 (14)	95,5±0,9 (97,3)	78,0±2,5 (82,7)

Note : Pour les méthodes nécessitant plusieurs essais (**Aléatoire**, **RGSS**, **AG**, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), le tableau donne la moyenne et l'écart type du taux de classification et (·) indique les performances optimisant la validation.

Tab. D.2 – Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Sonar*.

méthode de sélection	BNgauss			k-ppv		
	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)
évaluation sans processus de sélection						
pas de sélection	33	78,1	32,6	33	84,4	85,1
Aléatoire	15,0±9,9 (24)	78,0±7,9 (86,9)	67,7±24,8 (89,7)	15,0±9,2 (12)	83,3±3,1 (85,5)	85,2±3,7 (87,4)
sélection de variables par des processus séquentiels naïfs						
$\mathbf{SFS}_{Fisher}$	2	76,4	68,3	13	83,1	79,8
$\mathbf{SFS}_{RELIEF}$	16	72,3	65,4	15	78,6	75,0
sélection de variables par des processus séquentiels						
SFS	17	93,2	90,3	12	94,8	88,0
SBS	9	93,7	90,3	7	94,9	88,0
sélection de variables par des processus séquentiels avec retours en arrière						
$\mathbf{LRS}_{(L=3,\ R=2)}$	13	95,0	92,0	11	96,1	86,3
$\mathbf{LRS}_{(L=2,\ R=3)}$	7	94,1	89,1	11	95,9	86,3
SFFS	12	94,6	89,7	15	95,5	87,4
SFBS	12	94,1	89,7	14	93,4	86,9
sélection de variables par des processus non déterministes						
RGSS	10,2±6,5 (8)	91,0±4,8 (94,1)	83,7±18,0 (90,9)	13,7±5,2 (9)	92,4±1,6 (95,6)	86,7±1,8 (88,6)
AG	9,5±0,8 (10)	94,9±0,2 (95,1)	91,5±1,1 (93,1)	11,4±1,1 (10)	96,1±0,2 (96,3)	85,0±0,7 (86,3)
sélection de variables par des processus séquentiels avec retours en arrière par des AG						
$\mathbf{SFS}^*_{Fisher}$	8,7±2,2 (9)	93,9±1,4 (95,4)	89,9±1,7 (91,4)	4,5±2,5 (3)	90,8±4,1 (91,9)	87,9±5,8 (91,4)
$\mathbf{SFS}^*_{RELIEF}$	3,7±3,9 (2)	78,6±0,6 (79,1)	77,2±6,8 (81,2)	1,7±0,5 (2)	76,8±3,3 (77,5)	73,8±4,8 (81,2)
$\mathbf{SFS}^*$	4±0 (4)	87,3±0 (87,3)	70,7±0 (70,7)	15,4±4,3 (19)	91,3±1,2 (92,4)	75,6±2,2 (78,9)

Note : Pour les méthodes nécessitant plusieurs essais (**Aléatoire**, **RGSS**, **AG**, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), le tableau donne la moyenne et l'écart type du taux de classification et $(\cdot)$ indique les performances optimisant la validation.

TAB. D.3 – Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Ionosphere*.

méthode de sélection	BNgauss			k-ppv		
	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)
évaluation sans processus de sélection						
pas de sélection	44	72,1	67,7	44	72,8	67,7
Aléatoire	21,9±13,3 (35)	72,0±4,7 (70,5)	65,4±3,0 (67,2)	21,9±13,3 (8)	73,3±3,2 (75,1)	72,3±1,5 (74,4)
sélection de variables par des processus séquentiels naïfs						
$\mathbf{SFS}_{Fisher}$	3	79,2	66,6	12	83,3	72,2
$\mathbf{SFS}_{RELIEF}$	1	77,5	81,2	37	75,9	69,9
sélection de variables par des processus séquentiels						
SFS	4	87,3	70,7	16	88,6	78,9
SBS	7	84,0	69,2	14	85,1	73,7
sélection de variables par des processus séquentiels avec retours en arrière						
$\mathbf{LRS}_{(L=3,\ R=2)}$	5	86,7	69,9	17	88,6	78,9
$\mathbf{LRS}_{(L=2,\ R=3)}$	5	86,7	67,1	4	87,9	73,7
SFFS	5	86,7	65,4	16	92,3	76,7
SFBS	12	86,7	69,2	15	88,8	76,7
sélection de variables par des processus non déterministes						
RGSS	7,3±5,7 (4)	82,4±2,9 (84,4)	67,6±2,9 (71,4)	15,3±4,1 (12)	86,5±1,3 (87,5)	73,3±3,3 (74,4)
AG	9,9±2,2 (12)	86,5±0,6 (86,8)	68,3±2,6 (71,4)	17,6±2,5 (13)	91,3±1,1 (91,4)	74,4±3,1 (80,4)
sélection de variables par des processus séquentiels avec retours en arrière par des AG						
$\mathbf{SFS}^*_{Fisher}$	3,9±1,7 (2)	85,2±0,8 (86,2)	71,1±2,6 (74,4)	6,1±2,1 (7)	82,9±4,5 (86,3)	75,5±2,7 (79,7)
$\mathbf{SFS}^*_{RELIEF}$	3,7±3,9 (2)	78,6±0,6 (79,2)	77,2±6,8 (81,2)	1,7±0,5 (2)	76,8±3,3 (78,5)	73,8±4,8 (81,2)
$\mathbf{SFS}^*$	4±0 (4)	87,3±0 (87,3)	70,7±0 (70,7)	15,4±4,3 (19)	91,3±1,2 (92,4)	75,6±2,2 (78,9)

Note : Pour les méthodes nécessitant plusieurs essais (**Aléatoire**, **RGSS**, **AG**, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), le tableau donne la moyenne et l'écart type du taux de classification et (·) indique les performances optimisant la validation.

Tab. D.4 – Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Spectf*.

méthode de sélection	BNgauss			*k*-ppv		
	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)
évaluation sans processus de sélection						
pas de sélection	33	63,6	72,2	33	69,1	69,1
Aléatoire	20,0±8,6 (10)	63,3±2,2 (63,1)	74,7±2,0 (76,3)	20,0±8,6 (16)	68,4±3,6 (74,1)	67,2±3,3 (71,1)
sélection de variables par des processus séquentiels naïfs						
$\mathbf{SFS}_{Fisher}$	2	76,6	73,2	10	73,1	68,0
$\mathbf{SFS}_{RELIEF}$	13	63,6	72,2	11	70,3	64,9
sélection de variables par des processus séquentiels						
SFS	7	79,6	73,2	7	78,1	70,1
SBS	5	79,6	71,1	9	79,1	72,2
sélection de variables par des processus séquentiels avec retours en arrière						
$\mathbf{LRS}_{(L=3,\,R=2)}$	3	80,6	78,4	19	82,1	67,0
$\mathbf{LRS}_{(L=2,\,R=3)}$	4	80,6	78,4	17	81,1	70,1
SFFS	3	80,6	78,4	15	82,5	69,1
SFBS	3	80,6	78,4	11	79,6	73,2
sélection de variables par des processus non déterministes						
RGSS	3,1±1,8 (3)	75,7±4,3 (80,6)	77,3±2,7 (78,4)	17,5±3,5 (14)	79,2±1,9 (82,1)	70,2±3,3 (73,2)
AG	4,4±1,3 (3)	80,9±0,6 (81,2)	75,3±2,7 (78,3)	14,4±1,6 (13)	85,3±0,7 (86,1)	70,1±2,7 (74,2)
sélection de variables par des processus séquentiels avec retours en arrière par des AG						
$\mathbf{SFS}^*_{Fisher}$	3,2±0,4 (4)	79,3±1,5 (80,6)	75,7±2,9 (78,3)	2,0±0 (2)	69,1±0 (69,1)	71,1±0 (71,1)
$\mathbf{SFS}^*_{RELIEF}$	7,1±2,5 (8)	72,5±3,3 (74,6)	75,8±1,6 (78,3)	9,0±3,0 (9)	74,1±7,0 (81,6)	67,9±4,1 (76,3)
$\mathbf{SFS}^*$	7,1±0,3 (7)	79,6±0 (79,6)	73,2±0 (73,2)	14,7±3,1 (16)	82,0±2,7 (85,1)	69,0±2,0 (72,2)

Note : Pour les méthodes nécessitant plusieurs essais (**Aléatoire**, **RGSS**, **AG**, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), le tableau donne la moyenne et l'écart type du taux de classification et (·) indique les performances optimisant la validation.

Tab. D.5 – Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Wpbc*.

méthode de sélection	BNgauss nombre de variables sélectionnées	BNgauss performance de validation (%)	BNgauss performance de test (%)	k-ppv nombre de variables sélectionnées	k-ppv performance de validation (%)	k-ppv performance de test (%)
			évaluation sans processus de sélection			
pas de sélection	30	92,9	93,0	30	95,9	95,1
Aléatoire	16,7±8,8 (17)	92,8±1,1 (93,7)	94,2±1,3 (97,2)	16,7±8,8 (28)	94,2±2,0 (95,2)	92,2±2,1 (95,1)
			sélection de variables par des processus séquentiels naïfs			
$\mathbf{SFS}_{Fisher}$	21	93,3	94,0	20	96,2	95,1
$\mathbf{SFS}_{RELIEF}$	29	93,6	92,6	28	95,9	94,0
			sélection de variables par des processus séquentiels			
SFS	11	97,2	97,2	13	98,6	94,0
SBS	6	96,2	96,5	6	96,2	96,5
			sélection de variables par des processus séquentiels avec retours en arrière			
$\mathbf{LRS}_{(L=3,\ R=2)}$	9	97,9	95,4	9	98,6	91,2
$\mathbf{LRS}_{(L=2,\ R=3)}$	9	97,6	95,4	10	99,0	91,6
SFFS	9	97,2	96,5	14	99,0	92,3
SFBS	11	97,6	96,5	17	97,9	94,4
			sélection de variables par des processus non déterministes			
RGSS	12±5,1 (8)	95,6±1,0 (96,6)	95,7±0,7 (97,2)	14,2±4,5 (21)	97,6±0,3 (98,3)	92,8±1,5 (94,7)
AG	8,1±1,6 (10)	97,9±0,3 (98,1)	96,8±0,3 (97,5)	9,7±2,2 (8)	99,0±0,2 (99,2)	91,9±1,3 (94,4)
			sélection de variables par des processus séquentiels avec retours en arrière par des AG			
$\mathbf{SFS}^*_{Fisher}$	6,7±5,1 (6)	96,5±1,3 (96,9)	95,9±1,0 (97,5)	10,1±5,9 (21)	96,4±1,2 (97,4)	92,8±1,8 (95,1)
$\mathbf{SFS}^*_{RELIEF}$	11,9±9,2 (8)	96,0±1,6 (97,2)	95,2±1,7 (97,2)	20,8±10,1 (25)	96,0±0,8 (96,9)	90,0±6,4 (93,3)
$\mathbf{SFS}^*$	9,8±1,5 (8)	97,5±0,3 (97,6)	96,7±0,8 (97,8)	13,0±2,3 (19)	99,0±0,5 (99,7)	92,1±1,1 (95,1)

Note : Pour les méthodes nécessitant plusieurs essais (**Aléatoire**, **RGSS**, **AG**, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), le tableau donne la moyenne et l'écart type du taux de classification et $(\cdot)$ indique les performances optimisant la validation.

TAB. D.6 – Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Wdbc*.

méthode de sélection	BNgauss			k-ppv		
	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)	nombre de variables sélectionnées	performance de validation (%)	performance de test (%)
évaluation sans processus de sélection						
pas de sélection	166	72,8	58,4	166	86,6	79,8
Aléatoire	109,5±59,1 (8)	69,4±3,8 (62,2)	57,5±5,7 (62,6)	109,5±59,1 (113)	83,3±4,8 (86,6)	78,4±2,8 (81,9)
sélection de variables par des processus séquentiels naïfs						
$\mathbf{SFS}_{Fisher}$	80	78,6	58,4	79	85,8	75,2
$\mathbf{SFS}_{RELIEF}$	154	74,0	58,4	88	88,6	83,2
sélection de variables par des processus séquentiels						
SFS	43	84,5	79,8	78	93,3	83,2
SBS	64	82,8	78,1	77	94,6	82,8
sélection de variables par des processus séquentiels avec retours en arrière						
$\mathbf{LRS}_{(L=3,\ R=2)}$	24	85,8	78,1	34	93,7	78,6
$\mathbf{LRS}_{(L=2,\ R=3)}$	67	84,1	76,1	53	96,7	83,6
SFFS	66	84,1	78,2	26	95,0	81,9
SFBS	17	84,0	77,3	45	97,1	78,6
sélection de variables par des processus non déterministes						
RGSS	73,2±37,3 (41)	80,7±1,9 (81,5)	67,8±10,2 (81,1)	78,3±27,9 (111)	92,7±1,5 (93,3)	81,2±2,4 (84,0)
AG	76,7±5,3 (76)	83,2±0,3 (83,8)	57,0±4,5 (58,4)	76,4±5,6 (74)	94,0±0,6 (94,7)	83,7±1,2 (85,3)
sélection de variables par des processus séquentiels avec retours en arrière par des AG						
$\mathbf{SFS}^*_{Fisher}$	13,6±4,9 (11)	81,1±1,4 (83,2)	75,3±2,3 (78,2)	27,7±9,8 (22)	82,8±2,8 (86,2)	75,7±1,6 (79,0)
$\mathbf{SFS}^*_{RELIEF}$	7,1±3,5 (7)	68,3±1,7 (69,7)	65,1±2,2 (67,6)	14±8 (21)	77,7±8,6 (85,8)	72,2±6,5 (78,6)
$\mathbf{SFS}^*$	30,2±8,6 (26)	85,3±1,5 (86,9)	79,7±1,1 (81,5)	30,8±9,2 (33)	95,6±0,9 (96,2)	83,3±2,4 (86,6)

Note : Pour les méthodes nécessitant plusieurs essais (**Aléatoire**, **RGSS**, **AG**, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), le tableau donne la moyenne et l'écart type du taux de classification et $(\cdot)$ indique les performances optimisant la validation.

TAB. D.7 – Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Musk Clean1*.

méthode de sélection	BNgauss nombre de variables sélectionnées	BNgauss performance de validation (%)	BNgauss performance de test (%)	*k*-ppv nombre de variables sélectionnées	*k*-ppv performance de validation (%)	*k*-ppv performance de test (%)
évaluation sans processus de sélection						
pas de sélection	–	–	–	206	57,2	57,5
Aléatoire	–	–	–	101,9±74,4 (134)	55,3±3,3 (57,2)	55,2±3,3 (59,3)
sélection de variables par des processus séquentiels naïfs						
$\mathbf{SFS}_{Fisher}$	–	–	–	11	62,2	62,4
$\mathbf{SFS}_{RELIEF}$	–	–	–	22	59,0	59,3
sélection de variables par des processus séquentiels						
SFS	–	–	–	82	71,9	60,2
SBS	–	–	–	27	71,0	54,0
sélection de variables par des processus séquentiels avec retours en arrière						
$\mathbf{LRS}_{(L=3,\ R=2)}$	–	–	–	16	71,9	62,8
$\mathbf{LRS}_{(L=2,\ R=3)}$	–	–	–	20	73,6	55,3
SFFS	–	–	–	82	71,9	60,2
SFBS	–	–	–	25	76,8	58,8
sélection de variables par des processus non déterministes						
RGSS	–	–	–	45,0±38,3 (20)	68,5±1,9 (69,5)	57,9±3,0 (62,8)
AG	–	–	–	117,2±8,4 (123)	66,3±0,5 (66,7)	58,5±2,4 (61,5)
sélection de variables par des processus séquentiels avec retours en arrière par des AG						
$\mathbf{SFS}^*_{Fisher}$	–	–	–	6,0±1,9 (8)	63,7±4,1 (67,6)	62,0±2,8 (65,9)
$\mathbf{SFS}^*_{RELIEF}$	–	–	–	3,8±2,6 (6)	53,4±8,1 (57,6)	51,0±6,0 (59,3)
$\mathbf{SFS}^*$	–	–	–	25,2±6,7 (24)	72,9±1,8 (73,9)	62,9±1,6 (65,5)

Note : Pour les méthodes nécessitant plusieurs essais (**Aléatoire**, **RGSS**, **AG**, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), le tableau donne la moyenne et l'écart type du taux de classification et $(\cdot)$ indique les performances optimisant la validation.

TAB. D.8 – Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Arrhythmia*.

D.3.1 Analyses des résultats

Nous pouvons observer les résultats obtenus suivant deux aspects : le premier, en considérant les performances de classification et le second, en considérant le nombre de variables sélectionnées. Bien évidemment la méthode la plus efficace est celle dont les performances sont optimales tout en ayant sélectionné le minimum de variables. Les tableaux D.2 à D.8 donnent les résultats dans leur exhaustivité, il est alors difficile de pouvoir en extraire une synthèse. Le tableau D.9 et la figure D.2, tentent de résumer l'ensemble de ces résultats. Le tableau D.9 récapitule pour chaque ensemble de données et pour chacun des deux classifieurs, la méthode de sélection la plus efficace ; l'efficacité est alors évaluée suivant deux critères :

- par le taux de classification uniquement :

$$J_1 = taux\_de\_classification\,; \tag{D.2}$$

- par le taux de classification et le taux de variables éliminées :

$$\begin{aligned} J_2 = &\ 0,8 \cdot taux\_de\_classification + 0,2 \cdot \frac{nombre\_de\_variables\_éliminées}{nombre\_total\_de\_variables}\,, \\ = &\ 0,8 \cdot taux\_de\_class. + 0,2 \cdot taux\_variables\_éliminées\,. \end{aligned} \tag{D.3}$$

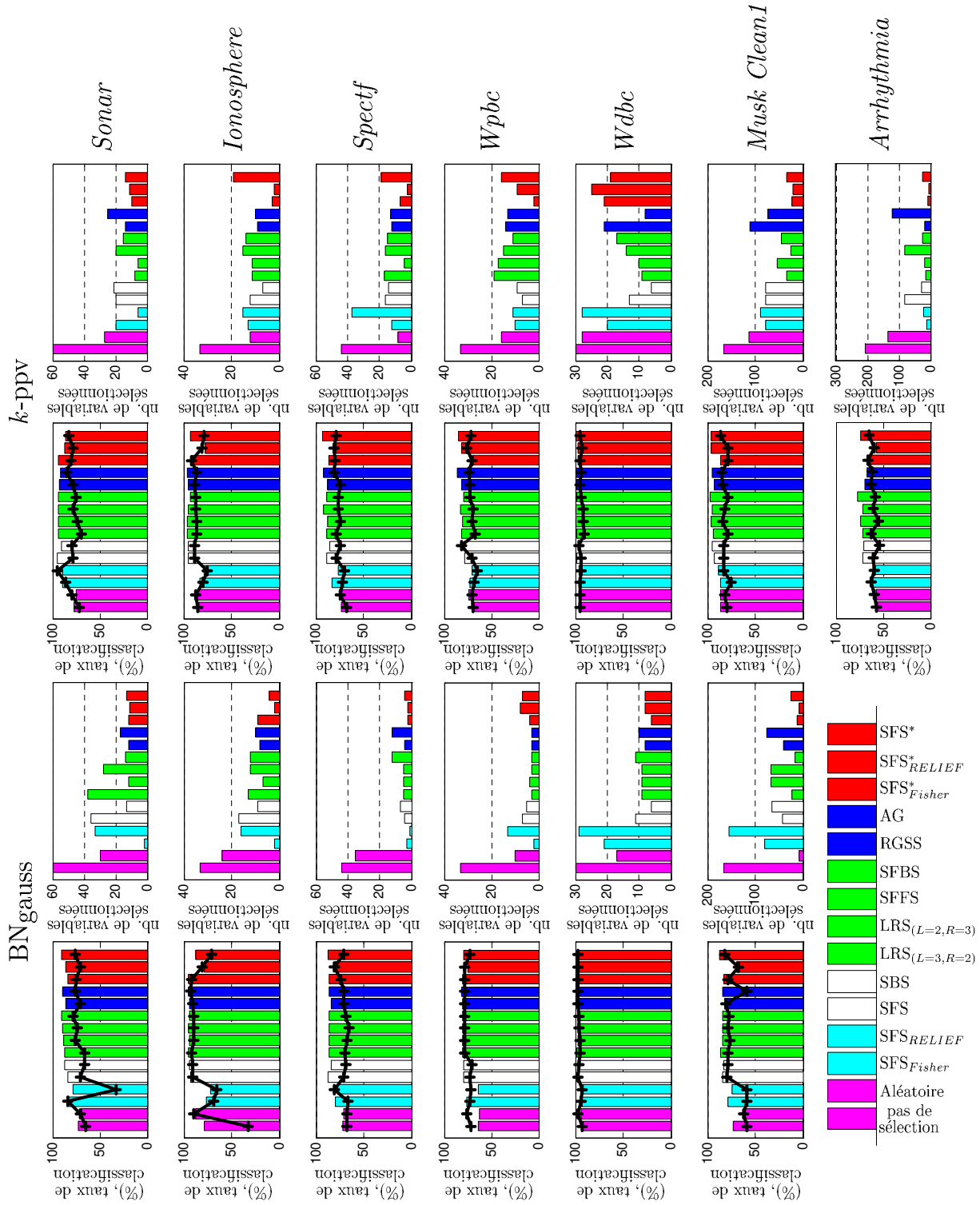
Notons que la forme du critère J_2 est parfois employée dans la fonction d'adaptation des algorithmes génétiques (*cf.* section 2.4.3.4, page 92), afin de privilégier entre deux individus, celui possédant le plus petit nombre de variables. Dans notre cas, nous avons légèrement détourné sa fonction, en augmentant l'importance de la contrainte liée au nombre de variables à éliminer.

ensemble de données	modèle de classification	$J_1 = taux\_de\_classification$ validation	$J_1 = taux\_de\_classification$ test	$J_2 = 0,8 \cdot taux\_de\_class. + 0,2 \cdot taux\_variables\_éliminées$ validation	$J_2 = 0,8 \cdot taux\_de\_class. + 0,2 \cdot taux\_variables\_éliminées$ test
Sonar	BNgauss	**SFBS**	$\mathbf{SFS}_{Fisher}$	$\mathbf{SFS}^*$	$\mathbf{SFS}_{Fisher}$
	k-ppv	$\mathbf{SFS}^*$	$\mathbf{SFS}_{RELIEF}$	$\mathbf{SFS}^*$	$\mathbf{SFS}_{RELIEF}$
Ionosphere	BNgauss	$\mathbf{SFS}^*_{Fisher}$	**AG**	$\mathbf{LRS}_{(L=2,\ R=3)}$	**AG**
	k-ppv	**AG**	$\mathbf{SFS}^*_{Fisher}$	$\mathbf{SFS}^*_{Fisher}$	$\mathbf{SFS}^*_{Fisher}$
Spectf	BNgauss	**SFS**	$\mathbf{SFS}_{RELIEF}$	$\mathbf{SFS}^*_{Fisher}$	$\mathbf{SFS}_{RELIEF}$
	k-ppv	$\mathbf{SFS}^*$	$\mathbf{SFS}^*_{RELIEF}$	$\mathbf{LRS}_{(L=2,\ R=3)}$	$\mathbf{SFS}^*_{RELIEF}$
Wpbc	BNgauss	**AG**	$\mathbf{LRS}_{(L=3,\ R=2)}$	**AG**	$\mathbf{LRS}_{(L=3,\ R=2)}$
	k-ppv	**AG**	**SBS**	**AG**	**SBS**
Wdbc	BNgauss	**AG**	$\mathbf{SFS}^*_{Fisher}$	$\mathbf{SFS}^*_{Fisher}$	$\mathbf{SFS}^*_{Fisher}$
	k-ppv	$\mathbf{SFS}^*$	**SBS**	**AG**	**SBS**
Musk Clean 1	BNgauss	$\mathbf{SFS}^*_{Fisher}$	$\mathbf{SFS}^*_{Fisher}$	$\mathbf{SFS}^*$	$\mathbf{SFS}^*$
	k-ppv	**SFBS**	$\mathbf{SFS}^*$	$\mathbf{SFS}^*_{RELIEF}$	$\mathbf{SFS}^*$
Arrhytmia	BNgauss	–	–	–	–
	k-ppv	**SFBS**	$\mathbf{SFS}^*_{Fisher}$	**SFBS**	$\mathbf{SFS}^*_{Fisher}$

Note : les cases grisées font apparaître les nouveaux processus de sélection fondés sur la combinaison de SFS et des AG.

TAB. D.9 – Observations, suivant les critères de classification et de réduction de dimension, des méthodes de sélection les plus pertinentes, pour chaque ensemble de données *UCI*.

L'analyse du tableau D.2 et de la figure D.2 montre que l'approche combinant la méthode SFS et des AG ($\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$) apparaît majoritairement comme l'approche la plus efficace pour réaliser la sélection de variables. Dans l'ensemble, nos méthodes obtiennent des performances de classification sensiblement supérieures aux méthodes habituellement utilisées dans la littérature, surpassant également les algorithmes génétiques, qui sont pourtant connus pour explorer plus largement l'espace des combinaisons de variables. En outre, nous pouvons noter que les performances obtenues sur le sous-ensemble de test sont majoritairement meilleures pour nos méthodes combinées, montrant alors leur capacité à extraire l'information privilégiant une certaine généralisation.



Note : Les taux de classification donnent les résultats optimisant le sous-ensemble de validation pour chacune des méthodes, où les barres affichent les performances de validation et les courbes, les performances de test.

FIG. D.2 – Récapitulatif des performances (classification et nombre de variables éliminées) pour les méthodes de sélection « classiques » et combinées aux AG, sur les ensembles de données *UCI*.

Il est alors intéressant de remarquer que les méthodes combinées ($\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$) utilisant l'indice de pertinence des variables (obtenu par Fisher et RELIEF) obtiennent des performances remarquables, en dépit de l'espace de recherche restreint par l'approche naïve. En effet, la variable ajoutée à chaque itération est dépendante de l'ordre induit par les critères de Fisher et de RELIEF. Dès lors, nous pouvons remarquer d'une part, la pertinence de ces critères et d'autre part, la capacité des AG à sélectionner judicieusement les variables à éliminer. Cette efficacité des AG peut s'expliquer par le fait que ces derniers sont exploités sur des petits sous-ensembles de variables, entraînant de ce fait, un coût calculatoire relativement faible. Ce coût est faible devant celui engendré par l'utilisation seule des AG qui obtiennent au détriment d'un coût calculatoire important, de bonnes performances. Or, dans la tâche de la sélection de variables, le coût calculatoire est un paramètre qui doit être pris en considération. En effet, plus le coût de cette tâche est réduit, plus les possibilités d'essais deviennent nombreux, laissant place par exemple aux changements et aux adaptations des modèles de classification. Le tableau D.10 montre pour chaque ensemble de données, le nombre de sous-ensembles de variables qui ont dû être évalués, afin d'obtenir le sous-ensemble final ; ce sous-ensemble optimise la classification, en sélectionnant le minimum de variables.

ensemble de données	*Sonar*	*Ionosphere*	*Spectf*	*Wpbc*	*Wdbc*	*Musk Clean 1*	*Arrhythmia*
pas de sélection	1	1	1	1	1	1	1
Aléatoire	20	20	20	20	20	20	20
$\mathbf{SFS}_{Fisher}$ $\mathbf{SFS}_{RELIEF}$	60	33	44	33	30	166	206
SFS et **SBS**	1 830	561	990	561	465	13 861	21 321
LRS	8 787	2 604	4 683	2 604	2 142	68 306	105 366
SFFS et **SFBS**	18 517*	4 410*	8 796*	4 410*	3 508*	212 940*	357 504*
RGSS	1 179 ± 262	379 ± 83	634 ± 145	379 ± 83	309 ± 68	8 806 ± 1 767	13 328 ± 2 331
AG	40 000	40 000	40 000	40 000	40 000	40 000	40 000
$\mathbf{SFS}^*_{Fisher}$ $\mathbf{SFS}^*_{RELIEF}$	3 838 ± 1 434	3 219 ± 1 682	4 416 ± 1 886	2 612 ± 1 662	3 110 ± 1 920	4 400 ± 1 411	4 400 ± 1 203
$\mathbf{SFS}^*$	7 830 ± 1 736	3 435 ± 2 352	6 216 ± 1 660	4 729 ± 2 400	3 425 ± 2 276	18 120 ± 1 768	20 423 ± 1 563

Note : * indique que le nombre de combinaisons donné est une estimation ([Kudo and Sklansky, 2000], *cf.* section 2.4.3.3).

TAB. D.10 – Nombre de combinaisons de variables évaluées par les méthodes « classiques » et combinées aux AG avant d'obtenir le sous-ensemble optimal, pour chaque ensemble de données de *UCI*.

Le tableau D.10 montre que les AG ont été, la plupart du temps, la méthode la plus coûteuse en évaluant 40 000 combinaisons. Pour autant, cela ne leur ont pas permis de dépasser largement les performances de nos approches.

L'extrapolation des données, révélées par le tableau D.10, a permis d'estimer d'une manière générale le coût calculatoire en fonction du nombre de variables présentes dans l'ensemble initial. La figure D.3 montre ainsi les capacités de nos méthodes combinées à trouver l'information, en évaluant, d'après nos estimations, moins de sous-ensembles de variables que la plupart des méthodes classiques de sélection séquentielle. Ces dernières sont pourtant connues dans la littérature comme des méthodes possédant un bon compromis : coût calculatoire/performances. Notons que cette estimation, par extrapolation, est donnée sans changer les paramètres des méthodes (LRS, AG, $\mathbf{SFS}^*_{Fisher}$, $\mathbf{SFS}^*_{RELIEF}$ et $\mathbf{SFS}^*$), qui pourraient être adaptés en fonction du nombre de variables initiales.

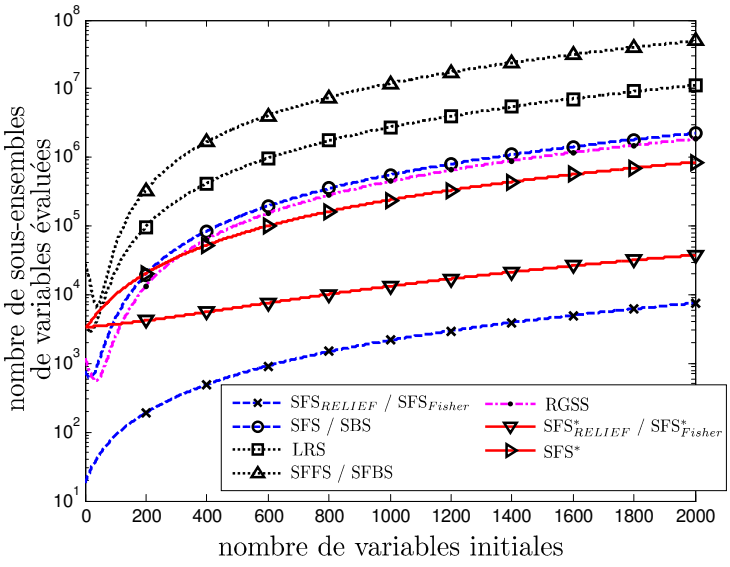


FIG. D.3 – Observation du coût calculatoire des méthodes de sélection de variables « classiques » et combinées aux AG en fonction du nombre de variables dans l'ensemble initial.

D.4 Discussions et conclusions

Les études présentées proposent une comparaison entre plusieurs types de méthodes de sélection de variables, en considérant essentiellement des approches de type *wrapper*. Nous avons ainsi comparé des techniques heuristiques, fondées sur des sélections séquentielles (naïves, montantes, descendantes, avec et sans retours en arrière), et non déterministes. Afin d'améliorer ces processus, nous avons proposé une nouvelle approche reposant sur un perfectionnement du processus de retours en arrière des méthodes séquentielles. Par cette approche, les retours en arrière réalisés habituellement de manière séquentielle, sont alors effectués par des AG. Cette approche conserve alors la rapidité de l'exploration séquentielle et la capacité des AG à explorer efficacement et plus globalement l'espace des combinaisons.

Utilisée à la section 5.4.2.2 et développée dans le cadre de la prédiction de la syncope, cette approche a montré ses capacités d'adaptation dans l'exploration de nouveaux ensembles de données. En effet, les résultats obtenus sur les ensembles provenant de la base *UCI*, ont montré des performances de classification équivalentes et parfois supérieures de nos méthodes face aux méthodes de sélection de variables habituellement utilisées dans la littérature. Aussi, l'intérêt de notre approche est d'obtenir des sous-ensembles pertinents dénombrant très peu de variables, en évaluant moins de combinaisons que les autres méthodes de sélection. En effet, comme le montre la figure D.3, l'approche de sélection séquentielle naïve (SFS$_{Fisher}$ et SFS$_{RELIEF}$) demeure être la seule à évaluer moins de combinaisons pour obtenir le sous-ensemble de variables optimal. Notons cependant que cette rapidité d'exécution a pour corollaire de mauvais résultats de classification, contrairement à nos méthodes combinées. Dès lors, l'approche proposée permet d'offrir certainement le compromis performance/coût calculatoire le plus intéressant.

Liste des figures

1 Schéma général d'un système de reconnaissance de formes [Belaïd and Belaïd, 1992]. . 3
2 Organigramme de lecture possible de cette thèse. 7

1.1 Illustration des domaines scientifiques apparentés à l'apprentissage artificiel. 12
1.2 Schéma d'un processus de fouille de données [Han and Kamber, 2006]. 13
1.3 Schéma général d'un système de reconnaissance de formes [Belaïd and Belaïd, 1992]. . 14
1.4 Règle de décision de Bayes pour un problème à deux classes. 17
1.5 Estimations de densités par la méthode des k-plus proches voisins. 19
1.6 Illustrations de frontières de décision obtenues par les fonctions discriminantes quadratique et linéaire. 21
1.7 Illustration d'une séparation linéaire en deux classes d'un ensemble de données. 23
1.8 Géométrie d'une fonction de discrimination linéaire. 24
1.9 Hyperplan discriminant de Fisher. 25
1.10 Influence d'observations aberrantes lors de la construction d'hyperplans discriminants par la méthode des moindres carrés. 27
1.11 Représentation du perceptron de Rosenblatt [Rosenblatt, 1962]. 29
1.12 Algorithme du perceptron : évolution de l'hyperplan discriminant de l'itération k à l'itération $k+1$. 31
1.13 Hyperplan discriminant maximisant les marges. 32
1.14 Hyperplan discriminant obtenu par les *support vector machines* en présence de données non linéairement séparables. 34
1.15 Comparaison d'approches pour séparer les observations d'un problème à trois classes : une classe contre toutes les autres et une classe contre une autre. 36
1.16 Séparation linéaire de trois classes, fondée sur un cas particulier de l'approche « une classe contre une autre ». 37
1.17 Principaux types d'architectures et de structures d'interconnexions des réseaux de neurones. 40
1.18 Représentation détaillée du perceptron [Rosenblatt, 1962]. 41
1.19 Fonctions d'activation usuelles pour les réseaux de neurones. 41
1.20 Synoptique de l'apprentissage supervisé et non supervisé d'un réseau de neurones [Amat and Yahiaoui, 2002]. 42
1.21 Illustration d'un réseau de neurones à une couche cachée (PMC) [Bishop, 1995]. . . . 43
1.22 Relations entre le neurone de sortie k et les M neurones cachés. 44
1.23 Influence de l'architecture d'un réseau de neurones sur les frontières de décision [Bishop, 1995]. 48
1.24 Influence de contraintes non pertinentes, résultant d'un modèle trop complexe (ajout de neurones cachés et surapprentissage), sur les frontières de décision. 48
1.25 Arrêt prématuré de l'apprentissage afin d'éviter le surapprentissage. 50
1.26 Réseau RBF. 51
1.27 Illustration de la représentation des classes par un réseau RBF. 52
1.28 Illustration de la transformation de l'espace par une fonction noyau (dans le cadre des *support vector machines*) sur un exemple de discrimination non linéaire. 53

2.1 Extraction de caractéristiques et sélection de variables [Webb, 2002]. 60
2.2 Détection d'observations aberrantes sur une variable aléatoire de distribution normale. 61

2.3 Influence de la présence de valeurs aberrantes sur la frontière de décision obtenue par les moindres carrés. 61
2.4 Remplacement de valeurs manquantes par la valeur moyenne. 63
2.5 Remplacement de valeurs manquantes par une valeur aléatoire. 63
2.6 Remplacement de valeurs manquantes par le plus proche voisin. 64
2.7 Remplacement de valeurs manquantes par une valeur prédite. 64
2.8 Illustration de la réduction de dimension par l'analyse en composantes principales. . . 66
2.9 Comparaison de critères pour le choix du nombre de composantes principales à conserver. 67
2.10 Projection par la méthode *multidimensional scaling* pour reconstituer le positionnement des villes sur une carte. 69
2.11 Comparaison de la projection par une analyse en composantes principales et par une analyse factorielle discriminante [Theodoridis and Koutroumbas, 2006]. 70
2.12 Comparaison des mesures de distances euclidienne et géodésique entre deux points $\mathbf{x}_1$ et $\mathbf{x}_2$ [Lee *et al.*, 2004]. 71
2.13 Algorithme de projection LLE [Roweis and Sam, 2000]. 72
2.14 Illustration du « graphe » et de la distance obtenue par la méthode *isomap* entre deux points $\mathbf{x}_1$ et $\mathbf{x}_2$ (pour $k = 3$). 73
2.15 Ensembles de données, du « petit suisse » (*swiss role*) et de deux anneaux imbriqués, utilisés pour comparer les méthodes de projection. 75
2.16 Comparaison de projections linéaires et non linéaires sur l'exemple du « petit suisse » (*swiss role*) et de deux anneaux imbriqués. 76
2.17 Illustration détaillée du problème *XOR* en deux dimensions. 78
2.18 Évaluation du nombre de combinaisons de variables possibles, en fonction du nombre de variables disponibles (p) et du nombre de variables à sélectionner (q). 78
2.19 Procédure traditionnelle de recherche d'un sous-ensemble de variables [Liu *et al.*, 1998; Liu and Yu, 2002]. 79
2.20 Approches à la sélection de sous-ensembles de variables (*filter* et *wrapper*) fondées sur l'intégration d'un algorithme d'apprentissage [Yang and Honavar, 1997]. 80
2.21 Observation de la mesure de divergence (distance probabiliste) en fonction du niveau de recouvrement des classes. 82
2.22 Observation de l'évolution des critères basés sur les matrices de covariances en fonction de différentes distributions des observations. 83
2.23 Illustration détaillée d'un problème binaire à trois variables dans le cadre d'une sélection naïve de variables. 86
2.24 Comparaison de la progression dans l'espace des variables de méthodes de sélection séquentielle ascendante. 91
2.25 Comparaison de la progression dans l'espace des variables de méthodes de sélection non déterministes. 92
2.26 Organigramme du fonctionnement classique d'un algorithme génétique. 94
2.27 Opérateurs génétiques (croisement et mutation). 95
2.28 Organigramme récapitulatif des approches et des méthodes de réduction de la dimension. 98

3.1 Illustration de l'analyse d'un problème [Denker *et al.*, 1987]. 102
3.2 Partitionnement de l'ensemble des observations en deux sous-ensembles pour effectuer les tâches d'apprentissage et de test. 103
3.3 Influence du nombre d'observations de l'échantillon sur la probabilité d'erreur (p_e) de l'apprentissage et de test [Tufféry, 2007]. 103
3.4 Illustration des frontières de décision après un sous-apprentissage et un surapprentissage. 104

3.5 Partitionnement de l'ensemble des observations en trois sous-ensembles pour effectuer les tâches d'apprentissage, de test et de validation (pour gérer par exemple l'arrêt prématuré). . . . 105
3.6 Illustration du partitionnement d'un ensemble d'observations pour une validation croisée. 106
3.7 Influence du nombre d'observations de test sur l'intervalle de confiance à $0,95$ pour une probabilité d'erreur de 5%. . . . 107
3.8 Illustration des quatre états possibles lors de la prédiction d'un modèle de classification binaire : vrai/faux positif et vrai/faux négatif. . . . 109
3.9 Exemple de courbes de ROC. . . . 112
3.10 Comparaison des courbes de ROC. . . . 113

4.1 Causes de la syncope [Linzer *et al.*, 1997]. . . . 118
4.2 Démarche diagnostique en cas de suspicion de syncope [Antonini-Revaz *et al.*, 2004]. . . 119
4.3 Test de la table d'inclinaison. . . . 120
4.4 Placement des électrodes d'acquisition des signaux de l'électrocardiogramme et de l'impédancemétrie thoracique. . . . 122
4.5 Caractéristiques extraites sur la dérivée du signal d'impédancemétrie thoracique durant un battement cardiaque. . . . 123

5.1 Illustration des deux données aberrantes supprimées de l'échantillon initial $\mathcal{E}_1$. . . . 128
5.2 Processus de sélection des sous-ensembles de variables pertinentes pour prédire le résultat du *tilt-test* en position couchée. . . . 132
5.3 Processus d'estimation des performances des modèles [Loughrey and Cunningham, 2005]. . . . 134
5.4 Comparaison des courbes de ROC des modèles de classification issus d'une sélection exhaustive des variables d'entrée pour prédire le résultat du *tilt-test* en position couchée. . . . 135
5.5 Comparaison de la distribution de la variable « hématocrite » pour les patients positifs et négatifs au *tilt-test*. . . . 136
5.6 Indice de corrélation entre chaque paire de variables pré-sélectionnées. . . . 137
5.7 Représentation du *biplot* et de l'inertie expliquée sur chacune des nouvelles composantes issues de l'analyse en composantes principales. . . . 138
5.8 Résultats de la projection en deux dimensions de l'analyse en composantes curvilignes. 139
5.9 Processus d'extraction des caractéristiques pertinentes pour prédire le résultat du *tilt-test* en position couchée. . . . 139
5.10 Qualité de la représentation des variables pré-sélectionnées dans les composantes principales utilisées pour prédire le résultat du *tilt-test* en position couchée. . . . 141
5.11 Comparaison des courbes de ROC des modèles de classification issus de processus de projection (linéaire et non linéaire) pour prédire le résultat du *tilt-test* en position couchée. . . . 143
5.12 Comparaison des courbes de ROC des modèles de classification issus de processus de réduction de dimension (sélection et extraction) pour prédire le *tilt-test* en position couchée. . . . 144
5.13 Comparaison des courbes de ROC de classifieurs (k-ppv) issus de méthodes de sélection dites « classiques » pour prédire le résultat du *tilt-test* en position couchée et basculée. 152
5.14 Comparaison des courbes de ROC de classifieurs (k-ppv) issus de méthodes de sélection combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée. 153
5.15 Comparaison des courbes de ROC de classifieurs (k-ppv) issus de méthodes de sélection « classiques » et combinées aux AG pour prédire le *tilt-test* en position couchée et basculée. . . . 153

5.16 Performances et nombre de variables sélectionnées pour les méthodes de sélection « classiques » et combinées aux AG pour prédire le *tilt-test* en position couchée et basculée. 156
5.17 Nombre de combinaisons de variables évaluées par les méthodes de sélection « classiques » et combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée. 157
5.18 Récapitulatif des caractéristiques extraites sur le signal d'impédancemétrie thoracique (Z) pour prédire le résultat du *tilt-test* [Schang *et al.*, 2003]. 159
5.19 Réponses impulsionnelle et fréquentielle du filtre utilisé pour réduire les nuisances lors de l'analyse des signaux ECG et Z. 159
5.20 Extraction d'un complexe sur les signaux ECG et Z et leur dérivée $dECG$ et dZ. . . 160
5.21 Extraction et sélection de complexes dZ par minimisation de l'erreur quadratique moyenne totale [Bellard *et al.*, 2003; Bellard, 2003]. 161
5.22 Nouvelles caractéristiques extraites du signal d'impédancemétrie thoracique (Z et dZ) dans le domaine temporel. 164
5.23 Illustration des paramètres liés aux méthodes de sélection automatique d'une fenêtre de complexes dZ. 166
5.24 Sélection automatique d'une fenêtre de complexes dZ par optimisation globale (OG) du rapport signal sur bruit. 167
5.25 Sélection automatique d'une fenêtre de complexes dZ par optimisation locale (OL) du rapport signal sur bruit. 168
5.26 Illustration de la probabilité d'erreur obtenue par la règle de Bayes. 169
5.27 Influence du RSB de la fenêtre (RSB_{w_j}) sur la pertinence des caractéristiques issues du signal dZ, lors de la sélection de fenêtre par optimisation globale et locale. . . . 170
5.28 Influence du RSB du complexe ($RSB_{c_i^j}$) sur la pertinence des caractéristiques issues du signal dZ, lors de la sélection de fenêtre par optimisation globale et locale. . . . 171
5.29 Influence de la sélection de fenêtres (aléatoire, optimisation globale et locale) sur la pertinence (P_{err}) des caractéristiques issues du signal dZ, lors d'extraction aléatoire de complexes. 172
5.30 Illustrations de la densité spectrale de puissance (DSP) obtenue par la méthode de Welch. 174
5.31 Analyse de la distribution des amplitudes de chaque fréquence pré-sélectionnée sur la DSP pour chaque groupe de patients (positif et négatif) de l'échantillon d'apprentissage. 175
5.32 Analyse de mesures statistiques des amplitudes de chaque fréquence pré-sélectionnée sur la DSP pour chaque groupe de patients (positif et négatif) de l'échantillon d'apprentissage. 176
5.33 Illustration des 14 fréquences sélectionnées et de leur amplitude pour un patient de chaque classe (positif et négatif au *tilt-test*). 176
5.34 Évolution de l'aire moyenne sous la courbe de ROC (en validation) des classifieurs PMC issus de la sélection naïve de fréquence (SFS_{Fisher}). 177

6.1 Illustration du résultats d'une analyse en composantes principales. 185
6.2 Interprétation de la qualité de la représentation des variables initiales dans les composantes principales sur un exemple en deux dimensions. 186
6.3 Ensemble de données, représentant une forme de « S », utilisé pour valider l'estimation de la représentation des variables dans les nouvelles composantes. 187
6.4 Résultats de l'analyse en composantes principales sur les données représentant une forme de « S ». 188

6.5 Illustration de l'extrapolation des « vecteurs propres locaux » ($\mathbf{u}^*_{jl}$, $l = 1, \cdots, h$) sur la courbe de projection. 190
6.6 Comparaison du résultat de l'extraction de la qualité de la représentation des variables dans les composantes principales et dans les composantes curvilignes. 191
6.7 Comparaison des projections par des méthodes de réduction linéaire (ACP) et non linéaire (ACC, LLE et NLM) sur les données représentant une forme de « S ». . . . 193
6.8 Ensemble de données, représentant une sphère en trois dimensions, dans l'espace initial et dans les espaces réduits par l'analyse en composantes principales et curvilignes. 194
6.9 Qualité de la représentation des variables initiales dans les composantes curvilignes pour l'exemple de la sphère en trois dimensions. 194
6.10 Qualité de la représentation des variables initiales, pondérées par leur distribution, dans les composantes curvilignes pour l'exemple de la sphère en trois dimensions. . 195
6.11 Qualité de la représentation des variables pré-sélectionnées dans les composantes curvilignes utilisées pour prédire le résultat du *tilt-test* en position couchée. 197

A.1 Illustration des notations utilisées pour la démonstration de l'algorithme de rétropropagation. 209
A.2 Illustration de la rétropropagation de l'erreur, des k neurones de sortie vers le neurone i de la couche cachée. 210
A.3 Nouvelles caractéristiques extraites du signal d'impédancemétrie thoracique dans le domaine temporel. 212
A.4 Illustration de l'erreur de classification obtenue par la règle de Bayes. 213
A.5 Représentation des paramètres utilisés dans l'estimation de la probabilité de l'erreur de classification. 214

B.1 Illustration et table de vérité du problème *XOR*. 217
B.2 Architecture du perceptron pour résoudre le problème *XOR*. 218
B.3 Architecture du perceptron multicouches pour résoudre le problème *XOR*. 218
B.4 Décomposition des frontières de décision associées à chaque neurone du perceptron multicouches (voir figure B.3) pour résoudre le problème *XOR*. 219
B.5 Architecture du perceptron multicouches fondé sur le perceptron (voir figure B.2) pour résoudre le problème *XOR*. 220
B.6 Architecture du réseau RBF pour résoudre le problème *XOR*. 220
B.7 Transformation de l'espace d'entrée par le réseaux RBF pour rendre le problème de discrimination linéaire et résoudre le problème *XOR*. 221
B.8 Normalisation du problème *XOR* pour les SVM (illustration et table de vérité. 222
B.9 Frontières de décision obtenues par les SVM pour résoudre le problème *XOR*. 223

C.1 Évolution des valeurs prédictives positive et négative en fonction de la prévalence de la maladie dans l'échantillon. 226
C.2 Exemple de construction de la courbe de ROC. 228

D.1 Comparaison de la progression dans l'espace des variables de la méthode de sélection de variables combinant la sélection séquentielle ascendante et les algorithmes génétiques. 230
D.2 Récapitulatif des performances (classification et nombre de variables éliminées) pour les méthodes de sélection « classiques » et combinées aux AG, sur les ensembles de données *UCI*. 241
D.3 Observation du coût calculatoire des méthodes de sélection de variables « classiques » et combinées aux AG en fonction du nombre de variables dans l'ensemble initial. . 243

Listes des tables

2.1 Comparaison et évaluation de l'impact du remplacement de valeurs manquantes sur la distribution d'un ensemble de données. 64
2.2 Matrice représentative de l'ensemble de données employé par la méthode MDS pour reconstituer le positionnement des villes sur une carte. 69

3.1 Tableau/matrice de confusion issu de la prédiction d'un modèle de classification binaire : vrai/faux positif et vrai/faux négatif. 109
3.2 Descriptions des indices d'évaluation obtenus à partir de la matrice de confusion (voir tableau 3.1). 111

4.1 Description des causes de la syncope et de leur degré de sévérité [Linzer *et al.*, 1997]. . 118
4.2 Récapitulatif des résultats significatifs de recherches sur la syncope inexpliquée. 125

5.1 Récapitulatif de l'ensemble des variables recueillies pour l'étude de l'apparition de la syncope lors d'un examen du *tilt-test* (partie 1/2). 129
5.2 Récapitulatif de l'ensemble des variables recueillies pour l'étude de l'apparition de la syncope lors d'un examen du *tilt-test* (partie 2/2). 130
5.3 Liste des variables pré-sélectionnées par les médecins susceptibles d'être pertinentes pour prédire l'apparition des symptômes de la syncope durant la position couchée du *tilt-test*. 131
5.4 Comparaison des performances des modèles de classification issus d'une sélection exhaustive des variables d'entrée pour prédire le résultat du *tilt-test* en position couchée. 134
5.5 Récapitulatif des variables sélectionnées par le processus exhaustif pour chaque modèle de classification, afin de prédire le résultat du *tilt-test* en position couchée. 136
5.6 Comparaison des performances des modèles de classification issus d'une analyse en composantes principales (ACP) pour prédire le résultat du *tilt-test* en position couchée. 140
5.7 Comparaison des performances des modèles de classification issus d'une analyse en composantes curvilignes (ACC) pour prédire le résultat du *tilt-test* en position couchée. 140
5.8 Comparaison des performances des modèles de classification issus d'un traitement, associant l'ACP et l'ACC, pour prédire le résultat du *tilt-test* en position couchée. 142
5.9 Récapitulatif des meilleures associations des méthodes de classification et de réduction (sélection et extraction) pour prédire le résultat du *tilt-test* en position couchée. . . 144
5.10 Comparaison des résultats de prédiction de la réponse du *tilt-test* en position couchée, avec les principales études analysant la syncope inexpliquée. 146
5.11 Récapitulatif des variables et de leur indice utilisées pour prédire l'apparition des symptômes de la syncope durant les deux positions du *tilt-test*. 147
5.12 Comparaison des performances de classifieurs (k-ppv) issus de méthodes de sélection dites « classiques » pour prédire le résultat du *tilt-test* en position couchée et basculée. 151
5.13 Comparaison des performances de classifieurs (k-ppv) issus de méthodes de sélection combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée. 153
5.14 Comparaison des performances (en test) de classifieurs (PMC) issus de méthodes de sélection combinées aux AG pour prédire le résultat du *tilt-test* en position couchée et basculée. 154

5.15 Comparaison des résultats de prédiction de la réponse du *tilt-test* en positions couchée et basculée, avec les principales études analysant la syncope inexpliquée. 155
5.16 Évolution des variables issues du signal d'impédancemétrie thoracique dZ durant les positions couchée et basculée du *tilt-test* [Bellard *et al.*, 2003]. 162
5.17 Comparaison des performances de classifieurs (SVM), associés aux nouvelles caractéristiques extraites du signal Z et dZ, pour prédire le résultat du *tilt-test* en position couchée. 164
5.18 Évaluation de la pertinence des nouvelles caractéristiques extraites du signal Z et dZ, pour prédire le résultat du *tilt-test* en position couchée [Schang *et al.*, 2006]. 165
5.19 Évaluation de la pertinence (P_{err}) des caractéristiques issues du signal dZ en fonction du processus de sélection de fenêtres (aléatoire, optimisation globale et locale) et de la méthode d'extraction des caractéristiques (choix aléatoire des complexes, moyenne des complexes et complexes optimisant le RSB). 173
5.20 Identification des indices correspondant aux 40 fréquences pré-sélectionnées sur la densité spectrale de puissance. 174
5.21 Comparaison des résultats obtenus lors de l'analyse exclusive du signal d'impédancemétrie thoracique, dans le cadre de la prédiction du résultat du *tilt-test* en position couchée, avec les principales études analysant la syncope inexpliquée. 178
5.22 Récapitulatif des études analysant l'apparition des symptômes de la syncope lors de l'examen du *tilt-test*. 181

6.1 Liste des variables pré-sélectionnées par les médecins susceptibles d'être pertinentes pour prédire l'apparition des symptômes de la syncope durant la position couchée du *tilt-test*. 196
6.2 Comparaison des performances des modèles de classification issus d'une sélection exhaustive des variables pour prédire le *tilt-test* en position couchée, extrait du tableau 5.4. 196
6.3 Récapitulatif des meilleures associations des méthodes de classification et de réduction (sélection et extraction) pour prédire le résultat du *tilt-test* en position couchée, extrait du tableau 5.9. 198

B.1 Récapitulatif des différents potentiels associés à chaque neurone du perceptron multicouches (voir figure B.3), en fonction des observations d'entrée du problème *XOR*. 219
B.2 Récapitulatif des différents potentiels associés à chaque neurone du perceptron multicouches (voir figure B.5), en fonction des observations d'entrée du problème *XOR*. 220
B.3 Récapitulatif des différents potentiels associés à chaque neurone du réseau RBF (voir figure B.6), en fonction des observations d'entrée du problème *XOR*. 221

C.1 Évaluation de l'influence du déséquilibre entre les classes (prévalence de la maladie) sur les indices de performances usuels. 226
C.2 Évaluation de l'influence du déséquilibre entre les classes (prévalence de la maladie) sur les valeurs prédictives, indépendamment des indices usuels. 227

D.1 Description des ensembles de données de *UCI* utilisés, pour évaluer les performances des méthodes de sélection de variables combinées aux AG. 231
D.2 Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Sonar*. 233
D.3 Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Ionosphere*. . . . 234
D.4 Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Spectf*. 235

D.5 Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Wpbc*. 236
D.6 Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Wdbc*. 237
D.7 Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Musk Clean1*. . . 238
D.8 Performances des méthodes de sélection de variables (méthodes dites « classiques » et méthodes combinées aux AG), sur l'ensemble de données nommé *Arrhythmia*. . . . 239
D.9 Observations, suivant les critères de classification et de réduction de dimension, des méthodes de sélection les plus pertinentes, pour chaque ensemble de données *UCI*. 240
D.10 Nombre de combinaisons de variables évaluées par les méthodes « classiques » et combinées aux AG avant d'obtenir le sous-ensemble optimal, pour chaque ensemble de données de *UCI*. 242

Listes des algorithmes

1.1 Règle d'adaptation de *Widrow-Hoff* 28
1.2 Version stochastique de l'algorithme du perceptron 30
2.1 Pseudo-code de l'algorithme d'évaluation RELIEF. 85
2.2 Pseudo-code d'une variante de l'algorithme de sélection « naïve » de variables. . . . 87
2.3 Pseudo-code de l'algorithme de sélection de variables SFS. 89
2.4 Pseudo-code de l'algorithme de sélection de variables SFFS. 90
D.1 Pseudo-code de l'algorithme de sélection de variables **SFS***, fondé sur une combinaison des algorithmes séquentiels et évolutionnaires. 230

Index

A

algorithmes génétiques, 88, 92–96, 229
  fonction d'adaptation, 94, 148, 232, 240
  opérateurs de croisement, 95
  opérateurs de mutation, 95
  remplacement, 95
  sélection, 94
analyse discriminante
  géométrique, voir Bayes, fonction discriminante
  probabiliste, voir Bayes, fonction discriminante
analyse en composantes curvilignes, 73–74
  contribution des variables, 187–191
  inertie, 189
analyse en composantes principales, 65–68
  biplot, voir biplot
  choix des composantes, 66–67
  contribution des variables, 68, 185–187
  inertie, 66, 67, 189
  par parties, 71, 73
analyse factorielle discriminante, 69–70
ANOVA, 84
apprentissage
  arrêt prématuré, voir arrêt prématuré
  artificiel, 3, 11–12, 59
  avec bruit, 45, 49
  avec régularisation, 49
  batch, 30, 45
  élagage, voir élagage
  non supervisé, 42
  off-line, 100
  on-line, 100
  par correction d'erreur, 42
  par renforcement, 42
  pas d', 28, 31, 45–46
  perceptron, 28
  perceptron multicouches, 44–50
  RBF, 51, 52
  séquentiel, 30, 45
  supervisé, 42
  terme d'inertie, 46
approximateur universel, 50
arrêt prématuré, 49, 105
AUC, Area Under the Curve, 112

B

backpropagation with momentum, 46
Bayes
  classifieur naïf, 22
  fonction discriminante, 20–21
  règle de décision de, 16
  théorème de, 16, 213
Bayes
  règle de décision de, 169
biais-variance (dilemme), 102, 104
biplot, 68
bootstrap, 106
boxplot, 62
Branch and Bound (B & B), 87
Broyden-Fletcher-Goldfarb-Shanno (BFGS), 47

C

cascade-correlation, 49
Cattell, voir test de l'éboulis
classification, voir tâche de
cluster, 14, 190
clustering, voir tâche de classification
complexe *dZ*
  extraction, 160
  prétraitement, 159
  sélection automatique, 165–167
  sélection manuelle, 161, 163
confusion (matrice de), 109
construction de caractéristiques, 98
contractibilité (indice de), 161
curvilinear distance analysis (ADC), 74

D

débit cardiaque, 122, 123
data mining, voir fouille de données
Delta, voir règle de
delta-bar-delta, 46
densité spectrale de puissance, 172, 173

discrimination, voir tâche de
distance
  curviligne, 74
  de Mahalonobis, 21
  euclidienne, 71
  géodésique, 71, 72
  probabiliste, 81–82
  variance covariance (fondée sur), 82–83
donnée
  aberrante, 27, 61–62
  manquante, 63–64, 100
  normalisation, 62

E

early stopping, voir arrêt prématuré
échantillon
  complexité, 59, 102–103, 225
  d'apprentissage, 102
  de test, 102
  de validation, 105
élagage, 49
électrocardiogramme, 121–124
erreur
  d'estimation, 104
  de classification, 108, 110, 213
  quadratique moyenne (EQM), 161
  quadratique moyenne totale (EQMT), 161
  risque d', 106, 169
extraction de caractéristiques, 60, 65, 183
  approche linéaire, 65
  approche non linéaire, voir projection non linéaire

F

feature ranking, 85
Fisher
  critère de, 25, 26, 83–85
  fonction discriminante, 70, 83
  fonction discriminante de, 21, 24
fitness function, voir algorithmes génétiques, fonction d'adaptation
fonction d'activation, 29, 41, 47
fonction discriminante
  de Fisher, 21, 24, 70, 83
  linéaire, 20
  marge, voir marge
  multiclasses, 36
  quadratique, 20
fonctions noyaux, 50, 54
fouille de données, 12, 97

G

généralisation, 99–101, 103, 106
generative topographic mapping (GTM), 73

H

Hamming (fenêtre de), 173
Hebb, voir règle de
hessian-based Locally Linear Embedding, 74
holdout, 102–103, 105, 106
HUTT, voir test de la table d'inclinaison
hyperplan séparateur, 23

I

inductive (technique), 15
intervalle de confiance, 107, 111
isomap, 72
isometric feature mapping, voir *isomap*

K

K-means, 71
k-plus proches voisins, 15, 19, 56, 150
Kaiser, voir règle de
kernel-PCA, 71
Kohonen (cartes auto-organisatrices), 42, 73
Kolmogorov-Smirnov, voir test de

L

leave-one-out, 106
Levenberg-Marquardt (méthode de), 47
Locally Linear Embedding, 72

M

Mahalonobis, voir distance
malédiction de la dimensionnalité, 17, 22, 59
marge (hyperplan), 31, 32
modèle
  évaluation du, 133
  complexité du, 101, 104
  erreur, voir erreur
  parcimonieux, voir parcimonie
  performance du, 101
  qualité du, 100
Monte-Carlo (méthode de), 214
multidimensional scaling (MDS), 68–69, 72, 73

N

Newton (méthode de), 46, 47
nonlinear mapping (NLM), 73
nonmetric MDS, 68

O

Optimal Brain Damage (OBS), voir élagage

optimal brain surgeon (OBS), voir élagage
outlier, voir donnée aberrante
overfitting, voir surapprentissage
overtraining, voir surapprentissage

P

parcimonie, 5, 50, 100, 104
perceptron
  algorithme du, 28
  apprentissage, 30–31
  fonction d'activation, 29
  limites du, 30, 39, 43, 45
perceptron multicouches, 43–50
  apprentissage, 44–50
  architecture, 48
  fonction d'activation, 47
phénomène de l'espace vide, voir malédiction de la dimensionnalité
phonocardiogramme, 122
Plus-L Minus-R Selection (LRS), 89, 90
prétraitement, voir tâche de
prévalence (maladie), 110, 225–226
projection non linéaire, 71
  approche algébrique, 72–73
  approche neuronale, 73–74
  qualité de la, 75, 183, 189
pruning, voir élagage
pseudo-valeurs propres, 188–189

Q

quasi-Newton (méthode de), 46
Quickprop, 46

R

random generation plus sequential selection (RGSS), 91
rapport signal sur bruit (RSB), 166
rapports de vraisemblance, 110
reconnaissance de formes, 12, 13
recouvrement (mesure du), 169, 213–215
rééchantillonnage, 106
règle
  Delta, 28, 42, 44, 52, 210
  de Hebb, 38
  de Kaiser, 67
  de Simpson, 212
  du perceptron, 28, 42
régression, voir tâche de
RELIEF, 84–85, 92
réseau de neurones, 38–52
  Adaline, 38
  approximateur universel, 50
  architecture, 39, 40
  auto-organisant, 42, 73
  bouclé, 39, 40
  connexions, 40
  dynamique, 39
  fonction d'activation, 41, 47
  Hopfield, 39
  Kohonen, voir Kohonen
  neurone, 40–42
  non bouclé, 39, 40
  perceptron, 28, 38, 217–218
  perceptron multicouches, 43–50, 218–220
  RBF, 50–52, 220–221
resilient backpropagation, 46
résistance vasculaire (indice de), 161
resubstitution (méthode de), 101
rétropropagation, 28, 39, 44, 45, 209–211
ROC (*Receiver Operating Characteristic*)
  courbe de, 112–114, 227
  indice de, 110
Rprop, voir *resilient backpropagation*

S

scalar feature selection, 85
scree test, voir test de l'éboulis
sélection de variable
  coût calculatoire, 78
  naïve, 150
sélection de variables, 60, 77
  avec retours en arrière séquentiels, 89–91, 148–149
  avec retours en arrière stochastiques, 149–150, 229–231
  embedded, 81
  filter, 80, 81, 85, 92
  heuristique, 87–92
  hybride, 81
  métaheuristique, 88, 92–93
  naïve, 85–87
  stratégie, 88
  wrapper, 80, 81, 85
self-organizing map (SOM), voir Kohonen
sensibilité, 110, 225–226
sequential backward selection (SBS), 89
sequential bidirectional selection (SBiS), 91
sequential backward forward selection (SFBS), 90
sequential floating forward selection (SFFS), 90, 92

sequential forward selection (SFS), 89, 92, 229
signal d'impédancemétrie thoracique, 121–124, 158–159, 212
    complexe dZ, voir complexe dZ
    dérivée dZ, 160
    filtrage, 159
sous-apprentissage, 104
spécificité, 110, 225–226
support vector machines, 31–35, 53–56, 221–223
    espace de redescription, 32, 53
    fonctions noyaux, 54, 55
    lagrangien, 33
    LS-SVMlab1.5 (*toolbox*), 132
    marge, 32, 33
    multiplicateur de Lagrange, 33
    variables ressort, 34, 53
    vecteurs de support, 33, 54
surapprentissage, 48, 49, 57, 100, 103–105
syncope, 117
    causes, 117–118
    diagnostic, 119
    prédiction, 124–125, 133–135, 139–142, 154, 162, 164–165, 177, 197–198

T

tâche
    de classification, 14, 39, 42
    de discrimination, 14
    de prétraitement, 13, 60
    de réduction de la dimension, 59–60
    de regression, 14
temps d'éjection ventriculaire (TEV), 123
test
    de Kolmogorov-Smirnov, 62
    de l'éboulis, 67
    de Schellong, 119
test de la table d'inclinaison, 120–121
tilt-test, voir test de la table d'inclinaison
transductive (technique), 15, 56

U

UCI Machine Learning Repository, 231

V

valeur prédictive, 110, 225–226
validation croisée, 105–106
variable
    redondante, 59, 65, 66, 77
    ressort, 34, 53
vecteurs de support, 33, 54
vecteurs propres locaux, 189–191
volume d'éjection systolique (VES), 122, 123

W

weight decay, 49
Welch (méthode de), 173

X

XOR (problème du), 77, 217–223

Bibliographie commentée

C. M. Bishop. *Neural networks for pattern recognition.* Oxford University Press, 1995.

> Il existe de nombreux ouvrages sur les réseaux de neurones, ce livre fournit selon moi, une remarquable compréhension des réseaux de neurones ; particulièrement pour les réseaux de type *feedforward*. L'une des particularités de cet ouvrage réside dans les nombreux exercices qui sont proposés tout au long des chapitres. Notons que C. M. Bishop a publié plus récemment un autre ouvrage[1], dans lequel il s'étend plus largement sur les méthodes issues de l'apprentissage artificiel. L'écriture de ce nouveau livre ressemble au précédent, où l'aspect théorique est prépondérant et illustré de manière remarquable.

A. Cornuéjols et L. Miclet. *Apprentissage artificiel : Concepts et algorithmes.* Eyrolles, 2002.

> La communauté française désigne cet ouvrage, comme étant la référence dans le domaine de l'apprentissage artificiel. Son écriture est claire et très pédagogique, le préambule de cet ouvrage a le mérite de nous faire rentrer en douceur dans ce domaine. Les auteurs décrivent un ensemble d'approches et d'algorithmes permettant de traiter tout type d'applications. La théorie, qui bien souvent, peut décourager les débutants, est abordée ici de manière claire et intelligible.

M. Huguier, A. Flahault. *Biostatistiques au quotidien.* Elsevier, 2nd édition, 2003.

> Cet ouvrage complété par un CD-Rom donne un état de l'art très complet sur les techniques biostatistiques, notamment les démarches diagnostiques, thérapeutiques, ainsi que leurs évaluations. Ce livre présente les méthodes en prenant soin de ne pas perdre le lecteur dans des développements mathématiques, les explications très claires sont agrémentées de nombreux exemples.

P. Siarry, J. Dréo, A. Pétrowski et É. Taillard. *Métaheuristiques pour l'optimisation difficile.* Eyrolles, 2003.

> Cet ouvrage présente les principales métaheuristiques telles que le recuit simulé, la recherche avec tabous, les colonies de fourmis, les algorithmes évolutionnaires et les algorithmes génétiques. Les descriptions théoriques sont accompagnées par des conseils méthodologiques, afin de permettre au lecteur de comparer et de choisir la méthode la plus adaptée à son problème. Ce livre est complété par de nombreuses études de cas détaillées.

[1] C. M. Bishop. *Pattern recognition and machine learning.* Springer, 2006.

Références bibliographiques

[Abdi, 1994] H. Abdi. *Les réseaux de neurones*. Presses universitaires de Grenoble, 1994.

[Amat and Yahiaoui, 2002] J.-L. Amat and G. Yahiaoui. *Techniques avancées pour le traitement de l'information*. Cépaduès - Editions, 2002.

[Andrews *et al.*, 1995] R. Andrews, J. Diederich, and A. B. Tickle. Survey and critique of technique for extracting rules from trained artificial neural networks. *Knowledge-Based Systems*, 8(6) :373–389, 1995.

[Antonini-Revaz *et al.*, 2004] S. Antonini-Revaz, F. P. Sarasin, and H. Stalder. Syncope. *Primary Care*, 4(23) :471–475, 2004.

[Asselin de Beauville and Kettaf, 2005] J.-P. Asselin de Beauville and F.-Z. Kettaf. *Bases théoriques pour l'apprentissage et la décision en reconnaissance des formes*. Cépaduès, 2005.

[Barnett and Lewis, 1994] V. Barnett and T. Lewis. *Outliers in statistical data*. John Wiley & Sons, 1994.

[Barron, 1993] A. Barron. Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, 39 :930–945, 1993.

[Baum and Haussler, 1988] E. B. Baum and D. Haussler. What size net gives valid generalization ? *Neural Computation*, 1 :151–160, 1988.

[Baux *et al.*, 1997] P. Baux, V. Dubreu, C. Kouackam, D. Dutoit, M. Goudemand, and S. Kacet. Syncopes inexpliquées présumées d'origine vagale ; relation avec les troubles psychiatriques. *L'Information Psychiatrique*, 73 :839–945, 1997.

[Belaïd and Belaïd, 1992] A. Belaïd and Y. Belaïd. *Reconnaissance des formes : Méthodes et applications*. InterEditions, 1992.

[Bellard *et al.*, 2003] E. Bellard, J. O. Fortrat, D. Schang, J. M. Dupuis, J. Victor, and G. Leftheriotis. Changes in the transthoracic impedance signal predict the outcome of 70° head-up tilt test. *Clinical Science*, 104(2) :119–126, 2003.

[Bellard, 2003] E. Bellard. *Détermination d'index prédictifs de la syncope neurocardiogénique : intérêt de l'analyse de l'hémodynamique centrale par impédancemétrie thoracique*. PhD thesis, Université d'Angers, 2003.

[Bellman, 1961] R. E. Bellman. *Adaptative control processes*. Princeton University Press, 1961.

[Benditt *et al.*, 1996] D. G. Benditt, D. W. Ferguson, B. P. Grubb, W. N. Kapoor, J. Kugler, B. B. Lerman, J. D. Maloney, A. Raviele, B. Ross, R. Sutton, M. J. Wolk, and D. L. Wood. Tilt table testing for assessing syncope. *Journal of the American College of Cardiology*, 28 :263–275, 1996.

[Bennani, 2001] Y. Bennani. Sélection de variables. *Revue d'intelligence artificielle*, 15(3-4) :303–316, 2001.

[Bernstein, 1986] D. P. Bernstein. A new stroke volume equation for thoracic electrical bioimpedance : theory and rationale. *Critical Care Medicine*, 14 :904–909, 1986.

[Bhatikar *et al.*, 2005] S. R. Bhatikar, C. DeGroff, and R. L. Mahajan. A classifier based on the artificial neural network approach for cardiologic auscultation in pediatrics. *Artificial Intelligence in Medicine*, 33(3) :251–260, 2005.

[Billat, 2003] V. Billat. *Physiologie et méthodologie de l'entraînement : de la théorie à la pratique*. De Boeck, 2003.

[Bishop *et al.*, 1997] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM : a principled alternative to the Self-Organizing Map. In *Advances in Neural Information Processing Systems*, volume 9, pages 354–360. Morgan Kaufmann, 1997.

[Bishop *et al.*, 1998] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM : the generative topographic mapping. *Neural Computation*, 10(1) :215–234, 1998.

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

[Bishop, 2006] C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[Blum and Langley, 1997] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97 :245–271, 1997.

[Boersma *et al.*, 2004] L. Boersma, L. Mont, A. Sionis, E. García, and J. Brugada. Value of the implantable loop recorder for the management of patients with unexplained syncope. *Europace*, 6 :70–76, 2004.

[Bonjer *et al.*, 1952] F. H. Bonjer, JW. Van Den Berg, and M. N. J. Dirken. The origin of the variations of body impedance occuring during the cardiac cycle. *Circulation*, 6 :415–420, 1952.

[Borg and Groenen, 2005] I. Borg and P. J. F. Groenen. *Modern multidimensional scaling : Theory and applications*. Springer, 2005.

[Bour *et al.*, 1994] J. Bour, D. Schang, and P. Notton. Apparatus for measuring and processing physiological signals and automatic method therefor. Patent PCT/FR94/00930, 1994.

[Brignole *et al.*, 2001] M. Brignole, P. Alboni, D. Benditt, L. Bergfeldt, J. J. Blanc, P. E. Bloch Thomsen, J. G. Van Dijk, A. Fitzpatrick, S. Hohnloser, J. Janousek, W. Kapoor, R. A, Kenny, P. Kulakowski, A. Moya, A. Raviele, R. Sutton, G. Theodorakis, and W. Wieling. Guidelines on management (diagnosis and treatment) of syncope. *European Heart Journal*, 22 :1256–1306, 2001.

[Brignole *et al.*, 2004] M. Brignole, P. Alboni, D. Benditt, L. Bergfeldt, J. J. Blanc, P. E. Bloch Thomsen, J. G. Van Dijk, A. Fitzpatrick, S. Hohnloser, J. Janousek, W. Kapoor, R. A, Kenny, P. Kulakowski, A. Moya, A. Raviele, R. Sutton, G. Theodorakis, and W. Wieling. Guidelines on management (diagnosis and treatment) of syncope - update 2004. *European Heart Journal*, 25 :2054–2072, 2004.

[Broyden, 1970] C. G. Broyden. The convergence of a class of double-rank minimization algorithms 2 : the new algorithm. *Journal of the Institue of Mathematics and its Applications*, 6 :1873–1896, 1970.

[Burden and Faires, 2001] R. L. Burden and J. D. Faires. *Numerical analysis*. 7th Ed, Brooks/Cole, 2001.

[Cattell, 1966] R. B. Cattell. The scree test for the number of factors. *Mulivariate Behavioral Research*, 1 :246–276, 1966.

[Charloux *et al.*, 2000] A. Charloux, E. Lonsdorfer-Wolf, R. Richard, E. Lampert, M. Oswald-Mammosser, B. Mettauer, B. Geny, and J. Lonsdorfer. A new impedance cardiograph device for the non-invasive evaluation of cardiac output at rest and during exercise : comparison with "direct" fick method. *Eur J Appl Physiol*, 82 :825–830, 2000.

[Christov, 2004] I. Christov. Real time electrocardiogram QRS detection using combined adaptive threshold. *BioMedical Engineering OnLine*, 3 :1–9, 2004.

[Coiera, 2003] E. Coiera. *A guide to health informatics*. Hodder & Stoughton Educational, UK, 2nd edition, 2003.

[Collette and Siarry, 2002] Y. Collette and P. Siarry. *Optimisation multiobjectif*. Eyrolles, 2002.

[Confais, 2003] N. Confais. *Statistique explicative appliquée*. Editions Technip, 2003.

[Cornuéjols and Miclet, 2002] A. Cornuéjols and L. Miclet. *Apprentissage artificiel : concepts et algorithmes*. Eyrolles, 2002.

[Cornuéjols, 2005] A. Cornuéjols. Apprentissage et circulation d'information. HDR, Université Paris-Sud, France, 2005.

[Cottet, 1997] F. Cottet. *Traitement des signaux et acquisition de données*. Dunod, 1997.

[Cottier, 2002] C. Cottier. Perte de connaissance de courte durée (syncopes) ; partie 1 : introduction à la problématique et stratégie d'investigation. *Forum Médical Suisse*, (18) :430–436, 2002.

[Cristianini and Shawe-Taylor, 2000] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[Cybenko, 1989] G. Cybenko. Approximation by superposition of a sigmoidal function. *Math. of Control, Signals and Systems*, 2(4) :303–314, 1989.

[Daigle, 2002] J.-M. Daigle. L'utilisation des courbes de ROC dans l'évaluation des tests diagnostiques de laboratoire clinique : application à l'étude de la pneumonite d'hypersensibilité. Master's thesis, Université de Laval, 2002.

[Darwin, 1859] C. Darwin. *On the Origin of Species*. John Murray, London, 1859.

[Das, 2001] S. Das. Filters, wrappers and a boosting based hybrid for feature selection,. In *Proceedings of International Conference on Machine Learning*, pages 74–81. Williams College, 2001.

[Delong *et al.*, 1988] E. R. Delong, D. M. Delong, and D. L. Clarke-Pearson. Comparing the areas under the two or more correlated receiver operating characteristic curves : a nonparametric approach. *Biometrics*, 44 :837–845, 1988.

[Demartines and Hérault, 1997] P. Demartines and J. Hérault. Curvilinear component analysis : A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1) :148–154, 1997.

[Demartines, 1992] P. Demartines. Mesures d'organization du réseau de kohonen. In *Congrès Satellite du Congrès Européen de Mathématiques : Aspects Théoriques des Réseaux de Neurones*, 1992.

[Demartines, 1994] P. Demartines. *Analyse de données par réseaux de neurones auto-organisés*. PhD thesis, Institut National Polytechnique de Grenoble, France, 1994.

[Denker *et al.*, 1987] J. Denker, B. Wittner, S. Solla, L. Jackel, and J. Hopfield. Large automatic learning, rule extraction, and generalization. *Complex Systems*, 1 :877–922, 1987.

[Diederich, 2008] Joachim Diederich, editor. *Rule Extraction from Support Vector Machines*, volume 80 of *Studies in Computational Intelligence*. Springer-Verlag, 2008.

[Do, 2006] T. T. Do. *Optimisation de forme en forgeage 3D*. PhD thesis, École des Mines de Paris, France, 2006.

[Doering *et al.*, 1995] L. Doering, E. Lum, K. Dracup, and A. Friedman. Predictors of between-method differences in cardiac output measurement using thoracic electrical bioimpedance and thermodilution. *Critical Care Medicine*, 23 :1667–1673, 1995.

[Domingos, 1997] P. Domingos. Context-sensitive feature selection for lazy learners. *Artificial Intelligence Review*, 11 :227–253, 1997.

[Donoho and Grimes, 2003] D. L. Donoho and C. Grimes. Hessian eigenmaps : locally linear embedding techniques for high-dimensional data. In *National Academy of Sciences of United States of America*, volume 100, pages 5591–5596, 2003.

[Dreyfus *et al.*, 2002] G. Dreyfus, J.-M. Martinez, M. Samuelides, M. B. Gordon, F. Badran, S. Thiria, and L. Hérault. *Réseaux de neurones : méthodologie et applications*. Eyrolles, 2002.

[Dréo *et al.*, 2003] J. Dréo, A. Pétrowski, P. Siarry, and E. D. Taillard. *Métaheuristiques pour l'optimisation difficile*. Eyrolles, 2003.

[Dubois, 2004] R. Dubois. *Application de nouvelles méthodes d'apprentissage à la détection précoce d'anomalies en électrocardiographie*. PhD thesis, Université Paris 6, 2004.

[Duda and Hart, 1973] R. O. Duda and P. E. Hart. *Pattern classification and scence analysis*. John Wiley and Sons, 1973.

[Duda *et al.*, 2001] R. O. Duda, P. E. Hart, and D. Stork. *Pattern classification*. Wiley, 2001.

[Dupont, 2002] W. D. Dupont. *Statistical modeling for biomedical researchers : a simple introduction to the analysis of complex data*. Cambridge University Press, 2002.

[Ebden *et al.*, 2004] M. J. Ebden, L. Tarassenko, S. J. Payne, A. Darowski, and J. D. Price. Time-frequency analysis of the ecg in the diagnosis of vasovagal syndrome in older people. In *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, pages 290–293, 2004.

[Edmunds *et al.*, 1982] A. T. Edmunds, S. Godfrey, and M. Tooley. Cardiac output measured by transthoracic impedance cardiography at rest, during exercise and at various lung volumes. *Clinical Science*, 63(2) :107–113, 1982.

[Efron, 1979] B. Efron. Bootstrap methods : another look at the jacknife. *The Annals of Statistics*, 7(1) :1–26, 1979.

[Fahlman and Lebiere, 1990] S. E. Fahlman and C. Lebiere. The cascade-correlation learning architecture. In *Advances in Neural Information Processing Systems*, volume 2, pages 524–532. Morgan Kaufmann, 1990.

[Fahlman, 1988] S. E. Fahlman. An empirical study of learning speed in backpropagation networks. Technical report, Pittsburgh, Carnegie Mellon University CMU-CS-88-162, 1988.

[Famili *et al.*, 1997] A. Famili, W.-M. Shen, R. Weber, and R. Simoudis. Data preprocessing and intelligent data analysis. *Intelligent Data Analysis*, 1(1) :3–23, 1997.

[Fawcett, 2005] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters, special issue on ROC analysis*, 27 :861–874, 2005.

[Fenton *et al.*, 2000] A. M. Fenton, S. C. Hammill, R. F. Rea, P. A. Low, and W. K. Shen. Vasovagal syncope. *Ann Intern Med*, 133 :714–725, 2000.

[Feuilloy *et al.*, 2005a] M. Feuilloy, D. Schang, J. O. Fortrat, and P. Nicolas. Early syncope prediction by a new neuronal approach. In *Proceedings of Computers in Cardiology*, 2005.

[Feuilloy *et al.*, 2005b] M. Feuilloy, D. Schang, J. O. Fortrat, S. Poggi, E. Bellard, and P. Nicolas. Prédiction précoce de la syncope chez l'homme par réseaux de neurones. In *Proceedings of the GRETSI*, 2005.

[Feuilloy *et al.*, 2005c] M. Feuilloy, D. Schang, P. Nicolas, J. O. Fortrat, and J. Victor. Dimension reduction methods for the early syncope prediction by artificial neural networks. In *Proceedings of the International Symposium on Signal Processing and its Applications*, 2005.

[Feuilloy *et al.*, 2006a] M. Feuilloy, D. Schang, and P. Nicolas. Comparison of feature selection methods for syncope prediction. In *Proceedings of the IEEE Congress on Evolutionary Computation*, 2006.

[Feuilloy *et al.*, 2006b] M. Feuilloy, D. Schang, and P. Nicolas. Optimization of the relevance of features extracted from transthoracic impedance signal. In *Proceedings of the 18th International EURASIP Conference BIOSIGNAL*, 2006.

[Feuilloy *et al.*, 2006c] M. Feuilloy, D. Schang, and P. Nicolas. A quick low cost method for syncope prediction. In *Proceedings of the 14th European Signal Processing Conference*, 2006.

[Feuilloy *et al.*, 2007] M. Feuilloy, D. Schang, and P. Nicolas. Extraction d'informations pour l'analyse en composantes curvilignes. In *Proceedings of the 39èmes Journées de Statistique*, 2007.

[Fisher, 1936] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 :179–188, 1936.

[Fitzpatrick *et al.*, 1991] A. P. Fitzpatrick, G. Theodorakis, S. Vardas, and R. Sutton. Methodology of head-up tilt testing in patients with unexplained syncope. *Journal of the American College of Cardiology*, 17 :125–130, 1991.

[Fletcher, 1987] R. Fletcher. *Practical methods of optimization.* John Wiley and Sons Ltd, second edition, 1987.

[Franzini *et al.*, 1990] M. Franzini, K. F. Lee, and A. Waibel. Connectionist viterbi training : a new hybrid method for continuous speech recognition. *IEEE Transaction on Neural Networks*, pages 425–428, 1990.

[Friesen *et al.*, 1990] G. M. Friesen, T. C. Jannett, M. A. Jadallah, S. L. Yates, S. R. Quint, and H. T. Nagle. A comparison of the noise sensitivity of nine QRS detection algorithms. *IEEE Transactions on Biomedical Engineering*, 37 :85–98, 1990.

[Fuller, 1992] H. D. Fuller. The validity of cardiac output measurement by thoracic impedance : a meta-analysis. *Clin Invest Med*, 15 :103–112, 1992.

[Furey *et al.*, 2000] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, 16(10) :906–914, 2000.

[Gabriel, 1971] K. R. Gabriel. The biplot graphical display of matrices with application to principal component analysis. *Biometrika*, 58 :453–467, 1971.

[Geman *et al.*, 1992] S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the biais/variance dilemma. *Neural Computation*, 4 :1–58, 1992.

[Georgin, 2002] J.-P. Georgin. *Analyse interactive des données (ACP, AFC) avec Excel 2000. Théorie et pratique.* Presses Universitaires de Rennes, 2002.

[Getchell *et al.*, 1999] W. S. Getchell, G. C. Larsen, C. D. Morris, and J. H. McAnulty. Epidemiology of syncope in hospitalized patients. *J Gen Intern Med*, 14 :677–687, 1999.

[Glover, 1989] F. Glover. Tabu search - part 1. *ORSA Journal on Computing*, 1 :190–206, 1989.

[Glover, 1990] F. Glover. Tabu search - part 2. *ORSA Journal on Computing*, 2 :4–32, 1990.

[Goldberg, 1991] D. E. Goldberg. *Genetic algorithms.* Addison-Wesley USA, 1991.

[Golub *et al.*, 1999] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Cliguiri, C. Bloomfield, and E. Lander. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286 :531–537, 1999.

[Goutte, 1997] C. Goutte. Note on free lunches and cross-validation. *Neural Computation*, 9 :1211–1215, 1997.

[Gower and Hand, 1996] J. C. Gower and D. J. Hand. *Biplots*. Chapman & Hall, 1996.

[Grandvalet *et al.*, 1997] Y. Grandvalet, S. Canu, and S. Boucheron. Noise injection : theoretical prospects. *Neural Computation*, 7 :1241–1256, 1997.

[Guérif, 2006] S. Guérif. *Réduction de dimension en apprentissage numérique non supervisé*. PhD thesis, Université Paris 13, 2006.

[Gutierrez-Osuna, 2002] R. Gutierrez-Osuna. Pattern analysis for machine olfaction : a review. *IEEE Sensors journal*, 2(3) :189–202, 2002.

[Guyon and Elisseeff, 2003] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3 :1157–1182, 2003.

[Hainsworth and Mark, 1993] R. Hainsworth and A. L. Mark. *Cardiovascular reflex control in health and disease*. W. B. Saunders, 1993.

[Hall, 2000] M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *International Conference on Machine Learning*, pages 359–366, 2000.

[Han and Kamber, 2006] J. Han and M. Kamber. *Data mining : concepts and techniques*. Morgan Kaufmann Publishers (second edition), 2006.

[Hanley and McNeil, 1982] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characterisitc (R.O.C.) curve. *Radiology*, 143(1) :29–36, 1982.

[Hanley and McNeil, 1983] J. A. Hanley and B. J. McNeil. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148 :839–843, 1983.

[Hassibi and Stork, 1993] B. Hassibi and D. G. Stork. Second order derivatives for netowrk pruning : optimal brain surgeon. In *Advances in Neural Information Processing Systems*, volume 5, pages 164–171. Morgan Kaufmann, 1993.

[Hassibi *et al.*, 1993] B. Hassibi, D.G. Stork, and G. J. Wolff. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, volume 1, pages 293–299, 1993.

[Hastie and Stuetzle, 1989] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84 :502–516, 1989.

[Hastie *et al.*, 2001] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics, 2001.

[Haykin, 1999] S. Haykin. *Neural networks : a comprehensive foundation*. Prentice Hall, 1999.

[Hebb, 1949] D. O. Hebb. *The organization of behavior*. New York, Wiley, 1949.

[Herault *et al.*, 1999] J. Herault, C. Jaussions-Picaud, and Guerin-Dugue. Curvilinear component analysis for high dimensional data representation : I. theoretical aspects and practical use in the presence of noise. In *International Work Conference on Artificial Neural Networks*, 1999.

[Hernández and Stolfo, 1998] M. A. Hernández and S. J. Stolfo. Real-world data is dirty : data cleansing and the merge/purge problem. *Data Mining Knowledge Discovery*, 2(1) :9–37, 1998.

[Holland, 1962] J. Holland. Outline for a logical theory of adaptative systems. *Journal of the Association of Computing Machinery*, 3, 1962.

[Holland, 1975] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, 1975.

[Holt, 2005] G. Holt. Clinical benchmarking for the validation of AI medical diagnostic classifiers. *Artificial Intelligence in Medicine*, 35 :259–260, 2005.

[Hopfield, 1982] J. J. Hopfield. Neural networks and systems with emergent collective computational abilities. In *Proceedings of the natural academy of sciences*, pages 2554–2558, 1982.

[Hornik *et al.*, 1989] K. Hornik, M. Stinchombe, and H White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2 :359–366, 1989.

[Hotteling, 1933] H. Hotteling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 :417–441, 498–520, 1933.

[Huguier and Flahault, 2003] M. Huguier and A. Flahault. *Biostatistiques au quotidien*. Elsevier, 2003.

[Illouz and Jardino, 2001] G. Illouz and M. Jardino. Analyse statistique et géométrique de corpus textuels. *Traitement automatique des langues et linguistique de corpus*, 42(2) :501–516, 2001.

[Jacobs, 1988] R. A. Jacobs. Increased rates of convergence through learning rate adaptation. *Neural Networks*, 1 :295–307, 1988.

[Jain and Zongker, 1997] A. Jain and D. Zongker. Feature selection : Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2) :153–158, 1997.

[Jain *et al.*, 2000] A. Jain, R. Duin, and J. Mao. Statistical pattern recognition : a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1) :4–37, 2000.

[Janikow, 1993] C. Z. Janikow. A knowledge-intensive genetic algorithm for supervised learning. *Machine Learning*, 13 :189–223, 1993.

[Jermyn *et al.*, 1999] P. Jermyn, M. Dixon, and B. J. Read. Preparing clean views of data for data mining. In *Proceedings of 12th ERCIM Workshop on Database Research*, 1999.

[Jodouin, 1994a] J.-F. Jodouin. *Les réseaux de neurones : principes et définitions*. Hermès, 1994.

[Jodouin, 1994b] J.-F. Jodouin. *Les réseaux neuromimétiques*. Hermès, 1994.

[Jolliffe, 2002] I. T. Jolliffe. *Principal component analysis*. Springer, 2002.

[Kaiser, 1961] H. F. Kaiser. A note on Guttman's lower bound for the number of common factors. *British Journal of Statistical Psychology*, 14 :1–2, 1961.

[Kapoor *et al.*, 1987] W. N. Kapoor, J. Peterson, H. S. Wieand, and M. Karpf. Diagnostic and prognostic implications of recurrences in patients with syncope. *Am J Med*, 83(4) :700–708, 1987.

[Kapoor, 1992] W. N. Kapoor. Evaluation and management of patients with syncope. *Journal of the American Medical Association*, 268 :2553–2560, 1992.

[Kapoor, 1995] W. N. Kapoor. Workup and management of patients with syncope. *The Medical clinics of North America*, 79 :1153–1170, 1995.

[Kapoor, 2000] W. N. Kapoor. Syncope. *N Engl J Med*, 343 :1856–1862, 2000.

[Karlis *et al.*, 2003] D. Karlis, G. Saporta, and A. Spinakis. A simple rule for the selection of principal components. *Communications in Statistics - Theory and Methods*, 32(3) :643–666, 2003.

[Karsai and Sztipanovits, 1999] G. Karsai and J. Sztipanovits. A model-based approach to self-adaptive software. *IEEE Intelligent Systems and their Applications*, 14(3) :46–53, 1999.

[Kay, 1993] S. M. Kay. *Fundamental of statistical signal processing - Estimation theory*. Prenctice Hall, pp 207–210, 1993.

[Kenny *et al.*, 1986] R. A. Kenny, J. Bayliss, A. Ingram, and R. Sutton. Head-up tilt : a useful test for investigating unexplained syncope. *Lancet*, 1 :1352–1354, 1986.

[Kirkpatrick *et al.*, 1983] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598) :671–680, 1983.

[Kohavi and John, 1997] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2) :273–324, 1997.

[Kohavi, 1995] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1995.

[Kohonen, 1982] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43 :59–69, 1982.

[Kohonen, 1995] T. Kohonen. *Self-organizing maps*. Springer-Verlag, 1995.

[Kononenko, 1994] I. Kononenko. Estimating attributes : analysis and extensions of relief. In *Proceedings of the European Conference on Machine Learning*, pages 171–182, 1994.

[Krahn *et al.*, 2003] A. D. Krahn, G. J. Klein, R. Yee, and A. C. Skanes. Use of the implantable loop recorder in patients with unexeplained syncope. *Minerva Cardionangiol*, 51 :21–27, 2003.

[Kubicek *et al.*, 1966] W. H. Kubicek, J. N. Karnegis, R. P. Patterson, D. A. Witsoe, and R. H. Matteson. Development and evaluation of an impedance cardiac output system. *Aerospace Med*, 37 :1208–1212, 1966.

[Kudo and Sklansky, 2000] M. Kudo and J. Sklansky. Comparison of algorithms that select features for pattern classifiers. *Pattern Recognition*, 33(1) :25–41, 2000.

[Kunt, 1981] M. Kunt. *Traitement numérique du signal*. Dunod (3 ème édition), 1981.

[Lababidi *et al.*, 1970] Z. Lababidi, D. A. Ehmke, R. E. Durnin, P. E. Leaverton, and R. N. Lauer. The first derivative thoracic impedance cardiogram. *Circulation*, 41 :651–658, 1970.

[Lachenbruch and Mickey, 1968] P. A. Lachenbruch and R. M. Mickey. Estimation of error rates in discriminant analysis. *Technometrics*, 10 :1–11, 1968.

[Landgrebe and Duin, 2007] T. C. W. Landgrebe and R. P. W. Duin. Approximating the multiclass ROC by pairwise analysis. *Pattern Recognition Letters*, (5), 2007.

[Landgrebe and Duin, 2008] T. C. W. Landgrebe and R. P. W. Duin. Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(5) :810–822, 2008.

[Langley, 1996] P. Langley. *Elements of machine learning*. Morgan Kaufmann, 1996.

[Lardeux *et al.*, 2006] F. Lardeux, F. Saubion, and J.-K. Hao. Gasat : a genetic local search algorithm for the satisfiability problem. *Evolutionary Computations*, 14(2) :223–253, 2006.

[Le Cun *et al.*, 1990] Y. Le Cun, J. S. Denker, and S. A. Solla. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2. Morgan Kaufmann, 1990.

[Lebart *et al.*, 2006] L. Lebart, M. Piron, and A. Morineau. *Statistique exploratoire multidimensionnelle : visualisation et inférence en fouille de données*. Dunod, 2006.

[Lee *et al.*, 2000] J. A. Lee, A. Lendasse, N. Donckers, and M. Verleysen. A robust nonlinear projection method. In *In European Symposium on Artificial Neural Networks*, pages 13–20, 2000.

[Lee *et al.*, 2004] J. A. Lee, A. Lendasse, and M. Verleysen. Nonlinear projection with curvilinear distances : Isomap versus curvilinear distance analysis. *Neurocomputing*, 57 :49–76, 2004.

[Levenberg, 1944] K. Levenberg. A method for the solution of certain nonlinear problems in least square. *Quarterly of Applied Mathematics*, 2 :164–168, 1944.

[Linzer *et al.*, 1997] M. Linzer, E. H. Yang, N.A. Estes, P. Wang, V. R. Vorperian, and W. N. Kapoor. Diagnosing syncope. Part 1 : Value of history, physical examination, and electrocardiography. *Annals of Internal Medicine*, 126(12) :989–996, 1997.

[Liu and Motoda, 1998] H. Liu and H. Motoda. Feature transformation and subset selection. *IEEE Intelligent Systems*, 13(2) :26–28, 1998.

[Liu and Yu, 2002] H. Liu and L. Yu. Feature selection for data mining. Technical report, Arizona State University, 2002.

[Liu and Yu, 2005] H. Liu and L. Yu. Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4) :491–502, 2005.

[Liu *et al.*, 1998] H. Liu, L. Huan, and H. Motoda. *Feature selection for knowledge discovery and data mining.* Springer, 1998.

[Loosli *et al.*, 2006] G. Loosli, S.-G. Lee, V. Guigue, S. Canu, and A. Rakotomamonjy. Perception d'états affectifs et apprentissage. *Revue d'intelligence artificielle*, 20 :553–582, 2006.

[Loughrey and Cunningham, 2005] J. Loughrey and P. Cunningham. Using early-stopping to avoid overfitting in wrapper-based feature selection employing stochastic search. In *Proceedings of the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 2005.

[Lu, 2005] C. Lu. *Probabilistic machine learning approaches to medical classification problems.* PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), 2005.

[Lukas, 2003] L. Lukas. *Least squares support vector machines classification applied to brain tumor recognition using magnetic resonance spectroscopy.* PhD thesis, Faculty of Engineering, K.U.Leuven (Leuven, Belgium), 2003.

[Mallat *et al.*, 1997] Z. Mallat, E. Vicaut, A. Sangaré, J. Verscueren, G. Fontaine, and R. Frank. Prediction of head-up tilt test result by analysis of early heart rate variations. *Circulation*, 96 :581–584, 1997.

[Malliani, 1999] A. Malliani. The pattern of sympathovagal balance explored in the frequency domain. *News in Physiological Sciences*, 14 :111–117, 1999.

[Malmivuo and Plonsey, 1995] J. Malmivuo and R. Plonsey. *Bioelectromagnetism.* Oxford University Press, 1995.

[Manolis *et al.*, 1990] A. S. Manolis, M. Linzer, D. Salem, and N. A. Estes. Syncope : current diagnostic evaluation and management. *Annals of Internal Medicine*, 112 :850–863, 1990.

[Marquardt, 1963] D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11 :431–441, 1963.

[Matheus, 1991] C. Matheus. The need for constructive induction. In *Machine Learning - Proceedings of the Eighth International Workshop*, pages 173–177, 1991.

[McCulloch and Pitts, 1943] W. S. McCulloch and W. S. Pitts. A logical calculus of ideas imminent in nervous activity. *Bulletin of Mathematical Biophysics*, 5 :115–133, 1943.

[McNeil and Hanley, 1984] B. J. McNeil and J. A. Hanley. Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making*, 4(1) :137–150, 1984.

[Mercier *et al.*, 2004] G. Mercier, N. Berthault, J. Mary, J. Peyre, A. Antoniadis, J.-P. Comet, A. Cornuéjols, C. Froidevaux, and M. Dutreix. Biological detection of low radiation doses by combining results of two microarray analysis methods. *Nucleic Acids Research*, 32(1), 2004.

[Metz, 1978] C. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 3(4), 1978.

[Miller, 1994] R. A. Miller. Medical diagnostic decision support systems–past, present, and future : a threaded bibliography and brief commentary. *Journal of the American Medical Informatics Association*, 1(1) :8–27, 1994.

[Minsky and Papert, 1969] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969.

[Moerland, 2000] P. Moerland. *Mixture models for unsupervised and supervised learning*. PhD thesis, Ecole Polytechnique de Lausanne, Suisse, 2000.

[Murata *et al.*, 1994] N. M. Murata, S. Yoshizawa, and S. Amari. Network information criterion-determining the number of hidden units for an artificial neural network model. *IEEE Transactions on Neural Networks*, 5(6) :865–872, 1994.

[Mustawi *et al.*, 1992] M. T. Mustawi, W. Ahmed, K. H. Chan, K. B. Faris, and D. M. Hummels. On the training of radial basis function classifiers. *Neural Networks*, 5 :595–603, 1992.

[Nagy, 1968] G. Nagy. State of the art in pattern recognition. *Proceedings of the IEEE*, 56(5) :836–862, 1968.

[Narendra and Fukunaga, 1977] P. Narendra and K. Fukunaga. A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, 26 :917–922, 1977.

[Nasser, 2007] A. Nasser. *Contribution à la classification non-supervisée par machines à noyaux*. PhD thesis, ULCO- Centre Universitaire de la Mi-Voix, Calais, 2007.

[Natale *et al.*, 1998] A. Natale, J. Sra, M. Akhtar, L. Kusmirek, G. Tomassoni, F. Leonelli, K. Newby, S. Beheiry, and A. Pacifico. Use of sublingual nitroglycerin during head-up tilt-table testing in patients $>$ 60 years of age. *The American Journal of Cardiology*, 82(10) :1210–1213, 1998.

[Newby and Grubb, 2006] D. E. Newby and N. R. Grubb. *Cardiologie*. traduit par N. Mansencal ; Elsevier, 2006.

[Newman and Calister, 1999] D. G. Newman and R. Calister. The non-invasive assessment of stroke volume and cardiac output by impedance cardiography : a review. *Aviation, Space, and Environmental Medicine*, 70 :780–789, 1999.

[Newman *et al.*, 1998] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz. UCI repository of machine learning databases : http ://www.ics.uci.edu/∼mlearn/mlrepository.html, 1998.

[Nyober *et al.*, 1940] J. Nyober, S. Bagno, and A. Barnet. Radiocardiograms - the electrical impedance changes of the heart in relation to electrocardiorganms and heart sounds. *Journal of Clinical Investigation*, 19 :963–966, 1940.

[Nyober, 1959] J. Nyober. *Electrical impedance Plethysmography*. Springfield, IL : C.C. Thomas, 1959.

[Oppenheim and Schafer, 1989] A. V. Oppenheim and R. W. Schafer. *Discrete-time signal processing*. Prentice-Hall, 1989.

[Pagani *et al.*, 1986] M. Pagani, F. Lombardi, S. Guzzetti, O. Rimoldi, R. Furlan, P. Pizzinelli, G. Sandrone, G. Malfatto, S. Dell'Orto, and E. Piccaluga. Power spectral analysis of heart rate and arterial pressure variabilities as a marker of sympatho-vagal interaction in man and conscious dog. *Circulation Research*, 59(2) :178–193, 1986.

[Pan and Tompkins, 1985] J. Pan and W. J. Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, 32(3) :230–236, 1985.

[Parzen, 1962] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33 :1065–1076, 1962.

[Pearlmutter, 1992] B. A. Pearlmutter. Gradient descent : second-order momentum and saturating error. In *Advances in Neural Information Processing Systems*, volume 4, pages 887–894. Morgan Kaufmann, 1992.

[Pearson, 1901] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, Sixth Series*, 2 :559–572, 1901.

[Perez-Paredes *et al.*, 1999] M. Perez-Paredes, F. Pico-Aracil, R. Florenciano, J. G. Sanchez-Villanueva, J. A. Ruiz Ros, and J. A. Ruiperez. Head-up tilt test in patients with high pretest likelihood of neurally mediated syncope : an approximation to the "real sensitivity" of this testing. *Pacing and clinical electrophysiology*, 22(8) :1173–1178, 1999.

[Personnaz and Rivals, 2003] L. Personnaz and I. Rivals. *Réseaux de neurones formels pour la modélisation, la commande et la classification*. CNRS Editions, 2003.

[Pitzalis *et al.*, 2002] M. Pitzalis, F. Massari, P. Guida, M. Iacoviello, F. Mastropasqua, B. Rizzon, C. Forleo, and P. Rizzon. Shortened head-up tilting test guided by sistolic pressure reduction in neurocardiogenic syncope. *Circulation*, 105 :146–148, 2002.

[Prechelt, 1998] L. Prechelt. Automatic early stopping using cross validation : quantifying the criteria. *Neural Networks*, 11(4) :761–767, 1998.

[Press *et al.*, 1992] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipies in C : The Art of Scientific Computing*. Cambridge University Press, 1992.

[Pyle, 1999] D. Pyle. *Data preparation for data mining*. Morgan Kaufmann Publishers, 1999.

[Ratle and Sebag, 2000] A. Ratle and M. Sebag. Genetic programming and domain knowledge : Beyond the limitations of grammar-guided machine discovery. In Springer Verlag, editor, *Proceedings of 6th International Conference on Parallel Problem Solving From Nature*, pages 211–220, 2000.

[Rennard, 2006] J.-P. Rennard. *Réseaux neuronaux : une introduction accompagnée d'un modèle Java*. Vuilbert, Paris, 2006.

[Riedmiller and Braun, 1993] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning : the RPROP algorithm. In *Proceedings of the IEEE International Conference on Neural Networks*, 1993.

[Robnik-Sikonja and Kononenko, 2003] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53(1-2) :23–69, 2003.

[Rosenblatt, 1958] F. Rosenblatt. The perceptron : a probabilistic model for information storage and organization in the brain. *Phychological Review*, 65 :386, 1958.

[Rosenblatt, 1962] F. Rosenblatt. *Principles of neurodynamics : perceptrons and the theory of brain mechanisms*. Washington D.C. : Spartan Press, 1962.

[Roweis and Sam, 2000] S. T. Roweis and L. K. Sam. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500) :2323–2326, 2000.

[Rudolph, 1997] S. Rudolph. On topology, size and generalization of non-linear feed-forward neural networks. *Neurocomputing*, 16 :1–22, 1997.

[Rumelhart *et al.*, 1986] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing*, volume 1, pages 318–362. Cambridge, MA : MIT Press, 1986.

[Sahai, 2000] H. Sahai. *The analysis of variance*. Birkhauser, Boston, MA, 2000.

[Salamé *et al.*, 2006] E. Salamé, R. Neemtallah, R. Azar, S. Antonios, C. Jazra, and R. Kassab. Quel agent pharmacologique utiliser pour la sensibilisation du test d'inclinaison dans l'évaluation des syncopes d'origine indéterminée ? *Annales de Cardiologie et d'Angéiologie*, 55(3) :135–139, 2006.

[Sammon Jr, 1969] J. W Sammon Jr. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computation*, 18 :401–409, 1969.

[Saporta, 2006] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, Paris, 2006.

[Saul and Roweis, 2003] L. K. Saul and S. T. Roweis. Think globally, fit locally : unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4 :119–155, 2003.

[Savage *et al.*, 1985] D. D. Savage, L. Corwin, D. L. Mcgee, W. B. Kannel, and P. A Wolf. Epidemiologic features of isolated syncope : The framingham study. *Stroke*, 16(4) :626–629, 1985.

[Schang *et al.*, 2003] D. Schang, G. Plantier, E. Bellard, and G. Leftheriotis. Prévention de la syncope chez l'homme. In *Proceedings of the GRETSI*, 2003.

[Schang *et al.*, 2006] D. Schang, E. Bellard, G. Plantier, J. M. Dupuis, J. Victor, and G. Leftheriotis. Comparison of computational algorithms applied on transthoracic impedance waveforms to predict head-up tilt table testing outcome. *Computers in Biology and Medicine*, 36(3) :225–240, 2006.

[Schang *et al.*, 2007] D. Schang, M. Feuilloy, G. Plantier, J. O. Fortrat, and P. Nicolas. Early prediction of unexplained syncope by support vector machines. *Physiological Measurement*, 28 :185–197, 2007.

[Scholkopf and Smola, 2002] B. Scholkopf and A. J. Smola. *Learning with kernels*. MIT Press, 2002.

[Scholkopf *et al.*, 1998] B. Scholkopf, A. J. Smola, and K. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10 :1299–1319, 1998.

[Sebban and Nock, 2002] M. Sebban and R. Nock. A hybrid filter/wrapper aproach of feature selection using information theory. *Journal of Pattern Recognition*, 35(4) :835–846, 2002.

[Secher *et al.*, 1979] N. J. Secher, P. Arnsbo, L. Heslet Andersen, and A. Thomsen. Measurements of cardiac stroke volume in various body positions in pregnancy and during caesarean section : a comparison between thermodilution and impedance cardiography. *Scandinavian Journal of Clinical and Laboratory Investigation*, 39(6) :569–576, 1979.

[Seidl *et al.*, 2000] K. Seidl, M. Rameken, S. Breunung, J. Senges, W. Jung, D. Andresen, A. Van Toor, A. D. Krahn, and G. J. Klein. Diagnostic assessment of recurrent unexplained syncope with a new subcutaneously implantable loop recorder. *Europace*, 2 :256–262, 2000.

[Semani *et al.*, 2004] D. Semani, C. Frélicot, and P. Courtellemont. Combinaison d'étiquettes floues/possibilistes pour la sélection de variables. In *RFIA*, 2004.

[Shawe-Taylor and Cristianini, 2004] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[Shepard, 1962] R. N. Shepard. The analysis of proximities : Multidimensional scaling with an unknown distance function. *Psychometrika*, 27(2) :125–140, 1962.

[Shoemaker *et al.*, 1994] W. C. Shoemaker, C. C. J. Wo, M. H. Bishop, P. L. Appel, J. M. Van de Water, G. R. Harrington, W. Xiang, and R. S. Patil. Multicenter trial of a new thoracic electrical bioimpedance device for cardiac output estimation. *Critical Care Medicine*, 22 :1907–1912, 1994.

[Smith and Cornell, 1993] W. F. Smith and J. A. Cornell. Biplot displays for looking at multiple response data in mixture experiments. *Technometrics*, 35(4) :337–350, 1993.

[Smith *et al.*, 1970] J. J. Smith, J. E. Bush, V. T. Wiedmeier, and F. R. Tristani. Application of impedance cardiography of the study of postural stress in the human. *Journal of Applied Physiology*, 29 :274–290, 1970.

[Srinivasan, 1999] A. Srinivasan. Note on the location of optimal classifiers in n-dimensional roc space prg-tr-2-99. Technical report, Oxford University Computing Laboratory Oxford, England, 1999.

[Stearns, 1976] S. D. Stearns. On selecting features for pattern classifiers. In *Proceedings of International Conference on Pattern Recognition*, pages 71–75, 1976.

[Stone, 1974] M. Stone. Cross-validatory choice and assessment of statistical prediction. *Journal of the Royal Statistical Society, B*, 36(1) :111–147, 1974.

[Stoppiglia *et al.*, 2003] H. Stoppiglia, G. Dreyfus, R. Dubois, and Y. Oussar. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3 :1399–1414, 2003.

[Stoppiglia, 1997] H. Stoppiglia. *Méthodes statistiques de sélection de modèles neuronaux; applications financières et bancaires*. PhD thesis, Université Pierre et Marie Curie, 1997.

[Strasberg *et al.*, 1989] B. Strasberg, E. Rechavia, A. Sagie, J. Kusniec, A. Mager, S. Sclarovsky, and J. Agmon. The head-up tilt table test in patients with syncope of unknown origin. *American Heart Journal*, 118 :923–927, 1989.

[Suykens *et al.*, 2002] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least squares support vector machines*. World Scientific Pub. Co., Singapore„ 2002.

[Tenenbaum *et al.*, 2000] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500) :2319–2323, 2000.

[Theodoridis and Koutroumbas, 2006] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 2006.

[Thiria *et al.*, 1997] S. Thiria, Y. Lechevallier, O. Gascuel, and S. Canu. *Statistique et méthodes neuronales*. Dunod, 1997.

[Torgeson, 1952] W. S. Torgeson. Multidimensional scalling : I : theory and method. *Psychometrika*, 17 :401–419, 1952.

[Tou and Gonzalez, 1974] J. T. Tou and R. C. Gonzalez. *Pattern recognition principles*. Addison-Wesley, 1974.

[Tufféry, 2007] S. Tufféry. *Data mining et statistique décisionnelle*. Editions Technip, 2007.

[Vapnik, 1995] V. N. Vapnik. *The nature of statistical learning theory*. Springer, 1995.

[Vapnik, 1998] V. N. Vapnik. *Statistical learning theory*. Wiley, 1998.

[Venturini, 1994] G. Venturini. *Apprentissage adaptatif et apprentissage supervisé par algorithme génétique*. PhD thesis, Université de Paris 11, Orsay, France, 1994.

[Webb, 2002] A. R. Webb. *Statistical pattern recognition*. John Wiley & Sons (2nd edition), 2002.

[Welch, 1967] P. D. Welch. The use of Fast Fourier Transform for the estimation of power spectra : a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15 :70–73, 1967.

[Widrow and Hoff, 1960] B. Widrow and M. E. Hoff. Adaptive switching circuits. In *IRE WESCON Convention Record*, volume 4, pages 96–104, 1960.

[Witten and Frank, 2005] I.H. Witten and E. Frank. *Data mining : practical machine learning tools and techniques*. Morgan Kaufmann Publishers (second edition), 2005.

[Wnek and Michalski, 1994] J. Wnek and R. S. Michalski. Hypothesis-driven constructive induction in aq17-hci : a method and experiments. *Machine Learning*, 14 :139–168, 1994.

[Xing *et al.*, 2001] E. P. Xing, M. I. Jordan, and R. M. Karp. Feature selection for high-dimensional genomic microarray data. In *International Conference on Machine Learning*, pages 601–608, 2001.

[Yammanouchi *et al.*, 1996] Y. Yammanouchi, S. Jaalouk, A. A. Shehadeh, F. Jaeger, H. Goren, and F. M. Fouadtarazi. Changes in left ventricular volume during head-up tilt in patients with vasovagal syncope : an echographic study. *American Heart Journal*, 131 :73–80, 1996.

[Yang and Honavar, 1997] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. In *Proceedings of the Genetic Programming : Proceedings of the Second Annual Conference*, 1997.

[Yang and Honavar, 1998] J. Yang and V. Honavar. Feature subset selection using a genetic algorithm. *IEEE Intelligent Systems*, 13(2) :44–49, 1998.

[Yu and Liu, 2003] L. Yu and H. Liu. Feature selection for high-dimensional data : a fast correlation-based filter solution. In *Proceedings of International Conference on Machine Learning*, pages 856–863, 2003.

[Yu and Liu, 2004] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5 :1205–1224, 2004.

[Zhou and Jiang, 2003] Z. H. Zhou and Y. Jiang. Medical diagnosis with c4.5 rule preceded by artificial neural network ensemble. *IEEE Transactions on Information Technology in Biomedicine*, 7(1) :37–42, 2003.

Étude d'algorithmes d'apprentissage artificiel pour la prédiction de la syncope chez l'homme

Résumé

La syncope, dont l'origine peut ne pas être clairement définie, est considérée comme une pathologie fréquente. Dans ce cas et lorsque les épisodes sont répétés, le patient peut être amené à réaliser le test de la table d'inclinaison. Cet examen appelé *tilt-test*, est une méthode reconnue pour recréer les conditions dans lesquelles le patient ressent les symptômes de la syncope. Cependant, le principal problème de ce test est sa durée, qui peut atteindre une heure. Dès lors, pour des raisons de coût et de bien-être des patients, il paraît important de pouvoir réduire sa durée. C'est dans cet objectif que s'inscrivent les travaux réalisés dans le cadre de cette thèse, qui tentent de prédire l'apparition des symptômes liés à la syncope, et ce, le plus tôt possible.

Durant nos recherches, deux axes sont ressortis naturellement : la fouille de données et le développement de modèles capables de prédire le résultat du *tilt-test*. Ces deux axes partagent des méthodes issues de l'apprentissage artificiel, qui permettent d'acquérir et d'extraire des connaissances à partir d'un ensemble d'observations significatif. La littérature propose tout un ensemble de méthodes, qui nous ont permis de mettre en évidence certaines caractéristiques pertinentes, de manière à construire des modèles parcimonieux et robustes. Ces derniers ont permis d'obtenir des résultats intéressants pour la prédiction du résultat du *tilt-test* au terme notamment, des dix premières minutes de l'examen. Ces performances ont pu être considérablement améliorées par le développement de nouvelles techniques de fouille de données, permettant d'extraire très efficacement de la connaissance. Les méthodes mises en place s'articulent autour de la sélection de variables et de l'interprétation de projections non linéaires. Ces méthodes, bien que développées autour de notre thématique, se sont montrées reproductibles lors de tests sur d'autres ensembles de données.

Mots-clés : apprentissage artificiel, fouille de données, signal d'impédancemétrie thoracique, syncope, test de la table d'inclinaison

Study of machine learning algorithms for syncope prediction

Abstract

Syncope is considered as a common pathology, although sometimes its cause cannot be clearly diagnosed. In this case and when syncope is frequently experienced, the patient can have a head-upright tilt test. This examination is based on the reproduction of symptoms of the syncope ; however, its major drawback is the duration of the examination, which can take up to one hour. Therefore, reducing the examination time would decrease its cost and improve the comfort of the patient. This is the challenge of this thesis, which tries to predict the appearance of the symptoms of the syncope before the end of the tilt test.

During the research, two areas of study became important : data mining and development of models used to predict the tilt-test result. Both areas use algorithms coming from machine learning, enabling the acquisition and extraction of relevant knowledge on data sets. Published works give many methods, which have enabled the extraction of some pertinent characteristics. With these, robust and efficient models have been constructed, which have enabled the prediction of the tilt-test results in the first ten minutes of the examination. Also, the performance has been improved by the development of new techniques of data mining, enabling more efficient analysis of data. These methods have been used for the selection of the variables and for the interpretation of the non-linear projection techniques. Even though these methods have been developed for this research, they have shown interesting performances during tests on other data sets.

Keywords : data mining, machine learning, syncope, transthoracic impedance signal, head-upright tilt test