N°ORDRE : 40105

Université des Sciences et Technologies Lille 1,
École Doctorale régionale Sciences Pour l'Ingénieur
Lille Nord-de-France

Thèse pour obtenir le grade de Docteur en Sciences de l'Université Lille 1

présentée par **Vincent Vandewalle**
le 09 décembre 2009

Discipline : **Mathématiques Appliquées**

Estimation et sélection en classification semi-supervisée

Directeurs de Thèse : Christophe Biernacki et Gilles Celeux

Jury :

Directeur	Christophe Biernacki	Professeur	Université de Lille 1
Directeur	Gilles Celeux	Dir. Recherche	INRIA
Rapporteur	Didier Chauveau	Professeur	Université d'Orléans
Rapporteur	Jean-Jacques Daudin	Professeur	AgroParisTech
Président	Gérard Govaert	Professeur	Université de Technologie de Compiègne

Résumé

Le sujet de cette thèse est la classification semi-supervisée qui est considérée d'un point de vue décisionnel. Nous nous intéressons à la question de choix de modèles dans ce contexte où les modèles sont estimés en utilisant conjointement des données étiquetées et des données non étiquetées plus nombreuses. Nous concentrons notre recherche sur les modèles génératifs où la classification semi-supervisée s'envisage sans difficulté, contrairement au cadre prédictif qui nécessite des hypothèses supplémentaires peu naturelles.

Après avoir dressé un état de l'art de la classification semi-supervisée, nous décrivons l'estimation des paramètres d'un modèle de classification à l'aide de données étiquetées et non étiquetées par l'algorithme EM. Nos contributions sur la sélection de modèles font l'objet des deux chapitres suivants. Au chapitre 3, nous présentons un test statistique où les données non étiquetées sont utilisées pour mettre à l'épreuve le modèle utilisé. Au chapitre 4 nous présentons un critère de sélection de modèles AIC_{cond}, dérivé du critère AIC d'un point de vue prédictif. Nous prouvons la convergence asymptotique de ce critère particulièrement bien adapté au contexte semi-supervisé et ses bonnes performances pratiques comparé à la validation croisée et à d'autres critères de vraisemblance pénalisée.

Une deuxième partie de la thèse, sans rapport direct avec le contexte semi-supervisé, présente des modèles multinomiaux pour la classification sur variables qualitatives. Nous avons conçu ces modèles pour répondre à des limitations des modèles multinomiaux parcimonieux proposés dans le logiciel MIXMOD. À cette occasion, nous proposons un critère type BIC qui prend en compte de manière spécifique la complexité de ces modèles multinomiaux contraints.

Mots clés : modèles de mélange, estimation par maximum de vraisemblance, données manquantes, algorithme EM, analyse discriminante, classification semi-supervisée, modèles parcimonieux, choix de modèle.

Abstract

Estimation and selection in semi-supervised classification

The subject of this thesis is the semi-supervised classification which is considered in decision-making perpective. We are interested in model choice issue when models are estimated using both labeled data and many unlabeled data. We focus our research on generative models for which the semi-supervised classification is considered without difficulty, unlike predictive framework that requires additional unnatural assumptions.

Having developed a state of the art of semi-supervised classification, we describe the parameters estimation of a classification model using labeled data and unlabeled data by the EM algorithm. Our contributions on models selection is closely-watched in the two following chapters. In Chapter 3, we present a statistical test where unlabeled data are used to test the model. In Chapter 4 we present a model selection criterion, AIC_{cond}, derived from the AIC criterion in a predictive point of view. We prove the asymptotic convergence of this criterion particularly well suited to semi-supervised setting and his good practical performance compared to the cross-validation and other penalized likelihood criteria.

A second part of the thesis, not directly connected with the semi-supervised setting, the multinomial models for classification of qualitative variables are considered. We designed these models to address the limitations of parsimonious multinomial models proposed in the program MIXMOD. For this setting, we propose a BIC-type criterion which takes into account specifically the complexity of the constrained multinomial models.

Keywords: mixture models, maximum likelihood estimation, missing data, EM algorithm, discriminant analysis, semi-supervised classification, parsimonious models, model choice.

Remerciements

Je tiens tout d'abord à remercier Christophe Biernacki pour tout ce qu'il m'a apporté durant ces trois années de thèse tant au niveau scientifique qu'au niveau humain. Je le remercie pour sa disponibilité et la confiance dont il a fait preuve. Dans les périodes difficiles, il m'a toujours offert du temps et de précieux conseils pour m'aider à avancer. Je remercie également Gilles Celeux pour son esprit critique et son soutien. Je n'oublierai pas nos réunions à l'Institut Henri Poincaré lieu de débats parfois houleux, souvent intenses mais toujours constructifs.

Je remercie très sincèrement Didier Chauveau et Jean-Jacques Daudin d'avoir consacré du temps à mes travaux en acceptant la mission de rapporteurs. Leurs deux lectures attentives de ma thèse m'ont permis de dégager de nouvelles perspectives intéressantes, que je ne manquerai pas d'exploiter dans ma nouvelle vie de jeune chercheur.

Un grand merci à Gérard Govaert pour avoir présidé mon jury de thèse mais aussi et surtout pour l'ensemble de ses remarques et pour son point de vue critique sur les problèmes de classification.

Je remercie la région Nord-Pas de Calais et l'INRIA pour le financement de ma thèse et les bonnes conditions matérielles dont j'ai bénéficié durant ces trois années. Je remercie en particulier les membres du projet Select de l'INRIA pour leur accueil et notamment pour la disponibilité et l'efficacité des assistantes de projet.

La période passée au Laboratoire Paul Painlevé m'a beaucoup apporté. Je ne démentirai pas la réputation des ch'tis en matière d'accueil. Je tiens à exprimer ma reconnaissance à l'équipe probabilités et statistique et j'en profite pour saluer tout le personnel administratif pour sa gentillesse.

Je tiens à remercier l'Ecole Doctorale des Sciences Pour l'Ingénieur pour la qualité des formations proposées.

Je remercie vivement Claire Bornais de Polytech'Lille et Fatma Bouali du DUT STID de Roubaix de m'avoir permis d'effectuer des vacations pendant ma thèse. Je remercie également l'IUT de Roubaix pour le poste d'ATER qui m'a permis de terminer ma thèse dans de bonnes conditions.

Enfin, je remercie les thésards du laboratoire Paul Painlevé pour leur soutien, Benoît, Alexis, Bénédicte, Alexandre, Shuyan, Eric, Anne, Martin, Julien, Léon, Chang et tous les autres. L'ambiance des fameux repas au RU me manquera.

Je remercie ma famille et mes amis pour leurs encouragements.

Cette thèse est avant tout dédiée à celle qui est devenue mon épouse. Je remercie Aurélie pour son soutien au quotidien, son amour et la patience dont elle a fait preuve dans sa croisade contre les fautes d'orthographe disséminées dans ce mémoire. Enfin bien sûr, je remercie mon petit Timéo d'avoir rapidement fait ses nuits pour faciliter la rédaction du mémoire de son papa.

Table des matières

Principales notations et abréviations

Notations

g	Nombre de classes
$\boldsymbol{z}$	Vecteur de la classe codée de façon disjonctive
$\boldsymbol{Z}$	Vecteur aléatoire de la classe
$\mathcal{Z}$	Espace de la classe
$\boldsymbol{x}$	Vecteur des covariables observé
$\boldsymbol{X}$	Vecteur aléatoire des covariables
$\mathcal{X}$	Espace des covariables
d	Dimension de l'espace $\mathcal{X}$
θ	Vecteur des paramètres
$p(\cdot, \theta)$	Modèle paramétré par θ à interpréter selon ses arguments
$\delta(\cdot; \theta)$	Règle de classement déduire à partir de θ
$\mathbf{x}_u$	Echantillon des données non classées
$\mathbf{x}_\ell, \mathbf{z}_\ell$	Echantillon des données classées
$\mathbf{x} = (\mathbf{x}_\ell, \mathbf{x}_u)$	Echantillon des covariables
n_ℓ, n_u, n	Nombres de données étiquetées, non étiquetées, total
Θ	Espace des paramètres
$\hat{\theta}_{\mathcal{D}}$	EMV de θ à partir de l'échantillon $\mathcal{D}$
S	Variable aléatoire qui indique si l'étiquette est observée
β	Fractions de données étiquetées
θ^*	Vraie valeur du paramètre θ

Abréviations

LDA	Analyse discriminante linéaire
QDA	Analyse discriminante quadratique
SVM	Supports à vastes marges
MAP	Maximum *a posteriori*
EMV	Estimateur du maximum de vraisemblance
i.i.d.	Indépendent et Identiquement Distribué
MCAR	Manquant complètement au hasard
MAR	Manquant au hasard
MNAR	Ne manquant pas au hasard

Introduction

Dans ce mémoire, nous traitons de la classification semi-supervisée. Dans Chapelle *et al.* (2006), elle est définie de la façon suivante :

Semi-supervised learning is half between supervised and unsupervised learning.

C'est-à-dire qu'elle se situe à mi-chemin entre la classification supervisée et la classification non supervisée. Cette définition nous laisse la liberté, ainsi que le devoir, de positionner le curseur plutôt vers la classification supervisée ou plutôt vers la classification non supervisée. La spécificité de la classification semi-supervisée est que l'échantillon d'apprentissage est constitué à la fois de données étiquetées et non étiquetées. Ce type de donnés pose alors des questions de classification supervisée, des questions de classification non supervisée, ainsi que des questions intermédiaires.

D'abord, nous illustrons les situations où des échantillons de données partiellement étiquetées peuvent être rencontrés. Ensuite, nous introduisons les questions suscitées par ce type de données. Enfin, nous annonçons le plan de la thèse.

Pourquoi des données partiellement étiquetées ?

Les méthodes modernes d'acquisition automatique permettent d'obtenir de nombreuses variables sur de nombreux individus pour un faible coût. Toutefois, la variable d'intérêt est souvent plus difficile à obtenir que les autres. Ceci est particulièrement vrai dans les problèmes de prédiction. Dans ce cas, il est souhaitable d'apprendre une règle qui permette de prédire la variable d'intérêt étant donné un ensemble d'autres variables obtenues à un coût réduit. Dans ce contexte, le praticien dispose souvent d'un grand échantillon de données non étiquetées et d'un plus petit échantillon de données étiquetées. Nous détaillons trois exemples.

Le premier exemple est celui de la lecture du code postal dans les centres de tri postaux. De nombreux codes postaux sont numérisés de manière automatique à très peu de frais. À partir de l'image numérisée, on souhaite classer les différentes images en fonction des dix chiffres possibles. Cette classification est longue et coûteuse si elle est effectuée par un opérateur humain ; de plus l'opérateur humain peut se fatiguer, augmentant au bout d'un certain temps le risque d'erreur. Pour ce type de données, le nombre de données classées manuellement est beaucoup plus petit que le nombre de codes postaux numérisés.

Le second exemple concerne l'indexation de contenu audiovisuel. L'Institut National de l'Audiovisuel (INA) dispose d'une collection très importante de vidéos. Il est alors souhaitable de classer cette information pour la retrouver plus facilement par la suite. L'indexation des vidéos par un expert est longue et coûteuse, tandis qu'une indexation automatique est moins longue et moins coûteuse. Ici encore, un grand échantillon de

données non étiquetées est disponible en plus d'un petit échantillon de données étiquetées.

Le troisième concerne la reconnaissance automatique de visages. Dans de nombreux sites Internet de partage de photos, il est maintenant possible de nommer les visages. L'objectif est de retrouver toutes les photos contenant une même personne. Ici encore le nombre de fois où la personne a été marquée peut être petit devant le nombre de fois où la personne apparaît.

Questions suscitées par ce type de données ?

En classification semi-supervisée, nous nous situons soit dans le cadre décisionnel soit dans le cadre exploratoire. Du point de vue de la classification non supervisée, on prend en compte l'information apportée par la présence d'étiquettes, comme le nombre minimal de classes présentes dans l'échantillon. Du point de vue de la classification supervisée, on cherche à améliorer les performances de la règle de classement apprise. Nous détaillons d'abord chacun de ces deux cadres standards, puis nous discutons ensuite de leur application en semi-supervisé.

Classification non supervisée

La classification non supervisée se situe dans un cadre exploratoire. Le nombre de classes et la signification de la variable qui explique l'hétérogénéité des données sont *a priori* inconnus. L'objectif de l'analyse est de déterminer des groupes les plus homogènes possibles entre eux et les plus différents les uns des autres. Nous développons ici deux exemples.

Premièrement en marketing, ce type de méthode est utilisé pour faire de la typologie clients. L'objectif est de partitionner les clients en un certain nombre de catégories, compte tenu des diverses données recueillies. Le but est d'élaborer des campagnes de publicité ciblées pour chaque catégorie de clients.

Deuxièmement en biologie, on cherche à structurer les êtres vivants en un nombre fini de classes ou d'espèces. On souhaite soit obtenir une classification hiérarchique en règnes, embranchements, classes ... Soit on souhaite uniquement une partition finale en espèces. Dans ce dernier cas, il convient de définir ce qu'est une espèce. La définition du dictionnaire Larousse indique :

> Ensemble d'individus animaux ou végétaux, vivants ou fossiles, à la fois semblables par leurs formes adultes et embryonnaires et par leur génotype, vivant au contact les uns des autres, s'accouplant exclusivement les uns aux autres et demeurant indéfiniment féconds entre eux.

La classification du vivant en espèces dépend donc des variables choisies : forme adulte, embryonnaire et génotype, contrainte d'accouplement exclusif, ... Les individus sont homogènes au sein d'une espèce, mais hétérogènes d'une espèce à l'autre.

Classification supervisée

L'objectif de la classification supervisée est de prédire la classe d'appartenance d'un nouvel individu, contrairement à la classification non supervisée qui vise à dégager une

structure présente dans les données. En classification supervisée nous nous situons dans un cadre décisionnel. Nous développons ici deux exemples.

Premièrement dans le domaine bancaire, un banquier veut pouvoir prédire le plus précisément possible si son client sera capable de rembourser son prêt. Pour cela il réalise une enquête sur son client : âge, profession, revenus, ... Puis en fonction de ces éléments, il accepte ou refuse l'octroi du prêt. La règle d'octroi du prêt, est apprise au préalable à partir des clients dont les caractéristiques ainsi que la variable prêt remboursé sont connus.

Deuxièmement en médecine, un médecin veut pouvoir différencier une hépatite d'une cirrhose du foie. Les symptômes de ces maladies sont effet relativement proches tandis que leur traitement est différent. Pour savoir avec exactitude si le patient souffre d'une hépatite ou d'une cirrhose du foie, il faut avoir recours à une biopsie. Cet examen est coûteux et comporte des risques. Or, la concentration de certains marqueurs sanguins est fortement liée au type d'infection du foie. La prédiction du type d'infection à partir de la mesure de ces marqueurs sanguins est donc importante. Pour des raisons de coût on cherchera à obtenir une bonne prédiction avec un nombre minimal de marqueurs, ce qui pose la question de la sélection de variables.

Classification semi-supervisée

La question de la classification non supervisée est selon nous mal posée dans le cadre semi-supervisé. En effet, une fois certaines classes observées la variable considérée comme latente ne l'est plus. Ainsi, on ne cherche plus vraiment à expliquer l'hétérogénéité des covariables par une variable latente. Cependant, la partition fournie par les données classées peut être utilisée comme une partition externe pour juger du bien fondé de la partition obtenue en classification non supervisée (Baudry & Celeux, 2009).

La question la plus développée est celle de l'amélioration des performances de la classification supervisée (Chapelle *et al.* , 2006). C'est sur cette question que nous nous focalisons dans ce travail.

Les données non étiquetées amènent aussi de nouvelles questions. D'une part, on peut uniquement souhaiter classer les données non étiquetées à disposition, et non pas apprendre une règle de classement pour toute nouvelle donnée. D'autre part, les données non étiquetées peuvent permettre une meilleure connaissance du domaine dans lequel on souhaite faire de la prédiction (Sokolovska *et al.* , 2008). Enfin, si le praticien a la possibilité de sélectionner les données à étiqueter parmi les données non étiquetées disponibles, l'enjeu sera de les choisir au mieux (Bach, 2007).

Une autre question à mi-chemin entre classification supervisée et non supervisée est celle de l'apparition de nouvelles classes dans l'échantillon de données non étiquetées (Bazell & Miller, 2005). En effet, il ne s'agit pas de classification supervisée puisque toutes les classes n'ont pas été observées dans l'échantillon de données étiquetées, mais il ne s'agit pas non plus de classification non supervisée puisque certaines variables latentes ont été observées.

Plan de thèse

Partie I : « Estimation et sélection en classification semi-supervisée »

Dans cette première partie, nous nous focalisons sur l'amélioration des performances de la classification supervisée en utilisant les données non étiquetées. Nous replaçons d'abord la classification semi-supervisée dans le cadre des problèmes de classification. Nous nous limitons ensuite à l'utilisation des modèles génératifs en classification semi-supervisée. En effet, il s'agit de la seule méthode qui permet la prise en compte rigoureuse et sans hypothèses supplémentaires des données non classées. Dans ce cadre les deux principales contributions de ce travail sont le traitement des questions suivantes : « Comment juger de la pertinence d'un modèle à l'aide des données non étiquetées ? » et « Comment prendre en compte l'objectif décisionnel dans le choix de modèle ? ».

Nous donnons maintenant le plan détaillé de cette partie chapitre par chapitre.

Chapitre 1 : « État de l'art »

Dans ce premier chapitre, nous dressons un état de l'art de la classification semi-supervisée. Ce chapitre offre une vision générale des questions posées dans le cadre semi-supervisé. Puis, nous verrons les différentes approches proposées pour y répondre. Nous montrons d'abord que diverses heuristiques ont été développées pour permettre aux méthodes prédictives de prendre en compte l'information apportée par les données non étiquetées. Cependant, seules les méthodes génératives permettent une prise en compte rigoureuse des données non classées c'est pourquoi, c'est cette approche que nous retenons dans les chapitres suivants.

Chapitre 2 : « Modèles génératifs en semi-supervisé »

Dans un second chapitre, nous détaillons l'utilisation des modèles génératifs en classification semi-supervisée. Nous montrons que ces derniers sont bien adaptés pour faire de la classification semi-supervisée. Nous détaillerons l'utilisation de l'algorithme EM qui est très bien adapté pour réaliser l'estimation des paramètres dans ce cadre. Nous discuterons aussi des questions d'initialisation et de vitesse de convergence qui s'y rapportent. Nous dressons un formulaire des modèles utilisés dans les cas continu et discret, l'utilisation de ces modèles étant ensuite illustrée sur des exemples variés. Ces expérimentations nous permettent de constater d'une part que le semi-supervisé peut améliorer les résultats, mais qu'il peut aussi les dégrader. En outre, ces améliorations ou dégradations dépendent fortement du modèle qui est choisi. Ceci nous amène à considérer les deux questions suivantes « Comment juger de la pertinence d'un modèle à l'aide des données non étiquetées ? » et « Comment prendre en compte l'objectif décisionnel dans le choix de modèle ? ».

Chapitre 3 : « Utilisation des données non étiquetées pour juger de la pertinence d'un modèle »

Dans ce troisième chapitre, nous mettons le modèle postulé à l'épreuve grâce aux données non étiquetées. Dans un premier temps, nous remarquons que si le modèle est bien spécifié les données non étiquetées apportent toujours des améliorations, dans cette situation les estimations supervisées, non supervisées et semi-supervisées sont proches. Partant de cette idée nous mettons en place un test statistique qui vérifie si ces différentes estimations sont suffisamment proches compte tenu de l'hypothèse que le modèle est bien spécifié. Si les paramètres estimés de ces différentes façons sont trop éloignés, il est alors préférable de proposer un autre modèle. Dans un second temps, nous reformulons la question précédente en termes de choix de modèle. Ce cadre nous permet de traiter de façon cohérente et simultanée les questions de choix de modèle et de la pertinence du modèle choisi. Cependant, l'approche proposée repose avant tout sur le critère BIC qui ne prend pas directement en compte l'objectif décisionnel.

Chapitre 4 : « Sélection prédictive d'un modèle génératif »

Dans ce quatrième chapitre, nous proposerons un critère de choix de modèle de type Akaike qui prend en compte l'objectif décisionnel. Pour cela, nous partons de l'idée que pour obtenir un classifieur avec de bonnes performances il faut bien approcher la distribution de la classe conditionnellement aux covariables. Cet objectif n'est pas directement pris en compte au niveau de l'apprentissage des paramètres. Nous proposons de le prendre en compte au niveau du choix de modèle. Ainsi, nous cherchons le modèle pour lequel la distribution de la classe conditionnellement aux covariables est la mieux approchée. Dans un cadre fréquentiste, cela nous amène à considérer la déviance à la distribution de la classe conditionnellement aux covariables. Par une série d'approximations de cette déviance nous obtenons un critère de choix de modèle, AIC_{cond}, facile à calculer et peu coûteux dans le cadre semi-supervisé. Des résultats théoriques assurent à AIC_{cond} des propriétés satisfaisantes et, en effet, le bon comportement de ce critère est ensuite mis en évidence expérimentalement sur données réelles et simulées.

Partie II : « Contribution à la modélisation multinomiale »

Dans une seconde partie, nous proposons une extension des modèles multinomiaux parcimonieux proposés dans Celeux & Govaert (1991). Le lien avec la partie précédente est que les modèles considérés sont encore des modèles génératifs. Ces modèles pourront donc aussi bien être utilisés dans le cadre supervisé, dans le cadre non supervisé, ainsi que dans le cadre semi-supervisé. Les modèles proposés contiennent cependant des paramètres discrets, ce qui pose alors la question de leur prise en compte au niveau du choix de modèle. Nous proposons alors un critère de type BIC qui prend en compte la complexité associée à l'estimation de ces paramètres particuliers.

Chapitre 5 : « Contribution à la modélisation multinomiale »

Dans ce cinquième chapitre, nous étendons d'abord les modèles de Celeux & Govaert (1991) proposés dans le cas des produits de distributions de Bernoulli au cas des produits de distributions multinomiales. Nous proposons d'une part une paramétrisation pour laquelle les contraintes sont automatiquement respectées lors de l'estimation, ceci même dans le cas où toutes les variables n'ont pas le même nombre de modalités, contrairement à l'extension proposée dans Biernacki *et al.* (2006). D'autre part, nous proposons une paramétrisation qui repose sur une permutation du vecteur des modalités selon la classe. Ensuite, nous posons la question du comptage des paramètres discrets dans l'approximation BIC, puisque les modèles précédents comprennent à la fois des paramètres continus et discrets. Enfin, nous illustrons sur divers exemples l'utilisation de ces modèles.

« Conclusion et perspectives »

Dans un dernier chapitre, nous dressons les principales conclusions de ce travail, et nous présentons les perspectives de recherches qui s'ouvrent à nous.

Première partie

Estimation et sélection en classification semi-supervisée

Chapitre 1

État de l'art

La classification semi-supervisée trouve ses racines dans les problèmes d'apprentissage en présence de données manquantes (Ganesalingam & McLachlan, 1978). De nombreux travaux y ont été dédiés dans les années 1970 (Hosmer, 1973; Dempster *et al.* , 1977; O'Neill, 1978). Ce thème de recherche a ensuite connu un regain d'intérêt à la fin des années 1990 dans la communauté du *Machine Learning*, avec la disponibilité croissante de grands jeux de données rendue possible par les nouvelles technologies. Il s'agit par exemple des travaux de Nigam *et al.* (2000) en classification de textes. Enfin, la récente parution d'un livre entièrement dédié à la classification semi-supervisée (Chapelle *et al.* , 2006) constitue la meilleure preuve de l'intérêt actuel suscité par ce sujet.

Dans ce chapitre, nous rappelons d'abord les fondements des méthodes de classification supervisée et non supervisée. Nous décrivons ensuite les spécificités des données dans le cadre semi-supervisé. Cette spécificité intervient aussi bien par la forme des données, les problèmes posés, que par les hypothèses de travail nécessaires à leur traitement. Enfin, le cadre semi-supervisé une fois posé nous décrivons les différentes approches utilisées. Nous montrons comment les données non étiquetées peuvent être utilisées de manière générale pour modifier la règle de classement. Puis, nous décrivons comment les méthodes prédictives et génératives font usage des données non étiquetées.

Ce chapitre résulte en grande partie de l'état de l'art sur la classification semi-supervisée que nous avons dressé dans la Revue Modulad[1] (Vandewalle, 2009a).

1.1 Introduction

1.1.1 Méthodes supervisées

En classification supervisée, l'objectif est de prédire au mieux la classe d'appartenance d'un nouvel individu à partir de l'observation d'un ensemble de covariables (Mclachlan, 2004). Nous noterons g le nombre de classes, $\boldsymbol{z}$ la classe qui sera codée de façon disjonctive, c.-à-d. $\boldsymbol{z} = (z_1, \ldots, z_g)$ avec

$$z_k = \begin{cases} 1 & \text{si l'individu appartient à la classe } k, \\ 0 & \text{sinon.} \end{cases}$$

[1] `http://www-roc.inria.fr/axis/modulad/archives/numero-40/vandewalle-40/Vandewalle-40.pdf`

Nous noterons $\mathcal{Z} = \{0,1\}^g$ l'espace auquel appartient la classe, et $\boldsymbol{x} = (x_1, \ldots, x_d)$ le vecteur des covariables à valeurs dans un espace mesurable $\mathcal{X}$ de dimension d, typiquement $\mathbb{R}^d$.

Nous cherchons à apprendre une règle de décision δ de $\mathcal{X}$ à valeur dans $\mathcal{Z}$. En pratique, cette règle est apprise à partir d'un échantillon constitué de données étiquetées

$$(\mathbf{x}, \mathbf{z}) = \{(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{z}_n)\}.$$

Les méthodes d'apprentissage supervisé découlent du paradigme suivant. Si les données sont issues du modèle paramétrique $p(\cdot;\theta^*)$, avec p la densité de probabilité qui sera interprétée en fonction de ses argument et θ^* le vrai vecteur des paramètres du modèle, on a

$$\underset{\boldsymbol{z}\in\mathcal{Z}}{\operatorname{argmax}} \underbrace{p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)}_{\text{Génératif}} = \underset{\boldsymbol{z}\in\mathcal{Z}}{\operatorname{argmax}} \underbrace{p(\boldsymbol{z}|\boldsymbol{x}; \theta^*)}_{\text{Prédictif}} = \delta(\boldsymbol{x}; \theta^*)$$

où $\delta(\cdot;\theta^*)$ est la règle de classement optimale (dite de Bayes) issue de θ^*. La valeur de θ^* étant bien sûr inconnue en général, les méthodes d'apprentissage cherchent à l'apprendre, ou tout du moins la partie de θ^* nécessaire à la règle de classement. Cet objectif se décline sous trois formes en partant des hypothèses les plus fortes et en allant vers les hypothèses les plus faibles :

- On apprend la distribution du couple $(\boldsymbol{X}, \boldsymbol{Z})$ puis on en déduit la règle de classement par maximum *a posteriori* (MAP).
- On apprend la distribution du couple $\boldsymbol{Z}|\boldsymbol{X}$ puis on en déduit la règle de classement de nouveau par MAP.
- On apprend directement la règle de classement.

La première s'inscrit dans ce qu'on appelle les méthodes génératives, et les deux suivantes s'inscrivent dans ce qu'on appelle les méthodes prédictives. Dans ce qui suit nous décrivons la mise en œuvre de ces deux méthodes.

Méthodes prédictives

Les méthodes prédictives modélisent directement la distribution $\boldsymbol{Z}|\boldsymbol{X}$, voire seulement la position de la frontière de classification. On distingue :

- les méthodes non paramétriques comme les k plus proches voisins (Dasarathy, 1990),
- les méthodes à base d'arbres comme CART (Breiman *et al.* , 1984),
- les méthodes paramétriques d'apprentissage de la distribution $\boldsymbol{Z}|\boldsymbol{X}$ comme la régression logistique (Anderson & Richardson, 1979),
- les méthodes de recherche d'un hyperplan optimal comme le perceptron de Rosenblat (1958) ou encore les séparateurs à vastes marges (SVM) (Vapnik, 1995).

Ces méthodes visent à faire le moins d'hypothèses possible sur la distribution des données, en ne modélisant que la distribution directement nécessaire à établir la règle de classement. Ici, la modélisation de la distribution de $\boldsymbol{X}$ est évitée puisqu'elle n'est pas nécessaire pour apprendre la règle de classement.

Nous présentons d'une part la régression logistique et d'autre part les SVM.

Régression logistique En régression logistique, l'objectif est d'apprendre au mieux la distribution de $\boldsymbol{Z}|\boldsymbol{X}$. La régression logistique linéaire sur variables continues se formule

de la façon suivante :

$$\begin{cases} p(Z_k = 1|\boldsymbol{x}) = \frac{e^{\beta_k^0 + \beta_k' \boldsymbol{x}}}{1+\sum_{j=1}^{g-1} e^{\beta_j^0 + \beta_j' \boldsymbol{x}}} & \text{pour } k \in \{1, \ldots, g-1\} \\ p(Z_g = 1|\boldsymbol{x}) = \frac{1}{1+\sum_{j=1}^{g-1} e^{\beta_j^0 + \beta_j' \boldsymbol{x}}} & . \end{cases}$$

Les paramètres $\theta = (\beta_1^0, \ldots, \beta_{g-1}^0, \beta_1, \ldots, \beta_{g-1})$ sont estimés par maximum de vraisemblance, le log de cette dernière s'écrivant

$$\sum_{i=1}^{n} \sum_{k=1}^{g} z_{ik} \log \frac{e^{\beta_k^0 + \beta_k' \boldsymbol{x}_i}}{1 + \sum_{j=1}^{g-1} e^{\beta_j^0 + \beta_j' \boldsymbol{x}_i}} + \sum_{i=1}^{n} z_{ig} \log \frac{1}{1 + \sum_{j=1}^{g-1} e^{\beta_j^0 + \beta_j' \boldsymbol{x}_i}}. \tag{1.1}$$

La maximisation de cette expression n'est pas explicite. Cependant, comme le problème est concave en θ, elle est facilement réalisée grâce à un algorithme de Newton. Nous illustrons la régression logistique figure 1.1 sur la variable longueur des sépales des iris de Fisher pour la distinction *setosa*/non-*setosa*.

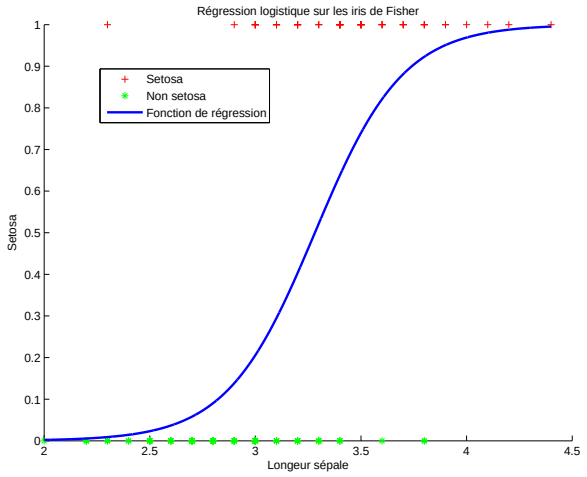

FIG. 1.1 – Régression logistique sur la variable longueur des sépales des iris de Fisher pour la distinction *setosa*/non-*setosa*.

Quand le nombre de données tend vers l'infini le paramètre estimé par maximum de vraisemblance (équation (1.1)) converge vers

$$\theta^*_{\boldsymbol{Z}|\boldsymbol{X}} = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}[\log p(\boldsymbol{Z}|\boldsymbol{X};\theta)],$$

cela implique une bonne approximation de la distribution de $\boldsymbol{Z}|\boldsymbol{X}$. Cette bonne approximation de la distribution conditionnelle est ensuite susceptible, de produire une règle de classement précise. Ici la distribution de $\boldsymbol{Z}|\boldsymbol{X}$ est modélisée, alors qu'en classification supervisée, seule la position de la frontière de classification est requise (Vapnik, 1995).

Séparateurs à vastes marges Les SVM s'affranchissent des hypothèses sur la distribution de $\boldsymbol{Z}|\boldsymbol{X}$ et recherchent directement la frontière de classification. Dans le cadre de

la classification binaire, les SVM cherchent la marge la plus grande qui sépare les deux classes. Nous noterons dans ce cas particulier la classe par $y \in \{-1; 1\}$. Plus la marge est grande, et plus la dimension de Vapnik-Chervonenkis (VC-dimension) est petite (Vapnik, 1995). Or, dans le cas des séparateurs linéaires, la VC-dimension intervient dans une borne en probabilité sur l'erreur produite par le classifieur obtenu. Cette borne est d'autant plus fine que la VC-dimension est petite, ce qui conduit au problème d'optimisation suivant : minimiser $\|\omega\|$, où $\omega \in \mathbb{R}^d$, sous la contrainte

$$\forall i \in \{1, \ldots, n\} : y_i[\omega' \boldsymbol{x}_i + b] \geqslant 1.$$

La résolution directe de ce problème d'optimisation est difficile. On le reformule en introduisant les multiplicateurs de Lagrange $\alpha_1, \ldots, \alpha_n$:

$$\frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(\omega' \boldsymbol{x}_i + b) - 1]. \tag{1.2}$$

Puis en annulant les dérivées en b et en ω, on obtient :

$$\sum_{i=1}^{n} \alpha_i^* y_i = 0, \alpha_i^* \geqslant 0, i = 1, \ldots, n$$

et

$$\omega = \sum_{i=1}^{n} y_i \alpha_i^* \boldsymbol{x}_i, \alpha_i^* \geqslant 0, i = 1, \ldots, n.$$

Enfin en remplaçant ω dans l'équation (1.2), on souhaite optimiser

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i' \boldsymbol{x}_j \tag{1.3}$$

sous la contrainte

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \text{ et } \forall i : \alpha_i \geqslant 0.$$

Remarquons ici que seuls les produits scalaires $\boldsymbol{x}_i' \boldsymbol{x}_j$ interviennent dans l'expression à optimiser. Ceci justifie alors l'utilisation de noyaux dans les SVM, c'est-à-dire qu'on projette les données dans des espaces de grande dimension $\phi(\boldsymbol{x}_i)$ où l'hypothèse de séparabilité est vérifiée. Mais on a juste besoin de calculer le produit scalaire $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)'\phi(\boldsymbol{x}_j)$. Dans de nombreux cas il suffit alors de calculer $\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$ sans avoir à expliciter $\phi(\boldsymbol{x}_i)$. Les noyaux habituellement utilisés sont les noyaux gaussiens, polynômiaux ou encore la tangente hyperbolique (Aronszajn, 1950).

Des algorithmes efficaces existent pour l'optimisation de l'équation (1.3) (Osuna *et al.* , 1997). Le résultat est un ensemble de cœfficients α_i^* utilisables pour la construction de l'hyperplan séparateur.

On illustre figure 1.2 l'utilisation des SVM sur les iris de Fisher en utilisant les variables longueur et largeur des sépales pour le problème de discrimination *setosa*/non-*setosa*.

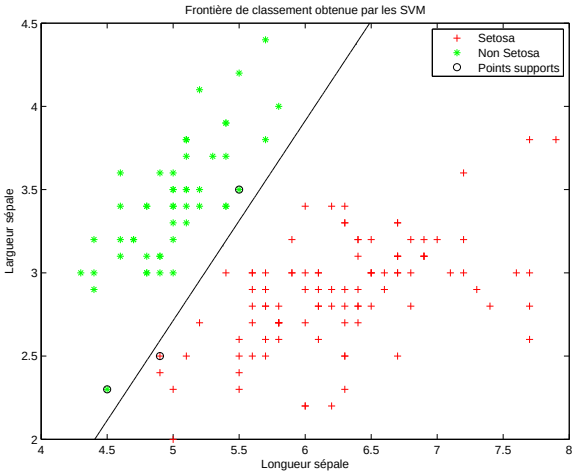

FIG. 1.2 – Régle de classement apprise par les SVM.

Méthodes génératives

Les méthodes génératives modélisent la distribution de $\boldsymbol{X}$ contrairement aux méthodes prédictives. Pour cela elles décomposent la distribution du couple $(\boldsymbol{X}, \boldsymbol{Z})$ sous la forme $\boldsymbol{X}|\boldsymbol{Z}$ et $\boldsymbol{Z}$. Puis, comme $\boldsymbol{Z}$ ne peut prendre que g valeurs distinctes, on modélise simplement cette distribution par $p(Z_k = 1) = \pi_k^*$. Ensuite on postule un modéle paramétrique sur la distribution de $\boldsymbol{X}|\boldsymbol{Z}$

$$\exists \theta^* \in \Theta / \forall (\boldsymbol{x}, k) \in \mathcal{X} \times \{1, \ldots, g\}, p(\boldsymbol{x}|Z_k = 1) = p(\boldsymbol{x}; \theta_k^*),$$

où $\theta^* = (\pi_1^*, \ldots, \pi_{g-1}^*, \theta_1^*, \ldots, \theta_g^*)$. Ces modèles sont appelés modèles génératifs puisqu'ils modélisent le processus de génération des données :

- On génère d'abord la classe : $\boldsymbol{Z} \sim \mathcal{M}(1, \pi_1^*, \ldots, \pi_g^*)$,
- puis on génére le vecteur des covariables conditionnellement à la classe : $\boldsymbol{X}|Z_k = 1$ a pour densité de probabilité $p(\cdot; \theta_k^*)$.

Dans le cas où $\mathcal{X} = \mathbb{R}^d$, la première approche à utiliser des modèles génératifs en classification supervisée est l'analyse discriminante linéaire (LDA) de Fisher (1936). Elle suppose que $\boldsymbol{X}|\boldsymbol{Z}$ suit une distribution gaussienne multivariée, et que les matrices de covariances sont identiques :

$$\boldsymbol{X}|Z_k = 1 \sim \mathcal{N}(\mu_k, \Sigma).$$

L'estimateur du maximum de vraisemblance de μ_k est alors la moyenne empirique des individus de la classe k, et celui de la matrice de covariance consiste simplement à agréger les matrices de variances intra-classe. La proportion de la classe k, π_k, est estimée par la fraction d'individus provenant de la classe k. Cela donne les formules classiques suivantes,

où on a noté $n_k = \sum_{i=1}^{n} z_{ik}$:

$$\hat{\pi}_k = \frac{n_k}{n}, \tag{1.4}$$

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n} z_{ik}\boldsymbol{x}_i, \tag{1.5}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik}(\boldsymbol{x}_i - \hat{\mu}_k)(\boldsymbol{x}_i - \hat{\mu}_k)'. \tag{1.6}$$

La règle de classement est ensuite déduite par MAP :

$$\hat{k} = \underset{k}{\operatorname{argmax}} \left(\boldsymbol{x} - \frac{\hat{\mu}_k}{2}\right)' \hat{\Sigma}^{-1}\hat{\mu}_k + \log \hat{\pi}_k.$$

La frontière de classement est alors un hyperplan séparateur comme illustré figure 1.3. Remarquons que pour classer un point, il faut être capable d'inverser la matrice $\hat{\Sigma}$. C'est-à-dire qu'en pratique il faut que $n \geqslant d + 1$. Cette approche fournit souvent de bons résultats (Hastie *et al.* , 2001), et permet de prendre facilement en compte des échantillons avec beaucoup de variables.

Quand les matrices de covariances sont supposées différentes on parle d'analyse discriminante quadratique (QDA) (illustrée figure 1.4). Celle-ci produit aussi de bons résultats, cependant elle se révèle être souvent moins robuste que la LDA quand le nombre de données est petit.

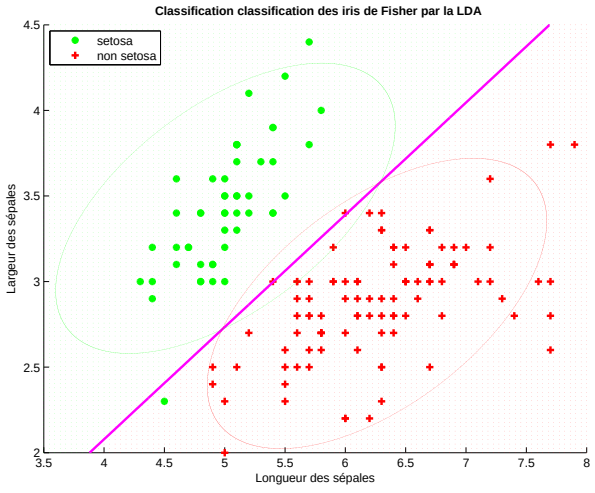

FIG. 1.3 – Illustration de la LDA sur les iris de Fisher.

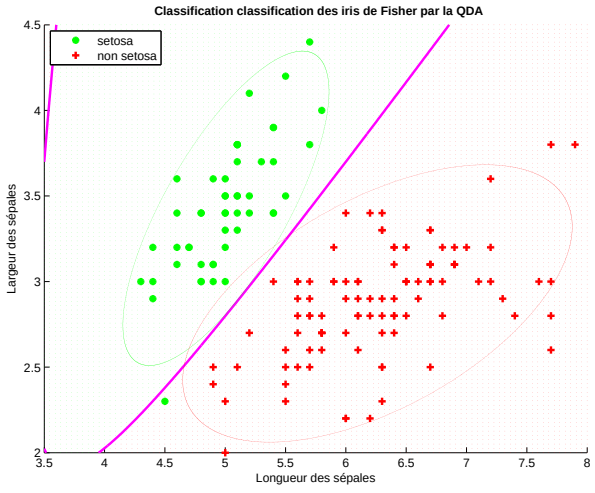

FIG. 1.4 – Illustration de la QDA sur les iris de Fisher.

Un intermédiaire entre LDA et QDA est l'analyse discriminante régularisée proposée par Friedman (1989). Cette méthode estime les matrices de covariance par une combinaison convexe des estimations dans les cas hétéroscédastique et homoscédastique

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}.$$

Le paramètre de régularisation $\alpha \in [0;1]$ peut être choisi par validation croisée du taux d'erreur. Quand le nombre de données est trop petit pour permettre une estimation correcte de $\hat{\Sigma}$, la régularisation scalaire est utilisée

$$\hat{\Sigma}(\gamma) = \gamma\hat{\Sigma} + (1-\gamma)\hat{\sigma}^2 I,$$

avec $\gamma \in [0;1]$. Une autre approche possible repose sur la décomposition en valeurs singulières de la matrice de covariance

$$\Sigma_k = \lambda_k D_k A_k D_k',$$

où $\lambda_k = |\Sigma_k|^{1/d}$, D_k est la matrice des vecteurs propres de Σ_k, et A_k est la matrice diagonale des valeurs propres dont le produit est normalisé à 1 et rangées dans l'ordre décroissant. En imposant à λ_k, D_k ou A_k d'être identiques ou non entre les classes, Bensmail & Celeux (1996) obtiennent 14 modèles.

Dans le cas discret, le modèle multinomial d'indépendance conditionnelle est le plus souvent utilisé (Hand & Yu, 2001). L'hypothèse d'indépendance conditionnelle évite la prise en compte des dépendances et permet de limiter fortement le nombre de paramètres à estimer. Ce modèle bien que très naïf permet dans de nombreuses situations d'intégrer efficacement la spécificité de la distribution discrète conditionnellement à la classe. En outre, la modélisation d'une classe par un mélange de produits de distributions multinomiales permet de prendre en compte facilement la dépendance entre les variables conditionnellement à la classe. Une autre possibilité est de rechercher un graphe de dépendance entre variables. La recherche d'un tel graphe est généralement difficile, la recherche du meilleur arbre au sens du maximum de vraisemblance restant cependant facile (Friedman, 1997).

Une critique émise à l'égard des modèles génératifs en classification supervisée est qu'ils ne prennent pas directement en compte l'objectif décisionnel. Notons

$$DK(p(\boldsymbol{X},\boldsymbol{Z}), p(\boldsymbol{X},\boldsymbol{Z};\theta)) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}\left[\log p(\boldsymbol{X},\boldsymbol{Z})\right] - \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}\left[\log p(\boldsymbol{X},\boldsymbol{Z};\theta)\right]$$

la divergence de Kullback entre la distribution jointe d'échantillonnage, et la distribution jointe paramétrée par θ. Dans le cas général le paramètre estimé par maximum de vraisemblance converge vers

$$\theta^*_{\boldsymbol{X},\boldsymbol{Z}} = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}[\log p(\boldsymbol{X},\boldsymbol{Z};\theta)].$$

On retrouve bien sûr $\theta^*_{\boldsymbol{X},\boldsymbol{Z}} = \theta^*$ si le modèle postulé est correct. Ainsi, l'estimation des paramètres du modèle par maximum de vraisemblance nous permet de trouver les paramètres qui minimisent la divergence de Kullback à la distribution de $(\boldsymbol{X},\boldsymbol{Z})$, ce qui n'implique pas nécessairement l'obtention d'une règle de classement précise.

Ceci dit, si les hypothèses formulées par les modèles génératifs sont vérifiées, ils améliorent la règle de classement (O'Neill, 1980) par rapport aux modèles prédictifs. Par exemple dans le cas de la comparaison entre LDA et régression logistique, plus de paramètres sont estimés dans le cadre de la LDA, mais comme l'information sur $\boldsymbol{X}$ est prise en compte dans l'estimation des paramètres, elle permet d'obtenir une règle de classement moins variable que la règression logistique. De plus, les méthodes génératives donnent souvent des résultats comparables aux méthodes prédictives en pratique (Hastie *et al.* , 2001).

1.1.2 Méthodes non supervisées

La distinction entre méthodes génératives et méthodes prédictives n'est plus faite en classification non supervisée. En effet, dans ce cas il est impossible de modéliser directement la distribution de $\boldsymbol{Z}|\boldsymbol{X}$, puisqu'aucune étiquette n'a été observée. L'équivalent des méthodes prédictives n'existe donc pas en classification non supervisée. Dans ce cadre, seuls les modèles génératifs restent utilisables. Ceux-ci modélisent la distribution du couple $(\boldsymbol{X}, \boldsymbol{Z})$, et par intégration sur la classe, la distribution marginale de $\boldsymbol{X}$. Dans ce cadre on parle souvent de classification à base de modèles par opposition aux méthodes de classification à base de distances.

L'objectif de la classification non supervisée est de trouver une structure intéressante dans les données à partir d'un échantillon $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Cette méthode, comme de nombreuses autres, repose en grande partie sur les choix de l'expérimentateur : variables prises en compte, métrique, ... Le récent développement de l'informatique a rendu possible l'utilisation d'algorithmes de classification sur de grands échantillons comportant beaucoup de variables. Le résultat de la classification est soit une partition, soit une hiérarchie (Gordon, 1981). On distingue deux types d'approches :

- **Les méthodes géométriques à base de distances :** classification hiérarchique, méthode des centres mobiles,
- **Les méthodes probabilistes :** classification à base de modèles.

En classification non supervisée, on considère qu'un individu provient d'un groupe *a priori* qu'il faut retrouver. L'approche probabiliste remonte à Pearson (1894) qui s'intéressa à la distribution de mesures biométriques sur une population de crabes de la baie de Naples. Il remarqua sur l'histogramme de ces mesures que leur distribution n'était pas gaussienne (voir figure 1.5). Il postula alors que l'histogramme observé ne résultait pas des mesures d'une seule espèce mais d'un mélange de deux sous espèces. Cela pourrait démontrer une évolution de l'espèce de crabes en deux sous espèces. L'analyse statistique donne alors une indication sur le nombre de classes présentes dans l'échantillon, ainsi que sur les paramètres respectifs de chaque classe.

La modélisation est ici la même que pour les modèles génératifs dans le cadre supervisé. La seule différence tient aux variables observées ; ici seulement les covariables sont observées alors que dans le cas supervisé nous disposions aussi des étiquettes. On n'observe donc plus que des réalisations de $\boldsymbol{X}$ qu'on appelle distribution mélange

$$\boldsymbol{x} \in \mathcal{X} \mapsto p(\boldsymbol{x}; \theta) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}; \theta_k).$$

Contrairement au cadre supervisé, nous ne pouvons plus choisir n'importe quelle famille pour paramétrer la distribution de $\boldsymbol{X}|\boldsymbol{Z}$. Il faut pouvoir reconstituer l'information sur $\boldsymbol{Z}$ à partir de la seule observation de $\boldsymbol{X}$. Ceci est le cas lorsqu'on considère des mélanges de distributions identifiables (Titterington *et al.* , 1985). Il s'agit par exemple des mélanges de gaussiennes, d'exponentielles, de Poisson et de Cauchy. Les mélanges de binomiales et d'uniformes ne sont quant eux pas identifiables.

Les paramètres sont souvent estimés par maximum de vraisemblance en utilisant l'algorithme EM. Cette approche est à relier à la question de l'approximation de densité, puisque les paramètres estimés sont ceux qui minimisent la divergence de Kullback-Leibler à la distribution marginale d'échantillonnage.

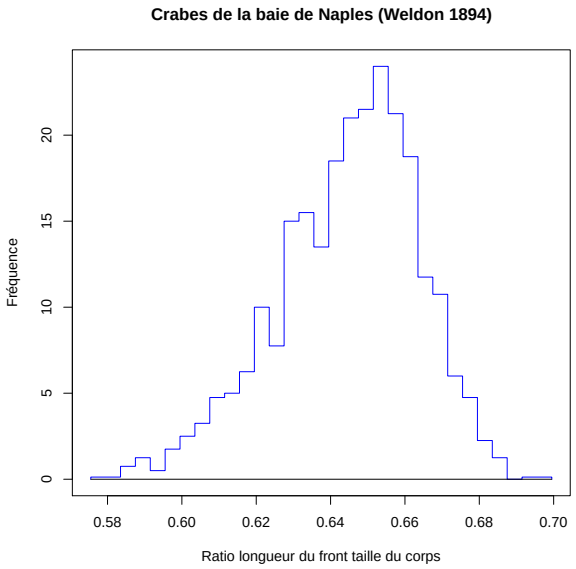

FIG. 1.5 – Histogramme de mesures biométriques sur des crabes de la baie de Naples.

Une question récurrente en classification non supervisée est le choix du nombre de classes présentes dans l'échantillon. Cette question reçoit souvent des réponses à l'aide des critères de choix de modèle comme AIC ou BIC. En pratique, le critère AIC a tendance à surestimer le nombre de classes présentes dans l'échantillon car il privilégie le point de vue approximation de densité. Si la vraisemblance reste bornée, et que le modèle postulé est correct, BIC sélectionne asymptotiquement le bon nombre de classes avec probabilité 1. Cependant en pratique, tous les modèles sont faux, et BIC peut surestimer le nombre de classes en privilégiant lui aussi l'aspect estimation de densité. Biernacki *et al.* (2000) ont alors proposé un critère de choix de modèle, le critère ICL, qui prend en compte l'objectif du praticien. Ces auteurs considèrent que le praticien ne cherche pas forcément le vrai modèle, mais celui pour lequel il va pouvoir bien interpréter les classes obtenues, c.-à-d. le modèle pour lequel les classes sont bien séparées. Le critère ICL s'interprète comme le critère BIC pénalisé par l'entropie de classification. Dans de nombreux cas la partition donnée par ce critère est plus utile que celle de BIC qui surestime souvent le nombre de classes.

Les méthodes génératives étant utilisables quand on dispose de données non étiquetées, cela permet d'entrevoir l'adaptation facile de ces méthodes au cadre semi-supervisé. Par contre, les méthodes prédictives sont elles incapables d'utiliser les données non étiquetées sans hypothèses supplémentaires, ce qui laisse entrevoir les difficultés de l'extension de ces dernières au cadre semi-supervisé. Dans ce qui suit nous posons plus précisément le cadre semi-supervisé.

1.2 Cadre semi-supervisé

Nous précisons ici les spécificités du cadre semi-supervisé. Le problème de la classification semi-supervisée se pose dès que des données partiellement étiquetées sont disponibles.

La disponibilité de ces données « mixtes » pose alors un certain nombre de questions, dont en tout premier lieu celle des hypothèses d'échantillonnage.

1.2.1 Hypothèses d'échantillonnage

Différentes hypothèses possibles

La question des données partiellement étiquetées se replace dans le cadre plus général des données manquantes. Les étiquettes des données non étiquetées, constituent ici ces données manquantes. Face à ce type de données, la première question qui vient à l'esprit est « Pourquoi les étiquettes des données non étiquetées sont-elles manquantes ? ». On suppose que l'échantillon de données partiellement étiquetées provient de la réalisation d'un n échantillon indépendant et identiquement distribué (i.i.d.) :

$$\{(\boldsymbol{X}_1, \boldsymbol{Z}_1, S_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Z}_n, S_n)\},$$

où

$$S_i = \begin{cases} 1 \text{ si la donnée } i \text{ est étiquetée} \\ 0 \text{ sinon.} \end{cases}$$

On distingue trois cas (Heitjan & Rubin, 1991) :

- Soit les données sont manquantes totalement au hasard (« *Missing Completely At Random* » : MCAR) : $p(s|\boldsymbol{x}, \boldsymbol{z}) = p(s)$. Dans ce cas la distribution de S est modélisée par une distribution de Bernoulli de paramètre $\beta \in]0; 1[$.
- Soit les données sont manquantes au hasard (« *Missing At Random* » : MAR) : $p(s|\boldsymbol{x}, \boldsymbol{z}) = p(s|\boldsymbol{x})$.
- Soit les données sont manquantes de manière non aléatoire (« *Missing Not At Random* » : MNAR) : $p(s|\boldsymbol{x}, \boldsymbol{z}) \neq p(s|\boldsymbol{x})$.

Par la suite, pour simplifier les notations, on notera les n_ℓ données étiquetées en premier puis les n_u données non étiquetées de telle sorte que notre échantillon de données partiellement étiquetées s'écrira

$$(\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \{(\boldsymbol{x}_1, \boldsymbol{z}_1), \ldots, (\boldsymbol{x}_{n_\ell}, \boldsymbol{z}_{n_\ell}), \boldsymbol{x}_{n_\ell+1}, \ldots, \boldsymbol{x}_n\}.$$

Nous noterons $\mathbf{z}_u$ les étiquettes des données non étiquetées qui n'ont ici pas été observées.

Il est en général difficile de tester les hypothèses d'échantillonnage. L'hypothèse MAR ne peut pas être testée, puisque ce test nécessiterait l'observation de l'étiquette des données non étiquetées. Sous l'hypothèse MCAR, les covariables des données étiquetées et non étiquetées ont la même distribution. Cette propriété peut donc être testée par le test de Kolmogorov-Smirnov. Remarquons que MCAR est réaliste dans la plupart des cas où on dispose initialement de données non étiquetées et où on choisit de manière aléatoire dans cet échantillon les données à étiqueter.

Les méthodes génératives et prédictives sont consistantes sous les hypothèses MCAR et MAR. Cependant, en cas de MAR la validation croisée du taux d'erreur n'évalue pas correctement le taux d'erreur sur les données à venir, puisque les données non étiquetées ont une distribution différente de celle qui est évaluée. Dans la situation MNAR, il faut modéliser $p(s|\boldsymbol{x}, \boldsymbol{z})$, ce qui est plus délicat. Dans ce qui suit, sauf exception, nous nous placerons sous l'hypothèse MCAR. Cette hypothèse est souvent faite implicitement,

puisque la règle de classement est généralement apprise pour classer des données supposées provenir de la même distribution que celles qui ont servi à l'apprendre.

Discussion dans le cas MNAR

Nous avons choisi de nous placer sous l'hypothèse MCAR. Cependant, l'objectif est ici de discuter des autres situations pour un objectif d'estimation. En classification supervisée, l'échantillon de données non étiquetées sur lequel on souhaite appliquer la règle de classement par la suite n'a pas été observé. Nous sommes donc obligés de faire l'hypothèse que les données à classer dans l'avenir sont issues de la même distribution que les données étiquetées qui ont servi à apprendre la règle de classement. En classification semi-supervisée, l'échantillon auquel on souhaite appliquer la règle de classement est généralement disponible, puisqu'il s'agit souvent de l'échantillon de données non étiquetées à disposition. Ainsi le cadre semi-supervisé permet de traiter des situations où les données à classer n'ont pas la même distribution que les données étiquetées.

Estimation des proportions Considérons le cas où l'échantillon de données étiquetées résulte d'un échantillonnage rétrospectif et l'échantillon de données non étiquetées résulte d'un échantillonnage mélange. C'est par exemple le cas dans les études cliniques où le nombre de patients sains et malades est fixé à l'avance. L'estimation du biais reste assez facile dans ce cas puisque $p(s|\boldsymbol{z}, \boldsymbol{x}) = p(s|\boldsymbol{z})$. Seules les proportions du mélange diffèrent entre les données étiquetées et non étiquetées. Les données non étiquetées supposées provenir de la distribution mélange permettent d'estimer les proportions des différentes classes. Ces proportions permettent de classer les individus si les distributions conditionnellement à la classe sont connues. Par ailleurs différents types d'information peuvent être extraits des données non étiquetées. Pour ce faire, on peut citer Hosmer (1973) pour les modèles génératifs et Anderson & Richardson (1979) pour la régression logistique. Dans un cadre plus général, des approches non paramétriques ont été proposées par Zou *et al.* (2004). Les auteurs proposent une méthode pour lever ce biais grâce aux données non étiquetées via une fonction de perte pondérée. Les principaux éléments sont les suivants : soit un problème à deux classes, où $y \in \{-1, 1\}$, et h une fonction de $\mathcal{X}$ à valeurs dans $\mathbb{R}$. Afin d'obtenir un estimateur consistant de π_1 la probabilité que $y = 1$, une méthode de type moment est proposée. Elle part de la décomposition

$$p(\boldsymbol{x}) = \pi_1 p(\boldsymbol{x}|y = 1) + (1 - \pi_1) p(\boldsymbol{x}|y = -1),$$

et en prenant l'espérance de $h(\boldsymbol{X})$, on a

$$\mathbb{E}_{\boldsymbol{X}}[h(\boldsymbol{X})] = \pi_1 \mathbb{E}_{\boldsymbol{X}|Y=1}[h(\boldsymbol{X})] + (1 - \pi_1)\mathbb{E}_{\boldsymbol{X}|Y=-1}[h(\boldsymbol{X})].$$

Si $\mathbb{E}_{\boldsymbol{X}|Y=1}[h(\boldsymbol{X})] \neq \mathbb{E}_{\boldsymbol{X}|Y=-1}[h(\boldsymbol{X})]$, on a

$$\pi_1 = \frac{\mathbb{E}_{\boldsymbol{X}}[h(\boldsymbol{X})] - \mathbb{E}_{\boldsymbol{X}|Y=-1}[h(\boldsymbol{X})]}{\mathbb{E}_{\boldsymbol{X}|Y=1}[h(\boldsymbol{X})] - \mathbb{E}_{\boldsymbol{X}|Y=-1}[h(\boldsymbol{X})]}.$$

En utilisant la distribution empirique, des données non étiquetées on obtient un estimateur consistant de $\mathbb{E}_{\boldsymbol{X}}[h(\boldsymbol{X})]$, tandis que les données étiquetées donnent des estimateurs consistants de $\mathbb{E}_{\boldsymbol{X}|Y=1}[h(\boldsymbol{X})]$ et $\mathbb{E}_{\boldsymbol{X}|Y=-1}[h(\boldsymbol{X})]$, on obtient par suite un estimateur consistant de π_1. Cet estimateur est ensuite utilisé dans un algorithme de classification via une fonction de coût pondérée.

Estimation générale du biais d'étiquetage Dans le cas où le biais d'étiquetage est plus général, une autre approche a été proposée par Rosset *et al.* (2004) et par Fan *et al.* (2005). L'approche proposée consiste encore en une méthode de type moment, celle-ci permet d'estimer le biais d'étiquetage aussi bien en régression qu'en classification. Pour cela postulons un modèle paramétré par γ sur le biais d'étiquetage $p(S = 1|\boldsymbol{x}, \boldsymbol{z}; \gamma)$. Les auteurs introduisent une fonction g et posent

$$f(\boldsymbol{x}, \boldsymbol{z}, s) = \begin{cases} \frac{g(\boldsymbol{x})}{p(S=1|\boldsymbol{x},\boldsymbol{z};\gamma)} & \text{si } s = 1 \\ 0 & \text{sinon.} \end{cases}$$

Si $p(S = 1|\boldsymbol{x}, \boldsymbol{z}; \gamma) > 0, \forall(\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}$ alors $\mathbb{E}[g(\boldsymbol{X})] = \mathbb{E}[f(\boldsymbol{X}, \boldsymbol{Z}, S)]$, ce qui conduit à l'équation suivante :

$$\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{z}_i, s_i) \approx \frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{x}_i). \tag{1.7}$$

Le paramètre γ est alors estimé en résolvant le système d'équations. Bien que cette méthode soit très générale, et qu'elle puisse s'appliquer à un grand nombre de situations, elle souffre d'instabilités comme toute méthode de type moment.

Une autre possibilité pour apprendre le biais d'étiquetage existe quand on considère des modèles génératifs et qu'on modélise les distributions de $\boldsymbol{X}, \boldsymbol{Z}|S = 1$ et $\boldsymbol{X}, \boldsymbol{Z}|\mathrm{S} = 0$. Ici l'objectif est d'apprendre au mieux la distribution de $\boldsymbol{Z}|\boldsymbol{X}, \mathrm{S} = 0$, c'est-à-dire d'apprendre la meilleure règle de classement possible pour les données non étiquetées, ce qui s'inscrit dans un contexte transductif. En établissant un lien entre ces deux distributions, Biernacki *et al.* (2002) ont montré qu'une meilleure classification pouvait être obtenue pour l'échantillon de données non étiquetées. L'application était la transposition d'une règle de classement mâle/femelle d'une population d'oiseaux tous étiquetés à une population pour laquelle aucun oiseau n'est étiqueté. En effet, si on considère la figure 1.6 on voit qu'il y a une relation entre les deux espèces considérées et qu'on doit pouvoir transposer la règle de classement de l'espèce *diomedea* à l'espèce *borealis*. Des modèles parcimonieux établissant un lien entre la distribution de l'espèce *diomedea* et la distribution de l'espèce *borealis* peuvent alors être utilisés.

Exemple de situation MNAR Le biais d'étiquetage survient typiquement en évaluation de risques clients (Thomas *et al.* , 2002). En effet, considérons des prêts octroyés selon une règle reposant sur un ensemble de variables $\boldsymbol{x}$ qu'on peut décomposer sous la forme $\boldsymbol{x} = (\boldsymbol{x}^1, \boldsymbol{x}^2)$. Si par la suite la règle de classement n'est plus apprise qu'à partir de $\boldsymbol{x}^1$ on sera alors dans le contexte MNAR. En effet, on a bien $p(s|\boldsymbol{z}, \boldsymbol{x}) = p(s|\boldsymbol{x})$, mais on n'a plus $p(s|\boldsymbol{z}, \boldsymbol{x}^1) = p(s|\boldsymbol{x}^1)$ sauf si $\boldsymbol{Z} \perp \boldsymbol{X}^2|\boldsymbol{X}^1$. Cette situation a lieu si pour diverses raisons certaines variables ne sont plus prises en compte dans le calcul du score. Pour éviter ce problème, il faudrait une période transitoire où les clients ne sont acceptés ou refusés qu'à partir de $\boldsymbol{x}^1$, puis ne réapprendre la règle de classement qu'à partir des individus acceptés ou refusés selon cette règle.

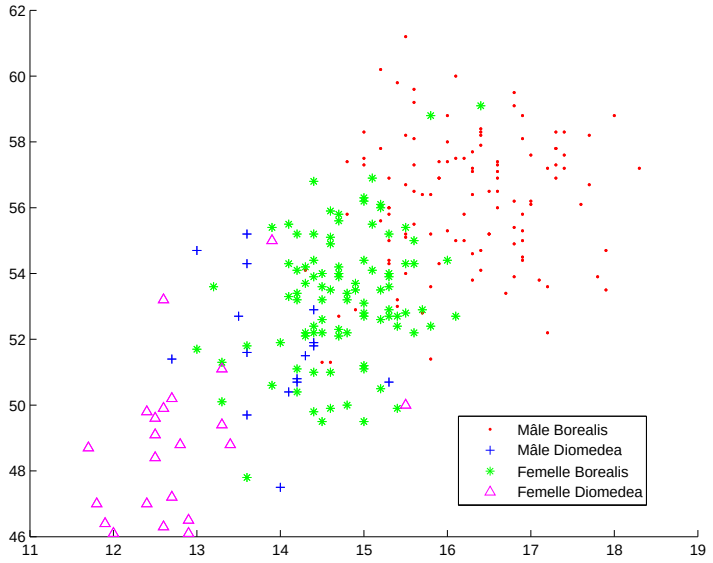

FIG. 1.6 – Analyse discriminante généralisée.

1.2.2 Variantes sur les données disponibles et l'objectif suivi

Autre type d'information sur la classe

Un autre type d'information s'inscrit dans le cadre semi-supervisé. Il s'agit de situations où l'on dispose de l'information suivante : on sait que l'individu A est dans la même classe que l'individu B (*Must-Link*) mais dans une classe différente de l'individu C (*Cannot-Link*). Cette information est appelée contrainte de paire. Il s'agit d'une connaissance partielle de la classe, puisque seules les appartenances relatives de certains individus sont connues. Ce type de données est facile à obtenir, il suffit pour cela de demander à un expert s'il pense que deux données (textes, images, ...) sont dans une même classes ou dans des classes différentes. Cette approche n'implique pas pour autant que l'expert ait une idée précise du nombre et de la signification des classes. On peut d'ailleurs voir cette approche également comme du non supervisé avec une information supplémentaire. Les contraintes aident alors à obtenir des classes plus en accord avec l'idée que s'en fait l'expert. Elle pose principalement deux questions. D'une part, une question de faisabilité : « Existe-t-il une solution qui vérifie l'ensemble des contraintes ? ». D'autre part, on pose la question de la prise en compte de ces contraintes dans l'algorithme de classification non supervisée.

Nous ne traiterons par la suite pas ce type de problème qui se ramène plutôt à du non supervisé avec de l'information supplémentaire, là où nous avons décidé de nous focaliser sur le cadre décisionnel.

Cadre transductif

Dans le cadre semi-supervisé, on souhaite généralement classer les données non étiquetées à disposition. On ne souhaite pas forcément apprendre une règle de classement de $\mathcal{X}$ dans $\mathcal{Z}$ (apprentissage inductif), mais seulement une application de $(\boldsymbol{x}_{n_\ell+1}, \ldots, \boldsymbol{x}_n)$ dans $\mathcal{Z}^{n_u}$ (apprentissage transductif). L'apprentissage transductif semble plus simple que l'apprentissage inductif puisqu'il ne requiert pas l'apprentissage d'une fonction définie sur $\mathcal{X}$ tout entier, mais uniquement celui d'une application à support discret fini. Cet aspect s'inscrit bien dans la philosophie de l'apprentissage statistique introduite par Vapnik (1995)

- *Do not estimate a density if you need to estimate a function.*
- *Do not estimate a function if you need to estimate values at given points.*
- *Do not estimate predictive values if your goal is to act well.*

La deuxième proposition correspond à la philosophie de l'apprentissage transductif. On peut parler d'algorithme transductif dès qu'une donnée non étiquetée est présente. Cependant comme remarqué dans la discussion de Chapelle *et al.* (2006, chap. 24), l'aspect transductif commence à avoir un réel impact quand le nombre de données non étiquetées est relativement grand. En effet, l'aspect transductif se manifeste avant tout lorsqu'on effectue simultanément le classement d'un grand nombre de données.

L'aspect transductif est également présent en classification bayesienne, puisque dans ce contexte pour classer un individu $\boldsymbol{x}_{n+1}$, on cherche à maximiser $p(\boldsymbol{z}_{n+1}|\mathbf{x}_\ell, \mathbf{z}_\ell, \boldsymbol{x}_{n+1})$ en $\boldsymbol{z}_{n+1}$. Les covariables de l'individu à classer interviennent dans la règle de classement. Dans le cadre semi-supervisé, on peut souhaiter classifier tous les individus simultanément, c'est-à-dire trouver $\mathbf{z}_u$ qui maximise $p(\mathbf{z}_u|\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u)$. Cependant dans certains cas, cette stratégie peut dégrader la règle de classement obtenue. En effet, quand le nombre de données est assez grand on a

$$p(\mathbf{z}_u, \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) \approx p(\mathbf{z}_u, \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u; \hat{\theta}_{\mathbf{z}_u, \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u}). \tag{1.8}$$

L'approche précédente revient asymptotiquement à apprendre le paramètre en maximisant la vraisemblance complétée. On substitue alors à une estimation des paramètres asymptotiquement sans biais une estimation biaisée des paramètres. Dans certaines situations, l'aspect transductif conduit donc à l'obtention d'une règle de classement biaisée. Dans le cadre bayésien l'aspect transductif est difficile à mettre en œuvre car il faut classer une observation à la fois. De même dans le cadre fréquentiste, l'aspect classifiant produit une estimation biaisée tandis que l'estimation par maximum de vraisemblance produit un résultat asymptotiquement sans biais.

Apprentissage actif

Dans un certain nombre de situations réelles, le praticien dispose d'un ensemble de données non étiquetées et il a la possibilité d'en étiqueter quelques unes. L'apprentissage actif consiste alors à choisir le plus judicieusement possible les données à étiqueter dans cet ensemble de données non étiquetées. Cette question se pose par exemple en indexation d'images (Grira *et al.* , 2005). Ce cadre est appelé apprentissage actif par opposition à l'apprentissage passif qui lui choisit les points à étiqueter au hasard. Prenons l'exemple

d'un étudiant qui déciderait de choisir au hasard les matières à réviser. Il risque de perdre du temps car il risque de tomber souvent sur des matières qu'il maitrise déjà bien et à l'inverse ne pas tomber assez souvent sur les points où il a des lacunes. Ainsi, il serait plus utile pour lui dans un premier temps de repérer les matières dans lesquelles il a des lacunes puis de se concentrer sur ces dernières. Il en est de même pour l'apprentissage actif. On connait quelque chose à partir des données étiquetées, les données non étiquetées nous permettent de juger de l'étendue de nos lacunes. Une fois cette étendue connue on sait dans quelle direction il faut apprendre.

D'un point de vue théorique, il est difficile de prouver qu'il est possible de choisir mieux que le hasard les exemples les plus pertinents à étiqueter. Pour cette tâche les modèles prédictifs semblent offrir de meilleures garanties que les modèles génératifs. En effet, dans certaines situations où les données sont séparables, Dasgupta *et al.* (2005) ont montré que le nombre de données nécessaires pour apprendre activement est très réduit par rapport au nombre de données obtenues passivement et nécessaires pour avoir le même résultat. Cependant dans les situations où ces hypothèses ne sont pas vérifiées, l'apprentissage actif peut causer la perte de la consistance de l'apprentissage passif. De telles situations font douter de l'utilité réelle de l'apprentissage actif. La question qui se pose alors est l'existence d'une approche permettant de préserver la consistance de l'estimation tout en limitant le nombre de données à étiqueter. Ce point est traité par Bach (2007) dans le cas des modèles linéaires généralisés, où les données étiquetées activement sont repondérées pour corriger le biais d'échantillonnage introduit par le choix des données à étiqueter. Plus récemment, ce problème a été traité dans un contexte plus général par Beygelzimer *et al.* (2009) dans le cas où les données non étiquetées arrivent de manière séquentielle. Leur approche permet alors de limiter le nombre de données à étiqueter pour avoir des résultats comparables à l'apprentissage passif. Notons toutefois que ces méthodes ne permettent pas de faire usage des données non étiquetées dans l'apprentissage, ce qui représente une perte d'information. D'autre part, la dernière méthode énoncée implique que les données arrivent séquentiellement ce qui est naturel dans certaines situations mais pas dans d'autres. Ainsi les méthodes prédictives peuvent faire un usage efficace de l'apprentissage actif.

Pour les méthodes génératives, l'avantage de l'apprentissage actif est moins évident. En effet l'information apportée par les données non étiquetées est déjà intégralement prise en compte. Toutefois cette approche a été utilisée par McCallum & Nigam (1998) en classification de textes où l'algorithme *Query By Commitee* (Freund *et al.* , 1997) a été utilisé. Cet algorithme choisit le point qui produit le plus grand désaccord pour différents classifieurs appris. La justification de cette approche est avant tout heuristique. Il n'est pas évident d'un point de vue théorique qu'une telle approche puisse nous aider. En effet, si le modèle postulé est correct, l'hypothèse MAR reste respectée et préserve donc la consistance de l'estimation par maximum de vraisemblance. Il n'est toutefois ni évident que cette approche réduise la variance des estimateurs ni qu'elle améliore systématiquement la règle de classement. D'autre part, contrairement aux approches prédictives, si le modèle postulé est incorrect, on ne peut rien dire d'un point de vue théorique sur cette approche.

Un autre problème relié à l'apprentissage actif est celui de la découverte de nouvelles classes dans l'échantillon de données non étiquetées. La présence de nouvelles classes peut avoir deux causes :

– soit l'échantillon de données étiquetées est petit et certaines classes sont en faibles

proportions,
- soit il y a un biais d'étiquetage c'est-à-dire que même si le nombre de données étiquetées augmentait certaines classes ne seraient jamais observées.

Le problème est alors plutôt un problème de classification non supervisée auquel les méthodes prédictives sont incapables répondre. Il s'agit principalement de trouver le nombre de classes dans un mélange de distributions. Ce problème trouve des solutions en classification non supervisée via des critères de choix de modèle comme les critères BIC (Schwarz, 1978) ou ICL (Biernacki *et al.* , 2000). Une question qui se pose une fois les nouvelles classes détectées est de les nommer, cela est possible quand on peut étiqueter de nouvelles données c'est-à-dire faire de l'apprentissage actif. Cette approche peut notamment être utilisée en astronomie pour la découverte de classes en classification de galaxies (Bazell & Miller, 2005).

1.2.3 Hypothèses nécessaires pour prendre en compte les données non étiquetées

L'absence d'hypothèses sur $\boldsymbol{X}$ qui pouvait être un atout des méthodes prédictives dans le cadre supervisé devient un handicap dans le cadre semi-supervisé. En effet, l'information apportée par les données non étiquetées ne porte que sur la distribution de $\boldsymbol{X}$. Pour la prise en compte de l'information apportée par $\boldsymbol{X}$, les méthodes prédictives nécessitent de rétablir le lien entre la distribution de $\boldsymbol{Z}|\boldsymbol{X}$ qui est modélisée, et la distribution de $\boldsymbol{X}$ qui ne l'est pas. Pour cela, elles essaient de mettre en œuvre des hypothèses implicites sur le lien entre la distribution de $\boldsymbol{Z}|\boldsymbol{X}$ et la distribution de $\boldsymbol{X}$.

D'un autre côté les méthodes génératives modélisent la distribution du couple $(\boldsymbol{X}, \boldsymbol{Z})$. Elles ne nécessitent donc pas d'hypothèses supplémentaires pour prendre en compte l'information apportée par les données non étiquetées.

Nous détaillons maintenant les hypothèses de « bon sens » qui permettent à toutes ces méthodes d'apprentissage de prendre en compte l'information sur les données non étiquetées.

Hypothèses de « bon sens »

Les hypothèses les plus courantes en classification semi-supervisée sont les suivantes (Chapelle *et al.* , 2006) :
- **Hypothèse de régularité :** si deux points dans des zones de forte densité sont proches alors il devrait en être de même pour leur classe.
- **Hypothèse de *cluster* :** si deux points sont dans le même *cluster* (groupe de points au sens non supervisé) alors il est probable qu'ils soient dans la même classe.
- **Hypothèse de séparation par zones de faible densité :** la frontière de classification se trouve dans des zones de faible densité.
- **Hypothèse de dimensionnalité :** les données en grande dimension appartiennent à des sous espaces de petite dimension.

Nous détaillons comment les hypothèses du semi-supervisé sont mises en œuvre par les méthodes prédictives et par les méthodes génératives.

Mise en place pour les méthodes prédictives

Hypothèse de régularité : cette hypothèse effectue un lien entre zones de forte densité et classes, elle justifie ainsi l'utilisation d'algorithmes de type propagation des étiquettes dans un graphe (Zhou *et al.* , 2004).
Hypothèse de *cluster* : cette hypothèse fait un lien entre la structure non supervisée des données et la structure supervisée du problème. Cette hypothèse est faite implicitement pour certains modèles prédictifs.
Hypothèse de séparation par zones de faible densité : cette hypothèse est à relier aux SVM transductifs (Vapnik, 1998; Joachims, 1999a), elle favorise les frontières de classement se situant dans des zones de faible densité. L'information apportée par les données non étiquetées permet alors une approximation plus précise de cette densité.
Hypothèse de dimensionnalité : cette hypothèse est utile notamment si les données en grande dimension appartiennent à des sous espaces de plus petite dimension. En effet, les données non étiquetées permettent alors une réduction efficace de la dimension. L'apprentissage de la règle de classement devient alors plus efficace dans cet espace de dimension moindre.

Mise en place pour les méthodes génératives

Hypothèse de régularité : Cette hypothèse n'est pas imposée dans les modèles génératifs. Cependant, les modèles à plusieurs composants par classe peuvent permettre le respect de ce type d'hypothèse, en effectuant une propagation des étiquettes des zones étiquetées vers les zones non étiquetées.
Hypothèse de *cluster* : Cette hypothèse de *cluster* est faite d'office dans les modèles génératifs, puisqu'on impose la correspondance entre les groupes obtenus d'un point de vue supervisé et les groupes obtenus d'un point de vue non supervisé.
Hypothèse de séparation par zones de faible densité : Cette hypothèse n'est pas formulée par les modèles génératifs, elle est cependant souhaitable lorsqu'on veut obtenir une règle de classement avec un faible taux d'erreur.
Hypothèse de dimensionnalité : Les modèles génératifs éprouvent parfois des difficultés à prendre en compte les données en grande dimension, c'est par exemple le cas des modèles gaussiens lorsque la matrice de covariance complète est estimée. Récemment des modèles génératifs haute dimension ont étés proposés par Bouveyron *et al.* (2007). Ceux-ci supposent que les données en grande dimension appartiennent intrinsèquement à des espaces de faible dimension. Ces modèles font donc l'hypothèse de dimensionnalité.

Dans ce qui suit nous montrons comment ces diverses hypothèses sont appliquées de manière générale. Puis nous verrons plus en détail leur application aux méthodes prédictives et aux méthodes génératives.

1.3 Différentes approches en classification semi-supervisée

Nous distinguons trois approches en classification semi-supervisée. D'abord les méthodes générales qui peuvent être appliquées à n'importe quelle méthode de classification supervisée. Ensuite, les méthodes spécifiques aux modèles prédictifs. Enfin les méthodes

spécifiques aux modèles génératifs.

1.3.1 Méthodes générales

Ici nous traitons des approches qui peuvent être appliquées à n'importe qu'elle méthode de classification supervisée pour faire usage des données non étiquetées.

Réduction de la dimension

Les méthodes de réduction de la dimension permettent souvent une régularisation des solutions obtenues. Une fois la réduction de dimension effectuée, les méthodes d'apprentissage supervisé usuelles sont utilisées sur l'espace de dimension réduite. Les données non étiquetées peuvent permettre une réduction efficace de la dimension. Cette réduction de la dimension peut être réalisée à partir de méthodes comme l'ACP (Pearson, 1901), le MDS (Torgerson, 1965) ou l'ISOMAP (Tenenbaum *et al.* , 2000). L'enjeu est de représenter dans un espace de plus petite dimension des données en grande dimension, ceci avec une perte minimale d'information. Les données non étiquetées permettent l'obtention de bonnes projections des données dans un sous-espace plus petit. Remarquons qu'en pratique, des méthodes de réduction de dimension de type analyse factorielle discriminante (AFD) recherchent efficacement les projections qui séparent au mieux les classes. Cependant, l'AFD n'utilise pas les données non étiquetées et se révèle donc inappropriée quand le nombre de données étiquetées est petit. Des méthodes de réduction de la dimension qui combinent les aspects réduction de la dimension et projection sur les axes les plus discriminants ont également été étudiées dans le cadre semi-supervisé (Wu *et al.* , 2000). Ces dernières essaient de recréer l'étiquette manquante des données non étiquetées à partir d'une approche de type EM, puis reprojettent selon les axes les plus discriminants.

Concernant l'ACP, l'amélioration est mise en évidence en utilisant les données Vin de l'UCI *Database Machine Learning Repository*[2]. Pour ces données, on dispose de 178 observations appartenant à 3 classes différentes et pour lesquelles 13 variables ont été observées. Initialement, aucune paire d'axes ne permet d'obtenir des classes bien séparées. On réalise une ACP sur ces données et on obtient la figure 1.7. Les classes sont très bien séparées uniquement avec les deux premiers axes de l'ACP.

Auto-apprentissage

L'auto-apprentissage permet d'adapter n'importe quelle méthode de classification supervisée à la classification semi-supervisée. Historiquement, l'auto-apprentissage est une des premières méthodes utilisées en classification semi-supervisée (Scudder, 1965; Fralick, 1967; Agrawala, 1970). L'algorithme est le suivant :

- Apprendre la règle de classement δ à partir de l'échantillon de données étiquetées $(\mathbf{x}_\ell, \mathbf{z}_\ell)$.
- Tant qu'il reste des données non étiquetées $\mathbf{x}_u$:
  - Appliquer δ à $\mathbf{x}_u^i$ une fraction des données non étiquetées restantes pour obtenir $\hat{\mathbf{z}}_u^i$;

[2]http ://archive.ics.uci.edu/ml/

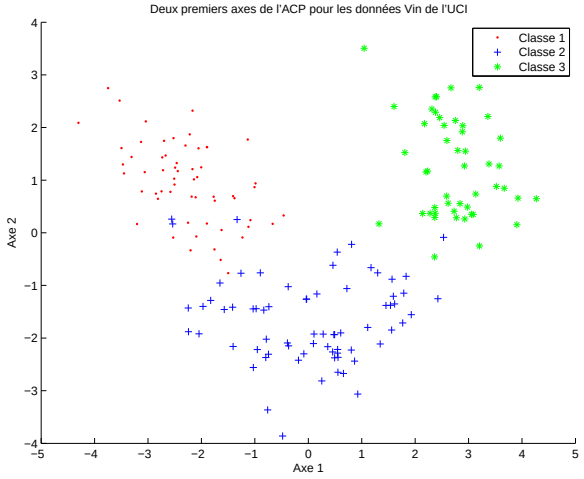

FIG. 1.7 – ACP pour les données Vin de l'UCI.

– Réapprendre δ à partir de $(\mathbf{x}_\ell, \mathbf{z}_\ell)$ et $(\mathbf{x}_u^1, \hat{\mathbf{z}}_u^1), \ldots, (\mathbf{x}_u^i, \hat{\mathbf{z}}_u^i)$.

En pratique, l'auto-apprentissage est facile à mettre en œuvre. A chaque étape, il fournit des données avec des étiquettes et permet ainsi une utilisation directe des méthodes de classification supervisée. Toutefois, son comportement dépend fortement de la méthode de classification supervisée utilisée.

Elle ne possède pas de justification théorique rigoureuse. Cependant quand les classes sont bien séparées, elle est susceptible d'améliorer les résultats. En effet, dans ce cas, l'hypothèse de séparation par zones de faible densité est mise en œuvre.

Nous illustrons maintenant son application à la régression logistique linéaire. Soit un problème à deux classes, on génère 20 données de la classe 1 et 20 de la classe 2. À ceci on ajoute 400 données dont l'étiquette est cachée et pour lesquelles 200 données appartiennent à la classe 1 et 200 données appartiennent à la classe 2. $\boldsymbol{X}|Z_1 = 1 \sim \mathcal{N}((-1,-1)', I_2)$ et $\boldsymbol{X}|Z_2 = 1 \sim \mathcal{N}((1,1)', I_2)$. L'auto-apprentissage est réalisé en incorporant successivement 10% des données non étiquetées au hasard. La figure 1.8 représente les données non étiquetées, la frontière optimale, la frontière apprise de manière supervisée, et les dix frontières successives résultant de l'auto-apprentissage. On remarque sur ces figures, que pour cet exemple, l'auto-apprentissage ne conduit pas au rapprochement de la frontière de classement vers la frontière optimale.

Co-training

Le *co-training*, suppose que les variables sont naturellement partitionnées en deux ensembles $\boldsymbol{x} = (\boldsymbol{x}^1, \boldsymbol{x}^2)$. Par exemple, pour les pages Web on considère l'ensemble liens hypertextes et l'ensemble contenu. Sous les hypothèses suivantes :

1. chaque composant est suffisant pour la classification,
2. les composants sont indépendants conditionnellement à la classe,

Blum & Mitchell (1998) démontrent des garanties de type *Probably Approximately Correct* (PAC) (Valiant, 1984) sur l'apprentissage en présence de données étiquetées et non étiquetées.

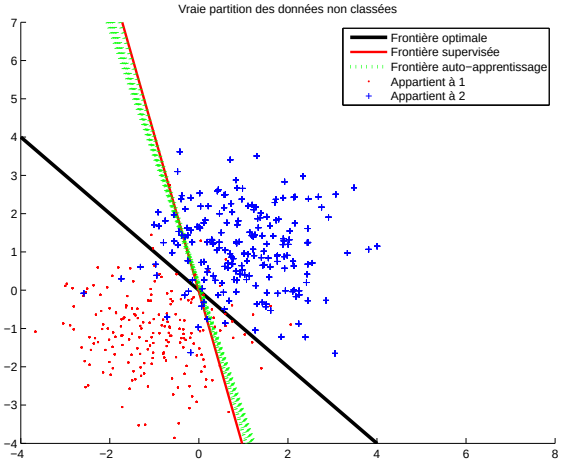

FIG. 1.8 – Vraie partition et frontières de classement : optimale, supervisée et auto-apprise.

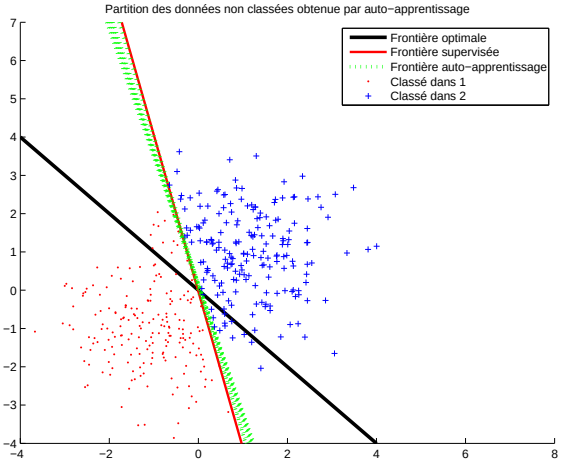

FIG. 1.9 – Partition auto-apprise et frontières de classement : optimale, supervisée et auto-apprise.

Un exemple d'algorithme de co-training est le suivant

- **Entrée :** une collection initiale de documents étiquetés.
- Boucler jusqu'à ce qu'il n'y ait plus de document sans étiquette.
  - Construire le classifieur δ_1 en utilisant la partie $\boldsymbol{x}^1$ de chaque document.
  - Construire le classifieur δ_2 en utilisant la partie $\boldsymbol{x}^2$ de chaque document.
  - Pour chaque classe k, ajouter à la collection de documents étiquetés le document non-étiqueté classé dans la classe k par le classifeur δ_1 avec la plus forte probabilité.
  - Pour chaque classe k, ajouter à la collection de documents étiquetés le document non-étiqueté classé dans la classe k par le classifeur δ_2 avec la plus forte probabilité.
- **Sortie :** Deux classifieurs, δ_1 et δ_2, qui prédisent l'étiquette des nouveaux documents. Ces prédictions peuvent ensuite être combinées.

1.3.2 Méthodes prédictives

Ici nous détaillons un certain nombre d'approches qui tentent d'étendre les méthodes de classification supervisée à la classification semi-supervisée. La diversité de ces méthodes montrera qu'il est difficile pour les méthodes prédictives de faire de la classification semi-supervisée dans un cadre cohérent et homogène, contrairement aux méthodes génératives qui sont détaillées par la suite.

SVM transductifs

En classification supervisée, les SVM maximisent la marge uniquement à partir des données étiquetées. Dans le cadre semi-supervisé, les SVM transductifs proposent de maximiser la marge sur l'ensemble des données, en choisissant l'étiquetage des données non

étiquetées le plus favorable :

$$\begin{aligned} \text{Minimiser :} \quad & \tfrac{1}{2}\|\omega\|^2, \\ \text{Sujet à :} \quad & \forall i \in \{1,\dots,n_\ell\} : y_i[\omega' \boldsymbol{x}_i + b] \geqslant 1, \\ & \forall j \in \{n_\ell+1,\dots,n\} : y_j^*[\omega' \boldsymbol{x}_j + b] \geqslant 1. \end{aligned}$$

Les SVM transductifs minimisent alors une borne en probabilité sur l'erreur commise sur les données non étiquetées. La précision de cette borne est améliorée quand les données non étiquetées permettent de structurer plus précisément l'espace des hypothèses (Vapnik, 1998).

Ce problème est difficile à résoudre exactement car il y a 2^{n_u} étiquetages possibles des données non étiquetées. Sa résolution exacte est même impossible lorsque le nombre de données non étiquetées excède la centaine. Des approches heuristiques permettent toutefois de faire face à ce problème. C'est par exemple le cas des SVM^{light} de Joachims (1999a), ces derniers nécessitant de fixer la proportion d'exemples étiquetés positivement et négativement afin d'éviter l'obtention de solutions dégénérées.

Une autre possibilité pour résoudre le problème des SVM transductifs est l'utilisation d'outils de programmation semi-définie positive (Bie & Cristianini, 2004). Notons $\Gamma = YY'$, cette matrice se réécrit

$$\Gamma = \begin{pmatrix} Y_\ell Y_\ell' & Y_\ell Y_u' \\ Y_u Y_\ell' & Y_u Y_u' \end{pmatrix}$$

où $Y_u \in \{-1;1\}^u$. Remarquons qu'on a

$$\text{diag}(\Gamma) = 1, \text{rang}(\Gamma) = 1, \Gamma \succ 0,$$

où diag représente les élément diagonaux de la matrice, rang représente le rang de la matrice, et $\Gamma \succ 0$ signifie que la matrice Γ est définie positive. Bie & Cristianini (2004) reformulent alors le problème d'optimisation comme un problème d'optimisation sur Γ, et relâchent les contraintes $Y_u \in \{-1;1\}^u$ et $\text{rang}(\Gamma) = 1$. Ce qui conduit à un problème d'optimisation semi-définie positive. Cette méthode permet le traitement de situations possédant jusqu'à 1000 données non étiquetées.

Dans la pratique, les SVM transductifs améliorent parfois la solution supervisée et la dégrade d'autres fois. L'amélioration a plutôt lieu quand l'hypothèse de séparation par zones de faible densité est vérifiée. Par exemple, Joachims (1999b) a constaté des améliorations en classification de textes.

Propagation des étiquettes dans un graphe

Les modèles à base de graphes reposent sur une matrice de voisinage W. Cette dernière est construite à partir d'une distance entre deux points de l'espace qui nécessite souvent le choix d'un paramètre de réglage. Par exemple

$$w_{ij} = e^{-\frac{\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2}{2\sigma^2}},$$

si on considère le noyau gaussien avec une fenêtre de largeur σ. On peut aussi considérer la matrice de voisinage W des k plus proches voisins. Ensuite, un algorithme itératif est

utilisé pour propager les étiquettes des points étiquetés vers les points non étiquetés à travers le graphe (Zhou *et al.* , 2004). Cette propagation repose principalement sur la méthode de Jacobi pour la résolution de systèmes linéaires (Saad, 2003). La complexité de cet algorithme est proportionnelle au nombre moyen de voisins. Ainsi dans certains cas, il est utile de seuiller les plus petites valeurs pour obtenir un graphe moins dense et donc un algorithme plus rapide. Quand les paramètres de réglage sont bien calibrés, de bons résultats sont obtenus (Zhou *et al.* , 2004).

Cette méthode repose sur l'hypothèse de régularité vue section 1.2.3. En effet, dans le cas où les données d'une même classe sont reliées par des zones de forte densité, on obtient un graphe dans lequel les données d'une même classe sont principalement reliées entre elles. Typiquement, cette méthode donne de bons résultats quand les classes sont séparées par des zones de faible densité et bien reliées entre elles. Pour illustrer ce propos, prenons l'exemple des deux lunes enchevêtrées (figure 1.10). Ici les méthodes linéaires sont incapables de trouver une bonne séparation des classes, tandis que les méthodes à base de graphe la trouvent efficacement. Dans ce contexte, les données non étiquetées permettent de relier des points de même classe par des chemins de haute densité.

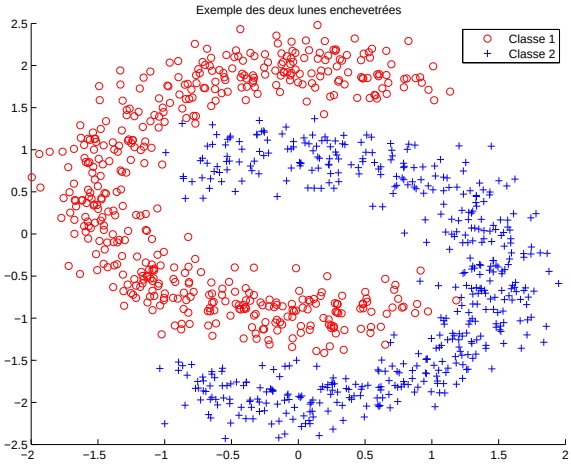

FIG. 1.10 – Exemple des deux lunes enchevêtrées.

Différents algorithmes fondés sur des graphes ont étés proposés (Szummer & Jaakkola, 2002; Zhu & Ghahramani, 2002b; Zhou *et al.* , 2004). En voici un exemple (Zhu & Ghahramani, 2002a) :

- Calculer la matrice de voisinage W
- Calculer la matrice diagonale D par $D_{ii} \leftarrow \sum_j w_{ij}$
- Initialiser $\widehat{Y}^{(0)} \leftarrow (y_1, \ldots, y_{n_\ell}, 0, \ldots, 0)$
- Boucler jusqu'à la convergence vers $\widehat{Y}^{(\infty)}$
  1. $\widehat{Y}^{(t+1)} \leftarrow D^{-1} W \widehat{Y}^{(t)}$
  2. $\widehat{Y}^{(t+1)} \leftarrow Y_\ell$
- Étiqueter les point $\boldsymbol{x}_i$ par le signe de $\widehat{y}_i^{(\infty)}$

Illustrons maintenant cet algorithme en conservant 20 points étiquetés sur les 1000 données sur l'exemple des deux lunes enchevêtrées. Le graphe des 5 plus proches voisins

est construit figure 1.11. En appliquant l'algorithme précédent, on obtient la figure 1.12. On remarque dans ce cas que l'algorithme de propagation des étiquettes produit de très bons résultats. Toutefois, les données réelles n'ont une distribution de ce type que dans des cas bien particuliers. En général, les classes sont chevauchantes par ailleurs, ce qui peut remettre en cause les performances de tels algorithmes.

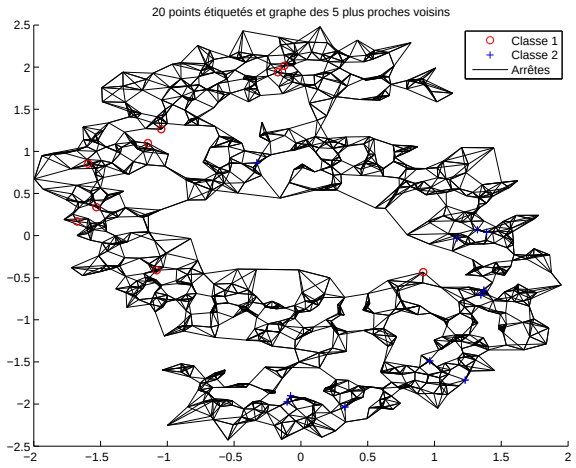

FIG. 1.11 – Graphe des cinq plus proches voisins.

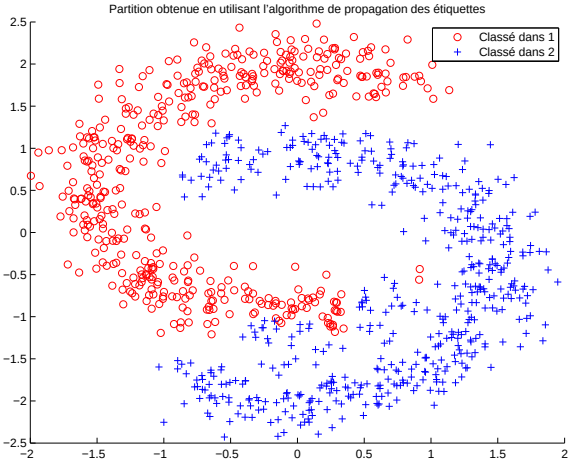

FIG. 1.12 – Etiquetage obtenu par propagation des étiquettes.

Ces méthodes mettent directement en avant l'aspect transductif dont nous avons discuté section 1.2.2. En effet, seuls les points non étiquetées appartenant au graphe peuvent être étiquetés par propagation des étiquettes. Pour étiqueter une nouvelle donnée, il faut donc reconstruire le graphe et appliquer à nouveau l'algorithme de propagation des étiquettes.

Pondération de la fonction objectif

Les données non étiquetées permettent également de pondérer de manière adéquate les observations pour les méthodes prédictives (Sokolovska *et al.* , 2008). Théoriquement, ceci est obtenu quand $\mathcal{X}$ est un dénombrable fini. Dans ce cadre, une connaissance quasi-parfaite de $p(\boldsymbol{x})$ est obtenue quand le nombre de données non étiquetées est assez grand. Supposons $p(\boldsymbol{x})$ connu, une pondération astucieuse de l'observation i est alors :

$$\frac{p(\boldsymbol{x}_i)}{\sum_{j=1}^{n_\ell} \mathbf{1}[\boldsymbol{x}_i = \boldsymbol{x}_j]}.$$

Ce qui conduit à l'optimisation de la fonction suivante :

$$\sum_{i=1}^{n_\ell} \frac{p(\boldsymbol{x}_i)}{\sum_{j=1}^{n_\ell} \mathbf{1}[\boldsymbol{x}_i = \boldsymbol{x}_j]} \log p(\boldsymbol{z}_i|\boldsymbol{x}_i;\theta).$$

Si les données étiquetées et non étiquetées ont la même distribution alors la fonction objectif sera asymptotiquement la même que sans pondération, c.-à-d. :

$$\mathbb{E}_{\mathbf{X},\mathbf{Z}}[\log p(\mathbf{Z}|\mathbf{X};\theta)]. \tag{1.9}$$

Dans ce cas, le semi-supervisé ne modifie en rien la solution obtenue asymptotiquement mais il est susceptible de l'atteindre plus rapidement. Effectivement Sokolovska *et al.* (2008) montrent que la variance asymptotique des estimateurs est alors réduite. Toutefois cette amélioration n'a lieu que si le modèle postulé est faux, hypothèse réaliste en pratique. Si $p(\boldsymbol{x})$ était exactement connu grâce à un échantillon de taille infinie de données non étiquetées, cette méthode aurait un intérêt certain. Cependant en pratique même si l'échantillon de données non étiquetées est grand, il ne fournit jamais une connaissance parfaite de $p(\boldsymbol{x})$ surtout si $\boldsymbol{x}$ appartient à un espace de grande dimension. Les auteurs proposent donc de partitionner $\mathcal{X}$ en un ensemble de taille raisonnable par rapport au nombre de données non étiquetées disponibles, cela en utilisant un algorithme de classification non supervisée. Les modifications apportées à la pondération sont d'une part le remplacement de $p(\boldsymbol{x}_i)$ par le poids du *cluster* dans lequel l'individu i est classé, et d'autre part celui de $\sum_{j=1}^{n_\ell} \mathbf{1}[\boldsymbol{x}_i = \boldsymbol{x}_j]$ par le nombre de données classées dans le même *cluster* que $\boldsymbol{x}_i$. Cette heuristique permet de traiter des problèmes réels tels que la classification de textes (Nigam *et al.* , 2000).

Si la distribution marginale des données non étiquetées est différente de la distribution marginale des données étiquetées, mais que la classe est manquante au hasard (MAR), cette méthode permet de repondérer la fonction objectif et ainsi de mieux prédire la classe des données non étiquetées. En effet, soit $q(\boldsymbol{x})$ la densité des données étiquetées et $p(\boldsymbol{x})$ la densité des données non étiquetées, sans pondération on obtient

$$\sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{z}\in\mathcal{Z}} q(\boldsymbol{x}) p(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{z}|\boldsymbol{x};\theta) \tag{1.10}$$

tandis qu'avec pondération on obtient

$$\sum_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{z}\in\mathcal{Z}} p(\boldsymbol{x}) p(\boldsymbol{z}|\boldsymbol{x}) \log p(\boldsymbol{z}|\boldsymbol{x};\theta). \tag{1.11}$$

L'expression (1.10) correspond à la fonction objectif qu'on cherche à optimiser pour bien classer des données provenant de la distribution marginale $q(\boldsymbol{x})$. Tandis que l'équation (1.11) correspond à la fonction qu'on cherche à optimiser pour bien classer les données provenant de la distribution marginale $p(\boldsymbol{x})$. Ainsi dans le cas où ce sont les données non étiquetées à disposition qu'on souhaite classer, on a intérêt à utiliser la pondération proposée. Notons pour cela qu'il faut $q(\boldsymbol{x}) > 0 \Rightarrow p(\boldsymbol{x}) > 0$, ce qui n'est pas systématique. Par exemple, en évaluation des risques clients, pour certaines valeurs de $\boldsymbol{x}$ le prêt est systématiquement refusé. Notons que cette possibilité met clairement en avant le cadre transductif, c'est-à-dire la prise en compte du fait que les données que nous souhaitons classer sont les données non étiquetées à disposition. De même que dans la situation manquant totalement au hasard (MCAR), l'intérêt de cette approche se manifeste principalement quand le modèle postulé est faux. Des approches similaires ont été utilisées dans le cadre de la régression (Shimodaira, 2000).

Remarquons que même dans une approche qui évite de modéliser l'information sur $\boldsymbol{X}$, cette dernière est plus ou moins prise en compte par la nécessité de regroupement des modalités par une approche non supervisée. Le point délicat de ce type de procédure est celui du choix du nombre de *clusters* à utiliser pour le découpage de $\mathcal{X}$. En effet, un découpage trop grossier donne une pondération imprécise alors qu'un découpage trop fin produit une estimation grossière. On voit ici la difficulté des modèles prédictifs à prendre

efficacement en compte l'information apportée par les données non étiquetées. En effet, ces dernières sont utiles uniquement si le modèle postulé est incorrect ou si les données sont manquantes au hasard. Cette perspective est certes réaliste, mais elle n'offre pas au modélisateur la possibilité de prendre en compte toute l'information disponible quand des modèles corrects existent.

Régularisation d'une solution supervisée

La solution supervisée est souvent instable quand le nombre de données étiquetées est petit. Il est alors souhaitable de la régulariser, en privilégiant une solution avec des propriétés qui nous semblent pertinentes. Dans le cadre semi-supervisé, les propriétés pertinentes correspondent aux hypothèses de la section 1.2.3. Par exemple, l'hypothèse de séparation par des zones de faible densité est utilisée d'une part dans la régression logistique régularisée par l'entropie, et d'autre part dans les méthodes de régularisation par l'information mutuelle.

La régression logistique régularisée optimise le critère suivant (Grandvalet & Bengio, 2006) :

$$\underbrace{\sum_{i=1}^{n_\ell} \log p(\boldsymbol{z}_i|\boldsymbol{x}_i;\theta)}_{\text{Régression logistique}} + \lambda \underbrace{\sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} p(Z_{ik}=1|\boldsymbol{x}_i;\theta) \log p(Z_{ik}=1|\boldsymbol{x}_i;\theta)}_{\text{Régularisation}}, \tag{1.12}$$

par rapport à θ, et où λ est un paramètre de régularisation fixé à l'avance. Le terme de régularisation repousse les frontières de classification vers les zones de faible densité. L'optimisation est réalisée à l'aide d'un algorithme de type recuit déterministe (Rose, 1998). Il consiste à alterner deux étapes :

– Pour chaque individu i et chaque classe k calculer :

$$t_{ik}^{(r+1)} = \frac{p(Z_{ik}=1|\boldsymbol{x}_i;\theta^{(r)})^{\frac{1}{1-\lambda}}}{\sum_{l=1}^{g} p(Z_{il}=1|\boldsymbol{x}_i;\theta^{(r)})^{\frac{1}{1-\lambda}}}.$$

– Optimiser en θ l'expression :

$$\sum_{i=1}^{n_\ell} \log p(\boldsymbol{z}_i|\boldsymbol{x}_i;\theta) + \lambda \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} t_{ik}^{(r+1)} \log p(Z_{ik}=1|\boldsymbol{x}_i;\theta),$$

cette optimisation pouvant être effectuée par un algorithme de Newton.

Remarquons que :

– si $\lambda = 0$, on retrouve le problème non régularisé,

– si $\lambda = 1$, on retrouve l'algorithme logistic-CEM proposé par Amini & Gallinari (2002).

Dans le cas où on prend la constante de régularisation λ égale à 1, et que $n_u >> n_\ell$ cette approche fonctionne mal. En effet, dans ce cas c'est le terme de régularisation qui l'emporte. Le choix du paramètre de régularisation λ est donc délicat. En pratique, on choisit la valeur de λ qui minimise l'erreur de classement estimée par 10-*fold* validation croisée. Cette méthode a montré de bonnes performances en classification d'images, notamment en classification d'expressions faciales (Grandvalet & Bengio, 2006).

Une autre possibilité est la régularisation par l'information mutuelle. Cette méthode repose principalement sur le découpage du domaine en petites régions pour lesquelles on calcule l'information mutuelle. L'information mutuelle globale est ensuite obtenue par une combinaison linéaire des informations mutuelles locales (Corduneanu & Jaakkola, 2004). De même que dans la régularisation par l'entropie de classification, l'opposé de l'information mutuelle joue un rôle de régularisation. Elle favorise les solutions où l'information mutuelle est importante.

Conclusion sur l'extension des méthodes prédictives au semi-supervisé

La diversité de ces méthodes doit nous faire remarquer deux principaux points. D'une part l'extension des méthodes prédictives au semi-supervisé ne va pas de soi. D'autre part, même si une amélioration a parfois lieu en utilisant la classification semi-supervisée, l'absence d'une approche unificatrice dans ce cadre décourage l'utilisation de ces méthodes.

Dans la partie suivante nous montrons que les méthodes génératives ne nécessitent pas d'hypothèses supplémentaires pour prendre en compte l'information apportée par les données non étiquetées. En effet ces dernières fonctionnent aussi bien dans le cadre non supervisé que dans le cadre supervisé et par suite dans le semi-supervisé. C'est pour ces méthode que le gain apporté par les données non étiquetées est le plus important. Elles constituent les seules méthodes capables de prendre en compte de façon rigoureuse l'information apportée par les données non étiquetées.

1.3.3 Méthodes génératives

Historique

Dans la communauté statistique, l'approche semi-supervisée a débuté avec la volonté d'actualiser la règle de classement de l'analyse discriminante linéaire à partir de données non étiquetées dans le cas gaussien homoscédastique (Ganesalingam & McLachlan, 1978; O'Neill, 1978). À notre connaissance, l'approche la plus ancienne est celle d'Hosmer (1973). Dans ses travaux, l'objectif était d'estimer la proportion de mâles et de femelles d'une population de flétans (grand poisson plat des mers froides). Pour ces poissons, la détermination du sexe n'est possible qu'après dissection. Des mesures biométriques de nombreux individus peuvent cependant aider à déterminer la fraction de mâles et de femelles. Hosmer avait à sa disposition de nombreuses données commerciales mentionnant la longueur des flétans pêchés. En parallèle, il disposait d'une étude scientifique, sur un nombre beaucoup plus petit de données, mentionnant également le sexe. L'utilisation des données commerciales ajoutées aux données de l'étude scientifique ont permis une estimation plus précise des paramètres. En effet, dans ces travaux l'estimation a été réalisée par maximum de vraisemblance grâce à un algorithme itératif. Plus tard, ce dernier sera formalisé sous le nom d'algorithme EM (Dempster *et al.* , 1977). Cet algorithme est bien adapté pour traiter les problèmes de données manquantes comme ceux des données non étiquetées pour lesquelles les étiquettes manquent. Hosmer distingue trois schémas d'échantillonnage différents :

(M1) : Seules des données non étiquetées sont disponibles.

(M2) : Des données étiquetées provenant d'un échantillonnage rétrospectif sont dispo-

nibles en plus des données non étiquetées.

(M3) : Des données étiquetées provenant d'un échantillonnage mélange sont disponibles en plus des données non étiquetées.

(M1) se situe dans le cadre non supervisé. (M2) est dans le cadre semi-supervisé avec les étiquettes des données non étiquetées qui ne manquent par au hasard. Ici, les données non étiquetées jouent déjà un rôle essentiel dans l'estimation des proportions, qui ne peuvent pas être estimées à partir des données étiquetées en raison de l'échantillonnage rétrospectif. (M3) se situe dans le schéma où les données manquent totalement au hasard. Ainsi les données non étiquetées et étiquetées sont utilisées pour estimer l'ensemble des paramètres du modèle. Les modèles utilisés consistent principalement en des mélanges de distributions gaussiennes et des mélanges de distributions multinomiales (Cooper & Freeman, 1970).

Principe

Sous les hypothèses présentées en classification supervisée (section 1.1.1), et en ajoutant l'hypothèse d'échantillonnage MCAR (section 1.2.1), la distribution de l'ensemble des données est alors totalement spécifiée, et aucune hypothèse supplémentaire n'est nécessaire. Ainsi, l'estimation des paramètres est réalisée en utilisant toutes les données (étiquetées et non étiquetées), les données non étiquetées nous permettant d'estimer plus précisément les paramètres du modèle et d'obtenir par suite une règle de classement par MAP plus précise. Cette cohérence nous amène à traiter plus en avant ces modèles dans les parties suivantes.

En pratique le continuum entre supervisé et non supervisé est établi si la variable latente en classification non supervisée correspond à l'étiquette observée en classification supervisée, c.-à-d. si le modèle postulé est correct. Si les covariables ont été correctement choisies pour la tâche de prédiction, il y a des fortes chances que l'hétérogénéité de $\boldsymbol{X}$ soit expliquée par la variable qu'on cherche à prédire $\boldsymbol{Z}$, le semi-supervisé est alors susceptible d'apporter des améliorations.

Nous détaillons plus précisément l'utilisation des modèles génératifs en classification semi-supervisée dans le chapitre 2.

1.3.4 Compromis entre prédictif et génératif

Avant de terminer ce chapitre, il nous semble intéressant de détailler une approche hybride entre modèle génératif et modèle prédictif. En effet, nous avons fait la distinction entre méthodes génératives et prédictives. Une question que nous pouvons nous poser est alors « Est-il possible de réconcilier génératif et prédictif ? ». En pratique, les modèles génératifs produisent souvent de meilleurs résultats quand le nombre de données est petit, tandis que les modèles prédictifs produisent de meilleurs résultats quand le nombre de données est grand (Ng & Jordan, 2002).

Soit un modèle génératif $p(\cdot;\theta)$. Afin d'allier faible variance des estimateurs génératifs et faible biais des estimateurs prédictifs, Bouchard & Triggs (2004) proposent de considérer

une combinaison convexe des critères génératifs et prédictifs, ce qui donne ici

$$\lambda \underbrace{\log p(\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u; \theta)}_{\text{Partie générative}} + (1-\lambda) \underbrace{\log p(\mathbf{z}_\ell | \mathbf{x}_\ell; \theta)}_{\text{Partie prédictive}} . \tag{1.13}$$

- Si $\lambda = 0$ on retrouve le modèle prédictif à une reparamétrisation identifiable du modèle près,
- si $\lambda = 1$ on retrouve le modèle génératif.

Par exemple, si le modèle génératif considéré est un modèle gaussien homoscédastique, alors le modèle prédictif correspondant sur $\boldsymbol{Z}|\boldsymbol{X}$ est la régression logistique linéaire. Pour les valeurs intermédiaires de λ l'estimateur obtenu est un compromis entre estimation prédictive et estimation générative. Quand la distribution d'échantillonnage est issue du modèle postulé, $\lambda = 1$ permet la meilleure estimation de la règle de classement (Bouchard & Triggs, 2004), tandis que si le modèle postulé est incorrect ce sera $\lambda = 0$. Cependant quand on considère des tailles d'échantillons modérées, un compromis intervient entre la variance et le biais dans l'estimation des paramètres. Ceci implique l'existence d'une valeur optimale $\lambda^* \neq 0$ qui permet la meilleure estimation possible de la distribution conditionnelle. En théorie, cette valeur peut s'exprimer sous forme de rapport de variances asymptotiques d'estimateurs (Bouchard & Triggs, 2004). Cependant en pratique, elle est obtenue par validation croisée de l'entropie de classification en faisant varier λ sur une grille discrète.

Cette approche a ensuite été reformulée dans un cadre bayesien par Lasserre *et al.* (2006). Cette dernière consiste à décomposer la vraisemblance jointe sous la forme

$$p(\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u; \theta, \theta') = p(\mathbf{z}_\ell | \mathbf{x}_\ell; \theta) p(\mathbf{x}_\ell, \mathbf{x}_u; \theta')$$

où $p(\mathbf{z}_\ell|\mathbf{x}_\ell;\theta)$ résulte d'un modèle génératif sur la distribution de (X, Z) avec $\theta \in \Theta$, et $p(\mathbf{x}_\ell, \mathbf{x}_u; \theta')$ resulte du même modèle génératif sur $(\boldsymbol{X}, \boldsymbol{Z})$ avec encore $\theta' \in \Theta$. On peut par exemple penser à un modèle gaussien homoscédastique qui implique un modèle de régression logistique linéaire sur la distribution de $\boldsymbol{Z}|\boldsymbol{X}$ et un modèle de mélange sur la distribution de $\boldsymbol{X}$. Dans le cadre bayesien, le lien entre la distribution de $\boldsymbol{Z}|\boldsymbol{X}$ et la distribution de $\boldsymbol{X}$ est rétabli à travers la distribution *a priori* $p(\theta, \theta')$ des paramètres du modèle. Dans le cas où θ et θ' sont supposés égaux on retrouve le modèle génératif, tandis que dans le cas où ils sont supposés différents et *a priori* indépendants on retrouve le modèle prédictif. Le cas intermédiaire est la situation où θ et θ' sont *a priori* dépendants. Les paramètres sont alors obtenus par MAP en optimisant

$$\log p(\mathbf{z}_\ell|\mathbf{x}_\ell;\theta) + \log p(\mathbf{x}_\ell, \mathbf{x}_u; \theta') + \log p(\theta, \theta'). \tag{1.14}$$

La difficulté à ce niveau est de choisir la distribution *a priori* $p(\theta, \theta')$. On peut effectuer la décomposition $p(\theta, \theta') = p(\theta|\theta')p(\theta')$, puis imposer $p(\theta|\theta') = p(\theta'|\theta)$ pour des raisons de symétrie. Un *a priori* du type

$$p(\theta|\theta') = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|\theta-\theta'\|^2}{2\sigma^2}}$$

peut ensuite être utilisé. Il faut alors choisir la largeur de la fenêtre σ ou alors définir un *a priori* sur cette dernière dans une perspective bayesienne. L'avantage d'une telle

approche est qu'une fois les distributions *a priori* choisies, le compromis s'obtient de manière automatique :

$$(\hat{\theta}, \hat{\theta}') = \underset{(\theta,\theta')\in\Theta^2}{\operatorname{argmax}} \log p(\mathbf{z}_\ell|\mathbf{x}_\ell;\theta) + \log p(\mathbf{x}_\ell, \mathbf{x}_u;\theta') + \log p(\theta,\theta').$$

La règle de classement pour classer une nouvelle donnée $\boldsymbol{x}_{n+1}$ est alors

$$\hat{\delta}(\boldsymbol{x}_{n+1}) = \underset{\boldsymbol{z}\in\mathcal{Z}}{\operatorname{argmax}}\, p(\boldsymbol{z}|\boldsymbol{x}_{n+1};\hat{\theta}).$$

Remarquons que l'obtention d'un bon compromis dépend souvent en pratique de l'*a priori* choisi. D'autre part, ces approches sont relativement coûteuses, puisque d'une part elles doublent le nombre de paramètres mis en jeu, et d'autre part elles nécessitent l'utilisation d'algorithmes de Newton numériquement instables en grande dimension. Cette approche, ne sera pas retenue par la suite, d'une part à cause de sa mise en œuvre difficile, et d'autre part par la nécessité de fixer des paramètres de réglage.

1.4 Conclusion

Nous avons détaillé comment les méthodes prédictives et génératives pouvaient être utilisées pour prendre en compte l'information apportée par les données non étiquetées. Nous retenons par la suite l'approche générative pour sa capacité à prendre de manière cohérente et rigoureuse l'information apportée par les données non étiquetées.

Chapitre 2

Modèles génératifs en semi-supervisé

Dans ce chapitre, nous traitons de l'utilisation des modèles génératifs dans le cadre semi-supervisé. Comme précisé dans le chapitre précédent, il s'agit de la seule méthode permettant de prendre en compte de façon rigoureuse l'information apportée par les données non étiquetées. Quand le modèle est correct, les données non étiquetées améliorent l'estimation des paramètres du modèle, et permettent par la suite d'obtenir une règle classement plus précise.

Dans un premier temps nous verrons plus en détail l'utilisation des modèles génératifs dans le cadre semi-supervisé. Nous détaillerons ensuite l'estimation des paramètres qui est généralement réalisée par maximum de vraisemblance en utilisant l'algorithme EM. Nous discuterons de l'initialisation de cet algorithme, ainsi que des stratégies permettant d'explorer efficacement l'espace des solutions, sans oublier les questions liées à sa vitesse de convergence car le mélange de données étiquetées ou non apporte de nouveaux éclairages pour ces questions. Nous détaillons ensuite quelques exemples de modèles utilisés dans le cas des données continues et discrètes. Nous illustrons enfin l'utilisation de ces modèles par des expériences sur des jeux de données variés. Une partie de ce chapitre résulte de l'état de l'art sur le semi-supervisé que nous avons dressé dans la revue Modulad (Vandewalle, 2009a)[1].

L'hypothèse MCAR définie section 1.2.1, et les hypothèses sur les modèles génératifs définies section 1.1.1 permettent de modéliser la distribution de l'ensemble des données. Il ne reste alors plus qu'à estimer les paramètres du modèle, ce dont nous parlons dans la partie suivante.

2.1 Estimation par maximum de vraisemblance

Diverses méthodes existent pour estimer les paramètres du modèle. Les deux approches les plus répandues dans le cadre des mélanges sont l'estimation de type moment d'une part et l'estimation par maximum de vraisemblance d'autre part. L'estimation de type moment consiste à partir d'un certain nombre d'équations de la forme $\mathbb{E}[h(\boldsymbol{X})] = g(\theta^*)$, à remplacer $\mathbb{E}[h(\boldsymbol{X})]$ par le moment empirique correspondant et à résoudre le système d'équations formé. Pearson (1894) utilisa cette approche pour estimer les paramètres d'un mélange de deux populations de crabes. Cependant, cette méthode se révèle souvent instable lorsque

[1] `http://www-roc.inria.fr/axis/modulad/archives/numero-40/vandewalle-40/Vandewalle-40.pdf`

des moments d'ordre élevé doivent être calculés et sa mise en œuvre nécessite la résolution de systèmes d'équations hautement non linéaires, ce qui limite son application à grande échelle. On lui préfère souvent l'estimation par maximum de vraisemblance plus facile à mettre en œuvre et possédant de meilleures propriétés sous des conditions de régularité standards. Nous détaillons maintenant l'estimation par maximum de vraisemblance.

2.1.1 Expression de la vraisemblance

L'estimation par maximum de vraisemblance consiste à maximiser la vraisemblance des paramètres conditionnellement aux données. Sous les hypothèses précédentes la log-vraisemblance du paramètre θ s'écrit

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \underbrace{\sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(\pi_k p(\boldsymbol{x}_i; \theta_k))}_{\text{Données étiquetées}} + \underbrace{\sum_{i=n_\ell+1}^{n} \log \left(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i; \theta_k) \right)}_{\text{Données non étiquetées}}. \tag{2.1}$$

Sous des conditions de régularité standards, l'estimateur du maximum de vraisemblance est asymptotiquement optimal puisqu'il atteint la borne de Cramér-Rao, ce qui justifie alors son utilisation. Nous noterons $\mathbf{x} = (\mathbf{x}_\ell, \mathbf{x}_u)$, et nous noterons en indice de $\hat{\theta}$ les données utilisées pour estimer les paramètres du modèle. Nous notons alors

$$\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u)$$

cette notation étant admissible puisque la vraisemblance est souvent bornée dans le cas semi-supervisé comme nous l'évoquons maintenant.

2.1.2 Bornitude de la vraisemblance

Dans le cadre de mélanges de distributions on recherche souvent un maximum local. Par exemple, dans le cas gaussien hétéroscédastique ($\mathcal{X} = \mathbb{R}^d$) on peut rendre la vraisemblance aussi grande qu'on le souhaite en choisissant comme centre d'un composant une donnée et en faisant tendre le déterminant de la matrice de covariance de ce composant vers zéro. Toutefois Redner & Walker (1984) montrent que sous des conditions de régularité standards il existe une racine convergente de la vraisemblance, ce qui justifie l'utilisation des modèles gaussiens hétéroscédastiques. Dans ce cas, il faudra rejeter les solutions dégénérées de la forme précédente. Dans le cadre semi-supervisé, ce problème est évité si on dispose d'au moins $d + 1$ données étiquetées par classe. Dans le cas où le problème d'estimation des paramètres est bien posé en classification supervisée, il en sera donc de même en classification semi-supervisée.

2.1.3 Convexité de la vraisemblance

Pour de nombreux modèles la vraisemblance est log-concave dans le cadre supervisé. On peut se demander quelle est la fraction de données non étiquetées nécessaire pour que ceci ne soit plus le cas, et à partir de quelle fraction la vraisemblance comporte plus d'un

maximum local. Remarquons que la non concavité de la log-vraisemblance n'implique pas qu'il y ait plusieurs maxima locaux. En pratique, un très petit ratio de données étiquetées peut suffire à ce que la log-vraisemblance n'ait qu'un seul maximum local, tandis que l'obtention de la concavité peut nécessiter un ratio de données étiquetées beaucoup plus grand.

2.1.4 Remarque sur la matrice d'information

D'autre part sous des conditions de régularité standards (van der Vaart, 2000, Théorème 5.39) l'estimateur du maximum de vraisemblance est asymptotiquement gaussien et a pour matrice de variance covariance l'inverse de la matrice d'information de Fisher.

Dans le cadre supervisé, on a

$$\sqrt{n_\ell}(\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_c^{-1}) \tag{2.2}$$

où J_c est la matrice d'information de Fisher J_c avec

$$J_c = -\mathbb{E}\left[\nabla^2 \log p(\boldsymbol{X}, \boldsymbol{Z}; \theta^*)\right] = \mathbb{E}\left[\nabla \log p(\boldsymbol{X}, \boldsymbol{Z}; \theta^*) \nabla \log p(\boldsymbol{X}, \boldsymbol{Z}; \theta^*)'\right].$$

Dans le cadre semi-supervisé, on a

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_\beta^{-1}) \tag{2.3}$$

avec $J_\beta = \beta J_c + (1-\beta)J$ avec β la fraction de données non étiquetées et où

$$J = -\mathbb{E}\left[\nabla^2 \log p(\boldsymbol{X}; \theta^*)\right] = \mathbb{E}\left[\nabla \log p(\boldsymbol{X}; \theta^*) \nabla \log p(\boldsymbol{X}; \theta^*)'\right].$$

Les deux propriétés précédentes sont valides si la distribution d'échantillonnage est incluse dans le modèle postulé. Si ce n'est pas le cas, on pourra se référer à White (1982).

On a donc $n\mathbb{V}[\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}] \to \beta J_c^{-1}$ et $n\mathbb{V}[\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}] \to J_\beta^{-1}$. Comme $\beta J_c^{-1} - J_\beta^{-1}$ est défini positif, on en déduit que pour n assez grand la variance de $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$ sera plus grande que celle de $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$. Cette différence sera d'autant plus grande que la fraction de données étiquetées sera petite. L'estimation semi-supervisée est alors à préférer à l'estimation supervisée pour des raisons de variance asymptotique plus faible. En pratique, pour des échantillons de petite taille le gain apporté par les données non étiquetées peut être plus grand que celui attendu asymptotiquement (Ganesalingam & McLachlan, 1979). Cette variance asymptotique plus faible conduit naturellement à une erreur de classification moyenne plus faible. Bien sûr, ceci n'est vrai que dans le cas où la distribution d'échantillonnage est incluse dans le modèle postulé. Dans le cas contraire, aucune conclusion générale sur l'avantage de l'estimation semi-supervisée des paramètres ne peut être dressée.

2.2 Algorithme EM en semi-supervisé

Nous avons justifié notre choix pour l'estimation par maximum de vraisemblance, il faut maintenant trouver le lieu du maximum de $\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u)$. Dans le cadre supervisé l'estimation des paramètres est souvent explicite. Cependant, en semi-supervisé, le logarithme d'une somme apparaît (équation (2.1)), rendant impossible la maximisation directe

de la vraisemblance. Ne pas observer la classe des données non étiquetées complexifie donc fortement l'estimation des paramètres. Pour ce type de problème où l'observation de certaines données simplifierait fortement l'estimation des paramètres, l'algorithme EM est bien adapté. À chaque étape on complète les données avec les données non observées, l'optimisation est ensuite aussi facile qu'avec des données complètes.

2.2.1 Principe

Avant sa formulation générale sous le nom d'algorithme EM, cet algorithme fut utilisé par Newcomb (1886), McKendrick (1926), Hartley (1958), Baum *et al.* (1970) et bien d'autres. L'algorithme EM part de l'idée qu'il est souvent plus facile d'optimiser la vraisemblance avec des données complètes, qu'avec des données incomplètes. Il reconstitue d'abord les données manquantes. Puis, une fois cette reconstitution effectuée, il est alors beaucoup plus facile de procéder à la recherche du maximum.

L'algorithme EM consiste à alterner les deux étapes suivantes (McLachlan & Krishnan, 1996) :

- **Étape E (*Expectation*)** : Calcul de $Q(\theta|\theta^{(r)})$, l'espérance de la vraisemblance complétée conditionnellement aux paramètres courants $\theta^{(r)}$ et aux données observées.
- **Étape M (*Maximization*)** : Maximisation de $Q(\theta|\theta^{(r)})$ en θ

$$\theta^{(r+1)} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, Q(\theta|\theta^{(r)}).$$

Remarquons qu'il y a en général plusieurs possibilités de définir les données manquantes. Lors de l'étape M, on peut se contenter d'une solution $\theta^{(r+1)}$ telle que

$$Q(\theta^{(r+1)}|\theta^{(r)}) > Q(\theta^{(r)}|\theta^{(r)})$$

et on obtient alors une variante de l'algorithme EM appelée algorithme EM généralisé (GEM). Cette variante est notamment utile quand la maximisation de $Q(\theta|\theta^{(r+1)})$ est difficile, mais que le maximisation par rapport à chaque composant de θ, les autres composants étant fixés à $\theta^{(r)}$, est facile.

En pratique on reproche parfois à l'algorithme EM une convergence lente (convergence linéaire) contrairement à l'algorithme de Newton qui a une convergence rapide (convergence quadratique). Toutefois l'algorithme EM a les propriétés intéressantes suivantes :

- il fait croître la vraisemblance à chaque étape,
- les contraintes sont naturellement vérifiées,
- il est peu coûteux en mémoire.

Ceci n'est pas le cas de l'algorithme de Newton qui, si la fonction à optimiser n'est pas convexe, ne permet pas de faire croître la vraisemblance à chaque étape et nécessite la recherche d'une solution améliorant la solution précédente le long de la ligne de plus fort gradient. De plus, les contraintes ne sont souvent pas automatiquement vérifiées, et il faut donc avoir recours à une reparamètrisation pour que ceci soit le cas. Enfin l'algorithme de Newton peut nécessiter le calcul du gradient et de la matrice jacobienne, ce qui est coûteux en espace mémoire, notamment en grande dimension.

2.2.2 Application à la classification semi-supervisée

Ici pour certains modèles la vraisemblance est non bornée. Toutefois de nombreux articles apportent des garanties pratiques et théoriques quant à l'intérêt d'estimer les paramètres de cette façon. Redner & Walker (1984) montrent que sous des conditions de régularité standards il existe une racine de la vraisemblance convergente et si l'initialisation de l'algorithme EM est suffisamment proche de cette dernière, alors EM y conduit.

Le premier terme dans l'équation (2.1) représente la log-vraisemblance associée aux données étiquetées et le second la log-vraisemblance associée aux données non étiquetées. Dans ce cadre l'algorithme EM est très bien adapté. Historiquement l'approche itérative d'estimation des paramètres par maximum de vraisemblance a été utilisée dans un cadre semi-supervisé par Hosmer (1973) avec un modèle gaussien homoscédastique avant même sa formulation plus générale sous le nom d'algorithme EM par Dempster *et al.* (1977).

La vraisemblance complètée par les données manquantes qui sont ici $\mathbf{z}_u = (\boldsymbol{z}_{n_\ell+1}, \ldots, \boldsymbol{z}_n)$ est alors

$$\mathcal{L}_c(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u, \mathbf{z}_u) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log\left(\pi_k p(\boldsymbol{x}_i; \theta_k)\right) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} z_{ik} \log\left(\pi_k p(\boldsymbol{x}_i; \theta_k)\right). \quad (2.4)$$

En prenant l'espérance de $\mathcal{L}_c(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u, \mathbf{z}_u)$ conditionnellement au paramètre courant $\theta^{(r)}$ et aux données observées $\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u$, on obtient

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log\left(\pi_k p(\boldsymbol{x}_i; \theta_k)\right) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} \mathbb{E}\left[Z_{ik}|\boldsymbol{x}_i, \theta^{(r)}\right] \log\left(\pi_k p(\boldsymbol{x}_i; \theta_k)\right). \quad (2.5)$$

Puis on remarque que

$$\mathbb{E}\left[Z_{ik}|\mathbf{x}_i, \theta^{(r)}\right] = p\left(Z_{ik} = 1|\boldsymbol{x}_i, \theta^{(r)}\right),$$

le théorème de Bayes en donnant une expression explicite simple

$$p(Z_{ik} = 1|\boldsymbol{x}_i, \theta^{(r)}) = \frac{\pi_k^{(r)} p(\boldsymbol{x}_i; \theta_k^{(r)})}{\sum_{l=1}^{g} \pi_l^{(r)} p(\boldsymbol{x}_i; \theta_l^{(r)})}.$$

Dans la suite on notera $t_{ik}^{(r+1)}$ cette quantité. On a alors

$$\boxed{Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log(p(\mathbf{x}_i; \theta_k)) + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} t_{ik}^{(r+1)} \log(p(\mathbf{x}_i; \theta_k)) + \sum_{k=1}^{g} n_k^{(r+1)} \log(\pi_k),} \quad (2.6)$$

en notant $n_k^{(r+1)} = \sum_{i=1}^{n_\ell} z_{ik} + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)}$. À ce stade, l'expression ne comporte plus le logarithme d'une somme. L'étape M consiste à maximiser $Q(\theta|\theta^{(r)})$ en θ. Pour l'actualisation des proportions on obtient la formule suivante

$$\boxed{\pi_k^{(r+1)} = \frac{n_k^{(r+1)}}{n}.} \quad (2.7)$$

Concernant l'actualisation des autres paramètres, cette étape dépend de la famille paramétrée choisie. Remarquons qu'en général cette étape n'est pas plus difficile que dans la cas supervisé. On la décline pour différents modèles génératifs plus loin dans ce chapitre.

2.2.3 Initialisation de EM

La vraisemblance comporte en général de nombreux maxima locaux, le résultat obtenu est donc sensible à l'initialisation utilisée. En semi-supervisé nous disposons d'une bonne initialisation de l'algorithme, qui est l'initialisation à partir des paramètres estimés uniquement à partir des données étiquetées. Nous discutons maintenant du bien fondé de cette initialisation.

Nous supposons ici que le nombre de données non étiquetées est très grand et que le modèle postulé est celui qui a généré les données. Nous nous intéressons alors à la probabilité que la solution supervisée se situe dans le domaine d'attraction du maximum global de la vraisemblance des données non étiquetées. Si cette probabilité est grande, l'initialisation de l'algorithme EM à partir de l'estimation supervisée est alors justifiée.

Or, quand le modèle est bien spécifié on sait que la solution supervisée converge avec probabilité 1 vers le maximum global de cette vraisemblance. Effectivement, si le nombre de données non étiquetées est très grand, la log-vraisemblance des données non étiquetées normalisée par le nombre de données converge vers $L_{ns}(\theta) = \mathbb{E}_{\boldsymbol{X}}[\log p(\boldsymbol{X};\theta)]$ d'après la loi des grands nombres. Soit θ^* le lieu du maximum de $L_{ns}(\theta)$ et $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$ le lieu du maximum de la vraisemblance des données étiquetées. Si le modèle et bien spécifié, et sous des conditions de régularité standards $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell} \xrightarrow{P} \theta^*$. Notons Θ^* le sous-ensemble de Θ qui contient toutes les initialisations conduisant l'algorithme EM dans un voisinage de θ^*. Dans le cas où l'estimateur du maximum de vraisemblance converge à une vitesse $\sqrt{n_\ell}$, on a

$$\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell} = \theta^* + \mathcal{O}_p(n_\ell^{-1/2}),$$

ce qui implique que si n_ℓ augmente, $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$ se rapproche de plus en plus de θ^* et par conséquent, la probabilité que $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$ appartienne à Θ^* augmente. Ce qui justifie alors l'initialisation de EM à partir des données étiquetées.

On peut s'attendre dans certains cas à une convergence beaucoup plus rapide de $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$ dans Θ^*. Par exemple, si on considère un mélange de distributions gaussiennes avec les moyennes inconnues et les variances connues, alors l'initialisation à partir des données étiquetées par les moyennes empiriques donne par un argument de grandes déviations (Dembo & Zeitouni, 1998)

$$\lim_{n_\ell \to \infty} \frac{1}{n_\ell} \log p(\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell} \notin \Theta^*) = -I \tag{2.8}$$

avec I une fonction de taux.

Ainsi si le modèle postulé est correct les données étiquetées devraient très vite conduire vers une initialisation permettant de trouver la racine consistante de la vraisemblance.

Les arguments qualitatifs précédents justifient le bien fondé de l'algorithme EM à partir des données étiquetées. Nous illustrons maintenant cette stratégie sur un exemple jouet de Gan & Jiang (1999). La première classe suit une distribution gaussienne avec pour moyenne $\mu_1 = -3$ et pour variance $\sigma_1^2 = 1$, la seconde classe suit une distribution gaussienne avec $\mu_2 = 8$ et $\sigma_2^2 = 16$, la première classe est en proportion $\pi_1 = 0,4$ et la seconde classe est en proportion $\pi_2 = 0,6$. Supposons que le nombre de données non étiquetées soit très grand et que tous les paramètres soient connus sauf μ_1. Alors, la log-vraisemblance admet deux maxima : un maximum local pour $\mu_1 \approx 8,4$ et le maximum

global pour $\mu_1 \approx -3$ (voir figure 2.1). Si l'algorithme EM est initialisé en prenant un centre au hasard dans les données disponibles, la probabilité de trouver le maximum global est de 45%. Si une donnée de la classe 1 est observée et qu'elle est utilisée pour initialiser μ_1, la probabilité de convergence vers le maximum global est de 89%. Au fur et à mesure que le nombre de données étiquetées dans la classe 1 augmente, la probabilité de convergence vers le maximum global de la vraisemblance converge très vite vers 1 (voir figure 2.2).

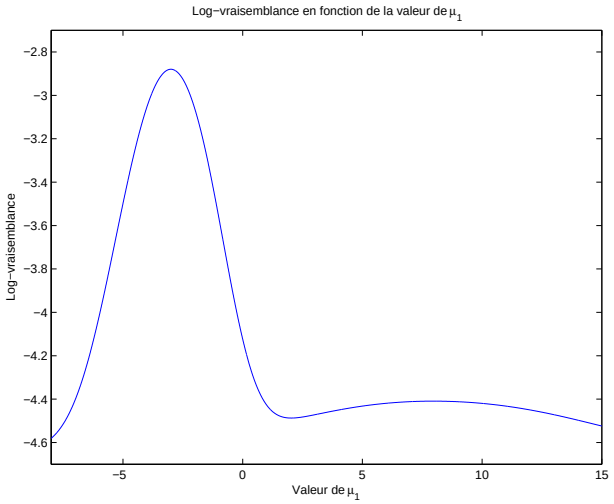

FIG. 2.1 – Log-vraisemblance en fonction de μ_1.

Ainsi, si le modèle est bien spécifié, l'algorithme EM peut dans la plupart des cas être initialisé efficacement à partir des données étiquetées. Cependant, si le nombre de données étiquetées est petit par rapport au nombre de paramètres, il peut être risqué de n'utiliser que l'initialisation à partir des données étiquetées. De plus, si le modèle postulé est incorrect, il n'y a pas de garantie de convergence vers le maximum global, ceci même dans le cas où le nombre de données étiquetées est grand. Cependant en pratique, si le modèle postulé est relativement correct, l'initialisation à partir des données étiquetées reste une initialisation de choix.

2.2.4 Effet associé des données étiquetées sur la convexité et l'initialisation de EM

Si le modèle est correct, d'après les arguments sur la convexité de la vraisemblance, et le rapprochement de $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$ vers θ^*, on a alors à mesure que le nombre de données étiquetées augmente une initialisation à partir des données étiquetées qui est de plus en plus fiable et de moins en moins de maxima locaux.

On illustre maintenant la combinaison de ces deux effets sur le jeu de données numéro 5 du livre sur l'apprentissage semi-supervisé[2] de Chapelle *et al.* (2006). Ces données proviennent d'un mélange de deux gaussiennes en dimension 241 avec matrices de covariances diagonales, 1500 individus ont été générés. Nous générons 100 partitions données non étiquetées/données étiquetées, le nombre de données étiquetées variant de 4 à 40. Le nombre de fois où l'initialisation à partir des données étiquetées conduit au maximum

[2] `http://www.kyb.tuebingen.mpg.de/ssl-book/`

global de la vraisemblance est représenté figure 2.3. Comme attendu, quand le nombre de données étiquetées augmente, la probabilité d'atteindre le maximum global à partir de l'initialisation supervisée augmente.

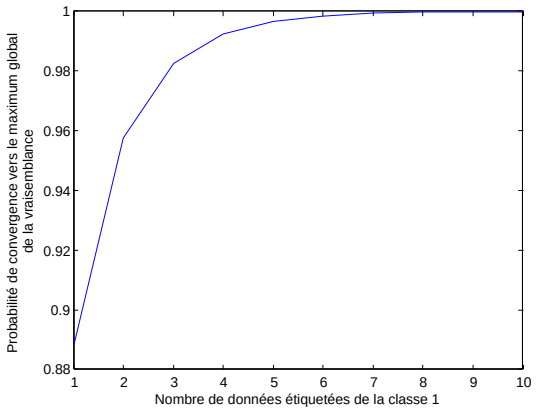

FIG. 2.2 – Probabilité de convergence vers le maximum global de la vraisemblance en fonction du nombre de données étiquetées de la classe 1.

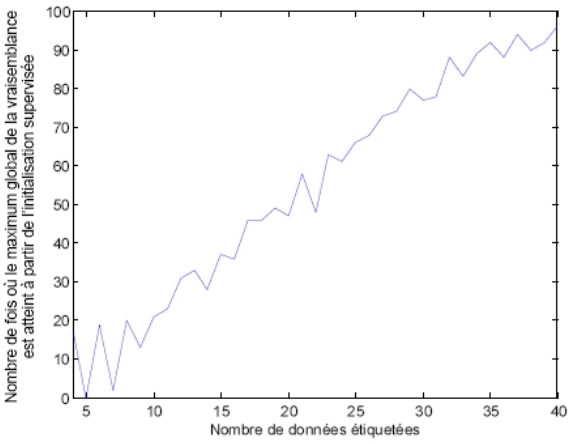

FIG. 2.3 – Nombre d'initialisations conduisant vers le maximum global de la vraisemblance en fonction du nombre de données étiquetées.

2.2.5 Vitesse de convergence

L'algorithme EM définit une application $M : \theta \rightarrow M(\theta)$ de Θ dans Θ, avec $\theta^{(r+1)} = M(\theta^{(r)})$. Un point fixe $\hat{\theta}$ de EM vérifie $\hat{\theta} = M(\hat{\theta})$. Pour $\theta^{(r)}$ dans le voisinage de $\hat{\theta}$, en utilisant une approximation de Taylor

$$\theta^{(r+1)} - \hat{\theta} \approx DM(\hat{\theta})(\theta^{(r)} - \hat{\theta}), \tag{2.9}$$

où $DM_{ij}(\theta) = \frac{\partial M_i(\theta)}{\partial \theta_j}$. Habituellement $DM(\hat{\theta})$ n'est pas nulle, et est positive ou semi-définie positive.

Soit

$$Q(\theta|\theta^{(r)}) = \mathbb{E}\left[\log p(\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u, \mathbf{z}_u; \theta)|\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u; \theta^{(r)}\right], \tag{2.10}$$

et

$$H(\theta|\theta^{(r)}) = \mathbb{E}\left[\log p(\mathbf{z}_u|\mathbf{x}_u; \theta)|\mathbf{x}_u; \theta^{(r)}\right], \tag{2.11}$$

on sait d'après Dempster *et al.* (1977) que

$$DM(\hat{\theta}) = \left[D^{20}Q(\hat{\theta}|\hat{\theta})\right]^{-1} D^{20}H(\hat{\theta}|\hat{\theta}), \tag{2.12}$$

où $D^{20}Q$ et $D^{20}H$ correspondent respectivement aux dérivées secondes de Q et H selon leur premier argument. $D^{20}Q(\hat{\theta}|\hat{\theta})$ est la matrice d'information complète et $D^{20}H(\hat{\theta}|\hat{\theta})$ est la matrice d'information manquante. Les calculs de $D^{20}Q(\hat{\theta}|\hat{\theta})$ et de $DM(\hat{\theta})$ sont souvent faciles, tandis que le calcul de $D^{20}H(\hat{\theta}|\hat{\theta})$ est souvent plus difficile. Le calcul de $DM(\hat{\theta})$

est alors utile lorsqu'on veut obtenir la variance asymptotique des estimateurs (Meng & Rubin, 1991). Nous reviendrons sur ce point dans le chapitre 4.

Des approches existent pour accélérer l'algorithme EM (McLachlan & Krishnan, 1996). Elles ne donnent en général que des améliorations mineures, tandis que leur mise en œuvre complexifie l'utilisation de EM.

2.2.6 L'algorithme λ-EM

Une variante de l'algorithme EM a été proposée dans Nigam *et al.* (2000). Il s'agit de l'algorithme $\lambda-$EM. Cet algorithme part du constat qu'il existe un compromis entre estimation supervisée et estimation semi-supervisée quand le modèle est mal spécifié. En effet, si le modèle postulé est incorrect et que les données non étiquetées sont en très grand nombre par rapport aux données étiquetées, les données non étiquetées dégradent souvent les performances en classification. D'un autre côté, quand le nombre de données étiquetées est petit l'estimation des paramètres du modèle comporte une grande variance. Ainsi la pondération de l'influence des données non étiquetées par un facteur λ bien choisi, doit permettre de trouver un compromis entre biais et variance. L'expression à optimiser en θ est alors

$$\mathcal{L}_\lambda(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \sum_{i=1}^{n_\ell}\sum_{k=1}^{g} z_{ik} \log\left(\pi_k p(\boldsymbol{x}_i;\theta_k)\right) + \lambda \sum_{i=n_\ell+1}^{n} \log\left(\sum_{k=1}^{g} \pi_k p(\boldsymbol{x}_i;\theta_k)\right). \quad (2.13)$$

Dans l'étape E de l'algorithme on obtient

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n_\ell}\sum_{k=1}^{g} z_{ik} \log\left(p(\boldsymbol{x}_i;\theta_k)\right) + \lambda \sum_{i=n_\ell+1}^{n}\sum_{k=1}^{g} t_{ik}^{(r+1)} \log\left(p(\boldsymbol{x}_i;\theta_k)\right) + \sum_{k=1}^{g} n_k^{(r+1)} \log(\pi_k), \quad (2.14)$$

en notant $n_k^{(r+1)} = \sum_{i=1}^{n_\ell} z_{ik} + \lambda \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)}$. Ce qui montre bien la pondération de l'influence des données non étiquetées par un facteur λ. L'étape M n'est ensuite pas plus difficile que dans l'algorithme EM. Comme EM, l'algorithme λ-EM possède la propriété de croissance monotone.

En pratique le paramètre λ est choisi par validation croisée du taux d'erreur optimisée sur une grille discrète. Cet algorithme conserve essentiellement les mêmes propriétés que l'algorithme EM. Dans le cas où λ est petit la vitesse de convergence de λ-EM est grande puisque la part d'information manquante est réduite. Une approche similaire à cette dernière est l'approche compromis entre génératif et prédictif (Bouchard & Triggs, 2004; Lasserre *et al.* , 2006) dont on a parlé dans la section 1.3.4. Cependant cette dernière nécessite le recours à un algorithme de Newton beaucoup plus instable comparativement à $\lambda-$EM. Remarquons que l'approche λ-EM peut être relativement coûteuse selon la finesse sur la grille en λ choisie.

2.2.7 Estimation avec étiquettes partielles

On dispose parfois de l'information suivante : l'individu i n'appartient ni à la classe 1, ni à la classe 2. Cette information entre alors facilement dans le cadre semi-supervisé puisqu'il s'agit d'une information partielle sur la classe. Sa prise en compte s'effectue

lors de l'étape E de l'algorithme EM en conditionnant sur les appartenances possibles de l'individu. En codant $\boldsymbol{z}$ de la façon suivante :

$$z_{ik} = \begin{cases} 1 & \text{si l'individu } i \text{ peut appartenir à la classe } k \\ 0 & \text{sinon,} \end{cases}$$

on a alors

$$t_{ik}^{(r+1)} = \frac{z_{ik}\pi_k p(\boldsymbol{x}_i;\theta_k)}{\sum_{l=1}^{g} z_{il}\pi_l p(\boldsymbol{x}_i;\theta_l)}.$$

L'étape M reste ensuite inchangée.

2.3 Exemples de modèles génératifs utilisés

2.3.1 Modèles pour données continues

Modèles gaussiens standards

Un modèle génératif très populaire quand $\mathcal{X} = \mathbb{R}^d$ est $\boldsymbol{X}|Z_k = 1 \sim \mathcal{N}(\mu_k, \Sigma_k)$ avec μ_k qui est le vecteur des moyennes et Σ_k la matrice de covariance. Pour ce modèle l'espérance de la vraisemblance complétée s'écrit à une constante près :

$$\begin{aligned} Q(\theta|\theta^{(r)}) &= -\frac{1}{2}\sum_{i=1}^{n_\ell}\sum_{k=1}^{g} z_{ik}[(\boldsymbol{x}_i-\mu_k)'\Sigma_k^{-1}(\boldsymbol{x}_i-\mu_k) + \log(|\Sigma_k|)] \\ &\quad -\frac{1}{2}\sum_{i=n_\ell+1}^{n}\sum_{k=1}^{g} t_{ik}^{(r+1)}[(\boldsymbol{x}_i-\mu_k)'\Sigma_k^{-1}(\boldsymbol{x}_i-\mu_k) + \log(|\Sigma_k|)] + \sum_{k=1}^{g} n_k^{(r+1)}\log(\pi_k). \end{aligned}$$

Ce qui en utilisant la trace et sa linéarité donne

$$\begin{aligned} Q(\theta|\theta^{(r)}) &= -\frac{1}{2}\sum_{k=1}^{g}\text{trace}\left[\left(\sum_{i=1}^{n_\ell} z_{ik}(\boldsymbol{x}_i-\mu_k)(\boldsymbol{x}_i-\mu_k)' + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)}(\boldsymbol{x}_i-\mu_k)(\boldsymbol{x}_i-\mu_k)'\right)\Sigma_k^{-1}\right] \\ &\quad -\frac{1}{2}\sum_{k=1}^{g} n_k^{(r+1)}\log(|\Sigma_k|) + \sum_{k=1}^{g} n_k^{(r+1)}\log(\pi_k). \end{aligned} \tag{2.15}$$

La maximisation de $Q(\theta|\theta^{(r)})$ est ici explicite voir par exemple Anderson (2003). Ainsi les formules d'actualisation pour les paramètres spécifiques au modèle sont :

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik}\boldsymbol{x}_i + \sum_{i=n_\ell+1}^{n} t_{ik}\boldsymbol{x}_i}{n_k^{(r+1)}} \tag{2.16}$$

$$\Sigma_k^{(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik}(\boldsymbol{x}_i-\mu_k^{(r+1)})(\boldsymbol{x}_i-\mu_k^{(r+1)})' + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)}(\boldsymbol{x}_i-\mu_k^{(r+1)})(\boldsymbol{x}_i-\mu_k^{(r+1)})'}{n_k^{(r+1)}}. \tag{2.17}$$

On retrouve le côté assez intuitif de l'algorithme EM, qui repondère pour chaque classe l'influence des données non étiquetées en fonction de leur probabilité d'appartenance à cette classe. Dans le cas où les matrices de covariance par classe sont supposées égales, on retrouve des formules similaires à l'analyse discriminante linéaire de Fisher (1936).

Comme détaillé section 1.1.1, une reparamètrisation de la matrice sous la forme $\Sigma_k = \lambda_k D_k A_k D_k'$ a été proposée par Bensmail & Celeux (1996). L'estimation des paramètres de ces modèles est facile, elle nécessite de repartir de l'équation (2.15) et de résoudre le problème d'optimisation en fonction des contraintes imposées. Pour la plupart des modèles l'estimation de paramètres est explicite. Cependant, pour certains d'entre eux elle nécessite le recours à une procédure d'optimisation alternée et éventuellement à l'algorithme de Flury (1988) pour la décomposition simultanée en valeurs propres. Le nombre de paramètres de chacun de ces modèles est représenté table 2.1, et on trace figure 2.4 les isodensités conditionnellement à la classe dans le cas $d = 2$ et $g = 2$ pour les différents modèles. Cette figure permet d'illustrer l'interprétation de la décompostion de Σ_k en termes de volume d'orientation et de forme.

Modèle	Nombre de paramètres	Étape M
Modèles généraux		
$[\lambda C]$	$\gamma + \eta$	CF
$[\lambda_k C]$	$\gamma + \eta + g - 1$	IP
$[\lambda D A_k D']$	$\gamma + \eta + (g-1)(d-1)$	IP
$[\lambda_k D A_k D']$	$\gamma + \eta + (g-1)d$	IP
$[\lambda D_k A D_k']$	$\gamma + g\eta - (g-1)d$	CF
$[\lambda_k D_k A D_k']$	$\gamma + g\eta - (g-1)(d-1)$	IP
$[\lambda C_k]$	$\gamma + g\eta - (g-1)$	CF
$[\lambda_k C_k]$	$\gamma + g\eta$	CF
Modèles diagonaux		
$[\lambda B]$	$\gamma + d$	CF
$[\lambda_k B]$	$\gamma + d + g - 1$	IP
$[\lambda B_k]$	$\gamma + gd - g + 1$	CF
Modèles sphériques		
$[\lambda I]$	$\gamma + 1$	CF
$[\lambda_k I]$	$\gamma + g$	CF

TAB. 2.1 – Modèles gaussiens parcimonieux et leur nombre de paramètres. $\gamma = gd + g - 1$ si les proportions sont libres et $\gamma = gd$ sinon. $\eta = \frac{d(d+1)}{2}$. On note CF si l'étape M est explicite, IP si elle nécessite une procédure itérative.

Modèles gaussiens haute dimension

Plus récemment, des modèles gaussiens capables de traiter des données en grande dimension ont été proposés (Bouveyron *et al.* , 2007; McLachlan *et al.* , 2003). Ils reposent sur la décomposition

$$\Sigma_k = D_k \Delta_k D_k'$$

avec

$$(\Delta_k)_{ii} = a_{ki} \text{ pour } i \in \{1, \ldots, p_k\}$$

avec p_k le nombre de plus grandes valeurs propres différentes pour la classe k,

$$(\Delta_k)_{ii} = b_k \text{ pour } i \in \{p_k + 1, \ldots, d\},$$

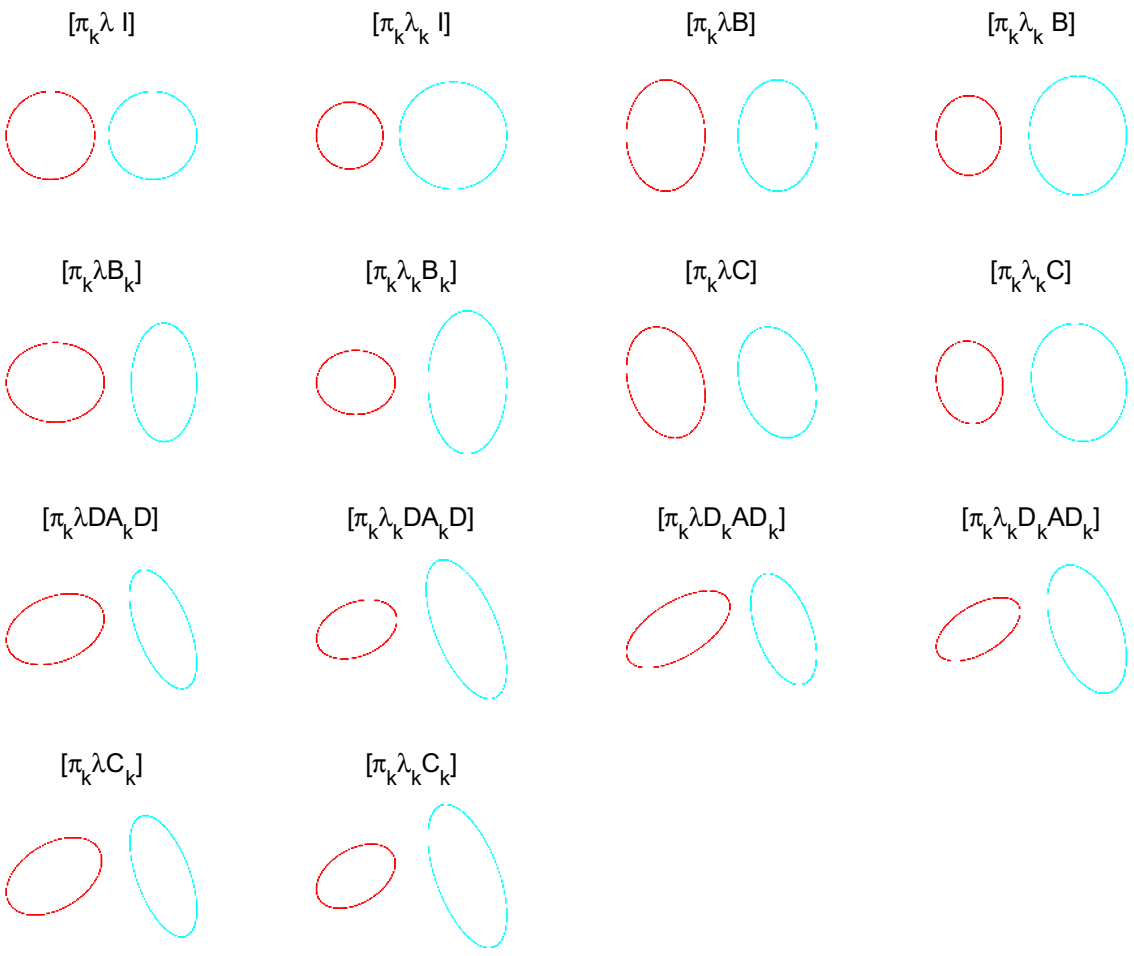

FIG. 2.4 – Illustration des 14 modèles parcimonieux.

sous la contrainte

$$a_{k1} \geqslant a_{k2} \geqslant \ldots \geqslant a_{kp} \geqslant b_k$$

et D_k la matrice des vecteurs propres. Ainsi on peut imposer différentes contraintes

$$\begin{array}{rcccccl} a_{ki} & = & a_{k2} & = & \ldots & = & a_{kp}, \\ a_{1i} & = & a_{2i} & = & \ldots & = & a_{gi}, \\ b_1 & = & b_2 & = & \ldots & = & b_g, \\ p_1 & = & p_2 & = & \ldots & = & p_g, \\ D_1 & = & D_2 & = & \ldots & = & D_g. \end{array}$$

En combinant ces différentes contraintes, Bouveyron *et al.* (2007) obtiennent une famille de modèles gaussiens haute dimension. Ces modèles permettent d'obtenir une matrice de covariance estimée inversible, même dans le cas où la dimension est supérieure au nombre de données. Dans un cadre supervisé ces modèles ont montré des performances comparables aux SVM (Bouveyron *et al.* , 2007) méthode de référence dans le contexte de la grande dimension.

Ce type de modèle a aussi montré de bonnes performances en classification non supervisée de données génomiques de type biopuces. Dans ce cas le nombre de variables observées (intensité de transcription pour un grand nombre de gènes) est de loin supérieur au nombre d'individus (patients dans l'étude clinique) (McLachlan *et al.* , 2003). Ces modèles se placent dans le cadre plus général des facteurs analysants.

2.3.2 Modèles pour données discrètes

Modèle d'indépendence conditionnelle

Supposons que d variables discrètes sont observées et que chaque variable $j \in \{1, \ldots, d\}$ admet m_j modalités, on a alors $\mathcal{X} = \prod_{j=1}^{d} \{0,1\}^{m_j}$. Un modèle très populaire dans ce cadre est le modèle d'indépendance des covariables conditionnellement à la classe. Ce modèle est aussi appelé modèle de Bayes naïf compte-tenu de l'hypothèse naïve qu'il fait par rapport à l'indépendance des covariables conditionnellement à la classe. Enfin dans le cadre non supervisé ce modèle est aussi appelé modèle à classe latente (Everitt, 1984). Ce modèle bien que rudimentaire donne de bons résultats dans de nombreuses situations réelles (Hand & Yu, 2001). Notons aussi que ce modèle n'est pas identifiable, mais que dans le cas où toutes les variables ont le même nombre de modalités m, si la condition $d \geqslant \lceil \log_m g \rceil + 1$ est vérifiée, alors le modèle est génériquement identifiable (Allman *et al.* , 2009). Ainsi, si le nombre de variables est suffisamment grand devant le nombre de classes, le modèle sera génériquement identifiable. Soit

$$x_i^{jh} = \begin{cases} 1 \text{ si l'individu } i \text{ présente la modalité } h \text{ de la variable } j, \\ 0 \text{ sinon.} \end{cases}$$

On note $\alpha_k^{jh} = p(X^{jh} = 1 | Z_k = 1)$, $\alpha_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$ et $\alpha_k = (\alpha_k^1, \ldots, \alpha_k^d)$. Ainsi $\theta = (\pi_1, \ldots, \pi_{g-1}, \alpha_1, \ldots, \alpha_g)$. Dans ce cas la vraisemblance pour une donnée i conditionnellement à la classe k, qui se résume ici à une probabilité discrète est

$$p(\boldsymbol{x}_i; \alpha_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}.$$

La log-vraisemblance s'écrit alors

$$\mathcal{L}(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \left[\log \pi_k + \sum_{j=1}^{d} \sum_{h=1}^{m_j} x_i^{jh} \log(\alpha_k^{jh}) \right] \tag{2.18}$$

$$+ \sum_{i=n_\ell+1}^{n} \log \left(\sum_{k=1}^{g} \pi_k \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}} \right). \tag{2.19}$$

Puis l'espérance de la vraisemblance complétée s'écrit

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} \sum_{j=1}^{d} \sum_{h=1}^{m_j} z_{ik} x_i^{jh} \log(\alpha_k^{jh}) \tag{2.20}$$

$$+ \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} \sum_{j=1}^{d} \sum_{h=1}^{m_j} t_{ik}^{(r+1)} x_i^{jh} \log(\alpha_k^{jh}) + \sum_{k=1}^{g} n_k^{(r+1)} \log \pi_k. \tag{2.21}$$

L'étape M de l'algorithme ne pose pas de difficulté, et donne la formule d'actualisation suivante

$$\boxed{\alpha_k^{jh(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik} x_i^{jh} + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} x_i^{jh}}{n_k^{(r+1)}}.} \tag{2.22}$$

Pour éviter que certains $\hat{\alpha}_k^{jh}$ soient nuls, l'estimation des paramètres peut être régularisée. Dans un cadre bayesien, cette régularisation peut être interprétée comme l'estimateur du maximum *a posteriori* où la distribution *a priori* sur les paramètres est une distribution de Dirichlet. En effet, on peut supposer que la distribution *a priori* du vecteur α_k^j a pour densité

$$p(\alpha_k^j) = \frac{\prod_{h=1}^{m_j} \Gamma(\beta_k^{jh})}{\Gamma(\sum_{h=1}^{m_j} \beta_k^{jh})} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{\beta_k^{jh}-1},$$

où β_k^j est le vecteur des hyperparamètres. Cette distribution est la distribution conjuguée de la distribution multinomiale. Par conséquent la distribution *a posteriori* de α_k^j reste une distribution de Dirichlet. Ainsi la formule d'actualisation lorsque le paramètre est estimé par maximum *a posteriori* est

$$\boxed{\alpha_k^{jh(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik} x_i^{jh} + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} x_i^{jh} + \beta_k^{jh}}{n_k^{(r+1)} + \sum_{h=1}^{m_j} \beta_k^{jh}}.} \tag{2.23}$$

Cette régularisation permet d'éviter l'estimation de certaines proportions à 0 et évite ainsi l'apparition de problèmes numériques. Comme dans le cas continu, des versions parcimonieuses de ces modèles existent dans le cas où les covariables sont des variables binaires (Celeux & Govaert, 1991). Ces versions parcimonieuses sont discutées en détail dans le chapitre 5.

Ici nous avons parlé des produits de distributions multinomiales d'ordre 1. Avec la progression des méthodes de génotypage automatique, l'observation d'allèles à certains loci et notamment l'observation de marqueurs micro-satellites devient de plus en plus fréquente. Ces zones sont très variables dans l'ADN comme par exemple la répétition du dinucléotide CA. Pour les individus diploïdes le nombre de versions de chaque micro-satellite suit

une distribution multinomiale d'ordre 2 sous les hypothèses de Hardy-Weinberg (Henry & Gouyon, 2008). Ainsi quand plusieurs loci indépendants sont considérés, on obtient un produit de distributions multinomiales d'ordre 2. Cette problématique a connu un essor important au cours des dix dernières années (Falush *et al.* , 2003; Corander *et al.* , 2004; Francois *et al.* , 2006; Toussile & Gassiat, 2008). Des problèmes de choix de variables se posent dans ce cadre (Toussile & Gassiat, 2008) puisque de nombreux marqueurs micro-satellites sont disponibles et ces derniers ne sont pas forcément indépendants conditionnellement à la classe. Pour l'instant des progrès restent à faire dans ce domaine puisque l'hypothèse d'indépendance des variables non informatives est trop restrictive, et l'utilisation des modèles de régression des variables non sélectionnées sur les variables sélectionnées est plus difficile que dans le cas continu (Raftery & Dean, 2006; Maugis *et al.* , 2008; Murphy *et al.* , 2008).

La principale difficulté dans l'utilisation de variables discrètes est la prise en compte des corrélations entre variables conditionnellement à la classe. L'utilisation de modèle à plusieurs composants par classe permet de lever cette limitation.

Modèle multinomial d'ordre quelconque

Un modèle assez similaire au produit de distributions multinomiales d'ordre 1 a été utilisé avec succès en classification de textes (Nigam *et al.* , 2000). Ce modèle assimile un texte à un sac de mots. Soit un dictionnaire de d mots $(w_1, \ldots, w_d)$. Soit le texte i de longueur ℓ_i et $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^d)$ avec x_i^j le nombre d'occurrences du mot j dans le texte i. On suppose que $\boldsymbol{X}|Z_k = 1 \sim \mathcal{M}(\ell_i, \alpha_k^1, \ldots, \alpha_k^d)$. α_k^j représente la fréquence du mot w_j dans les textes appartenant à la classe k. On retrouve des conditions d'identifiabilité génériques assez similaires à celles du modèle précédent. Cette condition est vérifiée lorsque la longueur des textes observés est suffisamment grande par rapport au nombre de classes, ce qui est souvent le cas en pratique. La formule d'actualisation est alors

$$\boxed{{\alpha_k^j}^{(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik} x_i^j + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} x_i^j}{\sum_{i=1}^{n_\ell} z_{ik} \ell_i + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} \ell_i}.} \tag{2.24}$$

Cette formule est assez proche de celle rencontrée dans le cas du modèle d'indépendance conditionnelle. Dans le cas supervisé il suffit simplement d'estimer la fréquence de chaque mot pour chaque type de texte. Dans le cas semi-supervisé on retrouve une interprétation similaire en utilisant l'algorithme EM. De même que pour le modèle précédent une régularisation de l'estimation des paramètres peut être effectuée en utilisant une distribution *a priori* de Dirichlet. La modélisation du texte comme un sac de mots est bien sûr assez irréaliste compte tenu du procédé d'écriture. On peut par exemple complexifier les modèles en prenant en compte la dépendance entre mots consécutifs en utilisant par exemple des modèles de type chaîne de Markov. Cependant, l'information sur la fréquence d'apparition des mots suffit dans de nombreux cas à classifier les textes avec une bonne précision. Toutefois, l'estimation semi-supervisée des paramètres est assez sensible aux mauvaises spécifications du modèle. Dans ce cas, les modèles à plusieurs composants par classe permettent de limiter ce risque de mauvaise spécification, et ainsi de prendre plus efficacement en compte l'information apportée par les données non étiquetées.

2.3.3 Modèles à plusieurs composants par classe

Dans certains cas la modélisation d'une classe par un seul composant peut se montrer trop peu flexible et mal s'adapter aux distributions spécifiques conditionnellement à la classe. Dans le cadre de modèles génératifs, une idée assez naturelle consiste à modéliser la distribution conditionnellement à la classe par un mélange (Ghahramani & Jordan, 1994; Hastie & Tibshirani, 1996; Miller & Uyar, 1997). Cette approche se justifie d'autant plus par les bonnes propriétés d'approximation des mélanges de distribution. Pour être utilisée cette approche nécessite un nombre de données relativement élevé ce qui peut parfois être impossible dans le cas supervisé en raison d'un nombre de données étiquetées trop petit. Cependant en semi-supervisé, de nombreuses données supplémentaires sont disponibles, améliorant ainsi l'estimation de la distribution marginale. Ces modèles sont utilisés dans le cadre semi-supervisé par Miller & Uyar (1997).

Deux hypothèses sont envisageables :

- Soit les composants sont communs aux classes ; sachant la classe k, la donnée est issue du composant h avec une probabilité $\tau_{hk} \in [0, 1]$.
- Soit chaque classe est modélisée par des composants différents ; sachant la classe k la donnée ne peut être issue que des composants spécifiques à la classe k.

L'intérêt de la première approche est qu'elle nécessite simplement de fixer le nombre total de composants et d'estimer les τ_{hk}, tandis que la seconde nécessite de fixer le nombre de composants pour chaque classe et peut nécessiter d'étudier un nombre de modèles relativement grand. Hastie & Tibshirani (1996) imposent aux nombres de composants par classe d'être identiques, ce qui évite tout problème combinatoire. Dans le cadre supervisé, Titsias & Likas (2002) montrent que l'approche à composants communs peut causer une moins bonne estimation de la règle de classement. En effet, le modèle à composants séparés conduit à une bonne estimation de la densité des covariables conditionnellement à la classe et peut dans certains cas produire de meilleurs résultats que le modèle à composants communs qui recherche avant tout un bonne approximation de la densité marginale. En semi-supervisé, cela est moins évident puisqu'il s'agit en grande partie de bien estimer la densité marginale de $\boldsymbol{X}$. Dans le cas à composants communs, la probabilité pour une donnée d'appartenir à une classe sachant son composant d'origine ne peut être estimée qu'à partir des données étiquetées. Ainsi en cas de nombreux composants et d'un faible nombre de données étiquetées, ce modèle peut être surajusté aux données étiquetées. En outre, remarquons que dans le cas des composants séparés, le choix d'affectation des composants aux classes peut être vu comme un problème d'optimisation discrète. Ce phénomène peut sembler quelque peu évité dans le cas à composants communs. Il reste cependant présent compte tenu de la sensibilité de l'algorithme EM à l'initialisation.

La log-vraisemblance s'écrit

$$\mathcal{L} = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log \left(\pi_k \sum_{h=1}^{H} \tau_{hk} p(\boldsymbol{x}_i; \theta_h) \right) + \sum_{i=n_\ell+1}^{n} \log \left(\sum_{k=1}^{g} \sum_{h=1}^{H} \pi_k \tau_{hk} p(\boldsymbol{x}_i; \theta_h) \right), \quad (2.25)$$

avec H le nombre de composants total.

Remarquons qu'il est possible de compléter les données de deux façons différentes :

- soit on considère que les variables manquantes sont les appartenances des données aux composants,

– soit on considère que les variables manquantes sont les étiquettes des données non étiquetées.

Notons

$$c_{ih} = \begin{cases} 1 \text{ si l'individu } i \text{ est issu du composant } h, \\ 0 \text{ sinon.} \end{cases}$$

La première complétion possible donne

$$\mathcal{L}_c = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} \sum_{h=1}^{H} z_{ik} c_{ih} \log\left(\pi_k \tau_{hk} p(\boldsymbol{x}_i; \theta_h)\right) + \sum_{i=n_\ell+1}^{n} \sum_{h=1}^{H} c_{ih} \log\left(\sum_{k=1}^{g} \pi_k \tau_{hk} p(\boldsymbol{x}_i; \theta_h)\right). \tag{2.26}$$

Dans ce cas il est plus simple de reparamétrer le problème sous la forme $\pi_k \tau_{hk} = \tau_h \pi_{kh}$ avec τ_h la probabilité du composant h et π_{kh} la probabilité de la classe k sachant le composant h. Cela donne alors

$$\mathcal{L}_c = \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} \sum_{h=1}^{H} z_{ik} c_{ih} \log\left(\tau_h \pi_{kh} p(\boldsymbol{x}_i; \theta_h)\right) + \sum_{i=n_\ell+1}^{n} \sum_{h=1}^{H} c_{ih} \log\left(\tau_h p(\mathbf{x}_i; \theta_h)\right). \tag{2.27}$$

Il faut maintenant calculer l'espérance de la vraisemblance complétée ce qui donne

$$\begin{aligned} Q(\theta|\theta^{(r)}) &= \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} \sum_{h=1}^{H} z_{ik} \mathbb{E}\left[c_{ih}|\boldsymbol{x}_i, \boldsymbol{z}_i; \theta^{(r)}\right] \log\left(\tau_h \pi_{kh} p(\boldsymbol{x}_i; \theta_h)\right) && (2.28) \\ &\quad + \sum_{i=n_\ell+1}^{n} \sum_{h=1}^{H} \mathbb{E}\left[c_{ih}|\boldsymbol{x}_i; \theta^{(r)}\right] \log\left(\tau_h p(\mathbf{x}_i; \theta_h)\right). && (2.29) \end{aligned}$$

On retrouve ensuite l'étape M standard en prenant en compte les poids précédents.

Pour la seconde approche, on complète dans un premier temps par la classe des données non étiquetées puis on calcule l'espérance de la classe conditionnellement aux données observées

$$\begin{aligned} Q(\theta|\theta^{(r)}) &= \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} z_{ik} \log\left(\pi_k \sum_{h=1}^{H} \tau_{hk} p(\mathbf{x}_i; \theta_h)\right) && (2.30) \\ &\quad + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} \mathbb{E}\left[z_{ik}|\boldsymbol{x}_i; \theta^{(r)}\right] \log\left(\pi_k \sum_{h=1}^{H} \tau_{hk} p(\boldsymbol{x}_i; \theta_h)\right). && (2.31) \end{aligned}$$

L'étape M est explicite pour les π_k mais pas pour les autres paramètres. Toutefois nous pouvons maintenant compléter par rapport au composant ce qui donne

$$\begin{aligned} Q_2(\theta|\theta^{(r+1/2)}) &= \sum_{i=1}^{n_\ell} \sum_{k=1}^{g} \sum_{h=1}^{H} z_{ik} \mathbb{E}\left[c_{ih}|\boldsymbol{x}_i, \boldsymbol{z}_i, \theta^{(r+1/2)}\right] \log\left(\pi_k^{(r+1)} \tau_{hk} p(\boldsymbol{x}_i; \theta_h)\right) \\ &\quad + \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{g} \sum_{h=1}^{H} \mathbb{E}\left[z_{ik}|\boldsymbol{x}_i, \theta^{(r)}\right] \mathbb{E}\left[c_{ih}|\boldsymbol{x}_i, z_{ik} = 1, \theta^{(r+1/2)}\right] \log\left(\pi_k^{(r+1)} \tau_{hk} p(\boldsymbol{x}_i; \theta_h)\right). \end{aligned}$$

L'étape M est alors une étape de maximisation classique en prenant en compte les poids précédents. On a donc deux algorithmes EM emboîtés. Il faut choisir le nombre d'itérations à l'intérieur du premier algorithme. Remarquons que ce schéma à deux niveaux

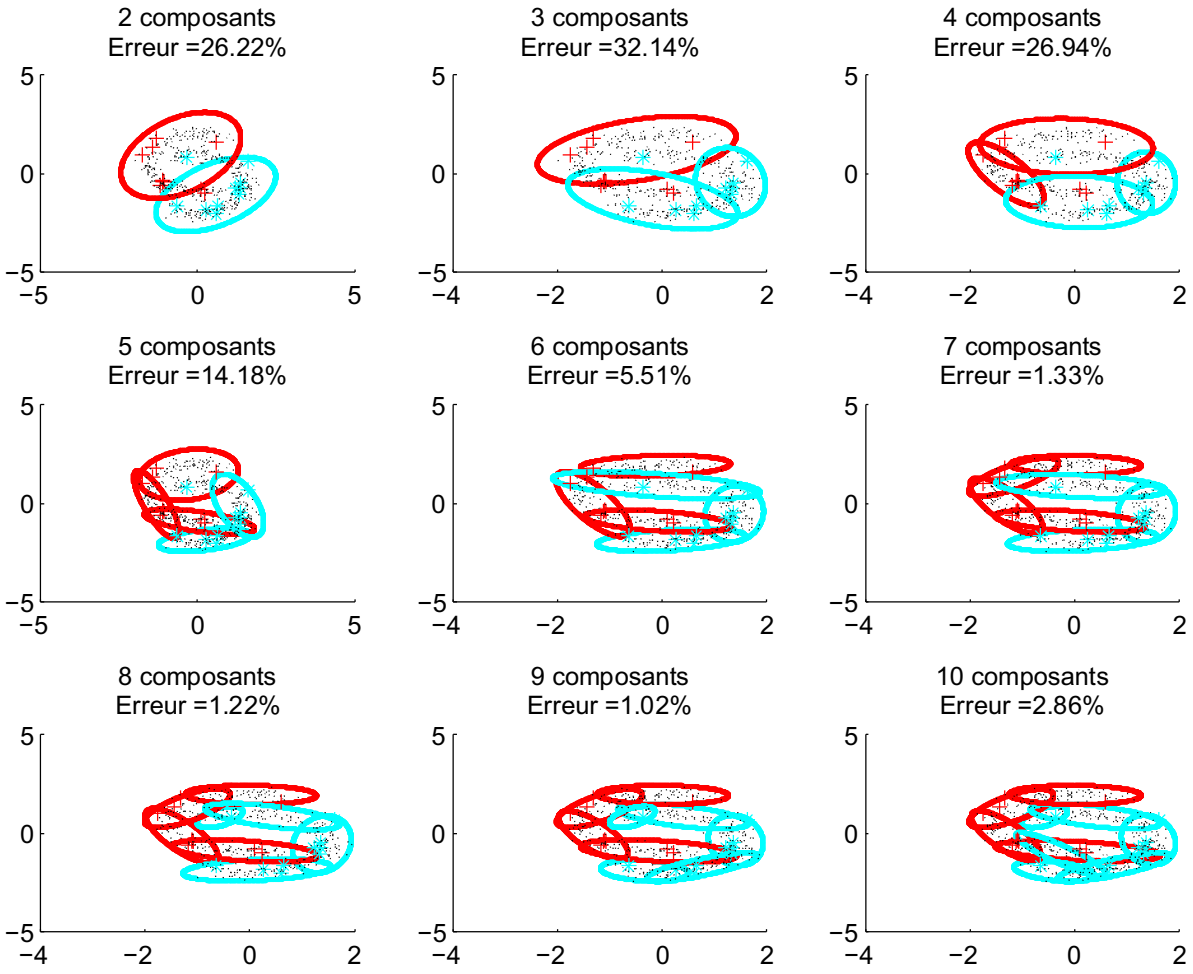

FIG. 2.5 – Isodensités estimées en fonction du nombre de composants totaux.

peut sembler être une complexification inutile, mais ils permettent deux interprétations différentes : on peut interpréter la première version comme se focalisant sur l'approche approximation de densité, tandis que la seconde se focalise sur l'approche estimation de la densité conditionnellement à la classe.

Pour l'approche à composants séparés on montre maintenant un exemple où l'utilisation de cette approche permet une amélioration radicale des résultats. Il s'agit de l'exemple des deux lunes enchevêtrées abondamment utilisé en classification semi-supervisée pour illustrer les bonnes performances des méthodes de type propagation des étiquettes dans un graphe et dont on a parlé dans la section 1.3.2. On représente figure 2.5 comment évolue l'estimation en fonction du nombre de composants totaux. On voit qu'au fur et à mesure que le nombre de composants totaux augmente, le taux d'erreur diminue, excpeté dans le cas de dix composants où elle commence à augmenter.

2.4 Expérimentations

Nous effectuons une série d'expériences sur données réelles et simulées. La mise en œuvre de l'estimation semi-supervisée des paramètres est implémentée dans le logiciel MIXMOD[3] (Biernacki *et al.* , 2006) pour les modèles gaussiens et le produit de distributions multinomiales.

2.4.1 Données simulées

Dans cette section nous mettons en évidence l'interêt de la classification semi-supervisée sur des données simulées. Soit deux classes gaussiennes en dimension 50, et en proportions identiques. On a

$$\boldsymbol{X}|Z_1 = 1 \sim \mathcal{N}(0, I_{50})$$

et

$$\boldsymbol{X}|Z_2 = 1 \sim \mathcal{N}(\mu, I_{50})$$

avec $\mu_i = \frac{1}{i}$ pour i allant de 1 à 50. Soit $n_\ell = 100$ données étiquetées et $n_u = 10000$ données non étiquetées. On trace figure 2.6 l'erreur moyenne en fonction du nombre de variables conservées dans les cadres supervisé et semi-supervisé. Cette figure nous permet de remarquer deux intérêts de l'approche semi-supervisée. Premièrement, le semi-supervisé permet d'obtenir un taux d'erreur plus bas (taux d'erreur moyen 27,79%) que le supervisé (taux d'erreur moyen 29,36%) dans le cas où on considère le nombre de variables optimal pour le supervisé. Deuxièmement, le semi-supervisé permet d'utiliser efficacement un plus grand nombre de variables et permet d'obtenir un taux d'erreur minimal de 26,82%. Ainsi deux effets participent à la réduction du taux d'erreur moyen. Tout d'abord pour des modèles de même complexité, on a une réduction du taux d'erreur moyen, ensuite le semi-supervisé permet de faire un meilleur usage des modèles plus complexes.

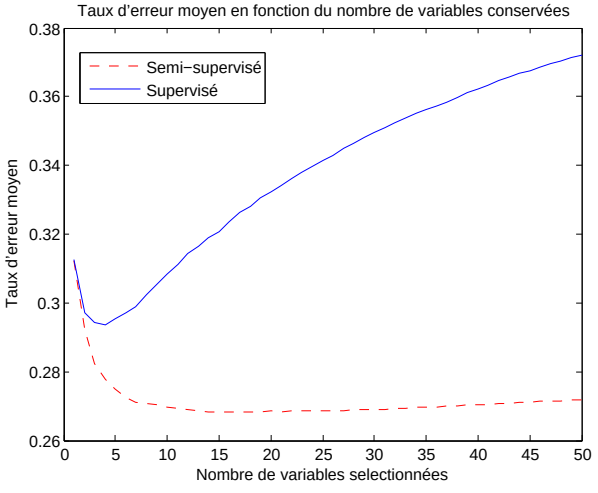

FIG. 2.6 – Taux d'erreur en fonction du nombre de variables conservées

[3] `http://www-math.univ-fcomte.fr/mixmod/`

2.4.2 Données de l'UCI

La plupart des jeux de données disponibles pour tester les performances des méthodes de classification sont des jeux de données totalement étiquetés. Ces jeux de données peuvent facilement être rendus partiellement étiquetés en cachant une partie des étiquettes. L'intérêt de cette approche, même si la problématique est à la base supervisée est qu'elle permet de vérifier l'intérêt de l'utilisation des données non étiquetées, ainsi que de valider les résultats. En effet, les étiquettes des données non étiquetées sont en fait connues. Ici on compare les performances des approches supervisées et semi-supervisées dans les cas gaussiens homoscédastiques et hétéroscédastiques. Le dispositif expérimental est illustré table 2.2. Pour chaque jeu de données, si un échantillon test est fourni, les étiquettes de ce dernier sont cachées et les données sont utilisées comme des données non étiquetées. Sinon, on génère aléatoirement 100 échantillons de n_u données non étiquetées et n_ℓ données étiquetées en cachant n_u étiquettes au hasard.

Jeu de données	n	d	G	Échantillon test	n_u	n_ℓ
Breast Cancer	569	30	2	non	500	69
Crabes	200	5	4	non	150	50
Iris	150	4	3	non	100	50
Parkinson	195	22	2	non	95	100
Pima	532	7	2	oui	332	200
Transfusion	748	4	2	non	548	200

TAB. 2.2 – Dispositif expérimental.

	Homoscédastique		Hétéroscédastique	
	Supervisé	Semi-supervisé	Supervisé	Semi-supervisé
Brest Cancer	9,79 (2,23)	9,38 (5,12)	59,69 (10,53)	7,66 (8,28)
Crabes	6,71 (2,33)	8,61 (2,25)	11,36 (4,76)	6,47 (3,46)
Iris	2,93 (1,34)	2,17 (1,04)	4,06 (1,93)	3,05 (1,35)
Parkinson	15,04 (3,46)	14,91 (4,03)	26,84 (19,23)	20,37 (9,00)
Pima	20,18	19,58	23,49	25,00
Transfusion	25,78 (9,25)	23,34 (2,72)	30,17 (17,11)	26,94 (10,56)

TAB. 2.3 – Taux d'erreur moyen dans différentes configurations.

Les résultats sont présentés table 2.3, l'écart-type du taux d'erreur est obtenu à partir des 100 séparations données étiquetées données non étiquetées et est écrit entre parenthèse. Dans la plupart de ces jeux de données le semi-supervisé produit de meilleurs résultats que le supervisé. D'autre part, on remarque certains cas où le modèle hétéroscédastique produit des résultats médiocres dans le cadre supervisé pour cause d'excès de variance. Le semi-supervisé réussit quant à lui à prendre efficacement en compte le cas hétéroscédastique. Pour les données Breast Cancer et Crabes il permet d'obtenir une erreur de classification plus faible que dans le cadre supervisé.

2.4.3 *Benchmarks* du livre de Chapelle *et al.* (2006)

Utilisation des modèles de classification en haute dimension

Dans le livre de Chapelle *et al.* (2006), des jeux de données sont fournis avec séparation données étiquetées données non étiquetées[4], et nous allons illustrer l'utilisation des modèles gaussiens en grande dimension vus section 2.3.1 sur ces derniers. On se limite à l'utilisation du modèle qui décompose la matrice de variance Σ_k en valeurs singulières et qui suppose que les $d-p$ plus petites valeurs propres sont identiques, ce qui correspond au modèle $[A_{ij}B_iQ_iD]$ de Bouveyron *et al.* (2007). Les jeux de données proposés comportent des données en grande dimension (241 variables pour 1500 données observées). Douze séparations entre 1400 données non étiquetées et 100 données étiquetées sont proposées. Celles-ci sont utilisées pour comparer les performances des différents modèles utilisés. Nous partons du modèle qui suppose que toutes les valeurs propres sont identiques et nous allons jusqu'au modèle qui suppose que les dix plus grandes sont différentes, et que les plus petites sont identiques.

Le premier jeu de données intitulé *g241c* est constitué de données artificielles qui respectent les hypothèses de modélisation, à savoir que conditionnellement à la classe la distribution des covariables est gaussienne. Les résultats supervisés et semi-supervisés sont illustrés figure 2.7. Les boîtes à moustache (Tukey, 1977) les plus larges représentent les résultats dans le cadre semi-supervisé, tandis que les boîtes à moustache les moins larges représentent les résultats dans le cadre supervisé. Le modèle le plus simple met en avant l'intérêt de l'utilisation des données non étiquetées pour améliorer la précision de la règle de classement apprise. Pour les modèles plus complexes, on voit que l'erreur de classement augmente dans les cadres supervisés et semi-supervisés, ce qui met en avant le phénomène de sur-apprentissage.

Le second jeu de données intitulé *g241n* illustre une situation où une classe est modélisée par deux composants gaussiens, le modèle postulé est donc faux. Figure 2.8 on remarque que les données non étiquetées contribuent tout de même à réduire le taux d'erreur moyen quand au moins une valeur propre est supposée différente des autres. Remarquons que ce n'est ni le modèle le plus complexe ni le modèle le plus simple qui produit les meilleurs résultats, mais qu'il existe un compromis entre la bonne approximation de la distribution des données par le modèle et la variance dans l'estimation des paramètres.

Le troisième jeu de données intitulé *Digit1* représente des données artificielles plus proches de la réalité. Pour l'exemple on s'est limité à 10 valeurs propres différentes au maximum. On voit figure 2.9 que pour les modèles les plus simples les données non étiquetées dégradent les performances de la règle de classement. Cependant au fur et à mesure que des modèles plus complexes sont proposés, les données non étiquetées améliorent les performances de la règle de classement. Remarquons qu'ici il aurait sûrement fallu laisser plus de valeurs propres libres pour obtenir de meilleures performances.

Enfin un quatrième jeu de données intitulé *USPS* représente un jeu de données réelles. Il s'agit de distinguer les chiffres 5 et 2 des autres chiffres. Ici on voit que pour les modèles considérés, les données non étiquetées dégradent la règle de classement apprise. Les résultats sont illustrés figure 2.10. Cependant cette dégradation semble diminuer au fur et à mesure que des modèles plus complexes sont utilisés.

[4] `http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html`

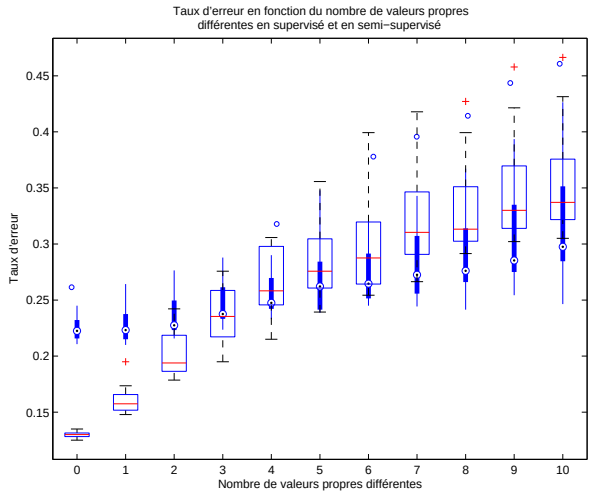

FIG. 2.7 – Jeu de données *g241c* (semi-supervisé : boîtes larges, supervisé : boîtes fines) avec les modèles haute dimension.

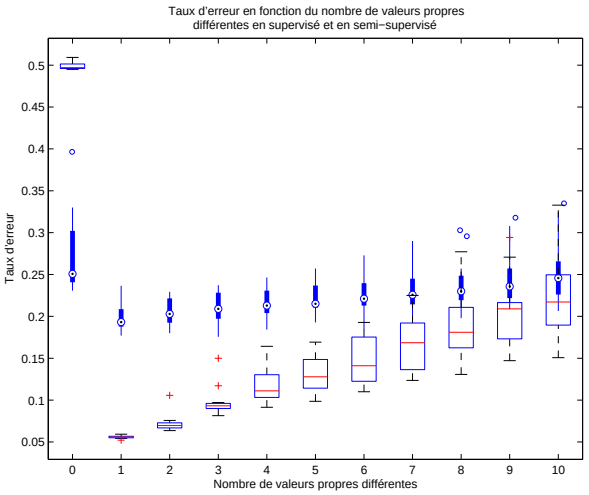

FIG. 2.8 – Jeu de données *g241n* (semi-supervisé : boîtes larges, supervisé : boîtes fines) avec les modèles haute dimension.

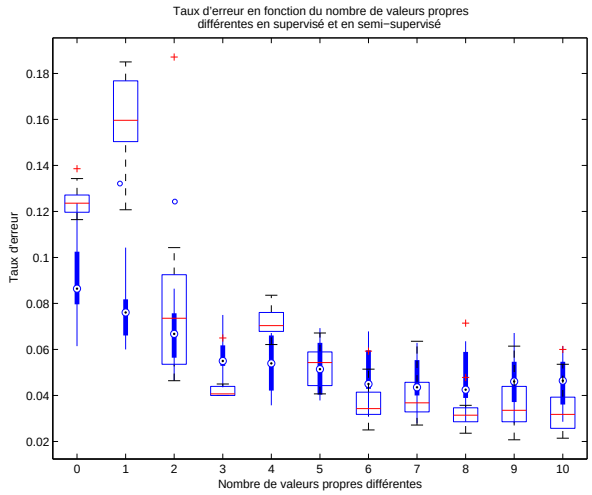

FIG. 2.9 – Jeu de données *Digit1* (semi-supervisé : boîtes larges, supervisé : boîtes fines) avec les modèles haute dimension.

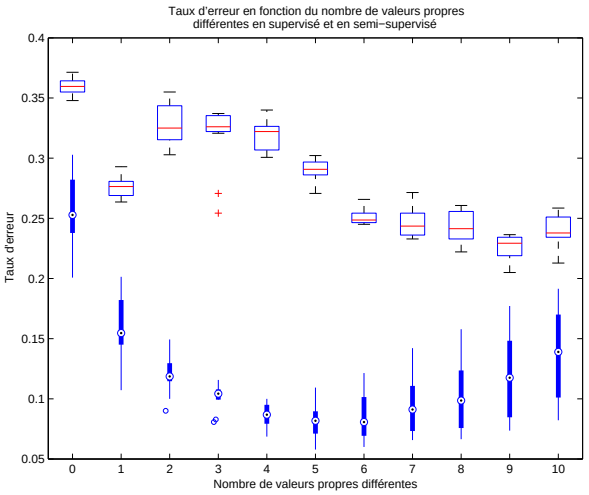

FIG. 2.10 – Jeu de données *USPS* (semi-supervisé : boîtes larges, supervisé : boîtes fines) avec les modèles haute dimension.

Utilisation des modèles à plusieurs composants par classe

Un autre type de modèle utile en grande dimension est le modèle d'indépendance conditionnelle qui évite les problèmes d'inversibilité de la matrice de covariance rencontrés en grande dimension. Il consiste à considérer des distributions gaussiennes avec des matrices de variance diagonales. Cependant ces modèles sont dans de nombreuses situations trop simplistes ; une possibilité est alors de considérer un modèle à plusieurs composants par classe. Nous illustrons les performances de ces modèles selon le nombre de composants

par classe choisis, sur les mêmes jeux de données que précédemment.

Pour le jeu de données *g241c*, le vrai nombre de composants par classe est 1. On voit figure 2.11 que c'est le modèle le plus simple qui produit les meilleures performances. On note encore une amélioration des performances dans le cadre semi-supervisé.

Pour le jeu de données *g241n*, les données sont issues d'un modèle à deux composants par classe. On voit figure 2.12 que c'est le modèle à deux composants par classe qui produit les meilleurs résultats. Un modèle à un composant par classe est trop simpliste et un modèle à plus de deux composants par classe est trop complexe. On remarque que dans le cas où les hypothèses de modélisation sont fausses, le semi-supervisé dégrade les performances du classifieur appris. Le semi-supervisé tire un très bon parti des données non étiquetées lorsque le modèle postulé est correct.

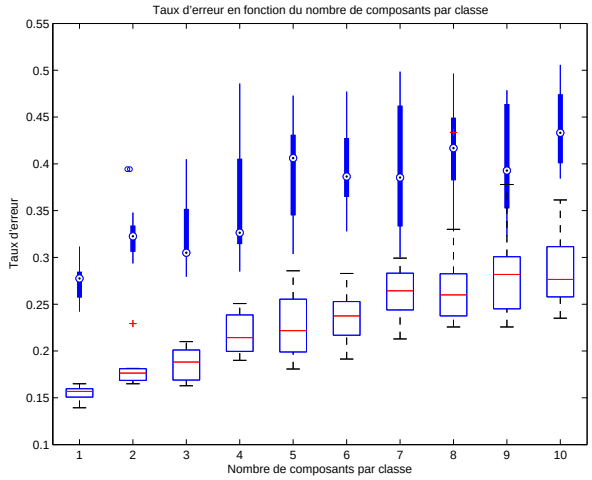

FIG. 2.11 – Jeu de données *g241c* (semi-supervisée : boîtes larges, supervisé : boîtes fines) avec les modèles à plusieurs composants par classe.

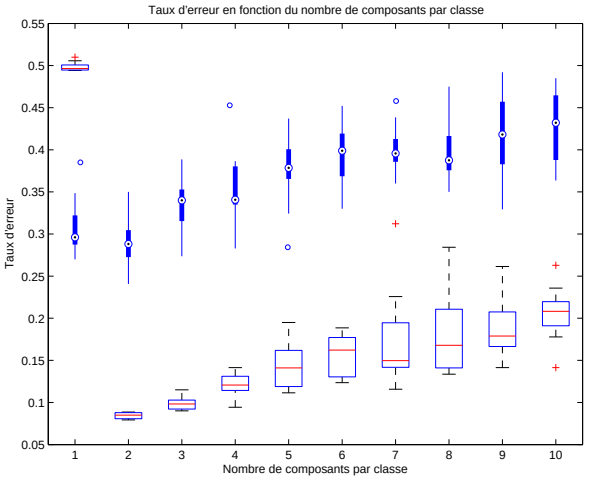

FIG. 2.12 – Jeu de données *g241n* (semi-supervisée : boîtes larges, supervisé : boîtes fines) avec les modèles à plusieurs composants par classe.

Pour le jeu de données *Digit1*, on voit figure 2.13 que les résultats obtenus sont à peu près les mêmes dans les cadres supervisés et semi-supervisés et qu'ils ne varient pas trop selon le nombre de composants choisis.

Pour le jeu de données *USPS*, on voit figure 2.14 que l'utilisation de plusieurs composants par classe permet d'améliorer fortement les résultats produits. Le faible nombre de données étiquetées ne permet pas d'estimer des modèles à plus de sept composants par classe dans le cadre supervisé.

Remarquons qu'on aurait pu combiner l'approche à plusieurs composants par classe et l'approche modèle en grande dimension. Dans ce cas il faudrait à la fois choisir le nombre de composants par classe et le nombre de plus grandes valeurs propres différentes. D'autre part nous avons opté pour l'approche composants séparés et nous avons imposé le même nombre de composants pour chaque classe, là où l'approche composants communs était aussi envisageable.

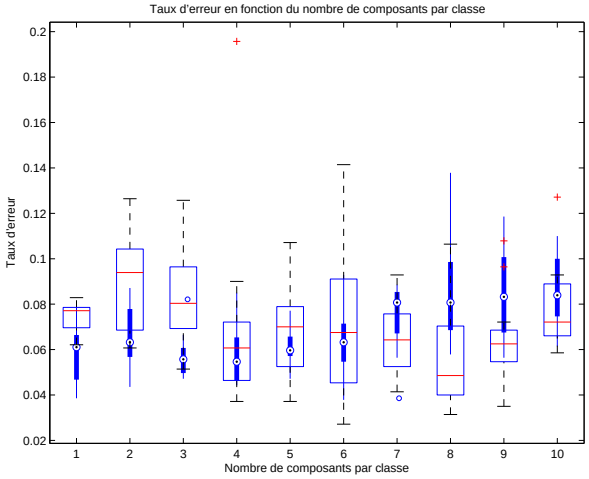

FIG. 2.13 – Jeu de données *Digit1* (semi-supervisée : boîtes larges, supervisé : boîtes fines) avec les modèles à plusieurs composants par classe.

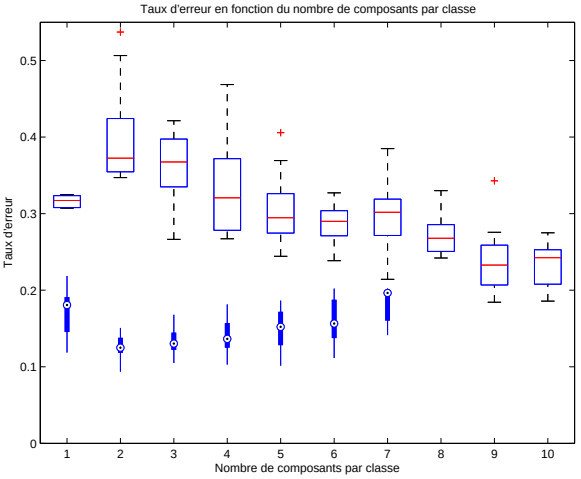

FIG. 2.14 – Jeu de données *USPS* (semi-supervisée : boîtes larges, supervisé : boîtes fines) avec les modèles à plusieurs composants par classe.

2.4.4 Données sur le syndrome de Cushing

L'intérêt du semi-supervisé se manifeste avant tout lorsque le nombre de données non étiquetées est grand devant le nombre de données étiquetées. Cependant dans certaines situations le nombre de données est tellement petit que l'utilisation de quelques données non étiquetées permet une amélioration des performances de l'analyse statistique. Il s'agit par exemple de l'étude des données d'Aitchison & Dunsmore (1975) sur le syndrome de Cushing. Le syndrome de Cushing est une maladie hypertensive associée à la sur-sécrétion de Cortisol par la glande surrénale. Il existe trois types différents de syndrome de Cushing : le type a (adénome), le type b (hyperplasie latérale), le type c (cancéreux). Les données d'Aitchison & Dunsmore sont constituées de 27 patients pour lesquels les concentrations urinaires de deux hormones stéroïdiennes ont été mesurées. Pour 6 des 27 patients le type est inconnu. Une transformation logarithmique des données permet d'obtenir des distributions relativement proches de la normalité conditionnellement à la classe. On trace figure 2.15 les isodensités conditionnellement à la classe quand les paramètres sont estimés dans le cadre supervisé et où le modèle est sélectionné en utilisant le critère BIC. Le modèle sélectionné est $[\pi\lambda D_k A D'_k]$ c'est-à-dire que les classes ont des proportions identiques, le même volume et la même forme mais des orientations différentes. Deux des points non étiquetées en noir n'appartiennent à aucune isodensité, ce qui tend à montrer que la distribution marginale des données est relativement mal approchée lorsque les données étiquetées sont utilisées seules. Maintenant si les données non étiquetées sont prises en compte dans l'estimation des paramètres on obtient la figure 2.16, et le modèle sélectionné par BIC est $[\pi\lambda_k D_k A D'_k]$ c'est-à-dire que les classes ont la même forme mais des volumes et des orientations différentes. On voit que les densités estimées « collent » maintenant mieux à la distribution de l'ensemble des données. La question qu'on se pose est « Dans quelle mesure cette amélioration en terme d'approximation de densité conduit-elle à une amélioration en terme de prédiction de la classe ? ». Les probabilités *a posteriori* pour les

individus non étiquetés sont présentées tables 2.4 et 2.5. Les classes d'appartenance des individus non étiquetées sont inconnues, cependant les 4 premiers individus sont suspectés d'appartenir aux classes b, c, b et a. On remarque donc que tous les individus sont biens classés en utilisant la règle de classement supervisée, tandis qu'un individu est mal classé en utilisant la règle de classement semi-supervisée.

Les appartenances de ces 4 individus ne sont pas certaines, en pratique on peut se demander si les deux cas qui modifient fortement l'analyse pour le type c ne serait en fait pas issus d'un nouveau type.

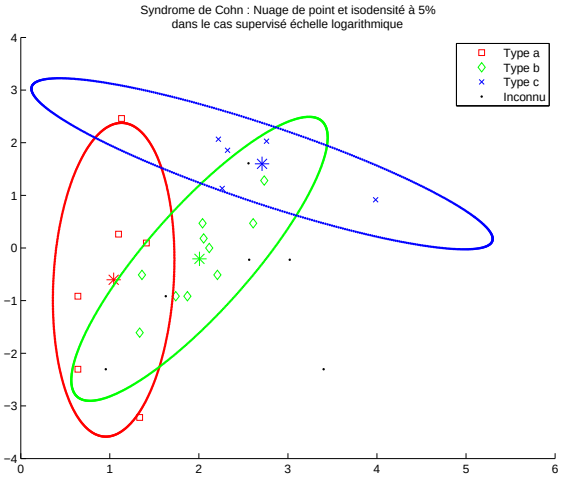

FIG. 2.15 – Syndrome de Cushing : situation supervisée.

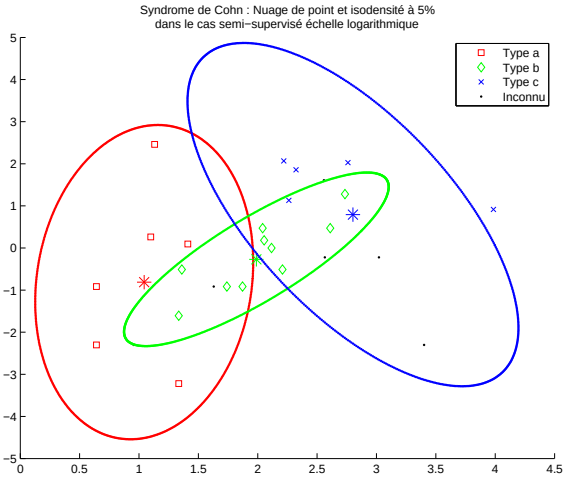

FIG. 2.16 – Syndrome de Cushing : situation semi-supervisée.

Type a	Type b	Type c
10,64	89,36	0,00
0,00	15,02	84,98
0,00	100,00	0,00
70,60	29,40	0,00
0,05	99,95	0,00
0,00	99,96	0,04

TAB. 2.4 – Syndrome de Cushing : probabilités a posteriori supervisée.

Type a	Type b	Type c
12,87	87,10	0,03
0,02	14,61	85,37
0,04	38,56	61,40
82,44	17,56	0,00
0,00	0,00	100,00
0,00	0,12	99,88

TAB. 2.5 – Syndrome de Cushing : probabilités à posteriori semi-supervisée.

2.5 Conclusion

Nous avons détaillé la richesse des modèles génératifs. Ceux-ci permettent de prendre en compte toute l'information disponible. On a pu voir l'importance du modèle choisi pour prendre efficacement en compte l'information apportée par les données non étiquetées. Deux questions se posent alors. Tout d'abord « Comment juger de la pertinence du modèle postulé à partir de données non étiquetées ». Ce qui fera l'objet du chapitre 3. Puis

« Comment choisir le meilleur modèle possible compte tenu de l'objectif décisionnel ? ». Ce qui fera l'objet du chapitre 4.

Chapitre 3

Utilisation des données non étiquetées pour juger de la pertinence d'un modèle

Comme mentionné dans Cozman & Cohen (2002) les données non étiquetées peuvent dans un certain nombre de situations dégrader la règle de classement. Sous l'hypothèse d'échantillonnage MCAR (section 1.2.1), ceci ne peut-être le cas que si le modèle postulé est mal spécifié. En effet, si le modèle postulé est bien spécifié l'information apportée par ces dernières est efficacement prise en compte par le modèle, et conduit à une amélioration de la règle de classement supervisée. Dans ce chapitre on cherche à répondre à la question « Le modèle utilisé est-il pertinent ? ». Pour cela, nous partons de l'idée que lorsque différentes méthodes peuvent être utilisées pour estimer un même paramètre il est intéressant de les comparer (Mclachlan, 2004). Nous comparons alors les estimations non supervisée, supervisée et semi-supervisée des paramètres.

Nous proposons alors la mise en place d'un test statistique qui permet de détecter, si les paramètres estimés de ces différentes façons sont suffisamment proches compte tenu de l'hypothèse que le modèle est bien spécifié. La mise en place de ce test à fait l'objet d'un communication lors des rencontres Franco-Italiennes SFC-CLADDAG Vandewalle *et al.* (2008)[1]. Ce test permet de détecter des situations où le semi-supervisé est susceptible d'améliorer les performances du supervisé ; dans le cas contraire il faudra proposer d'autres modèles. Nous abordons dans un second temps la question de la pertinence du modèle proposé sous l'angle du choix de modèle. Ainsi, nous proposons une procédure de choix de modèle utilisant le critère BIC et dont l'utilisation est pleinement justifiée compte tenu des propriétés de la statistique de test utilisée dans la partie précédente.

3.1 Introduction

L'apprentissage semi-supervisé des paramètres d'un modèle génératif consiste à utiliser ensemble les données étiquetées et non étiquetées pour estimer le même vecteur des paramètres. Or, ces paramètres pourraient très bien être estimés séparément en utilisant d'une part les données étiquetées et d'autre part les données non étiquetées. On note les

[1] `http://math.univ-lille1.fr/~vandewal/documents/vbcg.pdf`

log-vraisemblances théoriques suivantes

$$L_s(\theta) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Z}}[\log p(\boldsymbol{X},\boldsymbol{Z};\theta)], \quad (3.1)$$
$$L_{ns}(\theta) = \mathbb{E}_{\boldsymbol{X}}[\log p(\boldsymbol{X};\theta)], \quad (3.2)$$
$$L_{ss}(\theta) = \beta L_s(\theta) + (1-\beta) L_{ns}(\theta), \quad (3.3)$$

où, rappelons le, β indique la fraction de données étiquetées (voir définition précise au chapitre 1, section 1.2.1). Quand les paramètres sont estimés à partir des seules données étiquetées, les paramètres estimés $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$ convergent vers

$$\theta_s^* = \arg\max_{\theta\in\Theta} L_s(\theta).$$

Cela correspond aux paramètres du modèle qui minimisent la divergence de Kullback-Leibler à la distribution jointe de $(\boldsymbol{X},\boldsymbol{Z})$. Quand les paramètres sont estimés à partir des seules données non étiquetées, les paramètres estimés $\hat{\theta}_{\mathbf{x}_u}$ convergent vers

$$\theta_{ns}^* = \arg\max_{\theta\in\Theta} L_{ns}(\theta).$$

Cela correspond aux paramètres qui minimisent la divergence de Kullback-Leibler à la distribution marginale de $\boldsymbol{X}$. Ainsi d'un côté on cherche à approcher au mieux une distribution jointe, et de l'autre à approcher au mieux une distribution marginale. Ces deux objectifs sont réconciliés en pratique si l'hétérogénéité de la distribution de $\boldsymbol{X}$ est expliquée par la variable $\boldsymbol{Z}$. Plus précisément, ceci est le cas si le modèle est bien spécifié, puisqu'on a alors $\theta_s^* = \theta_{ns}^*$ à une permutation des classes près. Dans ce cas le fait d'utiliser toute l'information disponible pour estimer les paramètres permet une réduction de la variance des estimateurs, et par suite une réduction de l'erreur de classement moyenne, ce qui justifie pleinement l'utilisation de l'approche semi-supervisée. En semi-supervisé, les paramètres estimés $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$ convergent vers

$$\theta_{ss}^* = \arg\max_{\theta} L_{ss}(\theta).$$

Ainsi cette estimation correspond asymptotiquement à minimiser une combinaison convexe de la divergence de Kullback-Leibler à la distribution jointe de $(\boldsymbol{X},\boldsymbol{Z})$ et de la divergence de Kullback-Leibler à la distribution marginale de $\boldsymbol{X}$. Comme dit précédemment, si la distribution d'échantillonnage appartient au modèle postulé on a bien entendu :

$$\theta_s^* = \theta_{ns}^* = \theta_{ss}^*.$$

Si jamais ce n'est pas le cas, cela signifie que le modèle est mal spécifié et la valeur de θ_{ss}^* se rapprochera plus ou moins de θ_s^* ou θ_{ns}^* selon la fraction β de données étiquetées dans le cas où le modèle est mal spécifié.

Remarquons ici que l'hypothèse que les données sont exactement issues du modèle postulé est bien sûr irréaliste, cependant cette hypothèse peut fournir dans de nombreuses situations une approximation raisonnable de la réalité. Ainsi, nous allons chercher à détecter quand cette hypothèse est raisonnable, c.-à-d. quand les paramètres estimés de différentes manières sont suffisamment proches pour qu'une amélioration puisse être attendue quand ces derniers sont estimés ensembles.

3.2 Réponse par un test

3.2.1 Heuristique

Comme suggéré par Mclachlan (2004), quand différentes estimations des paramètres sont possibles, ici les estimations supervisée, non supervisée et semi-supervisée, il est alors intéressant de les comparer. Rappelons qu'on s'attend à ce que tous ces estimateurs soient relativement proches sous l'hypothèse que le modèle postulé est bien spécifié puisque dans ce cas on a $\theta_s^* = \theta_{ns}^* = \theta_{ss}^*$. En conséquence, si ces estimateurs sont trop éloignés cela indique que le modèle postulé est incorrect et qu'il faut alors proposer d'autres modèles. En revanche, le fait d'avoir $\theta_s^* = \theta_{ns}^* = \theta_{ss}^*$ n'implique pas nécessairement que le modèle postulé soit correct. Cependant, il est raisonnable de supposer que tel est le cas, hormis quelques situations pathologiques de peu d'intérêt *a priori* dans notre étude. Ainsi, en pratique notre problème est d'évaluer si les différence constatée entre $(\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}, \hat{\theta}_{\mathbf{x}_u})$ et $\hat{\hat{\theta}}_{\mathbf{x},\mathbf{z}_\ell}$ peuvent être expliquées par les seules fluctuations d'échantillonnage ou non. Voici deux exemple qui permettent d'illustrer cette heuristique.

Considérons les données « Crabes » étudiées section 2.4.2. Il s'agit d'un échantillon de 200 données en dimension cinq. Deux sous-espèces de crabes sont considérées, et pour chaque sous-espèce on distingue les mâles des femelles, ce qui conduit à un problème de classification à quatre classes. L'échantillon est constitué de 50 représentants de chaque classe. Pour rendre l'échantillon partiellement étiqueté 150 étiquettes sont cachées au hasard. Cette procédure est répétée 100 fois. Quand on postule un modèle gaussien homoscédastique le supervisé produit une erreur moyenne de 6,71% et d'écart type 2,33 et le semi-supervisé produit une erreur moyenne de 8,61% et d'écart type 2,25. Cette dégradation semble s'expliquer ici par le fait que le modèle postulé est mal spécifié. Pour s'en convaincre, on trace figure 3.1, les iso-densités estimées de manière supervisée, semi-supervisée et non supervisée ainsi que les données et leurs étiquettes, tout ceci selon les deux premiers axes de l'ACP. On voit que dans le cadre semi-supervisé deux classes du haut qui étaient peu séparées par l'estimation supervisée ne le sont maintenant plus du tout, ce qui implique une perte d'information au niveau de la règle de classement, et qui explique alors les moins bonnes performances de la classification semi-supervisée.

Considérons maintenant les iris de Fisher (1936). Ces données ont été récoltées par Fisher et ont été utilisées pour illustrer les bonnes performances de l'analyse discriminante linéaire. Il s'agit des mesures de la longueur et la largeur des pétales et des sépales sur 150 iris provenant de 3 variétés différentes. Comme remarqué dans O'Neill (1978) les classifications supervisée et non supervisée produisent des résultats similaires sur ces données. À nouveau on cache aléatoirement 100 étiquettes. La classification supervisée produit une erreur moyenne de 2,93% avec un écart type de 1,34 tandis que la classification semi-supervisée produit une erreur moyenne de 2,17% avec un écart type de 1,04. Ainsi une amélioration des résultats est obtenue ce qui va dans le sens de la remarque d'O'Neill (1978). On représente figure 3.2 les isodensitées estimées conditionnellement aux classes sur les deux premiers axes de l'ACP. Ici, contrairement aux données « Crabes » les isodensitées des données conditionnellement aux classes sont relativement proches dans les trois configurations. Ainsi les données non étiquetées ne biaisent pas la solution supervisée, et permettent en outre une réduction de la variance dans l'estimation des paramètres, ce qui explique alors que le semi-supervisé améliore les résultats de la classification supervisée.

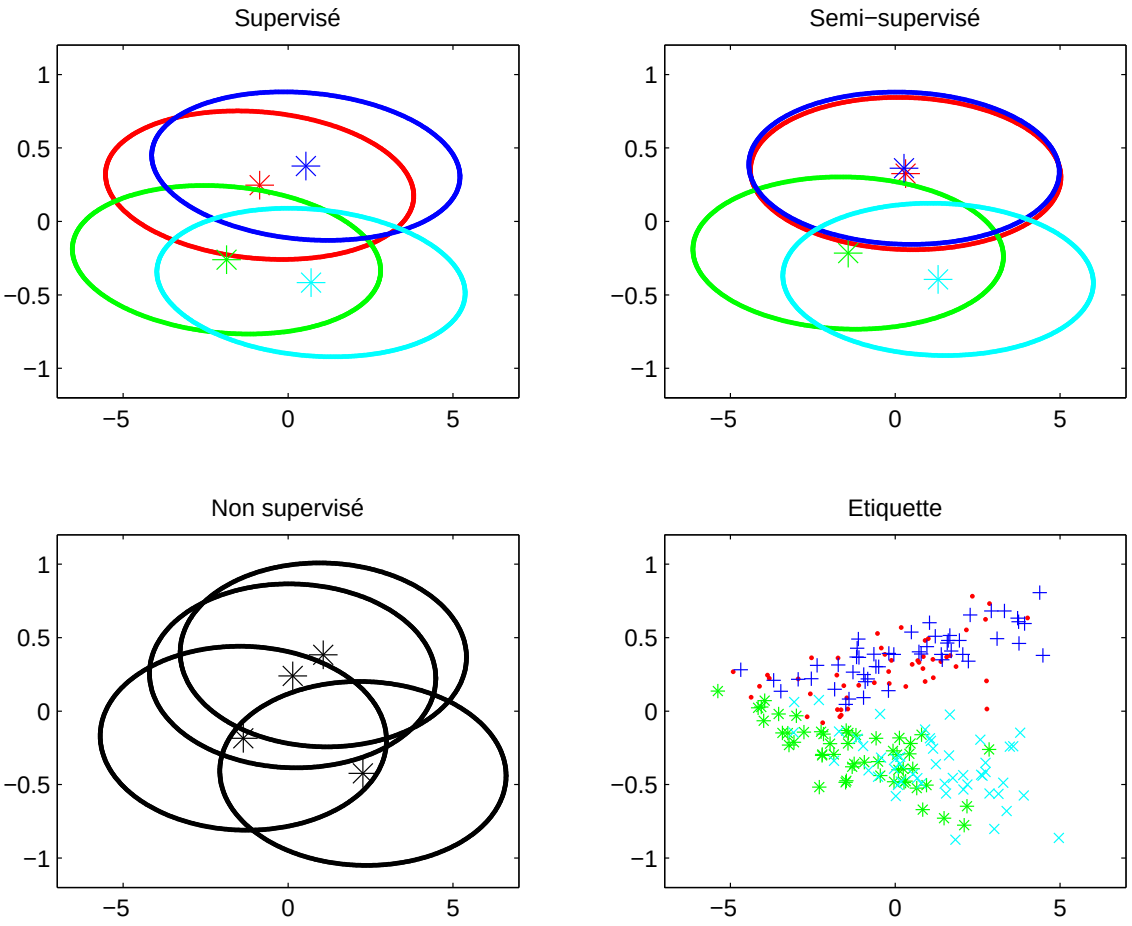

FIG. 3.1 – Densité conditionnellement aux classes estimée de différentes façons pour les données crabes.

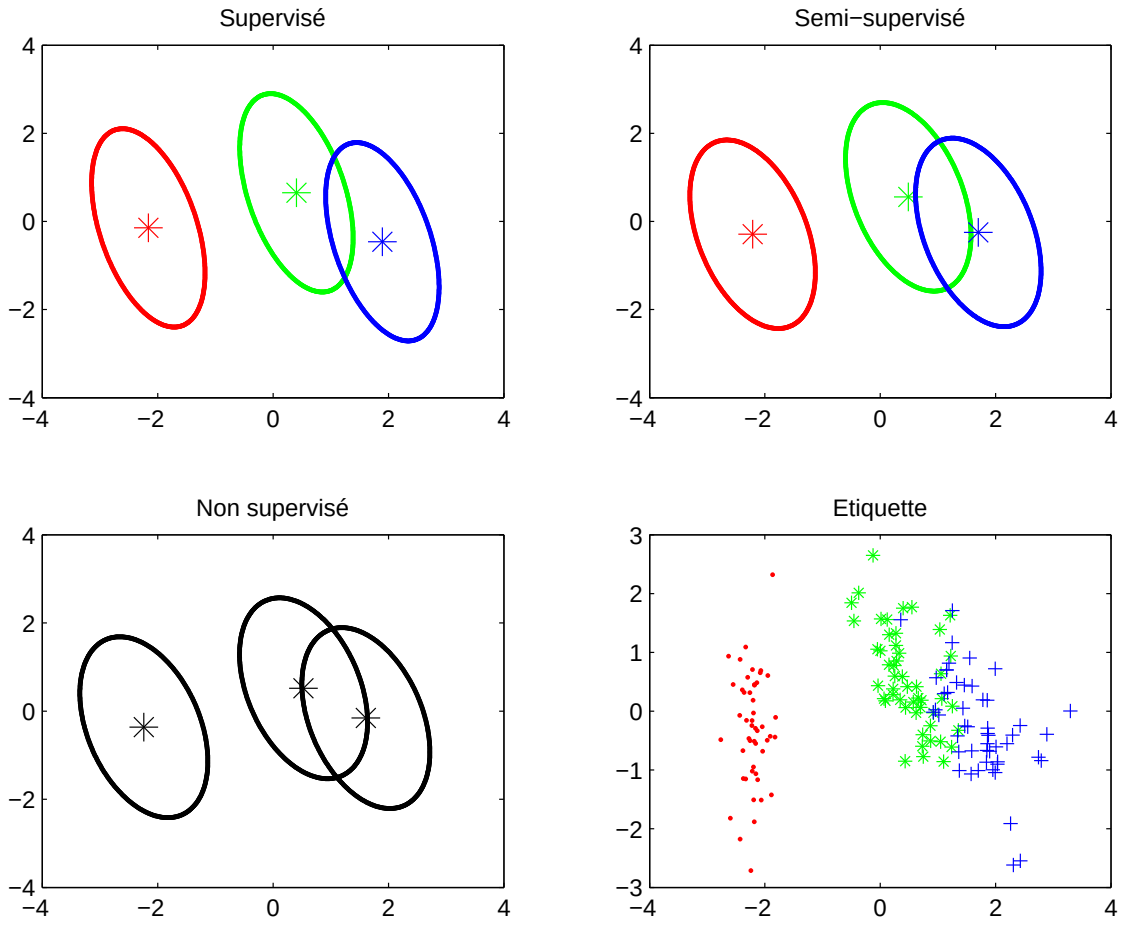

FIG. 3.2 – Densité conditionnellement aux classes estimée de différentes façons pour les données iris.

Ainsi on voit que le semi-supervisé est avant tout susceptible d'améliorer les performances du supervisé quand les paramètres estimés de différentes manières restent proches. La question qui se pose en pratique est alors « Comment détecter si ces paramètres estimés de différentes façons sont assez proches pour permettre une amélioration des performances ? ». Rigoureusement parlant, on souhaite tester si les paramètres estimés de ces différentes façons convergent vers la même valeur. Nous formalisons cette question dans la section suivante.

3.2.2 Test proposé

Soit $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$, $\hat{\theta}_{\mathbf{x}_u}$ et $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$ définis de la manière suivante

$$\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta;\mathbf{x}_\ell,\mathbf{z}_\ell) \tag{3.4}$$

$$\hat{\theta}_{\mathbf{x}_u} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta;\mathbf{x}_u) \tag{3.5}$$

$$\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathcal{L}(\theta;\mathbf{x}_\ell,\mathbf{z}_\ell,\mathbf{x}_u). \tag{3.6}$$

Si le modèle postulé est correct on sait maintenant qu'on doit avoir $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$, $\hat{\theta}_{\mathbf{x}_u}$ et $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$ qui convergent vers la même valeur θ^*. Par contre, si $\theta^*_s \neq \theta^*_{ns}$, il n'est pas souhaitable d'utiliser le modèle proposé. On teste alors $H_0 : \{\theta^*_s = \theta^*_{ns}\}$ contre $H_1 : \{\theta^*_s \neq \theta^*_{ns}\}$. Dans ce contexte, un test classique est le test du rapport des vraisemblances maximales (LRT). Soit $\mathcal{D}$ un échantillon, le LRT s'écrit

$$LRT = -2\log \frac{\sup_{\theta\in\Theta_0} p(\mathcal{D};\theta)}{\sup_{\theta\in\Theta_1\cup\Theta_0} p(\mathcal{D};\theta)}, \tag{3.7}$$

avec Θ_0 l'espace des paramètres sous H_0 et Θ_1 l'espace des paramètres sous H_1. Dans notre cas la statistique de test s'écrit

$$LRT = 2[\mathcal{L}(\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell};\mathbf{x}_\ell,\mathbf{z}_\ell) + \mathcal{L}(\hat{\theta}_{\mathbf{x}_u};\mathbf{x}_u) - \mathcal{L}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell};\mathbf{x}_\ell,\mathbf{z}_\ell,\mathbf{x}_u)]. \tag{3.8}$$

Soit ν le nombre de paramètres du modèles considéré. Sous H_0, ν paramètres sont estimés, tandis que sous H_1, 2ν paramètres sont estimés. Sous des conditions de régularité standards, la distribution asymptotique du LRT est un χ^2 dont le nombre de degrés de liberté est égal à la différence du nombre de degrés de liberté entre les deux paramètrisations considérées (ici ν). Plus précisément les conditions de régularité requises ici sont les mêmes que celles nécessaires à prouver la normalité asymptotique de l'estimateur du maximum de vraisemblance (van der Vaart, 2000, Théorème 16.7).

Le LRT permet d'apprécier de manière quantitative les différences entre $\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}$, $\hat{\theta}_{\mathbf{x}_u}$ et $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$. Ainsi, le test proposé revient à tester H_0 : { Le modèle postulé est pertinent (excepté des cas exotiques *a priori* sans intérêt).} contre H_1 : { Le modèle postulé n'est pas pertinent.}. Remarquons que pour les modèles pathologiques où on a $\theta^*_s = \theta^*_{ss} = \theta^*_{ns}$ sans pour autant que le modèle soit correct, il n'est pas aberrant d'utiliser l'estimation semi-supervisée des paramètres puisque les données non étiquetées contribuent alors à réduire la variance des paramètres estimés. L'approche proposée est donc de tester H_0 contre H_1. Si H_0 est rejetée, il faudra proposer de nouveaux modèles.

$a\backslash n_\ell$	20	50	100	200	500	1000
0,5	11	10	9	3	1	3
1	17	8	5	3	3	3
1,5	9	8	7	8	7	6
2	7	8	6	5	7	6
2,5	10	12	3	8	7	8
3	9	7	5	4	6	5

TAB. 3.1 – Nombre de rejets de H_0 en fonction du nombre de données et de la séparation des classes pour le modèle $[\pi_k \lambda B]$.

$a\backslash n_\ell$	20	50	100	200	500	1000
0,5	14	6	14	4	6	3
1	17	9	8	3	3	7
1,5	12	8	9	12	6	4
2	13	10	8	3	6	5
2,5	12	14	8	5	8	7
3	12	8	4	4	6	5

TAB. 3.2 – Nombre de rejets de H_0 en fonction du nombre de données et de la séparation des classes pour le modèle $[\pi_k \lambda C]$.

Reprenons l'exemple des données « Crabes » et « Iris » pour illustrer l'utilisation de ce test, et fixons son niveau asymptotique à 5%. Le test permet effectivement de rejeter l'utilisation du modèle gaussien homoscédastique sur le jeu de données « Crabes » (89% de rejet). Pour le jeu de données « Iris », les performances obtenues sont à nuancer puisque le semi-supervisé est souvent rejeté alors même qu'il permet une amélioration des résultats (54% de rejet).

3.2.3 Étude du test proposé à partir de simulations

Une question souvent posée par la réalisation d'un test à un certain niveau asymptotique est celle du nombre de données nécessaires pour que ledit niveau soit effectivement celui du test effectué. Une autre question est celle de la puissance du test proposé sous certaines hypothèses alternatives. Nous étudions ici plus en détail le comportement du test proposé sur des données simulées.

Pour $n_\ell = n_u$ à valeurs dans $\{20; 50; 100; 200; 500; 1000\}$, on simule un mélange de deux composants : $\mathcal{N}((0,0)', \text{diag}(1,2))$ et $\mathcal{N}((a,0)', \text{diag}(1,2))$, en proportions identiques et pour a à valeurs dans $\{0,5; 1; 1,5; 2; 2,5; 3\}$. Les modèles postulés sont $[\pi_k \lambda B]$ et $[\pi_k \lambda C]$. Dans les deux tableaux 3.1 et 3.2 on présente le nombre de rejets de H_0 en fonction de n_ℓ et n_u pour les deux modèles postulés sur 100 essais. Il apparaît que le niveau se stabilise pour un nombre raisonnable de données. Le niveau nominal est plus rapidement atteint pour le modèle $[\pi_k \lambda B]$, on pouvait s'y attendre puisque ce modèle comporte moins de paramètres.

Maintenant si la variance du second composant est $\text{diag}(2,1)$ et en conservant tous les

$a\backslash n_\ell$	20	50	100	200	500	1000
0,5	17	10	8	9	8	10
1	15	14	9	14	54	89
1,5	11	14	24	34	79	100
2	11	17	26	47	90	99
2,5	22	16	27	51	90	100
3	13	14	20	38	83	99

TAB. 3.3 – Nombre de rejets de H_0 en fonction du nombre de données et de la séparation des classes pour le modèle $[\pi_k \lambda B]$.

$a\backslash n_\ell$	20	50	100	200	500	1000
0,5	12	6	4	8	5	6
1	15	12	10	23	47	83
1,5	11	14	21	31	79	100
2	11	13	22	49	87	99
2,5	23	14	25	50	88	100
3	11	14	20	36	81	99

TAB. 3.4 – Nombre de rejets de H_0 en fonction du nombre de données et de la séparation des classes pour le modèle $[\pi_k \lambda C]$.

autres paramètres inchangés (c'est dire qu'on simule les données sous H_1), on s'intéresse à la puissance du test lorsque les deux modèles précédents sont utilisés. Les résultats sont représentés tables 3.3 et 3.4. La puissance du test est d'autant plus grande que le nombre de données est élevé et que les classes sont bien séparées.

Dans la plupart des situations, ce test fonctionne correctement. Cependant, il nécessite de fixer arbitrairement un niveau. Dans ce qui suit on propose de reformuler la question précédente sous la forme d'un choix de modèle, ce qui évite d'avoir à fixer un niveau.

3.3 Réponse par un choix de modèle

3.3.1 Pour un seul modèle paramétrique

Modèles joint et disjoint

Considérons un unique modèle paramétrique m. On distingue le modèle joint M_1 qui associe les données étiquetées et non étiquetées pour estimer les paramètres

$$M_1 = \{\forall(\boldsymbol{x}, \boldsymbol{z}, s) \in \mathcal{X} \times \mathcal{Z} \times \mathcal{S}\ \exists\theta \in \Theta_m / p(\boldsymbol{x}, \boldsymbol{z}, s) = p(\boldsymbol{x}, \boldsymbol{z})p(s) = p_m(\boldsymbol{x}, \boldsymbol{z}; \theta)p(s; \beta)\}, \tag{3.9}$$

et on définit le modèle disjoint M_2 qui dissocie les données étiquetées et non étiquetées pour estimer les paramètres

$$\begin{aligned} M_2 \;=\; & \{\forall(\boldsymbol{x},\boldsymbol{z}) \in \mathcal{X}\times\mathcal{Z}\ \exists(\theta,\theta') \in \Theta_m^2 / p(\boldsymbol{x},\boldsymbol{z},s=1) = p_m(\boldsymbol{x},\boldsymbol{z};\theta)\beta \\ & \text{et } p(\boldsymbol{x},\boldsymbol{z},s=0) = p_m(\boldsymbol{x},\boldsymbol{z};\theta')(1-\beta)\}. \end{aligned} \tag{3.10}$$

Expression du critère BIC associé

Dans ce contexte, les critères AIC et BIC sont les outils de choix de modèle standards. Le critère AIC, par une approximation asymptotique de la déviance, recherche le modèle qui minimise la divergence de Kullback-Leibler par rapport à la distribution des données. Le critère BIC, par une approximation asymptotique de la vraisemblance intégrée, recherche le modèle le plus probable conditionnellement aux données.

Nous nous intéressons ici au critère BIC, qui posséde des propriétés de consistance contrairement au criètre AIC.

Soit $p(\theta)$ une distribution *a priori* sur θ, la vraisemblance intégrée pour le modèle M_1 est

$$\log \int_{\Theta_m} p(\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u; \theta) p(\theta) d\theta. \tag{3.11}$$

L'approximation BIC appliquée à cette vraisemblance intégrée donne

$$BIC(M_1) = \log p(\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}) - \frac{\nu_m}{2}\log n. \tag{3.12}$$

Soit $p(\theta,\theta')$ la distribution a priori sur (θ,θ'). De plus supposons que θ et θ' sont a priori indépendants. On écrit alors $p(\theta,\theta') = p(\theta)p(\theta')$. Le logarithme de la vraisemblance intégrée pour le modèle M_2 s'écrit

$$\log \int_{\Theta_m^2} p(\mathbf{x}_\ell, \mathbf{z}_\ell; \theta) p(\mathbf{x}_u; \theta') p(\theta) p(\theta') d\theta d\theta', \tag{3.13}$$

ce qu'on peut réécrire sous la forme

$$\log \int_{\Theta_m} p(\mathbf{x}_\ell, \mathbf{z}_\ell; \theta) p(\theta) d\theta + \log \int_{\Theta_m} p(\mathbf{x}_u; \theta') p(\theta') d\theta'. \tag{3.14}$$

Puis en appliquant l'approximation BIC à chacun des deux termes on obtient

$$BIC(M_2) = \log p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}) + \log(\mathbf{x}_u; \hat{\theta}_{\mathbf{x}_u}) - \frac{\nu_m}{2}\log n_\ell n_u. \tag{3.15}$$

Quand $BIC(M_1) > BIC(M_2)$ le cadre semi-supervisé est choisi. Le modèle proposé peut donc être utilisé avec confiance puisqu'il doit permettre une amélioration par rapport au supervisé.

Quand $BIC(M_2) > BIC(M_1)$, l'approche la plus constructive à notre sens et de proposer de nouveaux modèles jusqu'à ce qu'un modèle de type M_1 soit sélectionné.

Remarquons que par construction le modèle M_2 permet une meilleure adéquation aux données puisqu'il utilise plus de paramètres

$$\mathcal{L}(\hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell}; \mathbf{x}_\ell, \mathbf{z}_\ell) + \mathcal{L}(\hat{\theta}_{\mathbf{x}_u}; \mathbf{x}_u) \geqslant \mathcal{L}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u).$$

Toutefois, l'estimation d'un plus grand nombre de paramètres entraîne également l'augmentation de sa pénalité car on a en général

$$\frac{\nu_m}{2}\log n_\ell n_u > \frac{\nu_m}{2}\log n.$$

Le modèle sélectionné est celui qui réalise le meilleur compromis entre adéquation aux données et moindre complexité. Cela est cohérent avec l'approche semi-supervisée qui tire particulièrement bien partie de l'information apportée par les données non étiquetées quand les données étiquetées ne suffisent pas à apprendre de manière assez précise les paramètres du modèle.

Propriétés asymptotiques de BIC

Le critère BIC est consistant dans de nombreux cas. Quand la distribution d'échantillonnage est incluse dans le modèle m postulé, l'utilisation semi-supervisée des données (modèle M_1) est justifiée car elle réduit la variance des paramètres estimés. Intéressons nous à $2\left[BIC(M_2) - BIC(M_1)\right]$:

$$\begin{aligned} 2\left[BIC(M_2) - BIC(M_1)\right] &= -2\log\frac{p(\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u; \hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell})}{p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell})p(\mathbf{x}_u; \hat{\theta}_{\mathbf{x}_u})} - \nu_m \log\frac{n_\ell n_u}{n} \\ &= LRT_m - \nu_m \log\frac{n_\ell n_u}{n}. \end{aligned}$$

Sous les conditions de régularité précédentes, on a la loi asymptotique

$$LRT_m \approx \chi^2_{\nu_m}.$$

De plus, comme $n \to +\infty$ et que $\frac{n_\ell}{n} \xrightarrow{P} \beta$ et $\frac{n_u}{n} \xrightarrow{P} 1-\beta$ alors

$$\nu_m \log\frac{n_\ell n_u}{n} \to +\infty.$$

Ainsi

$$p(BIC(M_2) - BIC(M_1) < 0) \to 1$$

si la distribution d'échantillonnage est incluse dans le modèle m postulé. Ce critère permet donc de choisir, quand cela est justifié, l'utilisation des données non étiquetées.

Quand la distribution d'échantillonnage n'est pas incluse dans m, on peut se demander si le critère BIC rejette le modèle M_2. Pour cela on réécrit

$$2\left[BIC(M_2) - BIC(M_1)\right] = -2\left[\log\frac{p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})}{p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell})} + \log\frac{p(\mathbf{x}_u; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})}{p(\mathbf{x}_u; \hat{\theta}_{\mathbf{x}_u})}\right] - \nu_m \log\frac{n_\ell n_u}{n}.$$

Comme $\theta^*_{ss} \neq \theta^*_s$ et $\theta^*_{ss} \neq \theta^*_{ns}$, on a

$$\frac{1}{n}\log\frac{p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})}{p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x}_\ell,\mathbf{z}_\ell})} \to \beta\mathbb{E}\left[\log\frac{p(\boldsymbol{X}, \boldsymbol{Z}; \theta^*_{ss})}{p(\boldsymbol{X}, \boldsymbol{Z}; \theta^*_s)}\right] < 0,$$

et

$$\frac{1}{n}\log\frac{p(\mathbf{x}_u; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})}{p(\mathbf{x}_u; \hat{\theta}_{\mathbf{x}_u})} \to (1-\beta)\mathbb{E}\left[\log\frac{p(\boldsymbol{X}; \theta^*_{ss})}{p(\boldsymbol{X}; \theta^*_{ns})}\right] < 0,$$

alors

$$P(BIC(M_2) - BIC(M_1) > 0) \to 1.$$

Le modèle m ne doit donc pas être utilisé et de nouveaux modèles doivent être proposés.

$n_\ell \backslash n_u$	20	40	80	100	200	500	1000	5000
20	87	91	95	96	98	97	99	99
40	97	100	95	99	99	98	100	98
80	92	93	95	99	95	99	99	100
100	93	96	95	98	96	94	96	97
200	93	84	88	84	93	86	91	83
500	81	77	74	64	52	30	18	10
1000	82	74	57	62	21	0	0	0
5000	84	77	50	41	6	0	0	0

TAB. 3.5 – Nombre de fois où le modèle M_1 est choisi pour n_ℓ données étiquetées et n_u données non étiquetées sur 100 simulations.

Lien avec le test

Comme on l'a vu dans la démonstration des propriétés asymptotiques de BIC, la distribution asymptotique de la statistique de test occupe un rôle important.

Le niveau du test effectué est

$$\alpha = p\left(\chi^2_{\nu_m} > \nu_m \log \frac{n_\ell n_u}{n}\right).$$

L'utilisation du critère BIC revient à effectuer un test dont le niveau varie avec le nombre de données disponibles. Dans les zones où le test est peu puissant, c.-à-d. lorsque le nombre de données est petit, un niveau plus élevé est choisi. Lorsque le nombre de données est plus grand, la puissance du test est plus élevée, un niveau plus faible peut donc être choisi. Le choix des données non étiquetées à partir du critère BIC s'interprète alors comme un test adaptatif.

Expérimentations

Dans la plupart des cas, une meilleure approximation de la distribution jointe est susceptible d'améliorer la règle de classement. On vérifie maintenant sur des simulations si le critère BIC permet de retrouver de telles situations.

Soit un mélange de deux gaussiennes : $\mu_1 = \begin{pmatrix} 0 \\ 3/2 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 3/2 \\ 0 \end{pmatrix}$, $\Sigma_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 4/5 \\ 4/5 & 1 \end{pmatrix}$, $\pi_1 = 3/5$. Utilisons à tort le modèle $[\pi_k \lambda C]$. On cherche à voir pour différentes valeurs de n_ℓ et n_u, quel modèle entre M_1 et M_2 approche au mieux la distribution jointe.

Quand le nombre de données des deux types augmente, on voit table 3.5 qu'on refuse le modèle M_1. En revanche, quand le nombre de données est petit on préfère le modèle M_1. On met en évidence table 3.6 l'existence d'un compromis biais variance, en regardant au niveau de la divergence de Kullback.

En comparant les tableaux 3.5 et 3.6, on voit que le critère de choix de modèle permet de choisir efficacement les situations où compte tenu du nombre de données disponibles, il faut choisir à juste titre de conserver le modèle.

$n_\ell \backslash n_u$	20	40	80	100	200	500	1000	5000
20	76	79	69	60	58	60	48	49
40	86	79	64	74	64	56	59	60
80	90	87	71	77	68	54	56	40
100	83	78	74	62	69	64	53	58
200	83	77	71	69	57	56	62	49
500	90	68	39	37	30	27	33	37
1000	89	63	32	23	8	10	9	26
5000	88	50	11	8	0	0	0	0

TAB. 3.6 – Nombre de fois sur 100 simulations, où la distribution jointe estimée à partir du modèle M_1 est plus proche de la distribution jointe d'échantillonnage (au sens de Kullback) que de la distribution jointe estimée à partir du modèle M_2.

3.3.2 Élargissement adaptatif de la collection de modèles

En théorie, si les hypothèses d'échantillonnage sont vérifiées, et que le modèle postulé est correct, le semi-supervisé doit toujours permettre une amélioration des résultats. Ainsi dans le cas où le critère BIC refuse le modèle joint M_1, cela nous fournit l'indication que le modèle proposé m n'est pas suffisamment bien spécifié. Une stratégie constructive consiste alors à proposer de nouveaux modèles m jusqu'à ce que BIC ne le refuse plus.

Dans le cadre de l'analyse discriminante à base mélange (MDA) (section 2.3.3), une complexification naturelle des modèles consiste à proposer des modèles avec un nombre de composants par classe croissant :

- commencer par un composant par classe,
- tant que le modèle est jugé non pertinent, proposer un modèle avec un composant par classe supplémentaire.

Cette stratégie doit permettre de trouver des modèles où le semi-supervisé permet d'améliorer les résultats du supervisé.

Nous illustrons cette stratégie sur l'exemple des deux lunes enchevêtrées (figure 3.3). Sur les 1000 données étiquetées, on cache les étiquettes de 800 d'entre-elles. On ajuste une modèle de MDA où tous les composants ont même variance.

On trace figure 3.4 les critères BIC pour M_1 et M_2 (respectivement BIC joint et BIC disjoint) en fonction du nombre de composants par classe, et figure 3.5 les erreurs dans les cadres supervisé et semi-supervisé en fonction du nombre de composants par classe. En appliquant la stratégie proposée on s'arrêterait quand les courbes M_1 et M_2 s'intersectent, c.-à-d. à trois composants par classe. Ce qui correspond à un nombre de composants qui produit une faible erreur de classement et où le semi-supervisé est aussi bon que le supervisé. En choisissant le nombre de composants par classes avec BIC, on aurait sélectionné quatre composants par classe, ce qui permet une erreur de classement encore plus basse. Ce résultat était prévisible car l'approche proposée répond avant tout à la question « À partir de quelle complexité sur m devient-il intéressant d'utiliser le modèle joint M_1 ? », et non pas à la question « Quel est le meilleur modèle en semi-supervisé ? ». Cet exemple nous montre que le nombre de composants optimaux doit en pratique être légèrement supérieur à celui obtenu à l'intersection des deux courbes.

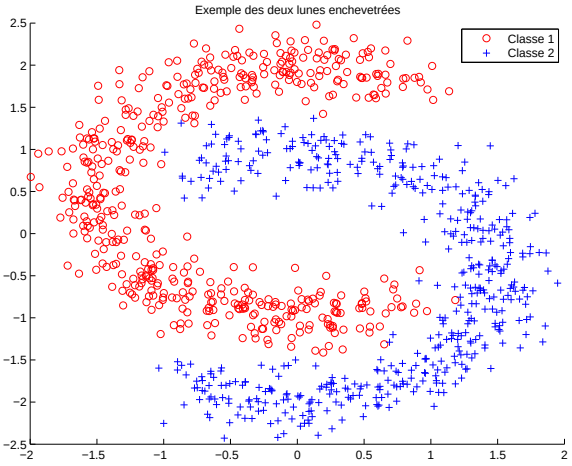

FIG. 3.3 – Exemple des deux lunes enchevêtrées.

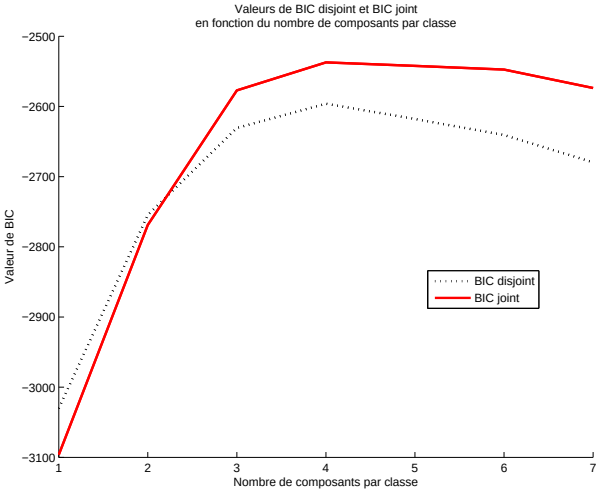

FIG. 3.4 – Critères BIC disjoint et BIC joint.

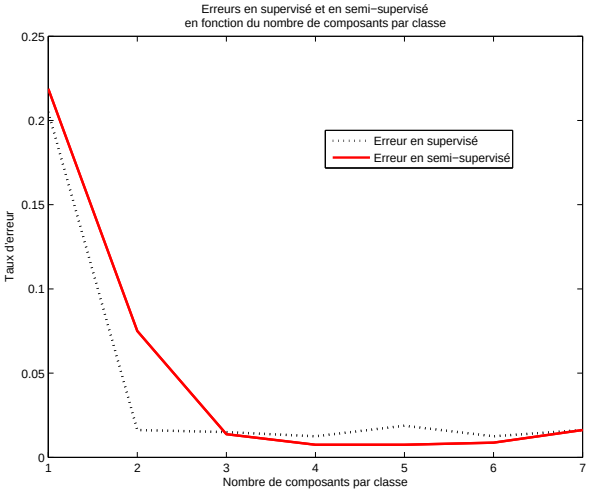

FIG. 3.5 – Taux d'erreur supervisé et semi-supervisé.

	Modèle M_1	Modèle M_2
$\pi\lambda I$	0	0
$\pi\lambda C$	0	0
$\pi\lambda C_k$	14	0
$\pi\lambda_k C_k$	80	0
$\pi_k\lambda I$	0	0
$\pi_k\lambda C$	0	0
$\pi_k\lambda C_k$	4	0
$\pi_k\lambda_k C_k$	1	1

TAB. 3.7 – Choix entre un modèle de type M_1 ou M_2.

La stratégie proposée fournit donc un nombre de composants par classe minimal noté K_{min} au-delà duquel le semi-supervisé est susceptible d'améliorer les résultats du supervisé. En pratique il sera alors raisonnable de fixer le nombre de composants par classe à K_{min}. Cette approche nous permet donc de trouver de façon adaptative le nombre de classes minimal à considérer sans avoir à le fixer à l'avance.

Comme en pratique, la valeur optimale pour K notée K^* est légérement supérieure à K_{min}. Une heuristique assez naturelle est de considérer les modèles pour $K \in \{K_{min}, K_{min}+1, \ldots, 2K_{min}\}$. Ce qui permettrait dans l'exemple précédent de trouver le nombre optimal de composants par classe.

3.3.3 Extension à plusieurs modèles

Principe

On considère maintenant la situation où plusieurs modèles $m_1, \ldots, m_H$ combinés aux cas joint/disjoint sont mis en compétition. L'objectif est de choisir à la fois le modèle et de vérifier le bien fondé du modèle choisi. Cela correspond à mettre les modèles $M_1^{m_1}, \ldots, M_1^{m_H}, M_2^{m_1}, \ldots, M_2^{m_H}$ en compétition via le critère BIC. Si un modèle m_h de type M_1 est choisi, alors le modèle m_h est pertinent. Dans le cas où c'est un modèle m_h de type M_2 qui est choisi, celui-ci ne fait pas sens puisque les estimations supervisées et non supervisées ne sont pas réconciliées. Il est alors nécessaire de proposer d'autres modèles jusqu'à ce qu'un modèle de type M_1 soit accepté.

Expériences sur des données simulées

Si le bon modèle est dans la liste. On génère un mélange de deux gaussiennes : $\mu_1 = \begin{pmatrix}0\\0\end{pmatrix}$, $\mu_2 = \begin{pmatrix}2\\0\end{pmatrix}$, $\Sigma_1 = \begin{pmatrix}1 & 7/8\\7/8 & 1\end{pmatrix}$, $\Sigma_2 = \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$, $\pi_1 = 1/2$. On génère 100 échantillons d'apprentissage avec $n_\ell = 20$ et $n_u = 100$. On représente table 3.7 le nombre de fois où chaque modèle de type M_1 et M_2 est choisi. On remarque ici la consistance de BIC, qui choisit un modèle de type M_1, qui plus est celui de la distribution d'échantillonnage.

	Modèle M_1	Modèle M_2
$\pi\lambda I$	0	0
$\pi\lambda C$	0	0
$\pi\lambda C_k$	0	0
$\pi\lambda_k C_k$	0	0
$\pi\lambda I$	0	0
$\pi_k\lambda C$	0	0
$\pi_k\lambda C_k$	0	90
$\pi_k\lambda_k C_k$	0	10

TAB. 3.8 – Choix entre un modèle de type M_1 ou M_2.

Si le bon modèle n'est pas dans la liste. Il s'agit d'un mélange de deux classes, où la première classe est constituée d'un mélange de deux gaussiennes et la seconde d'une seule : $\mu_{11} = \begin{pmatrix}2\\5\end{pmatrix}$, $\Sigma_{11} = \begin{pmatrix}1 & 0\\0 & 1/4\end{pmatrix}$, $\pi_{11} = 1/10$, $\mu_{12} = \begin{pmatrix}4\\0\end{pmatrix}$, $\Sigma_{12} = \begin{pmatrix}1/3 & 0\\0 & 3/2\end{pmatrix}$, $\pi_{12} = 4/10$, $\mu_2 = \begin{pmatrix}6\\0\end{pmatrix}$,$\Sigma_2 = \begin{pmatrix}1 & 0\\0 & 1/2\end{pmatrix}$. On simule 100 échantillons avec $n_\ell = 20$ et $n_u = 1000$. On représente la réalisation d'un échantillon figure 3.6. On voit que les estimations non supervisées et supervisées risquent d'être très différentes si un modèle à un seul composant par classe est utilisé. C'est en effet ce qui se passe en pratique, où sur les expériences réalisées, c'est toujours un modèle de type M_2 qui est choisi (voir table 3.8). Ce qui doit ensuite nous conduire à proposer des modèles de plus en plus complexes jusqu'à ce qu'un modèle de type M_1 soit choisi.

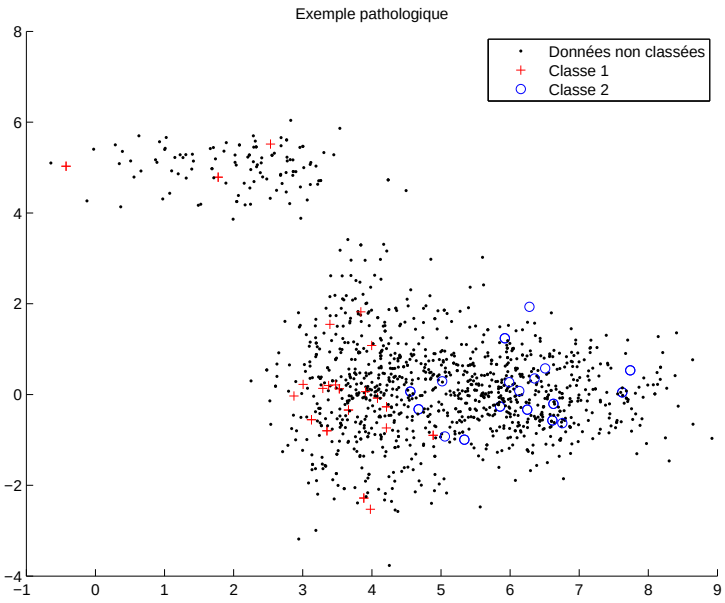

FIG. 3.6 – Cas où l'hétérogénéité des données reflète mal la classe.

3.4 Conclusion

Nous avons mis en place un test statistique permettant de détecter si les écarts entre estimations supervisée, non supervisée et semi-supervisée sont grands compte tenu des fluctuations d'échantillonnage et de la complexité du modèle. Cette indication nous permet de mettre à l'épreuve le modèle proposé et au besoin de proposer des modèles mieux adaptés. Cette approche a ensuite été reformulée en termes de choix de modèle via BIC et permet à la fois de choisir le modèle et de vérifier le bien fondé du modèle choisi.

Le critère BIC permet souvent de sélectionner un modèle pertinent. Cependant, en pratique il peut produire des résultats médiocres en classification puisqu'il ne prend pas en compte l'objectif décisionnel. Dans le chapitre 4 nous proposons un critère de choix de modèle bien adapté au cadre semi-supervisé et qui prend directement en compte cet objectif.

Chapitre 4

Sélection prédictive d'un modèle génératif

Dans ce chapitre nous nous focalisons sur le choix d'un modèle dans le cadre semi-supervisé avec un point de vue décisionnel. Nous rappelons tout d'abord les critères standards de choix de modèle tels que la validation croisée, AIC et BIC. Nous insistons tout particulièrement sur le critère BEC (Bouchard & Celeux, 2006) qui produit des résultats similaires à la validation croisée à un coût plus faible. Notre principale contribution dans ce chapitre est la proposition du critère AIC_{cond} critère dérivé du critère AIC d'un point de vue prédictif et s'interprétant ainsi comme une pénalisation particulière du critère BEC. Nous prouvons la convergence asymptotique de ce critère particulièrement bien adapté au contexte semi-supervisé et nous illustrons ses bonnes performances pratiques comparé aux critères de choix de modèle standards. Des extensions de ce critère au cadre supervisé seront aussi discutées.

4.1 Utilisation des critères standards en classification semi-supervisée

Dans ce qui suit nous détaillons les principaux critères de choix de modèle utilisés pour choisir un modèle génératif.

4.1.1 Validation croisée

Nous rappelons d'abord le principe de la validation croisée dans le cadre supervisé, puis discutons de son extension au cadre semi-supervisé.

4.1.2 Cadre supervisé

Dans le cadre décisionnel, l'objectif est de minimiser le taux d'erreur moyen

$$Err_s(m) = \mathbb{E}_{\mathbf{x}_\ell,\mathbf{z}_\ell}\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\mathbf{1}\{\delta(\boldsymbol{x};\hat{\theta}^m_{\mathbf{x}_\ell,\mathbf{z}_\ell}) \neq \boldsymbol{z}\}] \tag{4.1}$$

du classifieur $\delta(\cdot;\hat{\theta}^m_{\mathbf{x}_\ell,\mathbf{z}_\ell})$, où $\hat{\theta}^m_{\mathbf{x}_\ell,\mathbf{z}_\ell}$ provient de l'échantillon d'apprentissage $(\mathbf{x}_\ell,\mathbf{z}_\ell)$ en utilisant le modèle m, $\mathbf{1}$ est la fonction indicatrice et $(\boldsymbol{x},\boldsymbol{z})$ une nouvelle observation. Dans

tout ce chapitre, pour alléger les notations nous noterons de manière similaire les variables aléatoires et leur réalisation.

La validation croisée cherche à approcher au mieux $Err_s(m)$ à partir de l'échantillon disponible. Pour cela elle mime la double espérance présente dans l'équation (4.1) en séparant l'échantillon en deux parties, une partie d'apprentissage et une partie de test. Pour limiter la variabilité liée à la séparation aléatoire des données en un échantillon d'apprentissage et un échantillon test on utilise généralement la V-*fold* validation croisée dont le principe est le suivant :

- Couper au hasard l'échantillon des données étiquetées en V blocs $\{\mathcal{D}_\ell^{\{1\}}, \ldots, \mathcal{D}_\ell^{\{V\}}\}$ de tailles à peu près égales.
- **Pour** $i = 1$ **à** V
  $\hat{e}_i = \frac{1}{\text{card}(\mathcal{D}_\ell^{\{i\}})} \sum_{(\boldsymbol{x}_j, \boldsymbol{z}_j) \in \mathcal{D}_\ell^{\{i\}}} \mathbf{1}_{\{\delta(\boldsymbol{x}_j; \hat{\theta}^m_{\mathbf{x}_\ell^i, \mathbf{z}_\ell^i}) \neq \boldsymbol{z}_j\}}$ où $\hat{\theta}^m_{\mathbf{x}_\ell^i, \mathbf{z}_\ell^i}$ est l'estimateur du maximum de vraisemblance pour le modèle m calculé en utilisant toutes les données sauf le block $\mathcal{D}_\ell^{\{i\}}$, on note ces données $\mathbf{x}_\ell^i, \mathbf{z}_\ell^i$.
  fin
- Calculer $\widehat{Err}_s(m) = \frac{1}{V} \sum_{i=1}^V \hat{e}_i$.

4.1.3 Extension au cadre semi-supervisé

Approche naïve

En classification semi-supervisée on souhaite sélectionner le modèle qui minimise :

$$Err_{ss}(m) = \mathbb{E}_{\mathbf{x}, \mathbf{z}_\ell} \mathbb{E}_{\boldsymbol{x}, \boldsymbol{z}}[\mathbf{1}\{\delta(\boldsymbol{x}; \hat{\theta}^m_{\mathbf{x}, \mathbf{z}_\ell}) \neq \boldsymbol{z}\}] \tag{4.2}$$

où nous rappelons que $\mathbf{x} = (\mathbf{x}_\ell, \mathbf{x}_u)$. La différence avec 4.1 repose donc sur l'estimation de θ.

L'extension de la validation croisée au semi-supervisé n'est pas triviale. En effet, on peut se poser la question de l'intérêt du rééchantillonnage sur les données non étiquetées, puisque seules les données étiquetées servent à évaluer le taux d'erreur. Nous montrons que la stratégie qui consisterait à ne rééchantillonner que les données étiquetées conduit à une estimation biaisée du taux d'erreur.

En effet, notons

$$\theta^*_\beta = \underset{\theta \in \Theta}{\text{argmax}}\, \beta \mathbb{E}[\log p(\boldsymbol{x}, \boldsymbol{z}; \theta)] + (1 - \beta) \mathbb{E}[\log p(\boldsymbol{x}; \theta)],$$

ce qui correspond à la notation θ^*_{ss} du chapitre 3.

Les propriétés standards de l'estimateur du maximum de vraisemblance nous donnent

$$\hat{\theta}_{\mathbf{x}, \mathbf{z}_\ell} \xrightarrow{P} \theta^*_\beta.$$

Dans le cas où seule une partie des données étiquetées est extraite et que toutes les données non classées sont conservées on a

$$\hat{\theta}^m_{\mathbf{x}_\ell^i, \mathbf{z}_\ell^i, \mathbf{x}_u} \xrightarrow{P} \theta^*_{\frac{\beta(V-1)}{V-\beta}}.$$

La validation croisée n'estime donc pas correctement l'erreur de classification si $\theta^*_{\frac{\beta(V-1)}{V-\beta}} \neq \theta^*_\beta$, ce qui est généralement le cas si la distribution d'échantillonnage n'appartient pas au

modèle postulé (voir chapitre 3). Ici, pour estimer le taux d'erreur moyen correctement il faut enlever la même fraction de données étiquetées que de données non étiquetées de l'échantillon d'apprentissage lors du calcul de l'estimateur de θ. Les deux approches envisagées sont les suivantes :

- enlever une fraction $1/V$ des données non étiquetées,
- repondérer la log-vraisemblance des données non étiquetées par un facteur $1 - 1/V$.

Rééchantillonnage des données non classées

Nous pouvons étendre la validation croisée à la classification semi-supervisée en rééchantillonnant à la fois sur les données étiquetées et non étiquetées :

- Couper au hasard l'échantillon des données étiquetées et l'échantillon de données non étiquetées en V blocks respectivement $\{\mathcal{D}_\ell^{\{1\}}, \ldots, \mathcal{D}_\ell^{\{V\}}\}$ et $\{\mathcal{D}_u^{\{1\}}, \ldots, \mathcal{D}_u^{\{V\}}\}$ de tailles à peu prés égales.
- **Pour** $i = 1$ **à** V
  $\hat{e}_i = \frac{1}{\text{card}(\mathcal{D}_\ell^{\{i\}})} \sum_{(\boldsymbol{x}_j, \boldsymbol{z}_j) \in \mathcal{D}_\ell^{\{i\}}} \mathbf{1}\{\delta(\boldsymbol{x}_j; \hat{\theta}^m_{\mathbf{x}^i, \mathbf{z}^i_\ell}) \neq \boldsymbol{z}_j\}$ où $\hat{\theta}^m_{\mathbf{x}^i, \mathbf{z}^i_\ell}$ est l'estimateur du maximum de vraisemblance pour le modèle m calculé en utilisant $\mathcal{D} \backslash \{\mathcal{D}_\ell^{\{i\}}, \mathcal{D}_u^{\{i\}}\}$.
  fin
- Calculer $\widehat{Err}_{ss}(m) = \frac{1}{V} \sum_{i=1}^{V} \hat{e}_i$.

Remarquons que lorsque V augmente, la différence entre cette stratégie et celle qui consiste à ne retirer qu'une partie des données étiquetées diminue, le cas extrême étant le *leave-one-out*.

Repondération des données non étiquetées

Une alternative à l'extraction d'une fraction $1/V$ des données non étiquetées consiste à repondérer la log-vraisemblance des données non étiquetées par $1 - 1/V$, l'expression à optimiser étant alors

$$\log p(\mathbf{x}^i_\ell, \mathbf{z}^i_\ell; \theta) + \left(1 - \frac{1}{V}\right) \log p(\mathbf{x}_u; \theta). \tag{4.3}$$

Et conduisant à l'estimateur $\hat{\theta}^m_{\mathbf{x}^i_\ell, \mathbf{z}^i_\ell, \mathbf{x}_u^{1-1/V}}$. Dans cas on obtient encore

$$\hat{\theta}^m_{\mathbf{x}^i_\ell, \mathbf{z}^i_\ell, \mathbf{x}_u^{1-1/V}} \xrightarrow{P} \theta^*_\beta,$$

ce qui permet une estimation asymptotiquement sans biais du taux d'erreur. L'optimistation de l'expression (4.3) peut facilement être réalisée à l'aide de l'algorithme $\lambda-$EM décrit chapitre 2 section 2.2.6.

La fonction à optimiser est ici moins variable que pour le rééchantillonnage des données non étiquetées puisque la partie concernant les données non étiquetées reste la même pour chaque block. Cependant, comme l'objectif est de mimer la double espérance de l'équation (4.2), on recommande d'utiliser la première approche, la seconde revenant à travailler conditionnelle à $\mathbf{x}_u$ et ainsi à négliger la variabilité liée à l'échantillon de données non étiquetées.

Remarques sur l'extension de la validation croisée

La validation croisée produit de bons résultats en pratique. Cependant, comme remarqué dans Hastie *et al.* (2001), V est un paramètre à choisir avec attention pour obtenir une bonne estimation du taux d'erreur. Nous utiliserons le plus souvent la validation croisée avec $V = 3$ (CV3) et $V = 10$ (CV10). Remarquons que lorsque V augmente la différence entre les stratégies proposées et la stratégie consistant à ne retirer qu'une partie des données étiquetées diminue, le cas extrême étant le *leave-one-out*.

Dans le cadre semi-supervisé, un des principaux défauts de la validation croisée est son coût puisqu'elle nécessite d'avoir V fois recours à l'algorithme EM. Remarquons que même si $\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}$ représente une bonne initialisation de EM pour obtenir $\hat{\theta}^m_{\mathbf{x}^i,\mathbf{z}^i_\ell}$, suffisamment d'itérations doivent être effectuées pour limiter l'influence des données étiquetées cachées. Ce problème est absent en classification supervisée puisque l'estimation est souvent explicite. Remarquons qu'ici excepté pour n_ℓ petit la *leave-one-out* validation croisée n'est pas envisageable pour des raisons de temps de calcul.

Nous discutons alors des autres critères de choix de modèle envisagés.

4.1.4 Critère AIC

Cadre supervisé

Le critère AIC consiste en une approximation asymptotique de la déviance moyenne. La déviance moyenne du modèle m est :

$$\Delta = 2\mathbb{E}_{\mathbf{x}_\ell,\mathbf{z}_\ell,\mathbf{x}'_\ell,\mathbf{z}'_\ell}[\log p(\mathbf{x}'_\ell,\mathbf{z}'_\ell) - \log p(\mathbf{x}'_\ell,\mathbf{z}'_\ell;\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell})], \tag{4.4}$$

où $\mathbf{x},\mathbf{z}_\ell$ et $\mathbf{x}',\mathbf{z}'_\ell$ sont deux échantillons indépendants de même taille, et $\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}$ l'estimateur du maximum de vraisemblance de θ^m calculé à partir de l'échantillon d'apprentissage $\mathbf{x},\mathbf{z}_\ell$. Cette approximation asymptotique de la déviance donne sous des conditions de régularité standards :

$$AIC(m) = 2\log p(\mathbf{x}_\ell,\mathbf{z}_\ell;\hat{\theta}^m_{\mathbf{x}_\ell,\mathbf{z}_\ell}) - 2\nu_m, \tag{4.5}$$

où ν_m est le nombre paramètres à estimer pour le modèle m. Les conditions requises pour obtenir cette approximation sont les mêmes que celles nécessaires à l'obtention de la normalité asymptotique de l'estimateur du maximum de vraisemblance. Le critère AIC est V fois moins coûteux que la validation croisée, puisque le paramètre une fois estimé, le calcul du critère est direct. Les principales propriétés du critère AIC découlent des propriétés du test du rapport des vraisemblance maximales. En effet, si on considère deux modèles m et m' avec $m' \subset m$ on aura sous des conditions de régularité standards

$$AIC(m) - AIC(m') \xrightarrow{\mathcal{D}} \chi^2_{\nu_m-\nu'_m} - 2(\nu_m - \nu'_m). \tag{4.6}$$

On en déduit directement que pour des modèles corrects emboîtés, AIC peut sélectionner avec probabilité non nulle des modèles trop complexes, ceci même asymptotiquement. Autrement dit AIC n'est pas consistant. Remarquons toutefois que la probabilité de sélectionner un modèle trop complexe sera d'autant plus faible que la différence $\nu_m - \nu'_m$ sera grande. Notons d'autre part que dans le cadre de la régression AIC est minimax (Goldenshluger & Greenshtein, 2000).

Le critère AIC vise avant tout à minimiser (4.4), c'est-à-dire à choisir un modèle qui produit une bonne approximation de la distribution de $(\boldsymbol{x}, \boldsymbol{z})$ au sens de Kullback. Ainsi, dans le contexte de la classification, on s'attend à ce que AIC choisisse un modèle produisant un faible taux d'erreur si au moins un des modèles en compétition fournit une bonne approximation de la distribution de $(\boldsymbol{x}, \boldsymbol{z})$.

Extension au cadre semi-supervisé

L'extension du critère AIC au cadre semi-supervisé ne pose pas de difficulté particulière. En effet elle nécessite avant tout l'obtention de la normalité asymptotique de l'estimateur du maximum de vraisemblance, ce qui est le cas sous des conditions de régularité standards. On obtient alors le critère AIC suivant :

$$AIC(m) = 2 \log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}) - 2\nu_m, \tag{4.7}$$

Ses propriétés restent les mêmes que dans le cadre supervisé. Remarquons toutefois qu'ici le critère AIC ne recherche plus exactement la meilleure approximation possible de la distribution de $(\boldsymbol{x}, \boldsymbol{z})$, mais cherche plutôt un modèle qui minimise une combinaison convexe de la divergence de Kullback à la distribution jointe de $(\boldsymbol{x}, \boldsymbol{z})$ et de la divergence de Kullback à la distribution marginale de $\boldsymbol{x}$. De même qu'en supervisé, si au moins un des modèle en compétition fournit une bonne approximation de la distribution des données, on s'attend à ce que AIC sélectionne un modèle avec de bonnes performances en classification.

4.1.5 Critère BIC

Cadre supervisé

Comme nous l'avons déjà mentionné au chapitre 3, le critère BIC consiste en une approximation asymptotique du logarithme de la vraisemblance intégrée :

$$\log p(\mathbf{x}_\ell, \mathbf{z}_\ell; m) = \log \int_{\Theta^m} p(\mathbf{x}_\ell, \mathbf{z}_\ell; \theta^m) p(\theta^m) d\theta^m, \tag{4.8}$$

$p(\theta^m)$ étant la distribution *a priori* du paramètre θ^m qui est considéré aléatoire dans le cadre bayésien. Le critère s'écrit :

$$BIC(m) = \log p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}^m_{\mathbf{x}_\ell,\mathbf{z}_\ell}) - \frac{\nu_m}{2} \log n. \tag{4.9}$$

Ce critère, qui ne dépend pas de la distribution *a priori* $p(\theta^m)$ est consistant. Il cherche à sélectionner le modèle le plus probable conditionnellement aux données. Il sélectionne le vrai modèle s'il est dans la liste avec probabilité 1 à mesure que n tend vers l'infini sous des conditions de régularité standards. Comme pour AIC, aucun focus sur l'analyse discriminante n'est pris en compte et de bonnes performances en classification peuvent être attendues aussitôt qu'une bonne approximation de la distribution jointe est atteinte par au moins un modèle.

Extension au cadre semi-supervisé

L'extension du critère BIC au cadre semi-supervisé ne pose pas non plus de difficulté particulière puisqu'elle nécessite avant tout la normalité asymptotique de l'estimateur de paramètres (Lebarbier & Mary-Huard, 2006). On obtient

$$BIC(m) = \log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}) - \frac{\nu_m}{2} \log n. \tag{4.10}$$

De même que pour AIC le critère BIC reste peu coûteux en semi-supervisé, mais cherche également à minimiser asymptotiquement une divergence de Kullback et ne prend par conséquent pas en compte le point de vue décisionnel.

4.1.6 Critère BEC

Cadre supervisé

Le classifieur pour le modèle m est construit à partir de la distribution conditionnelle $p(\boldsymbol{z}|\boldsymbol{x}; \hat{\theta}^m_{\mathbf{x}_\ell,\mathbf{z}_\ell})$. Ainsi, pour sélectionner un modèle conduisant à un classifieur fiable il est nécessaire d'obtenir une bonne approximation de la distribution de $\boldsymbol{z}|\boldsymbol{x}$. Suivant cette idée dans une perspective bayesienne on souhaite sélectionner le modèle qui maximise $p(\mathbf{z}_\ell|\mathbf{x}_\ell, m)$. Partant de l'identité

$$\log p(\mathbf{z}_\ell|\mathbf{x}_\ell, m) = \log p(\mathbf{x}_\ell, \mathbf{z}_\ell|m) - \log p(\mathbf{x}_\ell|m),$$

et en effectuant l'approximation BIC pour chaque terme $\log p(\mathbf{x}_\ell, \mathbf{z}_\ell|m)$ et $\log p(\mathbf{x}_\ell|m)$, cela conduit à définir le critère d'entropie bayesienne (*Bayesian Entropy Criterion* : BEC) (Bouchard & Celeux, 2006) suivant

$$BEC(m) = \log p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}^m_{\mathbf{x}_\ell,\mathbf{z}_\ell}) - \log p(\mathbf{x}_\ell; \hat{\theta}^m_{\mathbf{x}_\ell}), \tag{4.11}$$

où $\hat{\theta}^m_{\mathbf{x}_\ell}$ est l'estimateur du maximum de vraisemblance de θ^m obtenu à partir de $\mathbf{x}_\ell$.

Ce critère sélectionne asymptotiquement le modèle maximisant le rapport $\frac{p(m|\mathbf{x}_\ell,\mathbf{z}_\ell)}{p(m|\mathbf{x}_\ell)}$. Ainsi, BEC sélectionne le modèle qui donne la plus grande probabilité aux étiquettes $\mathbf{z}_\ell$ sachant les covariables $\mathbf{x}_\ell$. D'un point de vue théorique, si la distribution d'échantillonnage appartient à un et un seul des modèles en compétition celui-ci sera asymptotiquement sélectionné (Bouchard & Celeux, 2006). Dans le cas de modèles corrects emboîtés, BEC peut sélectionner des modèles arbitrairement complexes. Pour éviter la sélection de modèles trop complexes, une règle simple mais fragile consiste à sélectionner le modèle le plus simple quand un plateau apparaît en traçant le critère BEC en fonction de la complexité du modèle (Bouchard & Celeux, 2006). BEC a montré un meilleur comportement que AIC et BIC sur de nombreux jeux de données et un comportement similaire à la validation croisée (Bouchard & Celeux, 2006).

Extension au cadre semi-supervisé

De même que pour BIC l'extension du critère BEC au semi-supervisé ne pose pas de difficulté particulière ; on obtient alors

$$BEC(m) = \log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}^m_{\mathbf{x}_\ell,\mathbf{z}_\ell}) - \log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}_\ell}). \tag{4.12}$$

Notons que dans le cadre semi-supervisé BEC ne réclame que le calcul supplémentaire de $\hat{\theta}^m_{\mathbf{x}}$ en utilisant l'algorithme EM, là où la *V-fold* validation croisée nécessite V fois l'utilisation de l'algorithme EM. Ainsi, l'utilisation de BEC est particulièrement appropriée dans le cadre semi-supervisé.

4.2 Proposition d'un critère spécifique : AIC$_{cond}$

Nous proposons maintenant un nouveau critère pour sélectionner un classifieur dans le cadre semi-supervisé. Ce critère vise à sélectionner un modèle avec de bonnes performances en classification. Plus précisément, il se focalise sur une bonne approximation de la distribution de la classe conditionnellement aux covariables. Nous montrerons que ce critère répond aux limitations de BEC.

4.2.1 Génèse et définition

Dans un cadre fréquentiste, une quantité d'intérêt pour sélectionner un modèle est sa déviance, ceci comme dans la construction du critère AIC. Pour sélectionner un modèle produisant un classifieur fiable, on considère la déviance de la vraisemblance du modèle associée à la distribution conditionnelle des étiquettes connaissant les covariables. On aimerait déterminer le modèle minimisant la divergence de Kullback-Leibler (KL) moyenne entre la vraisemblance conditionnelle de $\boldsymbol{z}|\boldsymbol{x}$ du modèle et la vraie distribution conditionnelle :

$$2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[\log p(\mathbf{z}'_\ell|\mathbf{x}') - \log p(\mathbf{z}'_\ell|\mathbf{x}';\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell})], \tag{4.13}$$

avec $(\mathbf{x},\mathbf{z}_\ell)$ et $(\mathbf{x}',\mathbf{z}'_\ell)$ deux échantillons indépendants et $\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}$ l'estimateur du maximum de vraisemblance de θ calculé à partir de l'échantillon d'apprentissage $(\mathbf{x},\mathbf{z}_\ell)$. Puisque le premier terme ne dépend pas du modèle, cela conduit à rechercher le modèle maximisant

$$E_{cond} = 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell} \log p(\mathbf{z}'_\ell|\mathbf{x}';\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}). \tag{4.14}$$

L'objectif est alors d'estimer cette quantité d'intérêt. On se place dorénavant sous l'hypothèse de la bonne famille de modèle c'est-à-dire que $p(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{x},\boldsymbol{z};\theta^{m*})$.

Résultat classique de convergence appliqué au semi-supervisé

Lemme 4.1.

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_\beta^{-1}) \tag{4.15}$$

PREUVE : Tout d'abord un développement de Taylor à l'ordre 1 au voisinage de θ^* donne :

$$\nabla \log p(\mathbf{x},\mathbf{z}_\ell;\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}) = 0 = \nabla \log p(\mathbf{x},\mathbf{z}_\ell;\theta^*) + \nabla^2 \log p(\mathbf{x},\mathbf{z}_\ell;\overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*) \tag{4.16}$$

où $\overline{\theta}$ est sur le segment joignant θ^* et $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$. Alors

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*) = \left[-\frac{1}{n}\nabla^2 \log p(\mathbf{x},\mathbf{z}_\ell;\overline{\theta})\right]^{-1} \frac{1}{\sqrt{n}} \nabla \log p(\mathbf{x},\mathbf{z}_\ell;\theta^*). \tag{4.17}$$

Par la loi des grands nombres et sous certaines conditions de régularité (Jennrich, 1969; Amemiya, 1973; White, 1982) $\left[-\frac{1}{n}\nabla^2 \log p(\mathbf{x}, \mathbf{z}_\ell; \overline{\theta})\right]^{-1} \xrightarrow{P} J_\beta^{-1}$. Le théorème central limite standard ne peut pas être appliqué directement sur $\frac{1}{\sqrt{n}}\nabla \log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*)$ puisque $\nabla \log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*)$ n'est pas une somme de variables i.i.d., cependant il peut être appliqué séparément sur les données classées et non classées ce qui conduit à :

$$\frac{1}{\sqrt{n}}\nabla \log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \beta K_c + (1-\beta)K), \tag{4.18}$$

avec $K_c = \mathbb{V}_{\boldsymbol{x},\boldsymbol{z}}[\nabla \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)]$ et $K = \mathbb{V}_{\boldsymbol{x}}[\nabla \log p(\boldsymbol{x}; \theta^*)]$. Maintenant sous l'hypothèse du bon modèle $K_c = J_c$ et $K = J$, ainsi

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_\beta^{-1}). \tag{4.19}$$

□

Première étape pour estimer E_{cond}

On s'appuiera sur le lemme 4.1

Proposition 4.1. *Si $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}, \boldsymbol{z}; \theta^{m*})$, et sous des conditions de régularité standards on obtient :*

$$E_{cond} \;=\; 2\mathbb{E}_{\mathbf{x},\mathbf{z}}[\log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}) - \log p(\mathbf{x}; \hat{\theta}^m_{\mathbf{x}})] - [\nu_m - \mathit{trace}(JJ_\beta^{-1})] + o(1), \tag{4.20}$$

où $\frac{n_\ell}{n} \xrightarrow{P} \beta$ quand $n \to +\infty$, J et J_β sont les matrices d'information de Fisher pour les données non étiquetées et partiellement étiquetées, $J_c = -\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\nabla^2 \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^{m*})]$, $J = -\mathbb{E}_{\boldsymbol{x}}[\nabla^2 \log p(\boldsymbol{x}; \theta^{m*})]$ et $J_\beta = \beta J_c + (1-\beta)J$.

PREUVE :

Pour simplifier on omettra de mentionner le modèle m dans les notations.

$\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$ sera l'estimateur de θ à partir de $(\mathbf{x}, \mathbf{z}_\ell)$ et $\hat{\theta}_{\mathbf{x}}$ l'estimateur de θ à partir de $\mathbf{x}$. Soit $\theta^* := \arg\max_{\theta\in\Theta} \beta\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\log p(\boldsymbol{x}, \boldsymbol{z}; \theta)] + (1-\beta)\mathbb{E}_{\boldsymbol{x}}[\log p(\boldsymbol{x}; \theta)]$. La consistance de l'estimation par maximum de vraisemblance implique $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} \xrightarrow{P} \theta^*$. Puisqu'on suppose que $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)$, on a alors $\hat{\theta}_{\mathbf{x}} \xrightarrow{P} \theta^*$ et $\mathbb{E}_{\boldsymbol{x}}[\nabla \log p(\boldsymbol{x}; \theta^*)] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\nabla \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)] = 0$. De plus, soit $K_c = \mathbb{V}_{\boldsymbol{x},\boldsymbol{z}}[\nabla \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)]$ et $K = \mathbb{V}_{\boldsymbol{x}}[\nabla \log p(\boldsymbol{x}; \theta^*)]$ alors $J = K$ et $J_c = K_c$.

On utilise maintenant les étapes suivantes :

$$\begin{aligned}
2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[\log p(\mathbf{x}', \mathbf{z}'_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*)] - \nu + o(1) && (4.21)\\
2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*)] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})] - \nu + o(1) && (4.22)\\
2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[\log p(\mathbf{x}'; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})] &= 2\mathbb{E}_{\boldsymbol{x}}[\log p(\boldsymbol{x}; \theta^*)] - \mathrm{trace}(JJ_\beta^{-1}) + o(1) && (4.23)\\
2\mathbb{E}_{\mathbf{x}}[\log p(\boldsymbol{x}; \theta^*)] &= 2\mathbb{E}_{\boldsymbol{x}}[\log p(\boldsymbol{x}; \hat{\theta}_{\mathbf{x}})] - \nu + o(1) && (4.24)
\end{aligned}$$

Les équations (4.22) et (4.24) découlent du même résultat.

On prouve d'abord l'équation (4.21). Par un développement de Taylor autour de θ^* :

$$\begin{aligned} 2\log p(\mathbf{x}', \mathbf{z}'_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}) &= 2\log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*) + 2(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)'\nabla_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*) && (4.25)\\ &\quad +(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)'\nabla^2_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*) && (4.26)\\ &= 2\log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*) + 2(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)'\nabla_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*) && (4.27)\\ &\quad +\mathrm{trace}[\nabla^2_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)'] && (4.28)\end{aligned}$$

avec $\overline{\theta}$ sur le segment joignant θ^* et $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$. En prenant l'espérance on obtient

$$\begin{aligned} 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[\log p(\mathbf{x}', \mathbf{z}'_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[\log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*)]\\ &\quad +2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)'\nabla_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*)]\\ &\quad +\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[\mathrm{trace}[\nabla^2_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)']].\end{aligned}$$

Puisque $(\mathbf{x}', \mathbf{z}'_\ell)$ est un réplicat indépendant de $(\mathbf{x}, \mathbf{z}_\ell)$,

$$\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[\log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*)] = \mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*)]$$

, et

$$\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)'\nabla_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*)] = \mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)']\mathbb{E}_{\mathbf{x}',\mathbf{z}'_\ell}[\nabla_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*)].$$

De plus $\mathbb{E}_{\mathbf{x}',\mathbf{z}'_\ell}[\nabla_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \theta^*)] = 0$ par définition de θ^*. Alors la loi des grands nombre nous donne $\frac{1}{n}\nabla^2_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \overline{\theta}) \xrightarrow{P} -J_\beta$. Ainsi en utilisant $\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J_\beta^{-1})$ on obtient $\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)' \xrightarrow{\mathcal{D}} \mathcal{W}_\nu(J_\beta^{-1}, 1)$. Puis à partir du lemme de Slustky (van der Vaart, 2000) on obtient

$$\nabla^2_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)' \xrightarrow{\mathcal{D}} -J_\beta \mathcal{W}_\nu(J_\beta^{-1}, 1). \tag{4.29}$$

Puis on déduit que

$$\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[\mathrm{trace}[\nabla^2_\theta \log p(\mathbf{x}', \mathbf{z}'_\ell; \overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)']] = -\nu + o(1) \tag{4.30}$$

ainsi l'équation (4.21) est prouvée.

On prouve maintenant l'équation (4.22). Un développement de Taylor de $\log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*)$ autour de $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$ nous donne

$$\begin{aligned} 2\log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*) &= 2\log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}) - 2(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)'\nabla_\theta \log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}) && (4.31)\\ &\quad +(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)'\nabla^2_\theta \log p(\mathbf{x}, \mathbf{z}_\ell; \overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*) && (4.32)\end{aligned}$$

où $\overline{\theta}$ est sur le segment joignant θ^* et $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$. On a $\nabla_\theta \log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}) = 0$. Puis, en prenant l'espérance

$$\begin{aligned} 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x}, \mathbf{z}_\ell; \theta^*)] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x}, \mathbf{z}_\ell; \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})] && (4.33)\\ &\quad +\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\mathrm{trace}[\nabla^2_\theta \log p(\mathbf{x}, \mathbf{z}_\ell; \overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)']] && (4.34)\end{aligned}$$

Ensuite, en utilisant les mêmes arguments que pour prouver l'équation (4.30),

$$\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\mathrm{trace}[\nabla^2_\theta \log p(\mathbf{x}, \mathbf{z}_\ell; \overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^*)']] = -\nu + o(1), \tag{4.35}$$

ce qui conclut la preuve de l'équation (4.22).

Nous pouvons maintenant prouver l'équation (4.23). Un développement de Taylor à l'ordre 2 autour de θ^* donne

$$\begin{aligned} 2\log p(\mathbf{x}';\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}) &= 2\log p(\mathbf{x}';\theta^*) + 2(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'\nabla_\theta \log p(\mathbf{x}';\theta^*) && (4.36)\\ &\quad +(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'\nabla^2_\theta \log p(\mathbf{x}';\overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*) && (4.37)\\ &= 2\log p(\mathbf{x};\theta^*) + 2(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'\nabla_\theta \log p(\mathbf{x}';\theta^*) && (4.38)\\ &\quad +\mathrm{trace}[\nabla^2_\theta \log p(\mathbf{x}';\overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'] && (4.39)\end{aligned}$$

avec $\overline{\theta}$ sur le segment joignant θ^* et $\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}$. Ainsi en prenant l'espérance

$$\begin{aligned} 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[\log p(\mathbf{x}';\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})] &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[\log p(\mathbf{x}';\theta^*)] + 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'\nabla_\theta \log p(\mathbf{x}';\theta^*)]\\ &\quad +\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'\nabla^2_\theta \log p(\mathbf{x}';\overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)]\\ &= 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[\log p(\mathbf{x}';\theta^*)] + 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'\nabla_\theta \log p(\mathbf{x}';\theta^*)]\\ &\quad +\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[\mathrm{trace}(\nabla^2_\theta \log p(\mathbf{x}';\overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)')].\end{aligned}$$

Remarquons que $\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[\log p(\mathbf{x}';\theta^*)] = \mathbb{E}_{\mathbf{x}}[\log p(\mathbf{x};\theta^*)]$ puisque $\mathbf{x}$ et $\mathbf{x}'$ ont la même distribution. Puisque $\mathbf{x}, \mathbf{z}_\ell$ et $\mathbf{x}'$ sont indépendants, on obtient

$$\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'\nabla_\theta \log p(\mathbf{x}';\theta^*)] = \mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)']\mathbb{E}_{\mathbf{x}'}[\nabla_\theta \log p(\mathbf{x}';\theta^*)],$$

puisque

$$\mathbb{E}_{\mathbf{x}'}[\nabla_\theta \log p(\mathbf{x}';\theta^*)] = 0$$

,

$$\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}'}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)'\nabla_\theta \log p(\mathbf{x}';\theta^*)] = 0.$$

Pour le troisième terme, en utilisant la loi des grands nombres on obtient $\frac{1}{n}\nabla^2_\theta \log p(\mathbf{x}';\overline{\theta}) \xrightarrow{P} -J$. Puis en utilisant le lemme 4.1, on a

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)' \xrightarrow{\mathcal{D}} \mathcal{W}_\nu(J_\beta^{-1},1). \quad (4.40)$$

Ainsi,

$$\nabla^2_\theta \log p(\mathbf{x}';\theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)' \xrightarrow{\mathcal{D}} -J\mathcal{W}_\nu(J_\beta^{-1},1). \quad (4.41)$$

Et par conséquence

$$\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell,\mathbf{x}',\mathbf{z}'_\ell}[\mathrm{trace}[\nabla^2_\theta \log p(\mathbf{x}';\overline{\theta})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}-\theta^*)']] = -\mathrm{trace}(JJ_\beta^{-1}) + o(1), \quad (4.42)$$

ce qui conclut la preuve de l'équation (4.23).

La preuve de l'équation (4.24) est la même que pour l'équation (4.22) puisque $\hat{\theta}_{\mathbf{x}} \xrightarrow{P} \theta^*$. □

Commentaires sur la pénalité obtenue

L'approximation de E_{cond} provenant de l'équation (4.20) est assez précise puisque le terme d'erreur est seulement en $o(1)$. Cela peut aussi s'écrire

$$E_{cond} = 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[BEC(m)] - [\nu - \mathrm{trace}(JJ_\beta^{-1})] + o(1), \quad (4.43)$$

ainsi on obtient deux fois l'espérance du critère BEC pénalisée par $(\nu - \mathrm{trace}(JJ_\beta^{-1}))$. On s'attend à ce que ce terme additionnel évite l'apparition d'un plateau quand on considère des modèles emboîtés, proposition que l'on discutera plus en détail dans la section 4.2.2. Décrivons maintenant son comportement en fonction de la séparation des classes. Quand les classes sont bien séparées, $J \approx J_c$ et par conséquent $J \approx J_\beta$ ce qui implique que $\nu - \mathrm{trace}(JJ_\beta^{-1}) \approx 0$. Au contraire, on s'attend à ce que $(\nu - \mathrm{trace}(JJ_\beta^{-1}))$ soit maximum quand les classes sont mal séparées. On montre par ailleurs que $(\nu - \mathrm{trace}(JJ_\beta^{-1}))$ est majorée par le nombre de paramètres algébriquement indépendants lorsque ces derniers sont obtenus en maximisant $\log p(\mathbf{z}|\mathbf{x}; \theta^m)$; c'est-à-dire lorsque les paramètres du modèle génératif sont obtenus d'un point de vue prédictif. Ce point sera discuté plus en avant dans la section 4.4.1.

Pour illustrer le comportement de $(\nu - \mathrm{trace}(JJ_\beta^{-1}))$ en fonction de la séparation des classes considérons l'exemple suivant. Supposons que les données ont été générées selon $\boldsymbol{X}|Z_1 = 1 \sim \mathcal{N}(0; 1)$, $\boldsymbol{X}|Z_2 = 1 \sim \mathcal{N}(\Delta; 1)$ et $\pi_1 = \pi_2 = 0,5$. Le modèle utilisé est un modèle gaussien hétéroscedastique, et on trace le comportement de la pénalité en fonction de la séparation des classes dans le cadre supervisé figure 4.1. On remarque que la pénalité est maximale si les classes ne sont pas séparées et qu'elle décroit vers 0 à mesure que la séparation des classes augmente. Remarquons que si $\Delta = 0$, la pénalité est égale à 3, ce qui correspond au nombre de paramètres impliqués dans la régression logistique quadratique. De plus remarquons que la régression logistique quadratique, est l'équivalent du modèle conditionnel résultant de l'analyse discriminante quadratique à une reparamétrisation identifiable près.

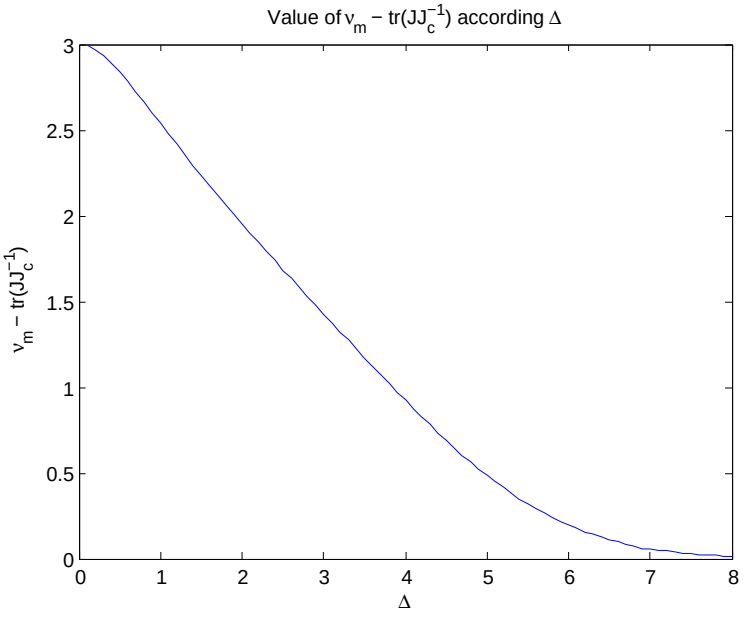

FIG. 4.1 – Valeur de la pénalité en fonction de la séparation des classes

Seconde étape pour estimer E_{cond} : estimation de la pénalité

La pénalité $(\nu - \text{trace}(JJ_\beta^{-1}))$ est difficile à calculer puisqu'elle nécessite le calcul des matrices d'information. La proposition 4.2 nous fournit alors une approximation de $(\nu - \text{trace}(JJ_\beta^{-1}))$.

Deux lemmes préliminaires

Lemme 4.2. *Soit* $\ell(\theta_1, \theta_2) = \log p(\mathbf{x}, \mathbf{z}_\ell; \theta_1) + \log p(\mathbf{x}; \theta_2)$.

$$\frac{1}{\sqrt{n}} \nabla \ell(\theta^*, \theta^*) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J_\beta & J \\ J & J \end{pmatrix}\right). \tag{4.44}$$

PREUVE : on a

$$\nabla \ell(\theta^*, \theta^*) = \sum_{i=1}^{n_\ell} \begin{pmatrix} \nabla \log p(\boldsymbol{x}_i, \boldsymbol{z}_i; \theta^*) \\ \nabla \log p(\boldsymbol{x}_i; \theta^*) \end{pmatrix} + \sum_{i=n_\ell+1}^{n} \begin{pmatrix} \nabla \log p(\boldsymbol{x}_i; \theta^*) \\ \nabla \log p(\boldsymbol{x}_i; \theta^*). \end{pmatrix} \tag{4.45}$$

En appliquant le théorème central limite sur les données étiquetées et non étiquetées, on obtient la relation (4.44).

On note $\nabla_{\theta(i)} g(\theta^*)$ la dérivée de $g(\theta)$ selon la $i$ème coordonnée de θ, évaluée en θ^*.

Nous avons $\mathbb{E}[\nabla_\theta \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)] = \mathbb{E}[\nabla_\theta \log p(\boldsymbol{x}; \theta^*)] = 0$.
Pour une donnée non classée : $\mathbb{E}[\nabla_{\theta(i)} \log p(\boldsymbol{x}; \theta^*) \nabla_{\theta(j)} \log p(\boldsymbol{x}; \theta^*)] = J_{ij}$.
Pour une donnée classée :

$$\mathbb{E}[\nabla_{\theta(i)} \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^*) \nabla_{\theta(j)} \log p(\boldsymbol{x}; \theta^*)] = \mathbb{E}_{\boldsymbol{x}}[\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[\nabla_{\theta(i)} \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)] \nabla_{\theta(j)} \log p(\boldsymbol{x}; \theta^*)] \tag{4.46}$$

$$\mathbb{E}_{\boldsymbol{z}|\boldsymbol{x}}[\nabla_{\theta(i)} \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^*)] = \sum_{k=1}^{g} \frac{\nabla_{\theta(i)} p(\boldsymbol{x}, z_k = 1; \theta^*)}{p(\boldsymbol{x}, z_k = 1; \theta^*)} p(z_k = 1|\boldsymbol{x}). \tag{4.47}$$

Puisque $p(z_k = 1|\boldsymbol{x}) = p(z_k = 1|\boldsymbol{x}; \theta^*)$, on a

$$\sum_{k=1}^{g} \frac{\nabla_{\theta(i)} p(\boldsymbol{x}, z_k = 1; \theta^*)}{p(\boldsymbol{x}, z_k = 1; \theta^*)} p(z_k = 1|\boldsymbol{x}) = \sum_{k=1}^{g} \frac{\nabla_{\theta(i)} p(\boldsymbol{x}, z_k = 1; \theta^*)}{p(\boldsymbol{x}; \theta^*)} = \frac{\nabla_{\theta(i)} p(\boldsymbol{x}; \theta^*)}{p(\boldsymbol{x}; \theta^*)} = \nabla_{\theta(i)} \log p(\boldsymbol{x}; \theta^*) \tag{4.48}$$

Ainsi

$$\mathbb{E}[\nabla_{\theta(i)} \log p(\boldsymbol{x}, \boldsymbol{z}; \theta^*) \nabla_{\theta(j)} \log p(\boldsymbol{x}; \theta^*)] = \mathbb{E}[\nabla_{\theta(i)} \log p(\boldsymbol{x}; \theta^*) \nabla_{\theta(j)} \log p(\boldsymbol{x}; \theta^*)] = J_{ij} \tag{4.49}$$

Le théorème central limite peut être utilisé sur les données étiquetées :

$$V_n = \sqrt{n_\ell} \left(\frac{1}{n_\ell} \sum_{i=1}^{n_\ell} \begin{pmatrix} \nabla \log p(\boldsymbol{x}_i, \boldsymbol{z}_i; \theta^*) \\ \nabla \log p(\boldsymbol{x}_i; \theta^*) \end{pmatrix} - 0 \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J_c & J \\ J & J \end{pmatrix}\right), \tag{4.50}$$

et il peut aussi être appliqué sur les données non étiquetées :

$$U_n = \sqrt{n - n_\ell} \left(\frac{1}{n - n_\ell} \sum_{i=n_\ell+1}^{n} \begin{pmatrix} \nabla \log p(\boldsymbol{x}_i; \theta^*) \\ \nabla \log p(\boldsymbol{x}_i; \theta^*) \end{pmatrix} - 0 \right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J & J \\ J & J \end{pmatrix}\right). \tag{4.51}$$

Ainsi en utilisant l'indépendance des données classées et non classées on obtient le résultat :

$$\frac{1}{\sqrt{n}}\nabla\ell(\theta^*,\theta^*) = \sqrt{\frac{n_\ell}{n}}V_n + \sqrt{\frac{n-n_\ell}{n}}U_n \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J_\beta & J \\ J & J \end{pmatrix}\right) \tag{4.52}$$

□

Nous pouvons maintenant prouver le lemme suivant

Lemme 4.3. $\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \hat{\theta}_{\mathbf{x}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J^{-1} - J_\beta^{-1})$

PREUVE : On a $\nabla\ell(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}, \hat{\theta}_{\mathbf{x}}) = 0$. Un développement de Taylor autour de θ^* donne

$$\nabla\ell(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}, \hat{\theta}_{\mathbf{x}}) = \nabla\ell(\theta^*,\theta^*) + \nabla^2\ell(\overline{\theta}_1,\overline{\theta}_2)\begin{pmatrix} \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^* \\ \hat{\theta}_{\mathbf{x}} - \theta^* \end{pmatrix} = 0 \tag{4.53}$$

avec $(\overline{\theta}_1, \overline{\theta}_2)$ sur le segment qui joint $(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell}, \hat{\theta}_{\mathbf{x}})$ et (θ^*,θ^*).

Ainsi

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^* \\ \hat{\theta}_{\mathbf{x}} - \theta^* \end{pmatrix} = \left[-\frac{1}{n}\nabla^2\ell(\overline{\theta}_1,\overline{\theta}_2)\right]^{-1}\frac{1}{\sqrt{n}}\nabla\ell(\theta^*,\theta^*). \tag{4.54}$$

Sous des conditions de régularités standards

$$\left[-\frac{1}{n}\nabla^2\ell(\overline{\theta}_1,\overline{\theta}_2)\right]^{-1} \xrightarrow{P} \begin{pmatrix} J_\beta^{-1} & 0 \\ 0 & J^{-1} \end{pmatrix}. \tag{4.55}$$

De plus en utilisant le lemme 4.2 on obtient

$$\sqrt{n}\begin{pmatrix} \hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \theta^* \\ \hat{\theta}_{\mathbf{x}} - \theta^* \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \begin{pmatrix} J_\beta^{-1} & J_\beta^{-1} \\ J_\beta^{-1} & J^{-1} \end{pmatrix}\right). \tag{4.56}$$

Enfin en utilisant le théorème de l'image continue

$$\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \hat{\theta}_{\mathbf{x}}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, J^{-1} - J_\beta^{-1}) \tag{4.57}$$

ce qui conclut la preuve du lemme 4.3. □

Proposition. Les deux lemmes précédents nous permettent de démontrer la proposition suivante.

Proposition 4.2. *si $p(\boldsymbol{x},\boldsymbol{z}) = p(\boldsymbol{x},\boldsymbol{z};\theta^{m*})$, et sous des conditions de régularité standards*

$$2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}}) - \log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell})] = [\nu - tr(JJ_\beta^{-1})] + o(1). \tag{4.58}$$

PREUVE : A partir d'un développement de Taylor à l'ordre deux autour de $\hat{\theta}_{\mathbf{x}}$,

$$\begin{aligned} 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}}) - \log p(\mathbf{x};\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})] &= -\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \hat{\theta}_{\mathbf{x}})'\nabla^2\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}})(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \hat{\theta}_{\mathbf{x}})] + o(1) \\ &= -\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}\left[\mathrm{trace}(\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \hat{\theta}_{\mathbf{x}})\sqrt{n}(\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell} - \hat{\theta}_{\mathbf{x}})'\frac{1}{n}\nabla^2\log p(\mathbf{x};\overline{\theta}))\right] \\ &\quad + o(1). \end{aligned}$$

Puis sous de conditions de régularité standards $-\frac{1}{n}\nabla^2\log p(\mathbf{x};\overline{\theta}) \xrightarrow{P} J$ et utilisant le lemme 4.3 on obtient

$$2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x};\hat{\theta}_{\mathbf{x}}) - \log p(\mathbf{x};\hat{\theta}_{\mathbf{x},\mathbf{z}_\ell})] = \mathrm{trace}(I - JJ_\beta^{-1}) + o(1) = \nu - \mathrm{trace}(JJ_\beta^{-1}) + o(1) \tag{4.59}$$

ce qui conclut la preuve. □

Dernière étape pour estimer E_{cond} : critère AIC_{cond}

En combinant les résultats des deux propositions précédentes, on obtient

$$E_{cond} = 2\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{z}_\ell|\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell})] - 4\mathbb{E}_{\mathbf{x},\mathbf{z}_\ell}[\log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}}) - \log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell})] + o(1). \quad (4.60)$$

Cette équation (4.60) est une approximation utile en pratique puisqu'elle ne n'implique que des quantités faciles à calculer. L'approximation de l'espérance dans E_{cond} pour une seule réalisation donne

$$E_{cond} = 2\log p(\mathbf{z}_\ell|\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}) - 4\log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}}) + 4\log p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}) + O_p(\sqrt{n}). \quad (4.61)$$

L'erreur d'approximation due aux fluctuation de l'échantillon est relativement élevée (en $O_p(\sqrt{n})$), cependant cette dernière est centrée en 0. Ainsi le critère proposé est :

$$\mathrm{AIC}_{cond}(m) = 2\log p(\mathbf{z}_\ell|\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell}) - 4\log \frac{p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}})}{p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell})}, \quad (4.62)$$

$$= 2BEC(m) - 2\log \frac{p(\mathbf{x};\hat{\theta}^m_{\mathbf{x}})}{p(\mathbf{x};\hat{\theta}^m_{\mathbf{x},\mathbf{z}_\ell})}. \quad (4.63)$$

4.2.2 Propriétés d'AIC_{cond}

Le critère AIC_{cond} peut être interprété comme un critère BEC surpénalisé. La pénalité additionnelle permet d'éviter l'apparition d'un plateau quand des modèles emboîtés sont considérés comme cela est montré dans les propriétés suivantes :

AIC_{cond} évite l'apparition d'un plateau

La preuve s'appuie sur un lemme suivi par la proposition principale.

Lemme 4.4. *Soit A et B deux matrices réelles de dimension $(d+p)$ symétriques définies positives. Soit $A = \begin{pmatrix} a_{11} & a_{12} \\ a'_{12} & a_{22} \end{pmatrix}$ et $B = \begin{pmatrix} b_{11} & b_{12} \\ b'_{12} & b_{22} \end{pmatrix}$ où a_{11} et b_{11} sont dans $\mathbb{R}^{p\times p}$, a_{12} et b_{12} sont dans $\mathbb{R}^{p\times d}$ et a_{22} et b_{22} sont des matrices symétriques de taille d. Alors $trace(BA^{-1}) > trace(b_{22}a_{22}^{-1})$.*

PREUVE : Commençons par prouver ce lemme pour $p = 1$. On peut montrer que comme A est symétrique et défini positif, a_{22} et $a_{11} - a_{12}a_{22}^{-1}a'_{12}$ sont des matrices symétriques et définies positives, ainsi l'inverse de A peut s'écrire

$$A^{-1} = \begin{pmatrix} c & cf' \\ cf & a_{22}^{-1} + cff' \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & a_{22}^{-1} \end{pmatrix} + c\begin{pmatrix} 1 \\ f \end{pmatrix}\begin{pmatrix} 1 & f' \end{pmatrix} \quad (4.64)$$

où $c = (a_{11} - a_{12}a_{22}^{-1}a'_{12})^{-1} \in \mathbb{R}$ et $f = -a_{22}^{-1}a'_{12}$.

On a

$$BA^{-1} = \begin{pmatrix} 0 & b_{12}a_{22}^{-1} \\ 0 & b_{22}a_{22}^{-1} \end{pmatrix} + cB\begin{pmatrix} 1 \\ f \end{pmatrix}\begin{pmatrix} 1 & f' \end{pmatrix}, \quad (4.65)$$

et donc

$$\mathrm{trace}(BA^{-1}) = \mathrm{trace}(b_{22}a_{22}^{-1}) + c\begin{pmatrix} 1 & f' \end{pmatrix} B\begin{pmatrix} 1 \\ f \end{pmatrix}. \quad (4.66)$$

Comme B est définie positive, $\begin{pmatrix}1 & f'\end{pmatrix} B \begin{pmatrix}1\\ f\end{pmatrix} > 0$, cela conclut la preuve du lemme pour $p = 1$. Pour obtenir le résultat pour p quelconque, il suffit d'appliquer le résultat précédent p fois. □

Nous montrons maintenant que AIC_{cond} préfère en moyenne le modèle le moins complexe quand deux modèles corrects emboîtés sont en compétition. Ainsi, contrairement à BEC, AIC_{cond} devrait éviter l'apparition d'un plateau.

Proposition 4.3. *Supposons que la distribution des données est incluse dans les modèles m et m' avec $m \subset m'$. Si le nombre de données est suffisamment grand :*

$$\mathbb{E}[AIC_{cond}(m) - AIC_{cond}(m')] > 0. \tag{4.67}$$

PREUVE : Supposons que la distribution d'échantillonnage appartienne à deux modèles emboîtés m et m' avec $m \subset m'$. En utilisant la proposition 4.2 :

$$\begin{aligned}\mathbb{E}[\text{AIC}_{cond}(m)] - \mathbb{E}[\text{AIC}_{cond}(m')] &= 2\mathbb{E}[BEC(m) - BEC(m')]\\ &\quad -[\nu_m - \text{trace}(J(m)J_\beta(m)^{-1})]\\ &\quad +[\nu_{m'} - \text{trace}(J(m')J_\beta(m')^{-1})] + o(1).\end{aligned}$$

Puisque $\mathbb{E}[BEC(m) - BEC(m')] \to 0$ (cf. Bouchard & Celeux (2006)), ainsi il nous suffit de prouver

$$-[\nu_m - \text{trace}(J(m)J_\beta(m)^{-1})] + [\nu_{m'} - \text{trace}(J(m')J_\beta(m')^{-1})] > 0.$$

En écrivant

$$[\nu_m - \text{trace}(J(m)J_\beta(m)^{-1})] = \text{trace}[(J_\beta(m) - J(m))J_\beta(m)^{-1}]$$

et

$$[\nu_{m'} - \text{trace}(J(m')J_\beta(m')^{-1})] = \text{trace}[(J_\beta(m') - J(m'))J_\beta(m')^{-1}].$$

On utilise alors la décomposition suivante : $J(m') = \begin{pmatrix}J^{11}(m') & J^{12}(m')\\ J^{21}(m') & J(m)\end{pmatrix}$ et on a aussi : $J_\beta(m') = \begin{pmatrix}J_\beta^{11}(m') & J_\beta^{12}(m')\\ J_\beta^{21}(m') & J_\beta(m)\end{pmatrix}$. Il est important de remarquer que $J_\beta(m) - J(m)$ et $J_\beta(m') - J(m')$ sont des matrices symétriques définies positives.

En prenant $A = J_\beta(m')$, $B = J_\beta(m') - J(m')$, $a_{22} = J_\beta(m)$, $b_{22} = J_\beta(m) - J(m)$, et en appliquant le lemme 4.4, on conclut que $-[\nu(m) - \text{trace}(J(m)J_\beta(m)^{-1})] + [\nu(m') - \text{trace}(J(m')J_\beta(m')^{-1})] > 0$, et par conséquent que pour n assez grand :

$$\mathbb{E}[\text{AIC}_{cond}(m)] - \mathbb{E}[\text{AIC}_{cond}(m')] > 0 \tag{4.68}$$

ce qui conclut la preuve. □

AIC_{cond} choisit un classifieur fiable

Tout comme BEC, AIC_{cond} sélectionne le vrai modèle quand ce dernier est unique. Si aucun modèle de la collection n'est vrai, il n'y a pas de garantie de sélectionner le meilleur classifieur. Cependant, puisque que l'objectif d'analyse discriminante a été pris en compte dans la construction d'AIC_{cond}, on peut s'attendre à ce qu'il choisisse une classifieur fiable. C'est ce qui nous montrons maintenant.

Proposition 4.4. *Si la distribution d'échantillonnage appartient à un seul des modèles en compétition m^* dans un collection finie de modèles $\{m_1, \ldots, m_M\}$, et sous des conditions standards sur la famille paramétrique, alors AIC_{cond} sélectionne m^* ou un modèle m' tel que :*

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\log p(\boldsymbol{z}|\boldsymbol{x}; \theta_0^{m^*})] = \mathbb{E}_{\boldsymbol{x},\boldsymbol{z}}[\log p(\boldsymbol{z}|\boldsymbol{x}; \theta_0^{m'})], \tag{4.69}$$

La preuve de cette propriété est analogue à la preuve pour BEC donnée dans Bouchard & Celeux (2006).

4.3 Évaluation numérique d'AIC$_{cond}$

Dans cette section nous évaluons les performances d'AIC$_{cond}$. Nous nous intéresserons d'abord à des expériences sur données simulées dans des contextes variés. Nous étudierons ensuite le comportement des différents critères sur des jeux de données réels.

4.3.1 Expériences sur données simulées

Sélection de variables

Dans de nombreuses applications d'apprentissage statistique, le nombre de variables est grand devant le nombre de données. Il est alors souvent souhaitable de sélectionner un sous échantillon de variables pertinentes du point de vue de la classification. Le problème de la sélection de variables peut être vu comme un problème de choix de modèle (Raftery & Dean, 2006; Maugis *et al.* , 2008; Murphy *et al.* , 2008). Le fait de sélectionner les variables à partir de la vraisemblance conditionnelle permet de sélectionner les variables en supposant que les variables non prises en compte sont indépendantes de la classe. Cette approche a l'avantage de ne pas nécessiter de faire d'hypothèses sur la distribution des variables non prises en compte. Ici nous décidons de ne pas utiliser AIC et BIC puisqu'ils nécessitent de modéliser la distribution des variables non conservées. Pour comparer les performances de CV3, CV10, BEC et AIC$_{cond}$, on a simulé les données selon un schéma où les variables apportent de moins en moins d'information pour finalement dégrader la précision de la règle de classement.

On génère deux classes $g = 2$ avec $P(z_1 = 1) = P(z_2 = 1) = 0,5$, les distributions conditionnellement à la classe sont gaussiennes, $\boldsymbol{x}|z_1 = 1 \sim \mathcal{N}(0_{50\times 1}, I_{50})$ et $\boldsymbol{x}|z_2 = 1 \sim \mathcal{N}(\mu, I_{50})$ avec $\mu_i = \frac{1}{i}\ \forall i \in \{1, \ldots, 50\}$ et I_d la matrice identité de rang d. Ainsi, les variables fournissent de moins en moins d'information sur la classification. L'ordre selon lequel les variables sont sélectionnées (de 1 à 50) est supposé connu. Le vrai modèle consiste à sélectionner toutes les variables. Cependant, les variables les moins informatives augmentent de façon dramatique la variance de la règle de classement apprise, et par suite l'erreur moyenne du classifieur résultant. Un échantillon test de 50000 données a été généré. Quatre combinaisons de n_ℓ données étiquetées et n_u données non étiquetées ont été considérées : $S_1 : n_\ell = 100$, $n_u = 0$; $S_2 : n_\ell = 1000$, $n_u = 0$; $SS_1 : n_\ell = 100$, $n_u = 1000$; $SS_2 : n_\ell = 1000$, $n_u = 10000$. Chaque combinaison a été répétée 100 fois.

Les taux d'erreur optimaux, réels et apparents en fonction du nombre de variables sélectionnées sont représentés pour SS_1 sur la figure 4.2. Les taux d'erreur optimaux et apparents décroissent quand le nombre de variables sélectionnées augmente, tandis le taux

d'erreur optimal estimé à partir de l'échantillon test diminue puis augmente. Soit NbVar* le nombre optimal de variables obtenu à partir du taux d'erreur réel et Err* son taux d'erreur.

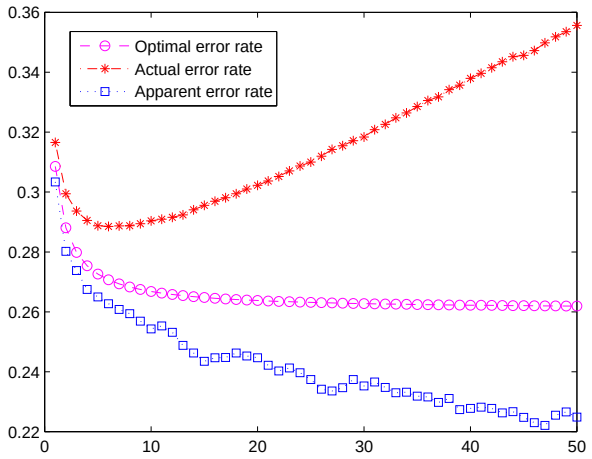

FIG. 4.2 – SS_1 **:** Taux d'erreur en fonction du nombre de variables sélectionnées.

	BEC	AIC$_{cond}$	CV3	CV10	NbVar*
S_1	10,5	**3,1**	7,8	7,8	3
S_2	21,7	**11,3**	12,2	14,0	11
SS_1	17,5	**9,2**	10,7	10,0	6
SS_2	33,8	**22,0**	21,1	21,4	23

TAB. 4.1 – Nombre moyen de variables sélectionnées par chaque critère (meilleur critère mis en gras).

	BEC	AIC$_{cond}$	CV3	CV10	Err*
S_1	31,53	**30,40**	31,08	31,08	29,68
S_2	27,90	**27,68**	27,77	27,78	27,55
SS_1	30,42	29,75	**29,70**	29,82	28,55
SS_2	27,18	**27,17**	**27,17**	27,21	27,03

TAB. 4.2 – Taux d'erreur pour les différents critères (meilleur critère mis en gras).

Les tableaux 4.1 et 4.2 montrent que AIC$_{cond}$ produit les meilleurs résultats puisqu'il sélectionne en moyenne le nombre de variables le plus adéquat et produit un taux d'erreur faible à la fois dans les configurations supervisées et semi-supervisées. La validation croisée produit elle aussi de bons résultats dans les différentes configurations. BEC produit des résultats peu satisfaisants puisqu'il sélectionne en moyenne trop de variables.

Ceci montre que sous l'hypothèse du bon modèle, AIC$_{cond}$ permet de trouver un bon compromis entre l'adéquation aux données et les performances en généralisation.

Choix d'un modèle en MDA

Cas ou le vrai modèle est dans la liste L'analyse discriminante à base de mélange (MDA) permet de s'adapter à de nombreuses situations (Hastie & Tibshirani, 1996). Comme mentionné section 2.3.3 un des paramètres délicat à choisir dans la MDA est le nombre de composants par classe. En classification semi-supervisée, de nombreuses données non étiquetés permettent d'estimer correctement la distribution marginale de $\boldsymbol{x}$ comme déjà remarqué dans Miller & Browning (2003). Des expériences numériques ont été réalisées pour comparer les comportement de BEC et AIC_{cond} en vue de choisir le nombre de composants. Ces résultats montrent que AIC_{cond} a un bon comportement et conduit souvent à des modèles plus parcimonieux que BEC sans nécessiter l'utilisation de la règle subjective du plateau. Pour éviter les problèmes combinatoires nous supposerons que la distribution conditionnellement à chaque classe est issue du même nombre de composants.

Exemple jouet : Ici on montre que AIC_{cond} ne sous-estime et ne surestime pas le nombre de composants. Un problème de classification à deux classes avec trois composants par classe est considéré. Pour la première classe on a

$$\begin{aligned}(\boldsymbol{X}, Z_1 = 1) \sim & \frac{1}{6}\mathcal{N}((0,0,0,0,0,0)', 0.15I_6) + \frac{1}{6}\mathcal{N}((1,1,0,0,0,0)', 0.15I_6) \quad (4.70)\\ & + \frac{1}{6}\mathcal{N}((2,0,0,0,0,0)', 0.15I_6). \quad (4.71)\end{aligned}$$

Pour la seconde classe on a

$$\begin{aligned}(\boldsymbol{X}, Z_2 = 1) \sim & \frac{1}{6}\mathcal{N}((1,0,0,0,0,0)', 0.15I_6) + \frac{1}{6}\mathcal{N}((2,-1,0,0,0,0)', 0.15I_6) \quad (4.72)\\ & + \frac{1}{6}\mathcal{N}((3,0,0,0,0,0)', 0.15I_6). \quad (4.73)\\ & \quad (4.74)\end{aligned}$$

Les isodensités des composants sont représentées figure 4.3.

Cent échantillons avec $n_\ell = 100$ et $n_u = 1000$ et un échantillon test de taille 50000 ont été générés. Les modèles mis en compétition sont de modèles gaussiens hétéroscedastiques diagonaux, dont le nombre de composants varie de 1 à 5 par classe. Le nombre de fois où chaque modèle a été choisi par BIC, AIC, BEC, AIC_{cond}, CV3 et CV10 est reporté dans le tableau 4.3 qui fournit aussi le taux d'erreur moyen produit par chaque modèle.

CC	AIC	BIC	AIC_{cond}	BEC	CV10	CV3	$\overline{Err}$
1	0	0	0	0	0	0	26,22
2	0	0	0	0	3	2	22,42
3*	64	97	53	32	45	42	15,61
4	15	3	27	31	25	31	16,07
5	21	0	20	37	27	25	16,62
$\overline{Err}$	15,97	15,80	15,93	16,08	16,07	16,07	15,60

TAB. 4.3 – Nombre de composants par classe (CC) sélectionné par chaque critère.

Dans cette situation, comme suggéré figure 4.3 les modèles à un et deux composants par classe ne suffisent pas à obtenir une erreur de classement faible. Le modèle à trois composants par classe produit bien évidemment le taux d'erreur le plus bas, puis pour

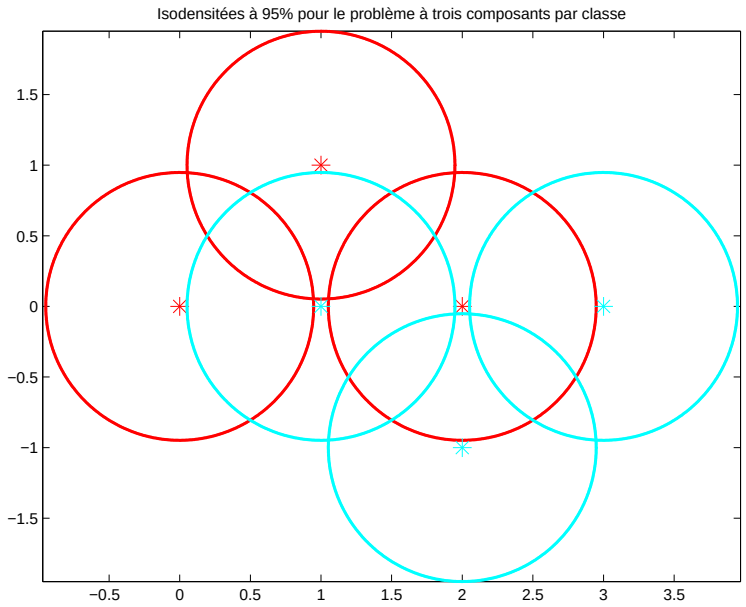

FIG. 4.3 – Configuration expérimentale.

quatre et cinq composants par classe, l'erreur augmente en raison de l'augmentation de la variance dans l'estimation des paramètres.

Cette expérience illustre la consistance de BIC, l'étalement de BEC, ainsi que la correction de l'étalement de BEC par AIC$_{cond}$.

Cas où le vrai modèle n'est pas dans la liste On génère deux classes de la façon suivante :

$$(\boldsymbol{x}, z_1 = 1) \sim 0.5\mathcal{N}((0,0)', (1, 0.5; 0.5, 2)) \text{ et } (\boldsymbol{x}, z_2 = 1) \sim 0.5\mathcal{N}((2,0)', (2, 0.5; 0.5, 1)).$$

On génère 100 échantillons où $n_\ell = 50$ et $n_u = 500$. On génère un échantillon test de taille $n_{test} = 50000$. Le modèle utilisé est un modèle diagonal homoscédastique, l'enjeu est de choisir le nombre de composants par classe le plus approprié possible. Les résultats sont présentés table 4.3.1.

Cet exemple montre l'avantage de BEC, de AIC$_{cond}$ et de la validation croisée quand le modèle postulé n'est pas dans la liste de modèles.

Choix de la forme de la matrice de covariance

Soit

$$(\boldsymbol{x}, z_1 = 1) \sim 0.5\mathcal{N}((2,0,0,0,0,0)', 2\text{diag}(2, 1.5, 1, 1, 1, 1))$$

$$(\boldsymbol{x}, z_2 = 2) \sim 0.5\mathcal{N}((0,0,0,0,0,0)', \text{diag}(2, 1.5, 1, 1, 1, 1))$$

On génère 100 échantillons avec $n_u = 2000$ et $n_\ell = 200$, et un échantillon test avec $n_{test} = 50000$. Les modèles mis en compétition sont les modèles $[\lambda I]$, $[\lambda B]$, $[\lambda C]$, $[\lambda_k I]$, $[\lambda_k B_k]$ et $[\lambda_k C_k]$. Les résultats sont exposés table 4.5. Ces résultats montrent l'avantage

	BIC	AIC	BEC	AIC_{cond}	CV3	CV10	$\overline{Err}$
1	32	4	0	0	5	1	24,52
2	40	16	5	9	12	16	21,18
3	18	11	5	8	12	10	19,80
4	9	12	17	19	15	15	20,20
5	1	7	21	18	18	15	19,57
6	0	9	9	8	7	10	20,55
7	0	14	14	13	10	8	20,60
8	0	13	9	9	10	10	21,14
9	0	9	11	9	7	10	21,51
10	0	5	9	7	4	5	22,14
Err^*	22,06	21,00	19,52	19,58	19,56	19,83	16,75

TAB. 4.4 – Nombre de composants par classe sélectionné par les différents critères.

	BIC	AIC	BEC	AIC_{cond}	CV3	CV10	$\overline{Err}$
λI	0	0	0	0	1	0	27,49
$\lambda_k I$	0	0	1	1	98	41	22,97
λB	0	0	0	0	0	0	27,80
$\lambda_k B$	100	98	49	62	1	34	20,60
λC	0	0	0	0	0	0	28,34
$\lambda_k C$	0	2	50	37	0	25	20,66
Err^*	20,60	20,60	20,67	20,66	23,00	21,61	20,58

TAB. 4.5 – Paramètrisation de la matrice de covariance choisie par les différents critères.

des critères d'information sur la validation croisée. Ici, on voit la consistance de BIC et le bon comportement de AIC, ainsi que les bonnes performances de AIC_{cond} et BEC. Remarquons qu'ici AIC_{cond} a tendance à sélectionner des modèles moins complexes que BEC.

4.3.2 Expériences sur données réelles

Dans cette section on étudie le comportement des différents critères sur des données de la base de données UCI[1], des jeux de données du livre *Pattern Recognition*[2], ainsi que sur des jeux de données sur des oiseaux utilisés dans Biernacki *et al.* (2002).

Choix de la paramétrisation de la matrice de covariance

Les performances des critères pour sélectionner parmi un des six modèles parcimonieux précédents est étudié sur des jeux de données de l'UCI et du livre *Pattern Recognition*. Les caractéristiques de ces jeux de données sont résumés tableau 4.6. Si un échantillon

[1]http ://archive.ics.uci.edu/ml/
[2]http ://www.stats.ox.ac.uk/pub/PRNN/

test est fourni, alors les covariables dans cet échantillon sont utilisées pour apprendre les paramètres dans un cadre semi-supervisé, puis les étiquettes sont utilisées pour estimer l'erreur de la règle de classement apprise. Dans le cas contraire 100 partitions aléatoires de n_u données non étiquetées et n_ℓ données étiquetées sont générées. Le tableau 4.7 montre que AIC$_{cond}$, BEC et la validation croisée ont des performances similaires et font beaucoup mieux que BIC et AIC sur le jeu de données Parkinson.

Jeu de données	n	d	g	Echantillon test	n_u	n_ℓ
Crab	200	5	4	non	150	50
Iris	150	4	3	non	100	50
Parkinson	195	22	2	non	95	100
Pima	532	7	2	oui	332	200
Wine	178	13	3	non	89	89

TAB. 4.6 – Configuration expérimentale.

	BIC	AIC	BEC	AIC$_{cond}$	CV3	CV10
Crab	**6,63**	6,75	6,80	6,77	7,81	7,78
Iris	2,98	2,98	**2,91**	**2,91**	3,25	3,21
Parkinson	26,45	30,68	15,43	**15,16**	18,20	16,38
Pima	25,00	25,00	**19,58**	**19,58**	22,53	**19,58**
Wine	3,24	**1,17**	1,45	1,47	1,73	1,70

TAB. 4.7 – Taux d'erreur produit par chaque critère sur les jeux de données de l'UCI (le critère produisant le plus faible taux d'erreur est mis en gras).

Choix du nombre de composants par classe

Considérons maintenant un exemple sur des puffins (oiseaux)(Biernacki *et al.* , 2002). Considérons les trois sous espèces *borealis*, *diomedea* et *edwardsii* (Biernacki *et al.* , 2002). Le jeu de données est constitué de 336 individus et 5 variables continues sont mesurées. L'objectif est de discriminer au mieux le sexe. On génère 100 échantillons d'apprentissage en cachant 80% des étiquettes au hasard. On propose des modèles allant de 1 à 10 composants par classe avec $[\pi_k \lambda C]$. Ici un modèle à plusieurs composants par classe doit être bien adapté car chaque sous espèce a une distribution spécifique, ce qui implique une hétérogénéité dans la distribution des covariables conditionnellement au sexe.

Les résultats sont illustrés table 4.8. Ici trois sous espèces différentes d'oiseaux sont considérées, le choix de BIC pour trois composants par classe est donc rationnel. Cependant les autres critères produisent de meilleurs résultats en prédiction.

4.3.3 Discussion

Comme illustré par des expériences numériques sur données réelles et simulées, AIC$_{cond}$ doit être préféré à BEC : Dans de nombreuses situations ils produisent les mêmes résultats, cependant quand ils donnent des résultats différents AIC$_{cond}$ fait mieux que BEC. Ceci

	BIC	AIC	BEC	AICcond	CV3	CV10	$\overline{Err}$
1	0	0	0	0	0	0	36,63
2	21	0	0	0	0	1	31,05
3	63	0	0	1	11	17	24,98
4	16	18	12	15	13	19	21,62
5	0	27	19	19	19	20	20,25
6	0	23	16	17	13	12	19,29
7	0	16	16	17	14	14	20,13
8	0	11	14	12	12	4	20,56
9	0	4	9	7	8	6	21,52
10	0	1	14	12	10	7	22,24
Err^*	24,18	19,14	19,84	19,62	20,38	20,78	17,63

TAB. 4.8 – Nombre de composants par classe sélectionnés par les différents critères.

est dû au fait que AIC_{cond} a de meilleures propriétés théoriques que BEC et ne nécessite pas d'utiliser une règle du plateau.

Le critère AIC_{cond} a avant tout été construit dans une perspective semi-supervisée, mais il peut aussi être utilisé dans un cadre supervisé. Dans le cadre semi-supervisé, le coût d'AIC_{cond} est faible puisque l'algorithme EM est déjà nécessaire pour estimer les paramètres. Dans le cadre supervisé, ce critère devient coûteux puisqu'il nécessite l'utilisation de EM pour le calcul de la pénalité, là où l'estimation des paramètres est explicite pour de nombreux modèles. Il est alors intéressant d'étudier les extensions possibles d'AIC_{cond} dans le cadre supervisé.

4.4 Extensions en classification supervisée

Dans cette section nous proposons deux extensions d'AIC_{cond} en supervisé. La première s'appuie sur le concept de dimension prédictive d'un modèle génératif, tandis que la seconde cherche à calculer la pénalité à partir de la vitesse de convergence de EM.

4.4.1 Critère AIC_p

Dans cette partie on propose une majoration de la pénalité présente dans AIC_{cond}. Cette partie a fait l'objet d'une communication aux 41es journées de la SFdS (Vandewalle, 2009b)[3].

Le calcul direct de $\text{trace}(I - JJ_c^{-1})$ est difficile et potentiellement instable puisqu'il nécessite l'évaluation de J qui résulte d'un mélange. Sous l'hypothèse du bon modèle, on prouve ci-dessous la majoration $\text{trace}(I - JJ_c^{-1}) \leqslant \overline{\nu}_m$, où $\overline{\nu}_m$ est une quantité qui ne dépend pas de θ^*, qu'on appelle la dimension prédictive du modèle m et qu'on définit plus précisément section 4.4.1.

On remplace alors $\text{trace}(I - JJ_c^{-1})$ par sa majoration $\overline{\nu}_m$ pour obtenir le critère suivant :

$$\text{AIC}_p(m) = \log p(\mathbf{z}_\ell|\mathbf{x}_\ell; \hat{\theta}^m) - \overline{\nu}_m. \tag{4.75}$$

[3] `http://hal.archives-ouvertes.fr/docs/00/38/66/78/PDF/p117.pdf`

On s'attend à ce que cette pénalité soit trop grande puisqu'il s'agit d'une majoration brutale de la pénalité idéale. Toutefois, dans le cas univarié hétéroscédastique gaussien, si les classes sont confondues, on prouve que cette majoration est atteinte. À ce stade du travail, on conjecture que cette borne supérieure pourrait également être atteinte par bien d'autres modèles génératifs quand les classes sont totalement confondues. On suppose en outre que la discrimination doit être réalisée dans le cas de classes peu séparées. Dans cette situation il est en effet crucial de choisir un modèle et l'utilisation de cette pénalité maximale est donc justifiée.

Définition de la dimension prédictive

Dans l'équation (4.75) on a remplacé la pénalité idéale par sa majoration $\overline{\nu}_m$ qualifiée de dimension prédictive du modèle m dont la définition est la suivante :

Définition 4.1. *Soit m un modèle génératif ayant pour espace des paramètres Θ_m. Soit r un modèle prédictif identifiable ayant pour espace des paramètres Ω_r et vérifiant :*

$$\{\forall\theta\in\Theta_m,\exists\omega\in\Omega_r/\forall(\boldsymbol{x},\boldsymbol{z})\in\mathcal{X}\times\mathcal{Z},p(\boldsymbol{z}|\boldsymbol{x};\theta)=p(\boldsymbol{z}|\boldsymbol{x};\omega), \tag{4.76}$$

$$et\ \forall\omega\in\Omega_r,\exists\theta\in\Theta_m/\forall(\boldsymbol{x},\boldsymbol{z})\in\mathcal{X}\times\mathcal{Z},p(\boldsymbol{z}|\boldsymbol{x};\theta)=p(\boldsymbol{z}|\boldsymbol{x};\omega)\}. \tag{4.77}$$

Alors la dimension prédictive de m est $\overline{\nu}_m=\dim(\Omega_r)$.

On interprète la dimension prédictive comme le nombre de paramètres algébriquement indépendants quand ceux-ci sont estimés en maximisant $p(\mathbf{z}_\ell|\mathbf{x}_\ell;\theta)$ (les problèmes de taille d'échantillon mis à part) ; c'est-à-dire quand les paramètres du modèle génératif sont estimés d'un point de vue prédictif. Ceci justifie le nom de dimension prédictive. Dans la section suivante nous détaillerons le calcul de $\overline{\nu}_m$ dans le cas gaussien. Montrons tout d'abord que $\overline{\nu}_m$ est une borne supérieure de $\nu_m-\text{trace}(JJ_\beta^{-1})$.

Proposition 4.5. *$\nu_m-tr(JJ_\beta^{-1})$ est borné supérieurement par la dimension de tout modèle prédictif identifiable m_p compatible avec m, et que l'on notera $\overline{\nu}_m$.*

PREUVE : D'abord puisque $J_\beta\prec J_c$, on a

$$\nu_m-tr(JJ_\beta^{-1})\leqslant\nu_m-\text{trace}(JJ_c^{-1}). \tag{4.78}$$

On peut maintenant écrire $\nu_m-\text{trace}(JJ_c^{-1})=\text{trace}((J_c-J)J_c^{-1})$. Puis en choisissant la paramétrisation du modèle telle que

$$J_c-J=\begin{pmatrix}\overline{J}_c-\overline{J}(\overline{\nu}_m\times\overline{\nu}_m) & 0(\overline{\nu}_m\times(\nu_m-\overline{\nu}_m))\\ 0((\nu_m-\overline{\nu}_m)\times\overline{\nu}_m) & 0((\nu_m-\overline{\nu}_m)\times(\nu_m-\overline{\nu}_m))\end{pmatrix},$$

avec $\overline{A}=(A_{ij})_{1\leqslant i,j\leqslant\overline{\nu}_m}$, on a

$$\text{trace}((J_c-J)J_c^{-1}) \;=\; \text{trace}((\overline{J}_c-\overline{J})\overline{J_c^{-1}}). \tag{4.79}$$

Puisque $J_c\succ 0$, $\overline{J_c^{-1}}\prec\overline{J}_c^{-1}$, et puisque $\overline{J}_c-\overline{J}\succ 0$, on obtient

$$\text{trace}((J_c-J)J_c^{-1}) \;\leqslant\; \text{trace}((\overline{J}_c-\overline{J})\overline{J}_c^{-1}) \tag{4.80}$$

$$\;=\; \overline{\nu}_m-\text{trace}(\overline{J}\overline{J}_c^{-1}). \tag{4.81}$$

Enfin puisque $\overline{J}$ et $\overline{J}_c^{-1}$ sont définies positives, on a

$$\text{trace}((J_c-J)J_c^{-1}) \;\leqslant\; \overline{\nu}_m. \tag{4.82}$$

Ce qui conclut la preuve du lemme 4.5. □

Modèle m	Dimension générative (ν_m)	Dimension prédictive ($\overline{\nu}_m$)
λC	$\eta + d + d(d+1)/2$	η
λB	$\eta + d + d$	η
λI	$\eta + d + 1$	η
$\lambda_k C_k$	$\eta + d + gd(d+1)/2$	$\eta + (g-1)d(d+1)/2$
$\lambda_k B_k$	$\eta + d + gd$	$\eta + (g-1)d$
$\lambda_k C$	$\eta + d + d(d+1)/2 - 1 + g$	$\eta + d(d+1)/2 - 1 + (g-1)$
$\lambda_k B$	$\eta + d + d - 1 + g$	$\eta + d - 1 + (g-1)$
$\lambda_k I$	$\eta + d + g$	$\eta + g - 1$
$\lambda_k D A_k D'$	$\eta + d + d(d-1)/2 + gd$	$\eta + d(d-1)/2 + (g-1)d$

TAB. 4.9 – Dimension générative et dimension prédictive pour certains modèles parcimonieux ($\eta = (g-1)(d+1)$).

Dimension prédictive dans le cas gaussien

Soit $\mathcal{X} = \mathbb{R}^d$ et supposons une distribution gaussienne ϕ conditionnellement à la classe. Prenons la classe g comme classe de référence. On a :

$$\log \frac{\pi_k \phi(\boldsymbol{x}; \mu_k, \Sigma_k)}{\pi_g \phi(\boldsymbol{x}; \mu_g, \Sigma_g)} = \eta_k + \beta_k' \boldsymbol{x} + \boldsymbol{x}' \Delta_k \boldsymbol{x} \ \forall k \in \{1, \ldots, g-1\}, \tag{4.83}$$

avec $\eta_k \in \mathbb{R}$, $\beta_k \in \mathbb{R}^d$ et Δ_k une matrice symétrique de $\mathbb{R}^{d \times d}$. Ainsi le modèle prédictif r vérifiant l'équation (4.76) est

$$p(z_k = 1 | \boldsymbol{x}; \omega) = \frac{e^{h(\boldsymbol{x}; \omega_k)}}{1 + \sum_{j=1}^{g-1} e^{h(\boldsymbol{x}; \omega_j)}} \ \forall k \in \{1, \ldots, g-1\}$$

où

$$h(\boldsymbol{x}; \omega_k) = \eta_k + \beta_k' \boldsymbol{x} + \boldsymbol{x}' \Delta_k \boldsymbol{x}.$$

Cela correspond à la régression logistique quadratique. On élague ensuite le modèle pr en fonction de la paramétrisation de Σ_k choisie pour qu'il vérifie (4.77) et qu'il soit identifiable. On montre que (η_k, β_k) est libre dans $\mathbb{R}^{d+1}$ quelque soit la paramétrisation de la matrice de variance Σ_k choisie. Il y a donc au moins $\eta = (g-1)(d+1)$ paramètres libres.

Pour l'obtention de modèles parcimonieux, la matrice de variance Σ_k est décomposée en valeurs singulières sous la forme $\Sigma_k = \lambda_k D_k A_k D_k'$, puis des contraintes d'égalité entre les classes pour λ_k, D_k ou A_k sont imposées, pour cela on se réfère à la section 2.3.1. Les résultats sont présentés table 4.9.

La plus grande différence entre $\overline{\nu}_m$ (dimension prédictive) et ν_m (dimension du modèle génératif) est obtenue pour $d >> g$ quand on considère le modèle λC. Dans ce cas, ν_m est quadratique en d alors que $\overline{\nu}_m$ est linéaire en d. Ceci explique la robustesse de l'analyse discriminante linéaire puisqu'un grand nombre de paramètres est estimé, mais seulement un petit nombre de combinaisons d'entre eux prend part à l'estimation de la distribution conditionnelle. Remarquons au passage que les modèles λI, λB et $\lambda DAD'$ ont la même dimension prédictive, tandis que leur dimension générative est différente. Le critère AIC_p risque donc de favoriser le modèle $\lambda DAD'$ par rapport aux modèles λB et λI. C'est certainement une conséquence de la majoration brutale de la pénalité idéale.

	d	g	n	n_{test}	AIC	BIC	AIC_p	CV3	CV10
Breast Cancer	30	2	400	169	**4,31**	**4,31**	4,52	4,73	4,76
Wine	13	3	89	89	4,89	2,99	**2,24**	2,94	2,72
Pima	7	2	200	332	23,49	**20,18**	**20,18**	24,10	**20,18**
Crab	5	4	100	100	6,57	5,60	**5,49**	5,84	5,84
Iris	4	3	75	75	2,81	2,93	**2,74**	3,86	3,76
Parkinson	22	2	146	49	**12,59**	12,79	12,89	13,44	13,16
Synt	2	2	250	1000	**10,20**	10,90	10,80	**10,20**	**10,20**
Transfusion	4	2	374	374	28,93	28,93	27,76	24,35	**24,34**
Ionosphere	32	2	175	176	**14,87**	**14,87**	16,14	16,34	16,34

TAB. 4.10 – Erreur produite par les différents critères de choix de modèle.

On compare AIC, BIC, la *3-fold* validation croisée (CV3), la *10-fold* validation croisée (CV10) et AIC_p sur des *Benchmarks* disponibles sur le site de l'UCI et *Pattern Recognition*. Quand un échantillon test est fourni, on l'utilise pour évaluer l'erreur produite par le modèle sélectionné. Si ce n'est pas le cas, on génère aléatoirement 100 jeux d'apprentissage/test de tailles n et n_{test}, et on moyenne l'erreur produite. Les modèles λC, λB, λI, $\lambda_k C_k$, $\lambda_k B_k$ et $\lambda_k I$ sont mis en compétition.

On remarque table 4.10 le bon comportement du critère AIC_p. Du fait de la variabilité dans l'estimation de l'erreur sur ces données, la validation croisée ne produit pas systématiquement les meilleurs résultats.

Bilan sur AIC_p

On a défini le critère de choix de modèle AIC_p en analyse discriminante focalisé sur la distribution des étiquettes conditionnellement aux covariables. Ce critère est construit à partir d'un critère idéal dont on a remplacé la pénalité par sa borne supérieure, ce qui a fait apparaître la notion de dimension prédictive d'un modèle génératif. On a calculé cette dimension pour certains modèles gaussiens et on a montré le bon comportement du critère AIC_p sur des données réelles.

4.4.2 Calcul de la pénalité à partir de la vitesse de convergence de EM

Cadre supervisé

Dans le cadre supervisé la pénalité peut être calculée en utilisant le lien entre le rapport des matrices d'information et la vitesse de convergence de l'algorithme EM.

On propose d'optimiser la fonction suivante

$$L_\beta(\theta) = \beta \log p(\mathbf{x}_\ell, \mathbf{z}_\ell; \theta) + (1-\beta) \log p(\mathbf{x}_\ell; \theta), \tag{4.84}$$

avec β proche de 1. Soit $\hat{\theta}_\beta = \arg\max_{\theta \in \Theta} L_\beta(\theta)$. Cette maximisation peut être réalisée en utilisant l'algorithme λ-EM décrit section 2.2.6.

On a alors la propriété suivante

Lemme 4.5. *Si la distribution d'échantillonnage appartient au modèle postulé*

$$\lim_{\beta \to 1} \frac{1}{1-\beta} tr(DM) \xrightarrow{P} tr(I - JJ_c^{-1}). \tag{4.85}$$

PREUVE : On sait d'après la section 2.2.5 que

$$DM(\hat{\theta}) = [D^{20}H(\hat{\theta}|\hat{\theta})]D^{20}Q(\hat{\theta}|\hat{\theta})^{-1}. \tag{4.86}$$

On peut alors appliquer cette propriété à $L_\beta(\theta)$. On a alors

$$D^{20}Q(\hat{\theta}_\beta|\hat{\theta}_\beta) = \beta\nabla^2 \log p(\mathbf{x}_\ell, \mathbf{z}_\ell; \hat{\theta}_\beta) + (1-\beta)\nabla^2 E[\log p(\mathbf{x}_\ell, \tilde{\mathbf{z}}_\ell; \hat{\theta}_\beta)|\hat{\theta}_\beta, \mathbf{x}_\ell]. \tag{4.87}$$

Remarquons d'abord que

$$\lim_{\beta \to 1} D^{20}Q(\hat{\theta}_\beta|\hat{\theta}_\beta) \to D^{20}Q(\hat{\theta}|\hat{\theta}) \tag{4.88}$$

si on suppose que $D^{20}Q$ est continue en β. Par conséquent

$$\frac{1}{n} D^{20}Q(\hat{\theta}_\beta|\hat{\theta}_\beta) \xrightarrow{P} J_c. \tag{4.89}$$

Puis en remarquant que

$$D^{20}H(\hat{\theta}_\beta|\hat{\theta}_\beta) = D^{20}Q(\hat{\theta}_\beta|\hat{\theta}_\beta) - \nabla^2 L_\beta(\hat{\theta}_\beta). \tag{4.90}$$

Et en normalisant $D^{20}H(\hat{\theta}_\beta|\hat{\theta}_\beta)$ par $(1-\beta)n$

$$\frac{1}{1-\beta}\frac{1}{n} D^{20}H(\hat{\theta}_\beta|\hat{\theta}_\beta) \xrightarrow{P} [J_c - J]. \tag{4.91}$$

De telle sorte que

$$\frac{1}{1-\beta} D^{20}H(\hat{\theta}_\beta|\hat{\theta}_\beta)D^{20}Q(\hat{\theta}_\beta|\hat{\theta}_\beta)^{-1} \xrightarrow{P} [J_c - J]J_c^{-1}. \tag{4.92}$$

Puisque

$$DM(\hat{\theta}_\beta) = [D^{20}H(\hat{\theta}_\beta|\hat{\theta}_\beta)]D^{20}Q(\hat{\theta}_\beta|\hat{\theta}_\beta)^{-1} \tag{4.93}$$

on obtient le résultat souhaité. □

Ainsi dans le cadre supervisé, la pénalité à calculer ne nécessite que le calcul des éléments diagonaux de DM. On propose alors la stratégie suivante

- Prendre β proche de 1,
- Trouver $\hat{\theta}_\beta$ en utilisant l'algorithme λ-EM initialisé avec $\hat{\theta}$,
- Perturber la solution en utilisant une étape de l'algorithme SEM (Celeux & Diebolt, 1992), on obtient ainsi $\theta_\beta^{(0)}$,
- Retourner dans le voisinage de $\hat{\theta}_\beta$ en utilisant deux étapes de λ-EM, on obtient $\theta_\beta^{(2)}$,
- Obtenir $\theta_\beta^{(3)}$ par une étape de λ-EM initialisé avec $\theta_\beta^{(2)}$,
- Obtenir $\theta_\beta^{(4)}$ par une étape de λ-EM initialisé avec $\theta_\beta^{(3)}$,

– Calculer

$$\frac{1}{1-\beta}\sum_{i=1}^{d}\frac{|(\theta_\beta^{(4)})_i-(\hat{\theta}_\beta)_i|}{|(\theta_\beta^{(3)})_i-(\hat{\theta}_\beta)_i|} \tag{4.94}$$

qui est la pénalité requise.

Illustrons maintenant cette stratégie sur un exemple jouet. Soit un problème de classification à deux classes, $\pi_1=\pi_2=1/2$, $\boldsymbol{X}|Z_1=1\sim\mathcal{N}(0,1)$ et $\boldsymbol{X}|Z_2=1\sim\mathcal{N}(\mu,1)$, où $\mu\in[0,10]$. Le modèle postulé est que chaque classe suit une distribution gaussienne avec une variance différente et les proportions de chaque classe sont inconnues. Les paramètres à estimer sont $\theta=(\pi_1,\mu_1,\mu_2,\sigma_1,\sigma_2)$ ainsi $\nu_m=5$. Prenons $\beta=0,9$, on trace la pénalité en fonction de la séparation des classes pour $n=\{100;1000;10000\}$ données sur la figure 4.4.

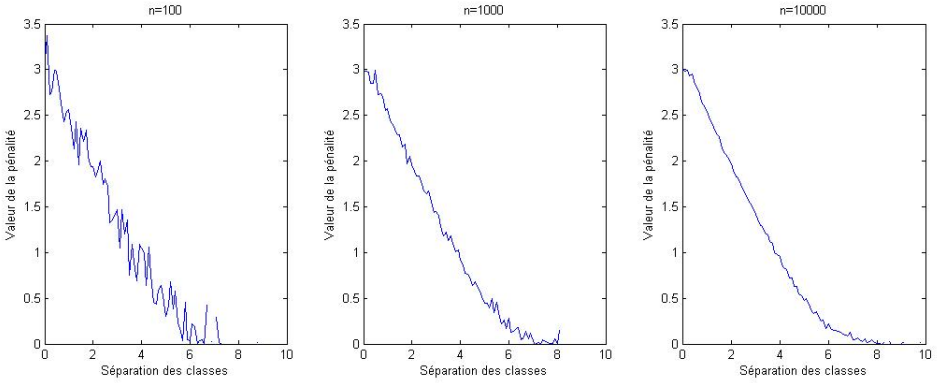

FIG. 4.4 – Pénalité en fonction de la séparation des classes pour diverses valeurs de n

Ainsi on peut remarquer que plus on a de données, plus le calcul de la pénalité est stable. D'autre part, plus les classes sont séparées et plus petite est la pénalité. L'interprétation est la suivante : si les classes sont bien séparées, le surajustement à la distribution conditionnelle n'a pas lieu. Remarquons que lorsque les classes sont mal séparées ($\mu=0$) la pénalité est proche de 3 ce qui correspond à la dimension prédictive du modèle.

Cadre semi-supervisé

Remarquons que dans le cadre semi-supervisé on a

$$AIC_{cond}(m)=\log p(\mathbf{z}_\ell|\mathbf{x}_\ell;\hat{\theta}_m)-[\nu_m-\mathrm{trace}(JJ_\beta^{-1})]. \tag{4.95}$$

Or,

Proposition 4.6. *Si la distribution d'échantillonnage est incluse dans le modèle*

$$\frac{\beta}{1-\beta}trace(DM[I-DM]^{-1})\xrightarrow{P}trace(I-JJ_\beta^{-1}), \tag{4.96}$$

où β est la fraction de données étiquetées $J_\beta=\beta J_c+(1-\beta)J$.

PREUVE : De même que précédemment on remarque que

$$DM(\hat{\theta}) \rightarrow (1-\beta)[J - J_c^{-1}]J_c^{-1}.$$

Puis en utilisant l'identité

$$J_\beta = \beta J_c + (1-\beta)J,$$

on vérifie facilement que

$$\frac{\beta}{1-\beta}\text{trace}(DM[I-DM]^{-1}) \xrightarrow{P} \text{trace}(I - JJ_\beta^{-1}).$$

Ce qui conclut la preuve. □

Remarquons que si β est proche de 0, la pénalité $\text{tr}(I - JJ_\beta^{-1})$, tend vers 0. Ce qui semble naturel puisque l'influence des données étiquetées sur l'estimation des paramètres est plus faible et par conséquent le risque de surajustement aux données disparaît. Cependant on peut pas supprimer ce terme de pénalisation puisqu'autrement aucun contrôle de la complexité ne serait fait.

Ici le calcul de chaque élément de DM est requis et pas seulement celui des éléments diagonaux. Ceci est possible en utilisant l'approche de Meng & Rubin (1991). Mais cette approche peut être coûteuse, surtout si le paramètre considéré est de grande dimension. Idéalement on voudrait calculer les valeurs propres de DM puisque $\text{trace}(DM[I-DM]^{-1}) = \sum_i \frac{\lambda_i}{1-\lambda_i}$ où λ_i sont les valeurs propres de DM. Dans ce contexte il peut sembler intéressant de trouver des stratégies efficaces pour trouver les valeurs propres de DM. Ici, le critère devient difficile à utiliser en pratique, c'est pourquoi on recommande plutôt d'utiliser le critère AIC_{cond}.

4.5 Conclusion

La contribution principale de ce travail est un nouveau critère fondé sur une approximation de la déviance prédictive, AIC_{cond}, pour sélectionner un modèle génératif en classification semi-supervisée. Dans ce but, le critère AIC_{cond} vise à minimiser la divergence de Kullback moyenne entre la distribution des étiquettes conditionnellement aux covariables et la vraie distribution conditionnelle. Par une série d'approximations, une formule qui ressemble à celle de BEC, mais qui comporte un terme de surpénalisation favorise le modèle le plus simple lorsque l'on considère deux vrais modèles emboîtés. Le calcul de ces deux critères est beaucoup plus rapide que la *V-fold* validation croisée dans le contexte semi-supervisé.

Deuxième partie

Contribution à la modélisation multinomiale

Chapitre 5

Contribution à la modélisation multinomiale

Dans le cas discret, des modèles multinomiaux parcimonieux ont été proposés par Celeux & Govaert (1991) pour les mélanges de produits de distributions de Bernoulli. Une extension de ces modèles au cas multinomial est utilisée dans le logiciel MIXMOD (Biernacki *et al.* , 2006). Cette extension pose cependant des difficultés quand on considère des variables avec des nombres de modalités différents. Nous proposons une reparamètrisation de ce modèle permettant de lever les difficultés précédentes. Puis, nous proposons un modèle qui impose l'égalité des paramètres entre classes, ceci à une permutation des modalités près. Enfin, on pose la question de la prise en compte des paramètres discrets dans l'approximation BIC.

5.1 Modèles parcimonieux standards

5.1.1 Présentation

Rappelons brièvement le principe du modèle d'indépendance conditionnelle présenté chapitre 2 section 2.3.2. Soit d variables discrètes, où chaque variable $j \in \{1, \ldots, d\}$ admet m_j modalités. Soit

$$x_i^{jh} = \begin{cases} 1 \text{ si l'individu } i \text{ présente la modalité } h \text{ de la variable } j, \\ 0 \text{ sinon.} \end{cases}$$

On note $\alpha_k^{jh} = p(X^{jh} = 1 | Z_k = 1)$, $\alpha_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$ et $\alpha_k = (\alpha_k^1, \ldots, \alpha_k^d)$. Dans ce cas la probabilité (discrète) pour une donnée i conditionnellement à la classe k s'écrit

$$p(\boldsymbol{x}_i; \alpha_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}.$$

Par ailleurs on note $\theta = (\pi_1, \ldots, \pi_{g-1}, \alpha_1, \ldots, \alpha_g)$ le paramètre du mélange.

Nous détaillons maintenant les modèles parcimonieux sur les mélanges de produits de distributions de Bernouilli proposés par Celeux & Govaert (1991) et étendus aux cas multinomial par Biernacki *et al.* (2006). Pour simplifier les formules d'actualisation nous

nous placerons dans le cadre supervisé, l'extension aux cadres non supervisé et semi-supervisé ne posant bien sûr pas de difficulté puisque nécessitant simplement le recours à l'algorithme EM.

Ces modèles procèdent à une reparamètrisation du vecteur α_k^j sous la forme $\alpha_k^j = (\beta_k^j, \ldots, \beta_k^j, \gamma_k^j, \beta_k^j, \ldots, \beta_k^j)$ avec $\gamma_k^j \geqslant \beta_k^j$. Autrement dit, il existe une modalité majoritaire et la masse de probabilité restante est équidistribuée entre les autres modalités. Puisque la somme des proportions est égale à 1, on a $\beta_k^j = (1-\gamma_k^j)/(m_j - 1)$. Pour que la modalité majoritaire soit effectivement majoritaire, il faut imposer la contrainte $\gamma_k^j \geqslant \frac{1}{m_j}$. Les paramètres sont la position de la modalité majoritaire pour la variable j dans la classe k notée $h(k,j)$ et la masse de probabilité ε_k^j restante pour les modalités minoritaires ($\varepsilon_k^j = 1 - \gamma_k^j$). Ainsi

$$\alpha_k^{jh} = \begin{cases} 1 - \varepsilon_k^j & \text{si } h = h(k,j) \\ \frac{\varepsilon_k^j}{m_j - 1} & \text{sinon.} \end{cases}$$

Ce modèle noté $[\varepsilon_k^j]$, permet de réduire l'estimation de $m_j - 1$ paramètres par variable et par classe, à l'estimation d'un paramètre continu et d'un paramètre discret par variable et par classe. Des modèles encore plus parcimonieux peuvent être obtenus en imposant des contraintes d'égalité entre classe et/ou entre variables pour le paramètre ε_k^j :

- $[\varepsilon^j]$: on impose aux classes de partager le même paramètre,
- $[\varepsilon_k]$: on impose aux variables de partager le même paramètre,
- $[\varepsilon]$: on impose les deux contraintes précédentes.

Par mimétisme, on note aussi $[\varepsilon_k^{jh}]$ le modèle général qui n'impose aucune modalité majorataire. L'étape M de l'algorithme EM est explicite pour ces modèles. Ils peuvent être reliés à des critères de classification non supervisée sur variables qualitatives (Celeux & Govaert, 1991). Dans le cas binaire l'écart identique pour les modalités majoritaires s'interprête comme une variance égale des variables de Bernoulli.

5.1.2 Limitations

Pour les modèles $[\varepsilon_k^{jh}]$, $[\varepsilon_k^j]$ et $[\varepsilon^j]$ la contrainte $\varepsilon_k^j \leqslant (m_j - 1)/m_j$ est automatiquement vérifiée puisqu'on agrège des estimations ayant le même nombre de modalités. Par contre, cette contrainte n'est plus systématiquement vérifiée pour les modèles $[\varepsilon_k]$ et $[\varepsilon]$.

Par exemple considèrons une seule classe avec $\alpha_1^1 = (0,37; 0,30; 0,33)$ et $\alpha_1^2 = (0,58; 0,42)$ et utilisons le modèle $[\varepsilon_k]$. Le paramètre minimisant l'écart de Kullback entre ce modèle et la vraie distribution sans la contraite $\varepsilon_k^j \leqslant (m_j - 1)/m_j$ est alors $\tilde{\varepsilon}_k = 0,52$, d'où $\tilde{\alpha}_1^1 = (0,48; 0,26; 0,26)$ et $\tilde{\alpha}_1^2 = (0,48; 0,52)$. La première modalité de la seconde variable qui devrait être majoritaire se retrouve minoritaire. Cette situation survient typiquement quand il y a une variable avec beaucoup de modalités et proche de l'uniformité et une autre variable avec moins de modalités et plus éloignée de l'uniformité. Une posibilité est d'imposer explicitement le respect de la contrainte $\varepsilon_k^j \leqslant (m_j - 1)/m_j$, ce qui est peu naturel puisque cela implique que les paramètres estimés sont au bord de l'espace des paramètres. Cependant, dans ce qui suit on propose une paramètrisation légèrement différente permettant d'obtenir des nouveaux modèles de type $[\varepsilon_k]$ et $[\varepsilon]$, mais pour lesquels les contraintes sont automatiquement vérifiées.

5.2 Modèles parcimonieux proposés

5.2.1 Reparamétrisation des modèles parcimonieux standards

Dans le cas des modèles $[\varepsilon_k]$ et $[\varepsilon]$ les contraintes sur ε_k^j diffèrent selon le nombre de modalités des variables. Définissons la paramétrisation suivante

$$\alpha_k^{jh} = \begin{cases} 1 - \frac{m_j - 1}{m_j}\varepsilon_k^j & \text{si } h = h(k,j), \\ \frac{\varepsilon_k^j}{m_j} & \text{sinon.} \end{cases}$$

La contrainte est alors $\varepsilon_k^j \leqslant 1$. L'espace dans lequel est défini ε_k^j ne dépend plus du nombre de modalités de la variable considérée. Ceci implique une certaine forme de stabilité quand on impose

$$\varepsilon_k^1 = \varepsilon_k^2 = \ldots = \varepsilon_k^d.$$

Les modèles obtenus sont équivalents aux modèles standards pour $[\varepsilon_k^{jh}]$, $[\varepsilon_k^{jh}]$ et $[\varepsilon^j]$, et sont différents pour $[\varepsilon_k]$ et $[\varepsilon]$ qu'on note respectivement $[\varepsilon_k]_{bis}$ et $[\varepsilon]_{bis}$.

Proposition 5.1. *La contrainte $\hat{\varepsilon}_k \leqslant 1$ est automatiquement vérifiée lors de l'estimation par maximum de vraisemblance des paramètres pour le modèle $[\varepsilon_k]_{bis}$.*

PREUVE : Notons

$$e_k^j = \sum_{i=1}^{n} z_{ik} x_i^{jh(k,j)},$$

La vraisemblance pour le modèle $[\varepsilon_k]_{bis}$ s'écrit à une constante près

$$g(\varepsilon_k) = \sum_{j=1}^{d} (n_k - e_k^j) \log(m_j - (m_j - 1)\varepsilon_k) + e_k^j \log(\varepsilon_k). \tag{5.1}$$

On vérifie facilement que

$$g_j(\varepsilon_k) = (n_k - e_k^j) \log(m_j - (m_j - 1)\varepsilon_k) + e_k^j \log(\varepsilon_k) \tag{5.2}$$

g_j est concave et que son maximum est atteint pour $\varepsilon_k \in [0, 1]$. Or

$$g(\varepsilon_k) = \sum_{j=1}^{d} g_j(\varepsilon_k),$$

donc g est concave et on obtient après un court raisonnement que

$$\hat{\varepsilon}_k = \underset{\varepsilon_k}{\operatorname{argmax}}\, g(\varepsilon_k) \in [\min_j \underset{\varepsilon_k}{\operatorname{argmax}}\, g_j(\varepsilon_k); \max_j \underset{\varepsilon_k}{\operatorname{argmax}}\, g_j(\varepsilon_k)] \subset [0; 1]$$

□

La maximisation de la vraisemblance g (équation (5.2)) ne peut pas être effectuée explicitement. Cependant puisqu'elle est concave en ε_k, la maximisation est facilement réalisée grâce à un algorithme de Newton. On pose

$$\eta_k = \log \frac{\varepsilon_k}{1 - \varepsilon_k},$$

de telle sorte que $\eta_k \in \mathbb{R}$. Notons $\overline{e}_k^j = (n - e_k^{jh(k,j)})$. On a

$$\begin{aligned} G_k(\eta_k) &= -\sum_{j=1}^{d} \frac{e^{-\eta_k}(nm_j - n - \overline{e}_k^j m_j - \overline{e}_k^j m_j e^{-\eta_k})}{(1+e^{-\eta_k})(m_j e^{-\eta_k}+1)}, \\ J_k(\eta_k) &= -\sum_{j=1}^{d} \frac{e^{-\eta_k}(n + 2\overline{e}_k^j m_j e^{-\eta_k} + \overline{e}_k^j m_j e^{-2\eta_k} + nm_j^2 e^{-2\eta_k} - e^{-2\eta_k} nm_j - nm_j + \overline{e}_k^j m_j)}{(1+e^{-\eta_k})^2 (m_j e^{-\eta_k}+1)^2}. \end{aligned}$$

L'algorithme de Newton est alors

$$\eta_k^{(r+1)} = \eta_k^{(r)} - \frac{G_k(\eta_k^{(r)})}{J_k(\eta_k^r)}.$$

De façon similaire, pour le modèle $[\varepsilon]_{bis}$ on a

$$\eta^{(r+1)} = \eta^{(r)} - \frac{\sum_{k=1}^{g} G_k(\eta^{(r)})}{\sum_{k=1}^{g} J_k(\eta^{(r)})}.$$

Pour le cas qui mettait le modèle $[\varepsilon_k]$ en défaut, on obtient maintenant $\tilde{\varepsilon}_k = 0,9$ d'où $\tilde{\alpha}_1^1 = (0,40; 0,30; 0,30)$ et $\tilde{\alpha}_1^2 = (0,55; 0,45)$. Cela est en accord avec l'interprétation en termes de modalité majoritaire.

La paramétrisation proposée permet donc d'obtenir deux nouveaux modèles multinomiaux parcimonieux qui s'ajoutent aux cinq modèles déjà proposés par Biernacki *et al.* (2006). Remarquons d'ailleurs que les modèles $[\varepsilon_k]_{bis}$ et $[\varepsilon]_{bis}$ se ramènent aux modèles $[\varepsilon_k]$ et $[\varepsilon]$ dans le cas où toutes les variables ont le même nombre de modalités.

5.2.2 Égalisation des paramètres à une permutation des modalités près

Remarquons que le modèle $[\varepsilon^{jh}]$ n'est pas utilisé puisqu'il suppose que toutes les classes ont la même distribution. Ceci ne comporte aucun intérêt ni en classification supervisée, ni en classification non supervisée.

Définissons le modèle $[\varepsilon^{j\sigma_k^j(h)}]$, où on suppose que tous les vecteurs α_k^j ont la même composition pour toutes les classes à une permutation des modalités près. On note :

- ε^j : le vecteur des fréquences rangées par ordre décroissant des modalités de la variables j,
- σ_k^j : la permutation de ε^j telle que $\alpha_k^{jh} = \varepsilon^{j\sigma_k^j(h)}$.

Par exemple si on a $\alpha_1^1 = (0,15; 0,25; 0,60)$ et $\alpha_2^1 = (0,60; 0,15; 0,25)$, alors $\varepsilon^1 = (0,60; 0,25; 0,15)$, $\sigma_1^1 = (3;2;1)$ et $\sigma_1^2 = (1;3;2)$.

Le modèle proposé se situe à mi-chemin entre le modèle $[\varepsilon_k^{jh}]$ et le modèle $[\varepsilon^j]$. On s'attend donc à ce que dans les situations où le modèle $[\varepsilon^j]$ produit des résultats trop biaisés, et le modèle $[\varepsilon_k^{jh}]$ des résultat trop variables, le modèle $[\varepsilon^{j\sigma_k^j(h)}]$ permette de trouver un bon compromis entre biais et variance.

La vraisemblance s'écrit

$$g(\varepsilon^j) = \sum_k \sum_h n_k^{jh} \log(\varepsilon^{j\sigma_k^j(h)}), \tag{5.3}$$

avec $n_k^{jh} = \sum_{i=1}^{n} z_{ik} x_i^{jh}$.

Proposition 5.2. *Pour estimer les paramètres du modèle* $[\varepsilon_k^{j\sigma_k^j(h)}]$ *on ordonne les différentes modalités pour chaque variable dans chaque classe, puis on agrège les estimations.*

PREUVE : Pour ε^j fixé, l'inverse de la permutation optimale consiste à permuter les composants du vecteur n_k^j par ordre décroissant. À permutation fixée la solution optimale pour ε^j consiste à agréger les estimations, c'est-à-dire à prendre

$$\hat{\varepsilon}^{jh} = \frac{\sum_k n_k^{j\sigma_k^{j^{-1}}(h)}}{n}.$$

□

Ce modèle permet de diviser par un facteur g le nombre de paramètres continus à estimer, ce à quoi il faut ajouter les paramètres discrets σ_k^j. L'idée sous-jacente de ce modèle est que c'est l'ordre de modalités qui apporte une information importante en classification.

Il est intéressant de constater que le modèle $[\varepsilon^j]$ effectue naturellement une sélection des modalités. En effet, les modalités minoritaires ou majoritaires pour toutes les classes n'influencent en aucune mesure la règle de classement. C'est aussi le cas du modèle $[\varepsilon^{j\sigma_k^j(h)}]$ si une modalité possède le même rang quelque soit la classe.

Dans le cadre bayesien, il est facile de proposer une loi *a priori* conjugée pour le modèle général $[\varepsilon_k^{jh}]$ mais difficile pour les modèles parcimieux classiques. Neanmoins, dans le cas du modèle $[\varepsilon^{j\sigma_k^j(h)}]$ il est facile de proposer une loi *a priori* conjuguée qui est ici une loi de Dirichlet sur ε^j. C'est-à-dire

$$p(\varepsilon^j) = \frac{\prod_{h=1}^{m_j}\Gamma(\beta^{jh})}{\Gamma(\sum_{h=1}^{m_j}\beta^{jh})}\prod_{h=1}^{m_j}(\varepsilon^{jh})^{\beta^{jh}-1},$$

où β^j est le vecteur des hyperparamètres.

5.2.3 Bilan

Trois nouveaux modèles parcimonieux ont donc été proposés. Deux d'entre-eux s'appuient sur une reparamétrisation des modèles $[\varepsilon_k]$ et $[\varepsilon]$ pour éviter que la solution obtenue ne viole les contraintes de modalité majoritaire. Le troisième égalise les paramètres pour toutes les classes à une permutation des modalités près. Ce modèle est à mi-chemin entre le modèle $[\varepsilon_k^{jh}]$ et $[\varepsilon^j]$. Le nombre de paramètres continus des différents modèles parcimonieux est indiqué table 5.1. Nous illustrons maintenant leur comportement sur quelques expériences.

5.3 Expériences sur les modèles parcimonieux

5.3.1 Illustration du modèle $[\varepsilon^{j\sigma_k^j(h)}]$ sur données simulées

Distribution d'échantillonnage incluse dans $[\varepsilon^{j\sigma_k^j(h)}]$

Le modèle d'indépendance conditionnelle permet de prendre en compte de nombreuses variables comportant de nombreuses modalités. Cependant, la prise en compte de nom-

Modèles standards	Nombre de paramètres continus
$[\varepsilon_k^{jh}]$	$(g-1)+\sum_{j=1}^{d}(m_j-1)g$
$[\varepsilon_k^{j}]$	$(g-1)+dg$
$[\varepsilon_k]$	$(g-1)+g$
$[\varepsilon^{j}]$	$(g-1)+d$
$[\varepsilon]$	$(g-1)+1$
Modèles proposés	**Nombre de paramètres continus**
$[\varepsilon_k]_{bis}$	$(g-1)+g$
$[\varepsilon]_{bis}$	$(g-1)+1$
$[\varepsilon^{j\sigma_k^j(h)}]$	$(g-1)+\sum_{j=1}^{d}(m_j-1)$

TAB. 5.1 – Nombre de paramètres continus pour les modèles multinomiaux parcimonieux.

Modèles	$[\varepsilon_k^{jh}]$	$[\varepsilon^{j\sigma_k^j(h)}]$	$[\varepsilon^{j}]$
$\overline{Err}_s$ (%)	22,19 (3,48)	13,85 (3,88)	22,83 (4,14)
$\overline{Err}_{ss}$ (%)	6,93 (0,21)	6,68 (0,13)	33,02 (2,91)

TAB. 5.2 – Erreur en classification supervisée ($\overline{Err}_s$) et en classification semi-supervisée ($\overline{Err}_{ss}$) pour les modèles $[\varepsilon_k^{jh}]$, $[\varepsilon^{j\sigma_k^j(h)}]$ et $[\varepsilon^{j}]$ quand la distribution d'échantillonnage est incluse dans $[\varepsilon^{j\sigma_k^j(h)}]$, les écarts-types sont représentés entre parenthèses.

breuses variables implique une augmentation de la variance des estimateurs. Dans de telles situations l'utilisation de modèles parcimonieux est à recommander. Considérons un problème de classification (semi-supervisé et supervisé) à deux classes, avec $d = 15$ variables discrètes comportant chacune cinq modalités. Les paramètres de simulation sont les suivants :

$$\pi_1 = \pi_2 = 0,5, \forall j \in \{1,\ldots,15\}, \alpha_1^j = (0,10; 0,35; 0,35; 0,10; 0,10),$$
$$\alpha_2^j = (0,35; 0,35; 0,10; 0,10; 0,10).$$

Dans le cas supervisé, on génère 100 échantillons de 50 données classées, et un échantillon test de taille 100000. Dans le cas semi-supervisé, on ajoute 500 données non classées aux 50 données classées. Remarquons, que seules deux des cinq modalités apportent une information sur la règle de classement. D'autre part, un modèle avec modalité majoritaire n'est pas adapté ici puisque deux modalités sont en fréquence plus importante que les autres. On veut comparer les résultats obtenus par les modèles $[\varepsilon_k^{jh}]$, $[\varepsilon^{j\sigma_k^j(h)}]$ et $[\varepsilon^{j}]$ dans les cas supervisé et semi-supervisé. Les résultats sont présentés table 5.2, résultats à comparer aussi à l'erreur de Bayes qui est égale à 6,36%. On remarque que le modèle parcimonieux est d'autant meilleur que le modèle complet que le nombre de données est petit.

En pratique aucun modèle n'est vrai. On se pose alors la question de la robustesse du modèle aux mauvaises spécifications dans les expériences suivantes.

Modèles	$[\varepsilon_k^{jh}]$	$[\varepsilon^{j\sigma_k(h)}]$	$[\varepsilon^j]$
$\overline{Err}_s$	22,35 (4,10)	14,56 (3,47)	19,95 (3,68)
$\overline{Err}_{ss}$	5,16 (0,16)	8,41 (0,45)	25,72 (4,30)

TAB. 5.3 – Erreur en classification supervisée ($\overline{Err}_s$) et en classification semi-supervisée ($\overline{Err}_{ss}$) pour les modèles $[\varepsilon_k^{jh}]$, $[\varepsilon^{j\sigma_k^j(h)}]$ et $[\varepsilon^j]$ quand la distribution d'échantillonnage n'est pas incluse dans $[\varepsilon^{j\sigma_k^j(h)}]$, les écart-types sont représentés entre parenthèses.

Robustesse du modèle $[\varepsilon^{j\sigma_k(h)}]$

Si la distribution d'échantillonnage est incluse dans $[\varepsilon_k^{jh}]$ et pas dans $[\varepsilon^{j\sigma_k(h)}]$. Prenons

$$\alpha_1^j = (0,07; 0,37; 0,40; 0,08; 0,08),\ \alpha_2^j = (0,34; 0,30; 0,15; 0,11; 0,10).$$

Tous les autres paramètres restent inchangés. L'objectif de cette simulation est d'évaluer la robustesse du modèle $[\varepsilon^{j\sigma_k^j(h)}]$ par rapport aux mauvaises spécifications. Les résultats sont présentés table 5.3 (l'erreur de Bayes est 5,11%). On voit qu'en supervisé le modèle $[\varepsilon^{j\sigma_k^j(h)}]$ améliore les résultats du modèle $[\varepsilon_k^{jh}]$, tandis qu'il les dégrade en semi-supervisé. Ceci est le résultat attendu puisque la distribution d'échantillonnage est incluse dans le modèle $[\varepsilon_k^{jh}]$, et pas dans le modèle $[\varepsilon^{j\sigma_k^j(h)}]$.

Si la distribution d'échantillonnage est n'incluse dans aucun modèle proposé. Considèrons une problème de classification à deux classes. Considérons 16 variables discrètes dépendantes deux par deux et comportant chacune trois modalités. Notons $\alpha_k^{j,j'}(h,h') = p(X^{jh}X^{j'h'} = 1|Z_k = 1)$. Les fréquences des couples sont $\forall j \in \{1,\ldots,8\}$

$$\alpha_1^{2j-1,2j} = \begin{pmatrix} 0,2 & 0,1 & 0,05 \\ 0,1 & 0,05 & 0,2 \\ 0,15 & 0,05 & 0,1 \end{pmatrix}$$

et

$$\alpha_2^{2j-1,2j} = \begin{pmatrix} 0,1 & 0,05 & 0,15 \\ 0,2 & 0,05 & 0,1 \\ 0,05 & 0,1 & 0,2 \end{pmatrix},$$

et conservons $\pi_1 = \pi_2 = 0,5$, $n_\ell = 50$, $n_u = 500$, $n_{test} = 10000$. Remarquons que l'hypothèse d'indépendance est violée, tous les modèles proposés sont donc faux, cependant la distribution de $X^{jh}|\boldsymbol{Z}$ est la même pour chaque classe à une permutation des modalités près, ce qui est en faveur de $[\varepsilon^{j\sigma_k(h)}]$. Les résultats sont présentés table 5.4 et l'erreur de Bayes est de 7,45%. On voit que, même si le modèle postulé n'est pas correct, le semi-supervisé permet d'améliorer les résultats pour les modèles $[\varepsilon_k^{jh}]$ et $[\varepsilon^{j\sigma_k^j(h)}]$.

Modèles	$[\varepsilon_k^{jh}]$	$[\varepsilon^{j\sigma_k(h)}]$	$[\varepsilon^j]$
$\overline{Err}_s$	21,84 (2,47)	21,49 (2,66)	23,50 (3,27)
$\overline{Err}_{ss}$	17,95 (0,93)	17,14 (0,88)	24,80 (1,62)

TAB. 5.4 – Erreur en classification supervisée ($\overline{Err}_s$) et en classification semi-supervisée ($\overline{Err}_{ss}$) pour les modèles $[\varepsilon_k^{jh}]$, $[\varepsilon^{j\sigma_k^j(h)}]$ et $[\varepsilon^j]$ quand la distribution d'échantillonnage n'est incluse dans aucun modèle, les écarts-types sont représentés entre parenthèses.

5.3.2 Illustration des modèles $[\varepsilon_k]_{bis}$ et $[\varepsilon]_{bis}$ sur données simulées

Données générées selon le modèle $[\varepsilon]_{bis}$

Soit un problème de classification à deux classes. On génère 100 données étiquetées avec $\pi_1 = \pi_2 = 0,5$ et selon le modèle $[\varepsilon]_{bis}$, avec $\varepsilon = 0,8$, $d = 9$ variables, et $\forall j \in \{1, \ldots, d\}$ la variable j comporte $j + 1$ modalités, les positions des modalités majoritaires étant $h(1, j) = 1$ et $h(2, j) = j$. On génère un échantillon test de taille 50000. On compare ici les trois modèles standards $[\varepsilon_k^{jh}]$, $[\varepsilon_k^j]$ et $[\varepsilon^j]$ aux deux modèles proposés $[\varepsilon_k]_{bis}$ et $[\varepsilon]_{bis}$ sur la base de l'erreur de classement.

Les résultats sont présentés table 5.5. Ces résultats montrent effectivement que ce sont les modèles $[\varepsilon]_{bis}$ et $[\varepsilon_k]_{bis}$ qui produisent les meilleures performances, l'écart avec les autres modèles est d'autant plus grand que le nombre de données est petit. Remarquons aussi que le modèle $[\varepsilon_k^{jh}]$ produit des résultats assez médiocres par rapport aux autres, ce qui est dû au trop grand nombre de paramètres estimés.

Modèles	$[\varepsilon_k^{jh}]$	$[\varepsilon_k^j]$	$[\varepsilon^j]$	$[\varepsilon_k]_{bis}$	$[\varepsilon]_{bis}$
$n = 100$	25,93 (1,92)	20,44 (1,15)	20,37 (1,57)	20,07 (1,56)	20,02 (1,69)
$n = 200$	22,28 (0,82)	19,34 (0,36)	19,20 (0,43)	18,94 (0,43)	18,88 (0,42)
$n = 300$	20,98 (0,50)	19,14 (0,21)	18,98 (0,19)	18,85 (0,20)	18,77 (0,14)

TAB. 5.5 – Erreur de classement en classification supervisée avec n données, écarts-types entre parenthèses (moyenne obtenue à partir de 100 réplicats).

Données générées selon le modèle $[\varepsilon_k]_{bis}$

Nous conservons tous les paramètres identiques à la simulation précédente excepté que maintenant, les données sont générées selon le modèle $[\varepsilon_k]_{bis}$ avec $\varepsilon = (0,8; 0,7)$.

Les résultats sont présentés table 5.6. On voit que c'est effectivement le modèle $[\varepsilon_k]_{bis}$ qui produit les meilleurs résultats. Le modèle $[\varepsilon]_{bis}$ produit maintenant des résultats un peu moins bons car il est mal spécifié, les résultats produits par ce modèle restent toutefois meilleurs que ceux des modèles $[\varepsilon_k^{jh}]$ et $[\varepsilon^j]$. En effet même si le modèle $[\varepsilon]_{bis}$ est mal spécifié, il est à préférer au modèle $[\varepsilon_k^{jh}]$, qui lui est bien spécifié, pour des raisons de plus faible variance des paramètres estimés.

Modèles	$[\varepsilon_k^{jh}]$	$[\varepsilon_k^j]$	$[\varepsilon^j]$	$[\varepsilon_k]_{bis}$	$[\varepsilon]_{bis}$
$n = 100$	20,15 (2,10)	14,57 (0,66)	15,45 (1,20)	14,17 (0,81)	15,12 (1,27)
$n = 200$	16,34 (0,72)	14,05 (0,24)	14,58 (0,42)	13,72 (0,23)	14,34 (0,46)
$n = 300$	15,15 (0,48)	13,74 (0,16)	14,23 (0,31)	13,54 (0,15)	14,08 (0,30)

TAB. 5.6 – Erreur de classement en classification supervisée avec n données, écarts-types entre parenthèses (moyenne obtenue à partir de 100 réplicats).

5.3.3 Analyse de séquences ADN

La reconnaissance d'espèces animales à partir de séquences ADN est une méthode de plus en plus répandue. Ainsi Hebert *et al.* (2003) remarque que la sous-unité 1 de la cytochrome c oxydase varie entre espèces proches. La cytochrome c oxydase est une protéine mitochondriale présente chez tous le animaux. La variabilité de cette protéine chez des organismes proches en fait une séquence de choix pour classer des individus. Ainsi en 2004 a été initié le projet *Barcoding of life*[1], qui consiste à séquencer cette protéine pour un grand nombre d'espèces animales, fournissant ainsi un grand nombre de séquences ADN. Ce type de données offre naturellement une perspective semi-supervisée. En effet imaginons que tout l'ADN d'un m^3 d'eau de mer soit séquencé, on obtient alors diverses séquences dont on ne connaît pas la provenance et d'un autre côté un petit nombre de données classées peut être obtenu en effectuant des prélèvements sur des individus d'espèces connues.

Travailler sur une même séquence pour tous les individus permet d'observer les mêmes variables pour chaque individu ; ceci simplifie l'étude statistique. Cependant comme dans toute séquence ADN, il peut y avoir des ajouts ou des délétions de bases azotées. Ainsi pour comparer différentes séquences il est nécessaire de les aligner. Pour cela des outils tels que BLAST (Altschul *et al.* , 1990) ou les profils HMM (Eddy, 1998) sont utilisés. Un exemple d'alignement multiple est présenté table 5.7. Les lettres « A », « C », « G » et « T » représentent chacune des quatre bases azotées, et le symbole « − » représente les blancs introduits pour aligner les séquences. Cette étape effectuée, chaque séquence a la même longueur, c'est-à-dire le même nombre de variables, et chaque variable admet 5 modalités différentes : {A,C,G,T,-}.

Séquence 1 :	G	A	G	C	-	C	C	A	G	T	T	C
Séquence 2 :	-	A	G	G	A	C	-	T	C	T	T	C
Séquence 3 :	A	A	T	C	A	C	C	C	G	A	T	-
Séquence 4 :	-	A	G	G	A	C	-	T	C	T	T	C

TAB. 5.7 – Exemple d'alignement multiple.

Les modèles multinomiaux parcimonieux sont bien adaptés pour résoudre ce type de problème. Ceci d'autant plus que le nombre de variables est grand devant le nombre de données. Ceci rend difficile la prise en compte de corrélations entre variables conditionnellement à la classe.

[1] `http://www.barcodinglife.org/`

Modèle	$[\varepsilon_k^{jh}]$	$[\varepsilon^{j\sigma_k(h)}]$	$[\varepsilon_k]$	$[\varepsilon]$
$\overline{Err}_s$	6,09(5,21)	5,24(5,50)	10,61(9,41)	7,11(11,18)
$\overline{Err}_{ss}$	11,43(8,81)	5,04(10,05)	15,24(10,42)	7,41(11,50)

TAB. 5.8 – Comparaison du supervisé et du semi-supervisé pour les modèles multinomiaux parcimonieux sur les données « Birds of North America, Canadian geese ».

Modèle	$[\varepsilon_k^{jh}]$	$[\varepsilon_k^{j}]$	$[\varepsilon_k]$	$[\varepsilon^{j}]$	$[\varepsilon]$	$[\varepsilon_k]$bis	$[\varepsilon]$bis	$[\varepsilon^{j\sigma_k^j(h)}]$
Taux d'erreur	5,80	7,25	4,35	18,84	18,84	4,35	18,84	7,25
BIC	-287,84	-280,75	-294,90	-301,70	-308,96	-300,41	-315,14	-297,19

TAB. 5.9 – Modèles multinomiaux parcimonieux pour la discrimination des sous espèces *lherminieri* et *subalaris*.

Intéressons nous à l'échantillon « *Birds of North America, Canadian geese* ». Il s'agit d'un échantillon de 141 oies du Canada dont on distingue les sous-espèces *Branta canadensis* (117 spécimens) et *Branta hutchinsii* (24 spécimens). Une fois l'ensemble des séquences alignées elles ont toutes une taille de 901 bases. On cache aléatoirement 90% des étiquettes en imposant qu'au moins un spécimen de chaque sous espèce soit observé, et on répète ce procédé 100 fois. On donne les erreurs moyennes et leurs écarts-types obtenus en % pour différents modèles dans les cas supervisés et semi-supervisés dans la table 5.8.

Le modèle $[\varepsilon^{j\sigma_k(h)}]$ produit les meilleurs résultats, le semi-supervisé permettant une diminution de l'erreur moyenne. Pour les autres modèles le semi-supervisé implique à chaque fois une augmentation de l'erreur de classement moyenne. Cet exemple permet de souligner deux aspects : tout d'abord le modèle $[\varepsilon^{j\sigma_k^j(h)}]$ permet une amélioration des performances en classification supervisée. D'autre part, ce modèle permet une amélioration de performances en utilisant le semi-supervisé.

Certes l'exemple précédent reste un exemple jouet concernant l'usage de la classification semi-supervisée, cependant il illustre bien un champ d'application où des problèmes de classification semi-supervisée vont se poser de plus en plus souvent.

5.3.4 Exemple sur des données provenant d'oiseaux

Un autre exemple de données discrètes consiste en la mesure d'attributs sur des oiseaux. Les données utilisées sont des données sur des oiseaux, et sont disponibles dans le logiciel MIXMOD. L'objectif est de discriminer deux sous-espèces d'oiseaux : *lherminieri* et *subalaris*, selon des caractéristiques morphologiques : sourcil allant de absent à très prononcé, collier allant de absent à continu, zébrures allant de absent à présence forte, sous caudales blanc ou noir ou noir & blanc ou noir & BLANC ou NOIR & blanc, liseré allant de absent à beaucoup. On supprime les modalités qui n'apparaissent jamais dans l'échantillon et la variable qui ne présente qu'une seule modalité.

On voit table 5.9 que les modèles $[\varepsilon_k]$ et $[\varepsilon_k]_{bis}$ produisent ici les meilleures performances en terme d'erreur de classement estimée par *leave-one-out*. Le critère BIC sélectionne le modèle $[\varepsilon_k^j]$ qui produit une faible erreur mais pas la plus faible possible.

Modèle	$[\varepsilon_k^{jh}]$	$[\varepsilon_k^j]$	$[\varepsilon_k]$	$[\varepsilon^j]$	$[\varepsilon]$	$[\varepsilon_k]_{bis}$	$[\varepsilon]_{bis}$	$[\varepsilon^{j\sigma_k^j(h)}]$
Taux d'erreur	13,07	28,10	34,64	39,87	39,87	32,68	35,95	15,69
BIC	-798,62	-848,78	-874,55	-876,84	-895,32	-875,05	-897,97	-789,85

TAB. 5.10 – Modèles multinomiaux parcimonieux pour la discrimination des sous espèces *lherminieri*, *subalaris* et *dichrous*.

Ajoutons aux deux sous-espèces précédentes la sous-espèce *dichrous*. On voit table 5.10 que le critère BIC sélectionne le modèle $[\varepsilon^{j\sigma_k^j(h)}]$. Celui-ci produit une erreur un peu plus élevée que le meilleur modèle.

Comme on vient de le voir sur cet exemple BIC sélectionne des modèles trop simples par rapport au modèle qui produit la plus faible erreur de classement possible. Jusqu'alors nous n'avons compté que les paramètres continus dans la pénalisation. Cependant les modèles parcimonieux comportent en plus des paramètres continus un certain nombre de paramètres discrets comme par exemple la position de la modalité majoritaire. Ainsi, dans la section suivante nous étudions la question « Comment prendre en compte les paramètres discrets dans l'approximation BIC ? ».

5.4 Comptage de paramètres

5.4.1 Introduction : $\text{BIC}_{standard}$ et BIC_{exact}

Pour simplifier le raisonnement prenons $g = 1$ et $d = 1$. Ici on supprime des notations les symboles k et j. On suppose $\boldsymbol{X} \sim \mathcal{M}(1; \alpha_1, \ldots, \alpha_m)$. Le modèle $[\varepsilon]$, suppose qu'il existe une modalité majoritaire notée h^*, telle que

$$\alpha_h = \begin{cases} 1-\varepsilon & \text{si } h = h^*, \\ \frac{\varepsilon}{m-1} & \text{sinon.} \end{cases}$$

Il y a ici un paramètre discret (h^*), et un paramètre continu (ε). On impose $\varepsilon \leqslant \frac{m-1}{m}$, dans le cas contraire h^* n'est pas la position de la modalité majoritaire. Soit $p(\varepsilon, h^*)$ la distribution *a priori* de (ε, h^*), qu'on peut réécrire sous la forme $p(\varepsilon|h^*)p(h^*)$. Supposons toutes les positions modales *a priori* équiprobables et ε indépendant de h^*, on a

$$p(\varepsilon, h^*) = \frac{1}{m} p(\varepsilon).$$

La vraisemblance intégrée est alors

$$\frac{1}{m} \sum_{h^*=1}^{m} \int_0^{\frac{m-1}{m}} p(\mathbf{x}|\varepsilon, h^*) p(\varepsilon) d\varepsilon. \tag{5.4}$$

On peut effectuer une intégration numérique par rapport à ε. Prenons $p(\varepsilon) \propto \varepsilon^{-1/2}(1-\varepsilon)^{-1/2}\mathbf{1}_{0\leqslant\varepsilon\leqslant\frac{m-1}{m}}$ (on notera $C = \int_0^{\frac{m-1}{m}} \varepsilon^{-1/2}(1-\varepsilon)^{-1/2} d\varepsilon$), qui est un *a priori* de Dirichlet tronqué. Notons $n_h = \sum_{i=1}^{n} x_h$ et $\overline{n}_h = n - n_h$.

$$\boxed{BIC_{exact} = \log\left(\frac{1}{m}\sum_{h=1}^{m}\int_0^{\frac{m-1}{m}} \frac{1}{C}(1-\varepsilon)^{n_h-\frac{1}{2}}\left(\frac{\varepsilon}{m-1}\right)^{\overline{n}_h}\varepsilon^{-\frac{1}{2}}d\varepsilon\right).} \tag{5.5}$$

On appelle le critère obtenu BIC_{exact}. Il nécessite une intégration numérique pour être calculé, ce qui peut être coûteux en grande dimension.

En notant $\hat{h}^* = \text{argmax}_h\, n_h$, et $\hat{\varepsilon} = 1 - \frac{n_{h^*}}{n}$, l'approximation BIC standard sans la prise en compte du paramètre discret dans la pénalisation s'écrit

$$\boxed{BIC_{standard} = \log\left(\frac{1}{m}\sum_{h=1}^{m}(1-\hat{\varepsilon})^{n_{h^*}}\left(\frac{\hat{\varepsilon}}{m-1}\right)^{\overline{n}_{h^*}}\right) - \frac{1}{2}\log n.} \tag{5.6}$$

Dans ce qui suit on cherche une stratégie moins coûteuse que BIC_{exact}, mais qui prend en compte le paramètre discret dans la pénalité.

5.4.2 Cas où la contrainte est saturée

Cherchons maintenant à approcher $\log\left(\frac{1}{m}\sum_{h^*=1}^{m}\int_0^{\frac{m-1}{m}} p(\mathbf{x}|\varepsilon, h^*)p(\varepsilon)d\varepsilon\right)$. Remarquons que si le maximum est atteint au niveau de la contrainte, l'approximation de Laplace habituellement utilisée dans l'approximation BIC n'est plus valide (Lebarbier & Mary-Huard, 2006). Nous traitons ici ce cas ($\varepsilon = \frac{m-1}{m}$). Pour cela on utilise le lemme suivant.

Lemme 5.1. *Soit* $L : [a,b] \to \mathbb{R}$ *telle que* L *soit différentiable une fois sur* $[a,b]$ *et atteint son maximum en* b *avec* $L'(b) > 0$. *Alors*

$$\int_a^b e^{nL(u)}du = \frac{e^{nL(b)}}{nL'(b)}[1 + o(n^{-1})]. \tag{5.7}$$

PREUVE : Puisque L est différentiable dans un voisinage de b, on a

$$L(u) = L(b) + (u-b)L'(b) + no((u-b)^2). \tag{5.8}$$

Ainsi,

$$\int_a^b e^{nL(u)}du = e^{nL(b)}\int_a^b e^{n(u-b)L'(b)}e^{no((u-b)^2)}du. \tag{5.9}$$

Puis en utilisant le développement limité de e^u au voisinage de 0 ; $e^u = 1 + O(u)$

$$\int_a^b e^{n(u-b)L'(b)}e^{no((u-b)^2)}du = \int_a^b e^{n(u-b)L'(b)}du + \int_a^b no((u-b)^2)e^{n(u-b)L'(b)}du. \tag{5.10}$$

On a alors

$$\int_a^b e^{n(u-b)L'(b)}du = \left[\frac{1}{nL'(b)}e^{n(u-b)L'(b)}\right]_a^b = \frac{1-e^{n(a-b)L'(b)}}{nL'(b)}. \tag{5.11}$$

Puisque $a - b < 0$ et $L'(b) > 0$, $e^{n(a-b)L'(b)} = o(n^{-c})$ avec c arbitrairement grand.

Il nous suffit maintenant de calculer $\int_a^b n(u-b)^2 e^{n(u-b)L'(b)}du$:

$$\int_a^b n(u-b)^2 e^{n(u-b)L'(b)}du = \mathcal{O}(n^{-2}), \tag{5.12}$$

et par conséquent

$$\int_a^b e^{nL(u)}du = \frac{e^{nL(b)}}{nL'(b)}[1 + o(n^{-1})], \tag{5.13}$$

ce qui conclut la preuve. □

Cela est exactement ce qui se passe dans le cas contraint. Ainsi on obtient

$$\log \int_a^b e^{nL(u)} du = nL(b) + \log n + O(1). \tag{5.14}$$

Là où l'approximation de Laplace donnerait

$$\log \int_a^b e^{nL(u)} du = nL(c) + \frac{1}{2}\log n + O(1), \tag{5.15}$$

dans le cas où L atteint son maximum en $c \in]a, b[$.

Ce résultat reste valable pour les fonctions L qui dépendent de n sous certaines conditions (Lebarbier & Mary-Huard, 2006). On voit donc que la pénalité est $-\log n$ au lieu de $-\frac{1}{2}\log n$. On interprète la saturation de la contrainte comme l'ajout d'un paramètre supplémentaire. Ceci est assez intuitif d'un point de vue choix de modèle puisqu'on souhaite surpénaliser les modèles pour lesquels les contraintes ne sont pas automatiquement satisfaites.

Ainsi, en appliquant le lemme 5.1 et sous certaines conditions de régularité supplémentaires

$$\log \int p(\mathbf{x}|\varepsilon, h^* = h)p(\varepsilon)d\varepsilon = \log p(\mathbf{x}|\hat{\varepsilon}_{|h^*=h}, h^* = h) - \frac{1+s_h}{2}\log n + \mathcal{O}(1), \tag{5.16}$$

avec $\hat{\varepsilon}_{|h^*=h}$ l'estimateur du maximum de vraisemblance de ε sous la contrainte que h est la position du mode, $s_h = 1$ si la contrainte est saturée ($\hat{\varepsilon}_{|h^*=h} = \frac{m-1}{m}$) et 0 sinon. Remarquons que $\hat{\varepsilon}_{|h^*=h} = \min\left(1 - \frac{n_h}{n}, \frac{m-1}{m}\right)$.

5.4.3 BIC comptant les deux types de paramètres : BIC$_{propose}$

On en déduit que

$$\log\left(\frac{1}{m}\sum_{h=1}^{m}\int p(\mathbf{x}|\varepsilon, h^* = h)p(\varepsilon)d\varepsilon\right) = \log\left(\frac{1}{m}\sum_{h=1}^{m} e^{\log p(\mathbf{x}|\hat{\varepsilon}_{|h^*=h}, h^*=h) - \frac{1+s_h}{2}\log n}\right) + \mathcal{O}(1). \tag{5.17}$$

On appelle le critère résultant BIC proposé.

$$\boxed{BIC_{propose} = \log\left(\frac{1}{m}\sum_{h=1}^{m}(1-\hat{\varepsilon}_{|h^*=h})^{n_h}\left(\frac{\hat{\varepsilon}_{|h^*=h}}{m-1}\right)^{\overline{n}_h} n^{-\frac{1+s_h}{2}}\right).} \tag{5.18}$$

On recommande d'utiliser ce critère en pratique puisque le calcul de la somme à l'intérieur du logarithme va permettre de prendre en compte la complexitée liée à l'estimation du paramètre discret.

5.4.4 Simplification de BIC$_{propose}$: BIC$_{surpenalise}$

BIC$_{propose}$ est coûteux à calculer, cependant dans la somme à l'intérieur du logarithme pour n assez grand un seul terme va dominer de telle sorte que

$$\log \int p(\mathbf{x}|\varepsilon, h^* = h)p(\varepsilon|h^* = h)d\varepsilon = \log\left(\frac{1}{m}e^{\log p(\mathbf{x}|\hat{\varepsilon}_{|h^*=h}, h^*=h) - \frac{1}{2}\log n}\right) + \mathcal{O}(1) \tag{5.19}$$

et par conséquent on obtient l'approximation BIC dite surpénalisée étant donnée la présence de $\log m$:

$$\boxed{BIC_{surpenalise} = BIC_{standard} - \log m.} \tag{5.20}$$

Remarquons que le terme $\log m$ est de l'ordre $O(1)$. Ceci est l'ordre de grandeur de l'approximation BIC. Il peut donc être négligé en théorie, ce qui donne alors le critère BIC standard.

Dans le cas du modèle libre, le critère BIC est :

$$BIC(M_1) = \log p(\mathbf{x}|\hat{\alpha}) - \frac{m-1}{2}\log n. \tag{5.21}$$

avec $\hat{\alpha}$ l'estimateur non contraint de α.

Dans le cas où le terme $-\log m$ est conservé pour le critère BIC du modèle contraint, il faut imposer

$$\frac{1}{2}\log n + \log m \leqslant \frac{m-1}{2}\log n, \tag{5.22}$$

pour rester cohérent. En effet, il serait paradoxal de pénaliser plus fortement un modèle plus simple. L'équation (5.22) est vérifiée du moment que n et/ou m sont assez grands.

5.4.5 Expériences

Comparaison du modèle parcimonieux et du modèle complet

Dans ce qui suit on notera, M_1 le modèle complet, M_2 le modèle parcimonieux avec modalité majoritaire, M_3 le modèle naïf qui suppose que toutes les modalités sont en proportions identiques.

Soit $\boldsymbol{X} \sim \mathcal{M}(1; 0,40; 0,35; 0.25)$. On met en compétition M_1 et M_2 et on s'intéresse au nombre de fois où le vrai modèle ici le modèle M_1 est sélectionné par les différents critères obtenus précédemment, ceci sur 1000 réplications et pour des valeurs de n allant de 50 à 500. Les résultats sont illustrés figure 5.1. On voit que le critère le plus proche de BIC exact est le critère définit équation (5.17). Le critère BIC usuel semble sous-pénaliser le modèle parcimonieux, tandis que le critère BIC surpénalisé le surpénalise. Cette surpénalisation est d'autant plus grande que le nombre de données est petit.

Soit $\boldsymbol{X} \sim \mathcal{M}(1; 0,40; 0,30; 0.30)$. On met en compétition M_1 et M_2 et on s'intéresse au nombre de fois où le vrai modèle ici le modèle M_2 est sélectionné par les différents critères obtenus précédemment sur 1000 réplications et pour des valeurs de n allant de 50 à 500. Les résultats sont illustrés figure 5.2. On voit que le critère le plus proche de BIC exact est le critère définit équation 5.17. Le critère BIC usuel sélectionne plus souvent le modèle parcimonieux puisque celui-ci est sous-pénalisé. Le critère BIC surpénalisé sélectionne moins souvent le modèle parcimonieux. Cette surpénalisation est d'autant plus grande que le nombre de données est petit.

Comparaison du modèle complet, du modèle parcimonieux, et du modèle d'équidistribution

On met maintenant les modèles M_1, M_2 et M_3 en compétition.

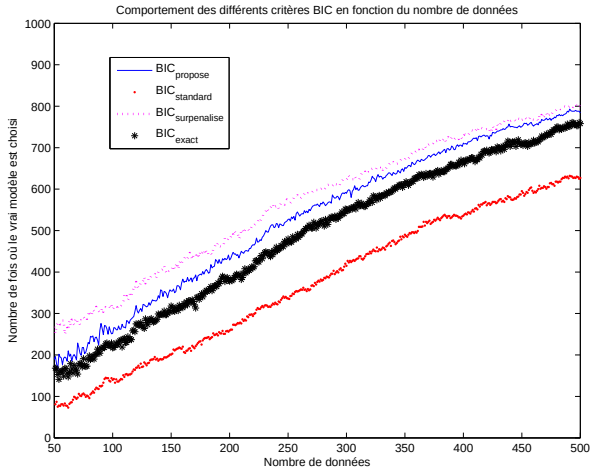

FIG. 5.1 – Nombre de fois où le modèle complet est choisi en fonction des valeurs n quand les données sont issues du modèle complet.

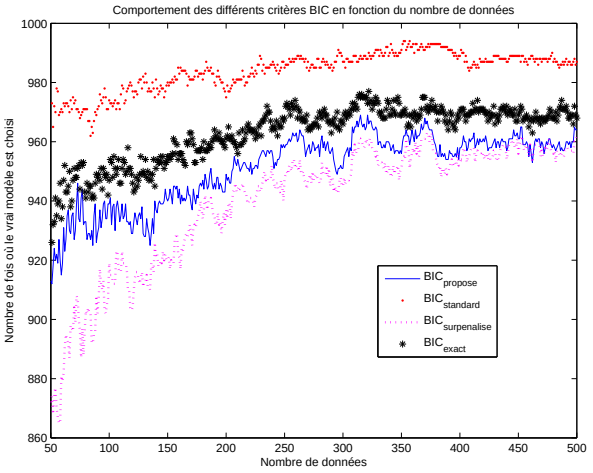

FIG. 5.2 – Nombre de fois où le modèle complet est choisi en fonction des valeurs n quand les données sont issues du modèle parcimonieux.

Si M_1 est correct. Le paramètre de simulation a pour valeur $\alpha = (0,40; 0,35; 0,25)$. Le nombre de fois où chaque modèle est choisi en fonction du nombre de données n et du critère de choix de modèle est représenté figure 5.3. On voit que le critère BIC standard sélectionne moins rapidement le bon modèle que les autres versions de BIC.

Si M_2 est correct. Le paramètre de simulation a pour valeur $\alpha = (0,40; 0,30; 0,30)$. Le nombre de fois où chaque modèle est choisi en fonction du nombre de données n et du critère de choix de modèle est représenté figure 5.4. Le critère BIC standard sélectionne plus rapidement le modèle M_2 que le autres critère, ceci bien évidemment puisqu'il sous-pénalise ce dernier.

Si M_3 est correct. Le paramètre de simulation a pour valeur $\alpha = (0,33; 0,33; 0,33)$. Le nombre de fois où chaque modèle est choisi en fonction du nombre de données n et du critère de choix de modèle est représenté figure 5.5. Le critère BIC standard sélectionne moins rapidement le bon modèle que les autres versions de BIC.

5.4.6 Conclusion

Remarquons que dans le cas où la contrainte est saturée le nombre de paramètres dans la pénalité est augmentée de 1. Ainsi, dans le cas où plusieurs contraintes sont saturées, il semblerait naturel d'augmenter le nombre de paramètres dans la pénalité par le nombre de contraintes saturées. Il serait par la suite intéressant d'étendre ce résultat à des situations plus complexes.

Remarquons que si la combinatoire sur les positions du mode est grande, le terme résiduel devrait être de l'ordre du log de cette combinatoire, ce qui peut poser des questions en pratique si la dimension est grande devant le nombre de variables. Il est souvent plus

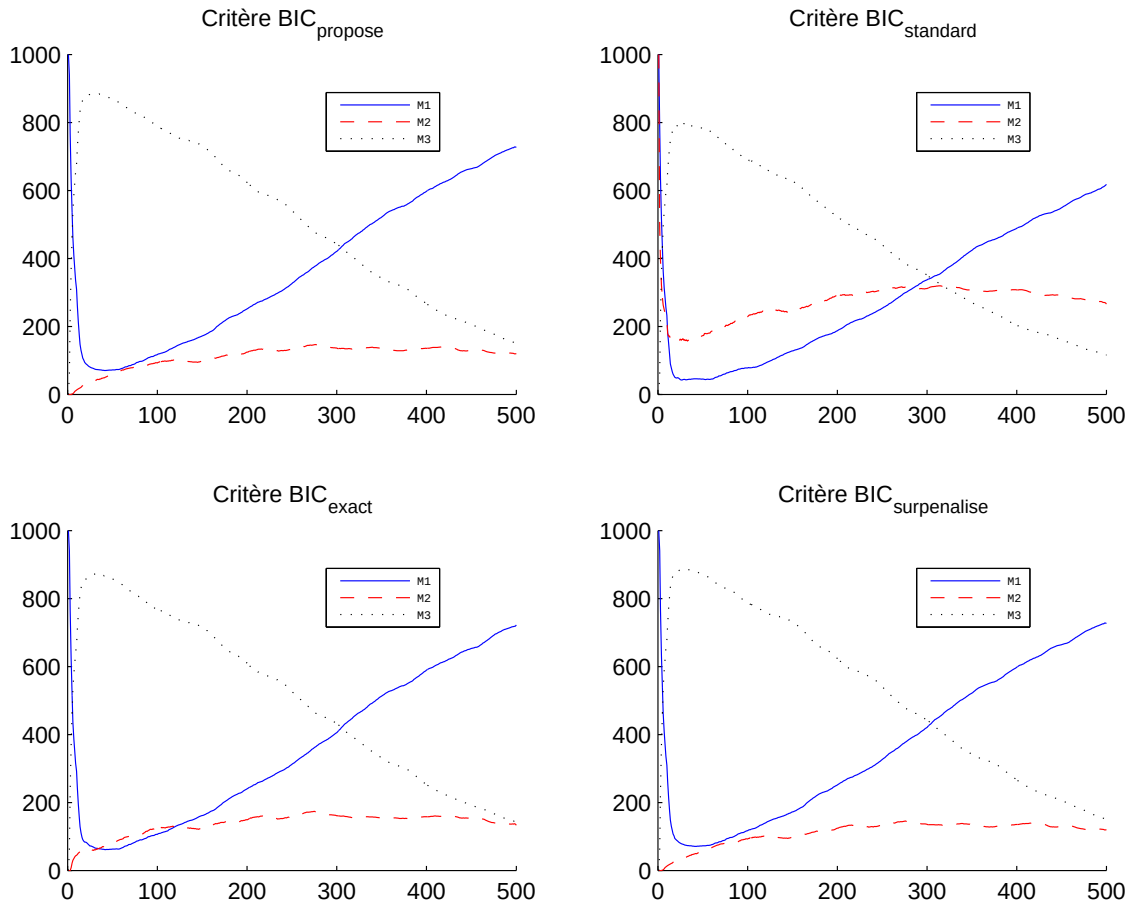

FIG. 5.3 – Choix entre les modèles $M1$, $M2$ et $M3$ selon le nombre de données et le critère utilisé, dans le cas où les données sont issues du modèle $M1$.

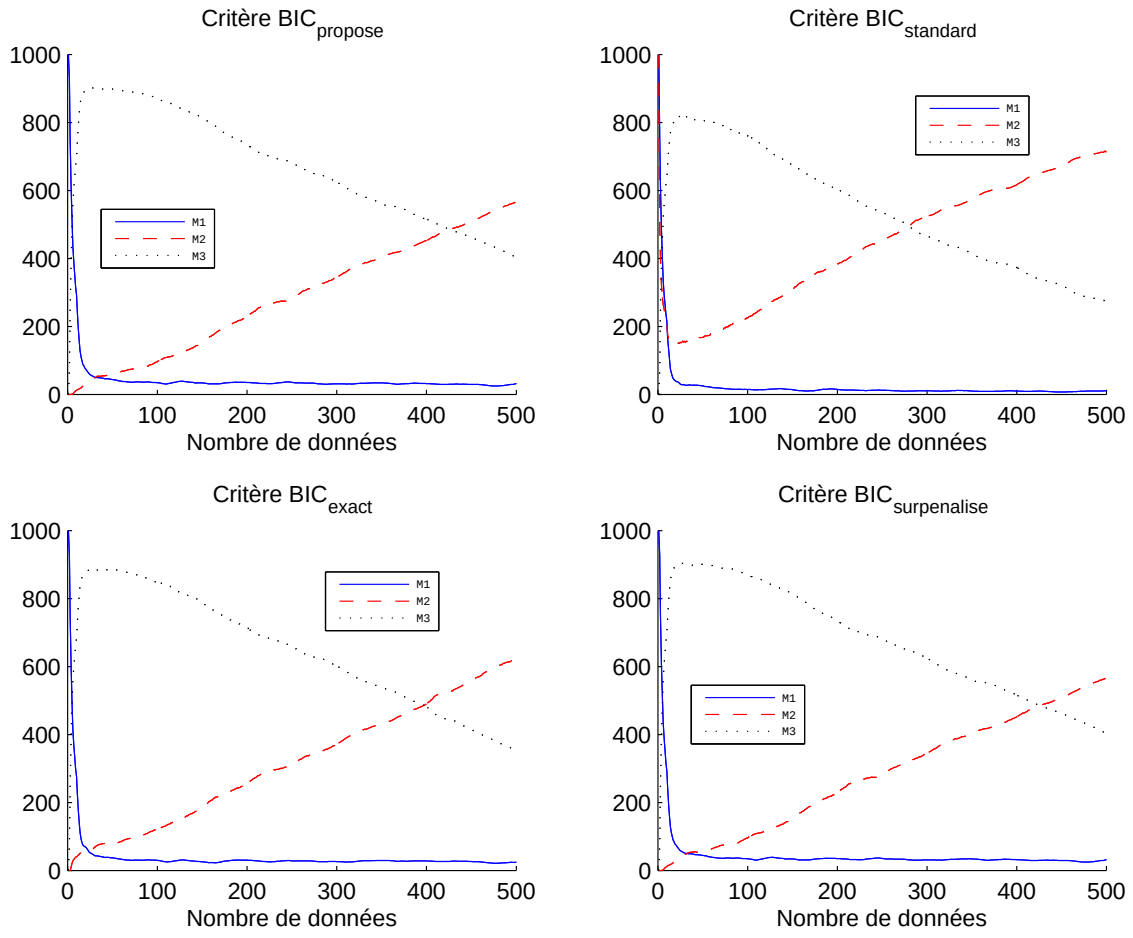

FIG. 5.4 – Choix entre les modèles $M1$, $M2$ et $M3$ selon le nombre de données et le critère utilisé, dans le cas où les données sont issues du modèle $M2$.

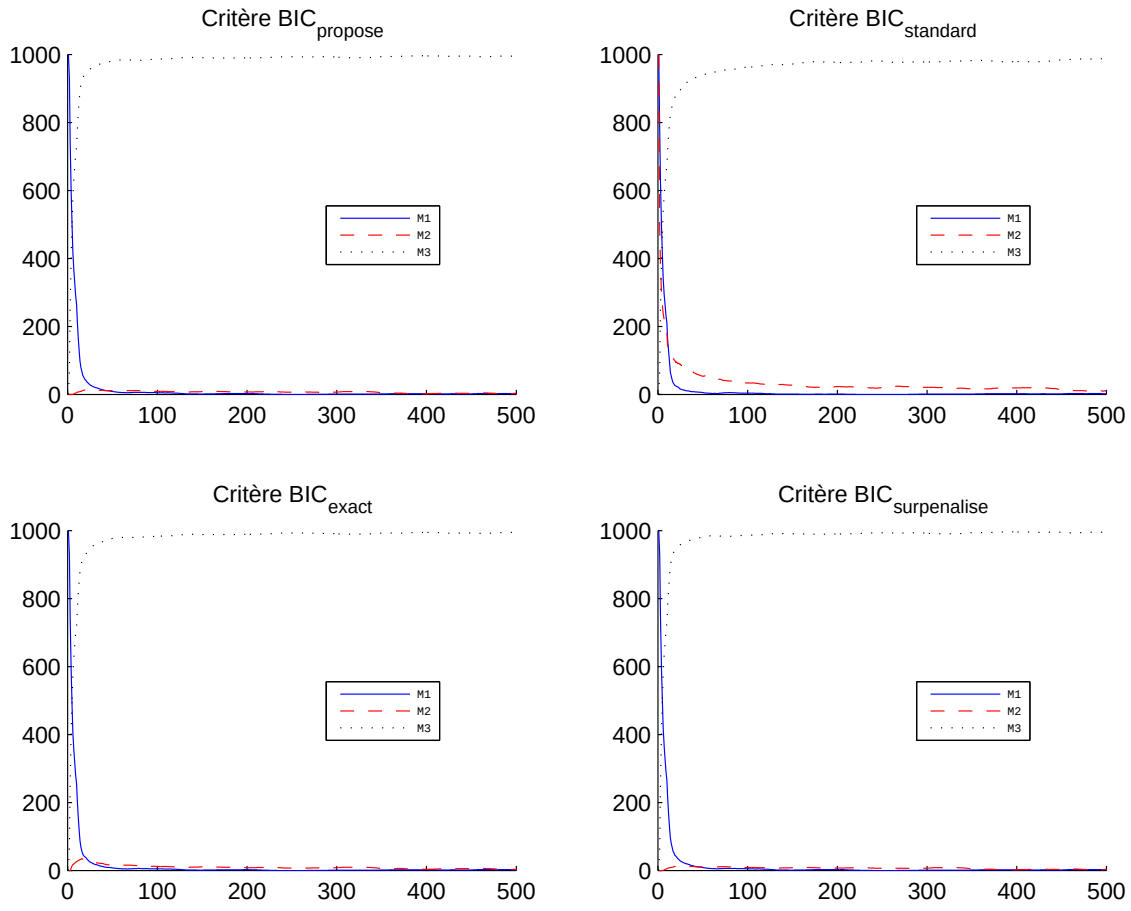

FIG. 5.5 – Choix entre les modèles $M1$, $M2$ et $M3$ selon le nombre de données et le critère utilisé, dans le cas où les données sont issues du modèle $M1$.

simple de ne pas compter le nombre de paramètres discrets, ce qui est asymptotiquement justifié, cependant cette approche peut favoriser à tort les modèles comportants des paramètres discrets. Dans les situations où le calcul de la somme à l'intérieur du logarithme est possible nous conseillons d'effectuer ce calcul qui permet une meilleure prise en compte de la complexité associée aux paramètres discrets.

5.5 Perspectives

5.5.1 Modèles parcimonieux

Extension à plusieurs modalités « majoritaires »

En considérant de 1 à m_j modalités en plus grandes proportions (modalités « majoritaires ») et en égalisant les modalités avec les plus faibles proportions, on passe du modèle $[\varepsilon_k^j]$ au modèle $[\varepsilon_k^{jh}]$. De plus, on peut autoriser le nombre de modalités « majoritaires » à varier selon la variable et la classe considérée. Si le nombre de variables est grand, il est impossible d'explorer tous les modèles intermédiaires. Cependant, en supervisé la factorisation de BIC pour les différentes variables des différentes classes peut aider à faire ce choix. En effet en notant $n_k^{jh} = \sum_{i=1}^n z_{ik} x_i^{jh}$, et en supposants $p(\alpha) = \prod_{k=1}^g \prod_{j=1}^{m_j} p(\alpha_k^j)$, la log-vraisemblance intégrée pour chaque variable dans chaque classe qu'on appelle par abus de language $BIC(j,k)$ est égale à

$$BIC(j,k) = \log\left(\int \prod_{h=1}^{m_j} {\alpha_k^{jh}}^{n_k^{jh}} p(\alpha_k^j) d\alpha_k^j\right), \tag{5.23}$$

pour chaque variable dans chaque classe. Le critère BIC global s'écrit alors

$$BIC = \sum_{k=1}^{g} \sum_{j=1}^{m_j} BIC(j,k). \tag{5.24}$$

Ainsi le problème se résume à optimiser $BIC(j,k)$ pour chaque variable dans chaque classe. Remarquons que l'extension d'une à plusieurs modalités « majoritaires » est possible pour tous les modèles parcimonieux précédents. Cependant ce problème reste ouvert puisqu'il faut dans ce cas redéfinir la façon de répartir la masse de probabilité restante entre les modalités minoritaires.

Extension du modèle $[\varepsilon^{j\sigma_k^j(h)}]$

Dans le cas où toutes les variables ont le même nombre de modalités, on étend facilement le modèle $[\varepsilon^{j\sigma_k^j(h)}]$ pour obtenir la famille de modèles suivante :

- $[\varepsilon^{\sigma_k^j(h)}]$: on suppose que le vecteur ε^j est identique pour toutes les classes et pour toutes les variables à une permutation des modalités près.
- $[\varepsilon_k^{\sigma^j(h)}]$: on suppose que le vecteur ε^j est identique pour toutes les variables dans une classe fixée à une permutation des modalités près.
- $[\varepsilon_k^h]$: on suppose que le vecteur α_k^j est identique pour toutes les variables dans une classe fixée.

L'étape M pour ces modèle est alors explicite.

Extension au cas continu

L'ordonnancement du vecteur des proportions et l'égalisation des plus petites valeurs est à relier aux modèles parcimonieux en haute dimension (Bouveyron *et al.* , 2007) vus chapitre 2 section 2.3.1. Pour ceux-ci, les valeurs propres sont ordonnées et des contraintes d'égalité sont imposées entre elles. Pour le modèle où p_k (le nombre de plus grandes valeurs propres différentes) est autorisé à varier entre les classes, mais où les b_k (valeurs des plus petites valeurs propres) sont contraintes à être identiques pour toutes les classes, des questions similaires de validité automatique des contraintes peuvent se poser.

Le modèle $[\varepsilon^{j\sigma_k^j(h)}]$, correspond à permuter les modalités en fonction de la classe. Dans le cas continu on peut envisager la situation où toutes les classes ont la même distribution à une permutation des variables près. Ce type de modèle peut éventuellement être utilisé dans les situations où des symétries entre variables existent (Bérard *et al.* , 2009).

5.5.2 Approximation BIC

Prise en compte de plusieurs contraintes

Nous nous sommes contentés d'un exemple jouet où une seule contrainte est considérée. Il serait intéressant d'étudier des situations où plusieurs contraintes sont saturées et de voir les conséquences sur l'approximation BIC. Dans ce cas des questions calculatoires peuvent apparaître puisque la somme effectuée peut comprendre un très grand nombre de termes.

Extension pour AIC

Nous nous sommes principalement posé la question de la sélection de modèle à partir du critère BIC. Cependant, il serait intéressant d'étudier ce qu'impliquent les paramètres discrets aux niveau de l'approximation AIC.

Conclusion et perspectives

Conclusion

Dans un premier chapitre nous avons dressé un état de l'art de la classification semi-supervisée. Celui-ci nous a permis de mettre en évidence la diversité des questions que pose ce cadre, ainsi que d'étudier les nombreuses méthodes développées pour y répondre. Cela justifie alors notre parti pris pour les modèles génératifs puisque ceux-ci permettent de prendre en compte de façon rigoureuse l'information apportée à la fois par les données étiquetées et non étiquetées. L'utilisation de ces modèles a fait l'objet d'un second chapitre dans lequel nous avons détaillé plus en avant les modèles génératifs utilisés ainsi que l'estimation de leurs paramètres à travers la mise en œuvre de l'algorithme EM. Les expérimentations réalisées dans ce chapitre nous ont permis de constater que les résultats obtenus pouvaient varier selon le modèle utilisé. Dans le cas où le semi-supervisé dégrade les performances du supervisé, c'est que le modèle postulé n'est pas pertinent. Ainsi, dans un quatrième chapitre nous avons proposé un test statistique qui fait usage des données non étiquetées pour juger de la pertinence d'un modèle. Ce test a ensuite été reformulé comme une choix de modèle et permet alors, dans le cas où plusieurs modèles sont considérés de choisir à la fois le modèle et de vérifier si celui-ci est bien fondé. Toutefois, la procédure de choix de modèle précédente repose sur l'utilisation du critère BIC, or ce dernier ne prend pas directement en compte l'objectif décisionnel et peut par conséquent conduire à des performances médiocres en classification. C'est pourquoi, dans un quatrième chapitre nous avons proposé un critère de choix de modèle basé sur une approximation de la déviance conditionnelle que nous avons nommé AIC_{cond}, et qui prend en compte l'objectif décisionnel contrairement à BIC. Celui-ci est comparable au critère BEC proposé par Bouchard & Celeux (2006), mais corrige les défauts de ce dernier lorsque des modèles corrects emboîtés sont considérés, puisqu'il privilégie en moyenne le modèle le plus simple, là où BEC n'en privilégie aucun. En outre, dans le cadre semi-supervisé il est moins coûteux que la validation croisée tout en produisant des résultats similaires. Dans un cadre supervisé, celui-ci devient nettement moins attractif pour des raisons de coût. Nous avons alors proposé dans ce cadre deux extensions d'AIC_{cond} peu coûteuses, la première reposant sur le comptage du nombre de paramètres du modèle génératif impliqués dans la prédiction, et la seconde reposant sur la vitesse de convergence de EM lorsque le problème de classification supervisée est artificiellement rendu semi-supervisé.

Dans un cinquième chapitre nous avons apporté notre contribution à l'étude des modèles multinomiaux. Nous avons proposé trois variantes des modèles parcimonieux pour les produits de distributions multinomiales utilisés dans le logiciel MIXMOD (Biernacki *et al.* , 2006). Les deux premières, évitent que les contraintes puissent être saturées lors de

l'estimation des paramètres. Le troisième modèle suppose que les paramètres des multinomiales sont les mêmes pour chaque classe à une permutation des modalités près. Celui-ci a été utilisé avec succès sur des données de séquences ADN. Ces modèles font intervenir un paramètre discret, ici la position de la modalité majoritaire, nous avons alors étudié la question de la prise en compte des paramètres discrets dans l'approximation BIC et montré que lorsque cela était possible il y avait tout intérêt à intégrer la vraisemblance par rapport à ce paramètre discret.

Perspectives

Les principales perspectives de notre travail se situent au niveau de la question du choix de modèle. Dans tout le chapitre 3 nous avons systématiquement eu recours à l'approximation BIC. Il est souvent souhaitable de ne pas avoir recours à l'approximation BIC mais d'intégrer effectivement la vraisemblance en utilisant des méthodes de Monte-Carlo. Par exemple dans le cas des modèles multinomiaux une perspective intéressante est le calcul de cette intégrale en choisissant un *a priori* de Dirichlet (Biernacki *et al.* , 2008). Dans ce chapitre 3 nous avons aussi ébauché une stratégie d'élargissement adaptatif de la liste des modèles mis en compétition. Il serait par la suite intéressant d'explorer plus avant cette stratégie aussi bien d'un point de vue pratique que d'un point de vue théorique. D'autre part tout au long de ce travail nous avons choisi le parti pris des modèles génératifs. Cependant, dans le cas où l'on dispose à la fois de modèles génératifs et prédictifs, des idées similaires à celles développées dans le chapitre 3 pourraient être mises en œuvre pour juger du bien fondé d'un modèle génératif par rapport à un modèle prédictif. Dans le chapitre 4 une extension du critère à AIC_{cond} a été proposée à partir de la vitesse de convergence de EM dans le cadre supervisé. Il serait intéressant de développer une stratégie permettant d'estimer cette vitesse de façon plus stable que la méthode proposée à l'heure actuelle. Dans le chapitre 5 nous avons considéré l'approximation BIC lorsque des paramètres discrets entraient en compte. Nous nous sommes limités dans ce cas à un exemple très simple. Il serait alors intéressant d'implémenter cette méthode pour des applications réelles où l'on dispose potentiellement de nombreuses variables.

D'autres perspectives concernent le traitement des questions spécifiques à la classification semi-supervisée. Les données étiquetées fournissent une partition externe qu'il serait intéressant d'utiliser de manière similaire à Baudry & Celeux (2009). En effet, du moment que des étiquettes sont observées, les prendre en compte dans l'estimation reviendrait à imposer une signification précise à la variable devant expliquer l'hétérogénéité des données. Par une approche similaire à celle utilisée au chapitre 3 on devrait pouvoir répondre à la question, « La variable observée explique-t-elle l'hétérogénéité des données ? ». Pour cela il suffit de définir une factorisation de la vraisemblance où la variable observée n'explique pas l'hétérogénéité des données. Cette factorisation étant ensuite mise en compétition avec le modèle où la variable latente coïncide avec la classe observée. Dans de nombreuses situations réelles l'hypothèse MCAR n'est pas vérifiée. Les données étiquetées et non étiquetées n'ont alors pas la même distribution. Dans ce cas, une bonne partie des résultats standards ne sont plus valides. Cette situation pose des questions pour lesquelles les méthodes génératives sont bien adaptées. Dans ce cas il serait intéressant de considérer plus précisément ce contexte et d'étudier les questions posées. Nous avons principalement abordé le problème sous un angle théorique, et testé nos méthodes sur des données simulées et

provenant de *Benchmarks*. Ceci est le cas de nombreuses méthodes d'apprentissage statistique qui n'interviennent qu'en fin de traitement des données. Par la suite on s'attachera à chercher des applications réelles motivant à la base les développements théoriques.

Enfin des perspectives s'offrent à nous concernant la proposition de modèles parcimonieux. Dans le chapitre 5 nous avons proposé de nouvelles paramétrisations parcimonieuses dans le cas multinomial. Ces dernières pourraient à terme être intégrées dans le logiciel MIXMOD (Biernacki *et al.* , 2006). Le modèle de permutation des modalités conditionnellement à la classe proposé au chapitre 5 pourrait d'une part générer toute un famille de modèles parcimonieux dans le cas où toutes les variables ont le même nombre de modalités, ce qui est par exemple le cas pour les séquences ADN. D'autre part ce modèle pourrait être étendu au cas gaussien en considérant une permutation des variables conditionnellement à la classe.

Bibliographie

Agrawala, A. 1970. Learning with a probabilistic teacher. *Information Theory, IEEE Transactions on*, **16**(4), 373–379.

Aitchison, J., & Dunsmore, I. R. 1975. *Statistical Prediction Analysis*. Cambridge University Press.

Allman, E. S., Matias, C., & Rhodes, J. A. 2009. Identifiability of latent class models with many observed variables. *Annals of Statistics*, 3099–3132.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of molecular biology*, **215**(3), 403–410.

Amemiya, T. 1973. Generalized least squares with an estimated autocovariance matrix. *Econometrica : Journal of the Econometric Society*, 723–732.

Amini, M., & Gallinari, P. 2002. Semi-Supervised Logistic Regression. *In : Proceedings of the 15th European Conference on Artificial Intelligence (ECAI'02)*.

Anderson, J. A., & Richardson, S. C. 1979. Logistic Discrimination and Bias Correction in Maximum Likelihood Estimation. *Technometrics*, **21**(1), 71–78.

Anderson, T. 2003. *An Introduction to Multivariate Statistical Analysis*. Wiley, 3rd edition.

Aronszajn, N. 1950. Theory of reproducing kernels. *Transactions of the American mathematical society*, **68**(3), 337–404.

Bach, F. 2007. Active learning for misspecified generalized linear models. *In :* Schölkopf, B., Platt, J., & Hoffman, T. (eds), *Advances in Neural Information Processing Systems 19*. Cambridge, MA : MIT Press.

Baudry, J.P., & Celeux, G. 2009. Sélection de modèle pour la classification en présence d'un classification externe. *In : Archives ouvertes 41es Journées de Statistique, SFdS, Bordeaux*.

Baum, L. E., Petriz, T., & Weiss, N. 1970. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

Bazell, D., & Miller, D.J. 2005. Class discovery in galaxy classification. *The Astrophysical Journal*, **618**(2), 723–732.

Bensmail, H., & Celeux, G. 1996. Regularized discriminant analysis. *Journal of the American Statistical Association*, **91**, 1743–1748.

Bérard, C., Martin-Magniette, M.L., To, A., Roudier, F., Colot, V., & Robin, S. 2009. Mélanges gaussiens bidimensionnels pour la comparaison de deux échantillons de chromatine immunoprécipitée. *Revue Modulad*, **41**, 53–68.

Beygelzimer, A., Dasgupta, S., & Langford, J. 2009. Importance Weighted Active Learning. *In : Proceeding of ICML 2009.*

Bie, T. De, & Cristianini, N. 2004. Convex methods for transduction. *Pages 73–80 of : Advances in Neural Information Processing Systems 16.* MIT Press.

Biernacki, C., Celeux, G., & Govaert, G. 2000. Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(7), 719–725.

Biernacki, C., Beninel, F., & Bretagnolle, V. 2002. A Generalized Discriminant Rule when Training Population and Test Population Differ on their Descriptive Parameters. *Biometrics*, **58**(2), 387–397.

Biernacki, C., Celeux, G., Govaert, G., & Langrognet, F. 2006. Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics and Data Analysis*, **51**(2), 587–600.

Biernacki, C., Celeux, G., & Govaert, G. 2008. Exact and Monte Carlo Calculations of Integrated Likelihoods for the Latent Class Model. *Rapport de Recherche INRIA N° 6609.*

Blum, A., & Mitchell, T. 1998. Combining Labeled and Unlabeled Data with Co-training. *Pages 92–100 of : COLT : Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers.*

Bouchard, G., & Celeux, G. 2006. Selection of Generative Models in Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **28**(4), 544–554.

Bouchard, G., & Triggs, B. 2004 (August). The Tradeoff Between Generative and Discriminative Classifiers. *Pages 721–728 of : IASC International Symposium on Computational Statistics (COMPSTAT).*

Bouveyron, C., Girard, S., & Schmid, C. 2007. High-Dimensional Discriminant Analysis. *Communications in Statistics-Theory and Methods*, **36**(14), 2607–2623.

Breiman, L., Friedman, J., Olsen, R.A., & Stone, C. J. 1984. *Classification and Regression Trees.* Chapman and Hall/CRC.

Celeux, G., & Diebolt, J. 1992. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics and Stochastics Reports*, **41**, 127–146.

Celeux, G., & Govaert, G. 1991. Clustering Criteria for Discrete Data and Latent Class Models. *Journal of Classification*, **8**, 157–17.

Chapelle, O., Schölkopf, B., & Zien, A. (eds). 2006. *Semi-Supervised Learning*. Cambridge, MA : MIT Press.

Cooper, D. B., & Freeman, J. H. 1970. On the asymptotic improvement in the outcome of supervised learning provided by additional nonsupervised learning. *IEEE Transactions on Computers*, **C-19**(11), 1055–1063.

Corander, J., Waldmann, P., Marttinen, P., & Sillanpaa, M. J. 2004. BAPS 2 : enhanced possibilities for the analysis of genetic population structure. *Bioinformatics*, **20**(15), 2363–2369.

Corduneanu, A., & Jaakkola, T. 2004. Distributed Information Regularization on Graphs. *In : Neural Information Processing Systems*.

Cozman, F.G., & Cohen, I. 2002. Unlabeled data can degrade classification performance of generative classifiers. *Pages 327–331 of : Fifteenth International Florida Artificial Intelligence Society Conference*.

Dasarathy, B. V. 1990. *Nearest neighbor (NN) norms : NN pattern classification techniques*. Los Alamitos : IEEE Computer Society Press.

Dasgupta, S., Tauman Kalai, A., & Montreleoni, C. 2005. Analysis of perceptron-based active learning. *Pages 249–263 of : Proceedings of the Annual Conference on Learning Theory*.

Dembo, A., & Zeitouni, O. 1998. *Large deviations techniques and applications, second edition*. Springer, Application of Mathematics, vol. 38.

Dempster, A. P., Laird, N. M., & Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Proceedings of the Royal Statistical Society*, **B-39**, 1–38.

Eddy, S.R. 1998. A review of the profile HMM literature from 1996-1998. *Bioinformatics*, **14**, 755–763.

Everitt, B. 1984. *A Introduction to Latent Variable Models*. Chapman and Hall.

Falush, D., Stephens, M., & Pritchard, J. K. 2003. Inference of Population Structure Using Multilocus Genotype Data : Linked Loci and Correlated Allele Frequencies. *Genetics*, **164**(4), 1567–1587.

Fan, W., Davidson, I., Zadrozny, B., & Yu, P. S. 2005. An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias. *Pages 605–608 of : ICDM*. IEEE Computer Society.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

Flury, B. 1988. *Common Principal Components and Related Multivariate Models*. Wiley, New York.

Fralick, S. 1967. Learning to recognize patterns without a teacher. *Information Theory, IEEE Transactions on*, **13**(1), 57–64.

Francois, O., Ancelet, S., & Guillot, G. 2006. Bayesian Clustering using Hidden Markov Random Fields in Spatial Population Genetics. *Genetics*, genetics.106.059923.

Freund, Y., Seung, H.S., Shamir, E., & Tishby, N. 1997. Selective sampling using the query by committee algorithm. *Machine Learning*, **28**(2), 133–168.

Friedman, J. H. 1989. Regularized Discriminant Analysis. *Journal of the American Statistical Association*, **84**(405), 165–175.

Friedman, N. 1997. Bayesian Network Classifiers. *Machine Learning*, **29**, 131–163.

Gan, L., & Jiang, J. 1999. A test for global maximum. *Journal of the American Statistical Association*, **94**(447), 847–854.

Ganesalingam, S., & McLachlan, G. J. 1978. The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika*, **65**(3), 658–665.

Ganesalingam, S., & McLachlan, G. J. 1979. Small sample results for a linear discriminant function estimated from a mixture of normal populations. *Journal of Statistical Computation and Simulation*, **9**, 151–158.

Ghahramani, Z., & Jordan, M. I. 1994. Supervised learning from incomplete data via an EM approach. *Pages 120–127 of : Advances in Neural Information Processing Systems 6*. Morgan Kaufmann.

Goldenshluger, A., & Greenshtein, E. 2000. Asymptotically minimax regret procedures in regression model selection and the magnitude of the dimension penalty. *The Annals of Statistics*, **28**, 1620–1637.

Gordon, A. D. 1981. *Classification : Methods for the Exploratory Analysis of Multivariate Data*. London : Chapman & Hall Ltd.

Grandvalet, Y., & Bengio, Y. 2006. Entropy Regularization. *Pages 151–168 of :* Chapelle, O., Schölkopf, B., & Zien, A. (eds), *Semi-Supervised Learning*. MIT Press.

Grira, N., Crucianu, M., & Boujemaa, N. 2005. Active semi-supervised fuzzy clustering for image database categorization. *Pages 9–16 of : MIR '05 : Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*. New York, NY, USA : ACM.

Hand, David J., & Yu, Keming. 2001. Idiot's Bayes - Not So Stupid After All ? *International Statistical Review*, **69**(3), 385–398.

Hartley, H.O. 1958. Maximum likelihood estimation from incomplete data. *Biometrics*, **14**, 174–194.

Hastie, T., & Tibshirani, R. 1996. Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 155–176.

Hastie, T., Tibshirani, R., & Friedman, J. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics.

Hebert, P.D.N., Ratnasingham, S., & deWaard, J.R. 2003. Barcoding animal life : cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B*, **270**.

Heitjan, D.F., & Rubin, D.B. 1991. Ignorability and coarse data. *The Annals of Statistics*, 2244–2253.

Henry, J.P., & Gouyon, P.H. 2008. *Précis de génétique des populations - Cours, exercices et problèmes résolus*. Dunod.

Hosmer, D. W. 1973. A comparison of iterative maximum likelihood estimates of the parameters of a mixture of two normal distributions under three different types of samples. *Biometrics*, **29**, 761–770.

Jennrich, R.I. 1969. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 633–643.

Joachims, T. 1999a. *Making large-scale support vector machine learning practical*. Cambridge, MA, USA : MIT Press. Pages 169–184.

Joachims, T. 1999b. Transductive Inference for Text Classification using Support Vector Machines. *Pages 200–209 of :* Bratko, Ivan, & Dzeroski, Saso (eds), *Proceedings of ICML-99, 16th International Conference on Machine Learning*. Bled, SL : Morgan Kaufmann Publishers, San Francisco, US.

Lasserre, J.A., Bishop, C.M., & Minka, T.P. 2006. Principled Hybrids of Generative and Discriminative Models. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, **1**(June), 87–94.

Lebarbier, E., & Mary-Huard, T. 2006. Une introduction au critère BIC : fondements théoriques et interprétation. *JSFdS*, **147**(1), 39–57.

Maugis, C., Celeux, G., & Martin-Magniette, M.-L. 2008. Variable selection for Clustering with Gaussian Mixture Models. *(to appear) Biometrics*.

McCallum, A. K., & Nigam, K. 1998. Employing EM in pool-based active learning for text classification. *Pages 350–358 of :* Shavlik, Jude W. (ed), *Proceedings of ICML-98, 15th International Conference on Machine Learning*. Madison, US : Morgan Kaufmann Publishers, San Francisco, US.

McKendrick, A. G. 1926. Application of mathematics to medical problems. *In : Proceedings of the Edinburgh Mathematical Society*, vol. 44.

Mclachlan, G. 2004. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Statistics. Wiley-Interscience.

McLachlan, G., & Krishnan, T. 1996. *The EM Alogorithm and Extensions*. Wiley-Interscience.

McLachlan, G. J., Peel, D., & Bean, R. 2003. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, **41**, 379–388.

Meng, X.L., & Rubin, D. B. 1991. Using EM to Obtain Asymtptic Variance-Covariance Matrices : The SEM Algorithm. *Journal of the American Statistical Association*, **86**(416), 899–909.

Miller, D. J., & Browning, J. 2003. A Mixture Model and EM-Based Algorithm for Class Discovery, Robust Classification, and Outlier Rejection in Mixed Labeled/Unlabeled Data Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(11), 1468–1483.

Miller, D. J., & Uyar, H. 1997. A Mixture of Experts Classifier with Learning Based on Both Labelled and Unlabelled Data. *Pages 321–328 of : Proceedings in Neural Information Processing Systems Conference*, vol. 9.

Murphy, T. B., Dean, N., & Raftery, A. E. 2008. *Variable Selection and Updating In Model-Based Discriminant Analysis for High-Dimensional Data.* Tech. rept. Department of Statistics, University of Washington.

Newcomb, S. 1886. A generalized theory of the combination of observations so as to obtain the best result. *American Journal of Mathematics*, **8**, 343–366.

Ng, A., & Jordan, M. 2002. *On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes.*

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, **39**(2-3), 103–134.

O'Neill, T. 1978. Normal Discrimination with Unclassified Observations. *Journal of the American Statistical Association*, **73**(364), 821–826.

O'Neill, T. 1980. The General Distribution of the Error Rate of a Classification Procedure with Application to Logistic Regression Discrimination. *Journal of the American Statistical Association*, **75**(369), 154–160.

Osuna, E., Freund, R., Girosi, F., CBCL, MIT, & Cambridge, MA. 1997. An improved training algorithm for support vector machines. *Pages 276–285 of : Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop.*

Pearson, K. 1894. Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society A*, 71–110.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, **2**(6), 559–572.

Raftery, A. E., & Dean, N. 2006. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, **101**(473), 168–178.

Redner, R., & Walker, H. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, **26**(2), 195–239.

Rose, K. 1998. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Pages 2210–2239 of : Proceedings of the IEEE.*

Rosenblat, F. 1958. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, **65**, 386–408.

Rosset, S., Zhu, J., Zou, H., & Hastie, T. 2004. A Method for Inferring Label Sampling Mechanisms in Semi-Supervised Learning. *Advances in Neural Information Processing Systems*, **17**(Dec.), 1161–1168.

Saad, Y. 2003. *Iterative Methods for Sparse Linear Systems, Second Edition*. Society for Industrial and Applied Mathematics.

Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Scudder, III, H. J. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, **IT-11**, 363–371.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, **90**(2), 227 – 244.

Sokolovska, N., Cappé, O., & Yvon, F. 2008. The asymptotics of semi-supervised learning in discriminative probabilistic models. *Pages 984–991 of : ICML '08 : Proceedings of the 25th international conference on Machine learning*. New York, NY, USA : ACM.

Szummer, M., & Jaakkola, T. 2002. Partially labeled classification with Markov random walks. *Pages 945–952 of : Advances in Neural Information Processing Systems*. MIT Press.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**(5500), 2319–2323.

Thomas, L. C., Crook, J., & Edelman, D. 2002. *Credit Scoring and Its Applications*. Philadelphia, PA, USA : Society for Industrial and Applied Mathematics.

Titsias, M. K., & Likas, A. 2002. Mixture of experts classification using a hierarchical mixture model. *Neural Computation*, **14**(9), 2221–2244.

Titterington, D.M., Smith, A.F.M., & Makov, U.E. 1985. *Statistical analysis of finite mixture distributions*. Wiley.

Torgerson, W. 1965. Multidimensional scaling of similarity. *Psychometrika*, **30**(4), 379–393.

Toussile, W., & Gassiat, E. 2008. Model-Based Clustering using multilocus data with loci selection. *Preprint submitted to Computational Statistics & Data Analysis*.

Tukey, J. W. 1977. *Exploratory data analysis*.

Valiant, L. 1984. A theory of the learnable. *Communications of the ACM*, **27**.

van der Vaart, A. W. 2000. *Asymptotic Statistics*. Cambridge University Press.

Vandewalle, V. 2009a. Les modèles de mélange, un outil utile pour la classification semi-supervisée. *Revue Modulad*, **41**, 121–145.

Vandewalle, V. 2009b. Sélection prédictive d'un modèle génératif par le critère AIC_p. *In : Archives ouvertes 41es Journées de Statistique, SFdS, Bordeaux.*

Vandewalle, V., Biernacki, C., Celeux, G., & Govaert, G. 2008. Are unlabeled data useful in semi-supervised model-based classification ? Combining hypothesis testing and model choice. *Pages 433–436 of : Proceedings of the first joint meeting of the Société Francophone de Classifcation and the CLassification And Data Analysis Group of SIS.*

Vapnik, V. 1995. *The Nature of Statistical Learning Theory.* Springer-Verlag.

Vapnik, V. 1998. *Statistical Learning Theory.* Wiley.

White, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica*, **51**, 1–25.

Wu, Y., Tian, Q., & Huang, T. 2000. Discriminant-EM Algorithm with Application to Image Retrieval. *Pages 222–227 of : Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-00).* Los Alamitos : IEEE.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. S. 2004. Learning with local and global consistency. *Pages 321–328 of : Advances in Neural Information Processing Systems*, vol. 16.

Zhu, X., & Ghahramani, Z. 2002a. *Learning from labeled and unlabeled data with label propagation.*

Zhu, X., & Ghahramani, Z. 2002b. *Learning from labeled and unlabeled data with label propagation, Technical report, Carnegie Mellon University.*

Zou, H., Zhu, J., & Hastie, T. 2004 (June 21). *Automatic Bayes Carpentry Using Unlabeled Data In Semi-Supervised Classification.*