Les modèles de mélange, un outil utile pour la classification semi-supervisée

Vincent Vandewalle[1,2]

[1] Laboratoire Paul Painlevé UMR CNRS 8524
Université Lille I
59655 Villeneuve d'Ascq Cedex, France
vincent.vandewalle@math.univ-lille.fr
[2] INRIA

Résumé En classification supervisée, la règle de classement est apprise à partir d'un échantillon d'apprentissage généralement constitué de données classées. Dans la plupart des cas l'obtention de la classe est plus coûteuse que l'obtention de covariables associées à la classe d'où l'intérêt d'apprendre une règle de prédiction de la classe à partir de ces covariables. Ainsi dans de nombreuses situations beaucoup de données non classées, obtenues à un coût relativement faible, sont disponibles en plus des données classées. Au cours des dernières années la classification semi-supervisée, qui fait usage des données non classées pour améliorer la précision de la règle de classement apprise, a connu un essor important, ceci notamment dans la communauté du Machine Learning. Les modèles génératifs, qui modélisent la distribution jointe de la classe et des covariables, permettent de prendre naturellement en compte l'information apportée par les données non classées dans l'apprentissage de la règle de classement. Dans cet article nous dressons un panorama de la classification semi-supervisée et nous détaillons sa mise en oeuvre dans le cadre des modèles génératifs.
Mots-clés : données manquantes, modèles de mélange, algorithme EM, analyse discriminante, validation croisée.

Abstract In supervised classification, the classification rule is learnt from a learning sample generally composed of labeled data. In most settings obtaining the label is more expensive than obtaining covariates linked with the label, hence the interest to learn a prediction rule of the label given these covariates. So, in many settings a lot of unlabeled data, obtained at a relatively low cost, are available in addition to labeled data. Over past years the semi-supervised classification, which uses unlabeled data in order to improve the classification rule accuracy, has known a great development, especially in Machine Learning community. Generative models, which model the joint distribution of the label and of the covariates, allow to naturally take into account information contained in unlabeled data when learning the parameters of the model. In this article we give a survey of semi-supervised classification and we detail how to use it with generative models.
Keywords: missing data, mixture models, EM algorithm, discriminant analysis, cross-validation.

1 Introduction

En analyse discriminante l'objectif est de prédire la classe d'appartenance d'un individu à partir de l'observation de covariables [38]. Cette règle de classement doit la plupart du temps être apprise à partir d'un échantillon d'apprentissage constitué de données classées. Dans ce cadre un grand nombre de méthodes ont été développées. La grande majorité d'entre-elles modélise directement la distribution de la classe conditionnellement aux covariables, ou même uniquement la position de la frontière de classification. On parlera de ces méthodes sous le nom générique de méthodes prédictives. Il s'agit entre autres des méthodes non paramétriques comme les k plus proches voisins [21] ou les méthodes de classification à base d'arbres comme CART [13], ou encore des méthodes d'apprentissage de la distribution de la classe conditionnellement aux covariables comme la régression logistique [3]. Enfin des méthodes de recherche d'un hyperplan optimal comme le Perceptron de Rosenblat [49] ou les Support Vecteur Machines (SVM) [59] sont utilisées. Ces méthodes ont pour avantage de prendre directement en compte l'objectif de prédiction. Cependant elles ne font pas d'hypothèses sur la distribution des covariables. Ainsi sans modifications il est impossible pour elles de prendre en compte l'information contenue dans les données non classées. En effet, c'est justement de l'information sur les covariables qu'apportent les données non classées. Une autre catégorie de méthodes consiste à modéliser la distribution jointe des étiquettes et des covariables. Il s'agit par exemple de la célèbre analyse discriminante linéaire (ADL) de Fisher [13]. On regroupera de manière générique l'ensemble de ces méthodes sous le nom de méthodes génératives. On peut leur reprocher de ne pas assez prendre en compte l'objectif de prédiction dans l'apprentissage de la règle de classement puisqu'elles modélisent la distribution jointe de la classe et des covariables, alors qu'en prédiction seule la distribution de la classe conditionnellement aux covariables est requise [59]. Cependant, ces méthodes donnent dans de nombreuses situations des résultats comparables aux méthodes prédictives [32]. En outre, dans le cadre semi-supervisé elles ont l'avantage de naturellement prendre en compte l'information apportée par les données non classées, ceci puisqu'elles modélisent la distribution jointe de la classe et des covariables et par conséquent la distribution des covariables en marginalisant sur la classe.

Le plan de cet article est le suivant : dans une première partie on dresse un panorama de la classification semi-supervisée. Dans une seconde partie, on détaille sa mise en oeuvre dans le cadre des modèles génératifs. Dans une troisième partie, on illustre son comportement par des exemples sur des données réelles et simulées.

2 Panorama de la classification semi-supervisé

L'apprentissage semi-supervisé trouve ses racines dans les problèmes d'apprentissage en présence de données manquantes [27]. Ainsi de nombreux travaux on étés effectués à ce sujet à la fin des années 1970 [33, 22, 45]. Cependant l'utilisation des données non étiquetées pour améliorer la précision de la règle de classement apprise connaît un regain d'intérêt depuis la fin des années 1990 où la communauté du Machine Learning a commencé à s'intéresser à ce sujet. Ceci suite à la disponibilité d'un grand nombre de données acquises de manière automatique grâce aux nouvelles technologies. Ainsi des travaux en classification de texte [43] ont contribué à relancer l'intérêt de l'utilisation des données

non classées en vue d'améliorer la précision de la règle de classement apprise. La parution récente d'un livre entièrement dédié à la classification semi-supervisée [17] est la preuve de l'intérêt actuel pour ce sujet. Dans cette section on détaille les principales approches de classification semi-supervisée.

2.1 L'auto-apprentissage

La première approche à effectuer de l'apprentissage semi-supervisé est l'auto-apprentissage. Elle consiste à apprendre une règle de classement à partir de l'une des méthodes décrites précédemment uniquement sur les données classées. Ensuite une fraction des données non classées est classée à partir de la règle apprise. La règle est ensuite réapprise à partir des données classées à la base et des données classées à l'étape précédente qui sont maintenant considérées comme classées. Ces étapes sont itérées jusqu'à ce que toutes les données non classées soient classées [52, 26, 1]. L'intérêt pratique de cette méthode est qu'elle permet d'adapter n'importe qu'elle méthode de classification supervisée au cadre semi-supervisée. Cependant, le comportement de l'auto-apprentissage ainsi que ses conséquences dépendent fortement de la méthode d'apprentissage supervisée utilisée. De plus l'absence de résultat théorique sur ce à quoi correspond l'auto-apprentissage rend difficile la compréhension de ce qui est effectivement fait et des améliorations qui peuvent en être attendues.

2.2 Débat apprentissage inductif/apprentissage transductif

L'approche semi-supervisée fait apparaître une différence entre apprentissage inductif et apprentissage transductif. En effet la plupart des méthodes de classification supervisée consistent à apprendre une règle de classement qui permet de classer tout nouveau point du domaine. C'est ce que qu'on appelle l'apprentissage inductif ; à partir d'un nombre limité d'exemples on en déduit une règle de classement pour tout nouvel exemple. Cependant en classification semi-supervisée l'objectif se résume souvent à classer les données non étiquetées à disposition ; c'est à dire à un problème d'apprentissage transductif. Il n'est alors pas nécessaire de partitionner tout le domaine, mais il suffit de classer les données non étiquetées à disposition. Ainsi d'un certain point de vue ce problème peut sembler plus simple puisqu'il ne nécessite que de classer un ensemble fini de points, et non pas d'apprendre une règle pour classer tout nouvel exemple à venir. On pourra retrouver ce débat au Chapitre 25 du livre d'O. Chapelle [17]. Cependant, l'apprentissage transductif se révèle être plus complexe à mettre en oeuvre, c'est par exemple le cas des SVM transductifs.

2.3 Les SVM transductifs

Le principe des SVM transductifs reste toujours de trouver la séparation avec la plus vaste marge possible. Cependant, le principe d'apprentissage transductif implique que la règle de classement est uniquement apprise pour classer les données non étiquetées. Les SVM tranductifs, recherchent la frontière de classement avec la plus vaste marge possible une fois les données non étiquetées classées. Ainsi l'optimisation se fait à la fois sur la taille de la marge et sur les étiquettes des données non classées. D'un point de vue théorique ceci consiste à minimiser une borne sur l'erreur de l'échantillon test [59]. Contrairement aux SVM standards, le problème d'optimisation est non convexe et pose par conséquent

des problèmes combinatoires. Il ne peut pas être résolu exactement quand le nombre de données non classées excède 100. Des approches heuristiques permettent de faire face à ce problème d'intractabilité. C'est par exemple le cas des SVM^{light} [34], qui nécessitent en pratique de fixer la proportion d'exemples étiquetés positivement et négativement afin d'éviter l'obtention de solutions dégénérées. Une autre possibilité qui permet de traiter le problème des SVM transductifs est d'utiliser des outils de programmation semi-définie positive [6]. Celle-ci consiste à relaxer les contraintes imposées. Cette méthode permet de traiter des situations allant jusqu'à 1000 données non étiquetées.

Les SVM transductifs, comme les SVM classiques sont particulièrement bien adaptés pour résoudre des problèmes de classification de données en grande dimension. Les SVM transductifs ont notamment montré de bonnes performances en classification de texte [35].

2.4 Régularisation de la solution supervisée

Une approche pour prendre en compte la distribution des données non étiquetées dans l'apprentissage de la règle de classement consiste à pénaliser par l'entropie de classification. Elle permet par exemple de régulariser la solution obtenue par la régression logistique linéaire [28]. La pénalisation par l'entropie conduit à favoriser les solutions pour lesquelles l'entropie de classification est petite. Ainsi, les frontières de classification sont repoussées dans des zones de faible densité. Elle nécessite le choix d'un paramètre de régularisation qui est généralement effectué par validation croisée du taux d'erreur [53]. Cette méthode a montré de bonnes performances en classification d'images et notamment en classification d'expressions faciales [28].

Une autre possibilité de régularisation consiste en la régularisation par l'information mutuelle. Le principe de 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 [19]. De même que dans la régularisation par l'entropie de classification, l'opposé de l'information mutuelle joue un rôle de régularisation en favorisant les solutions pour lesquelles l'information mutuelle est importante.

2.5 Propagation des étiquettes

Les modèles à base de graphes reposent sur une matrice de voisinage. Cette dernière est construite à partir d'une distance entre deux points de l'espace qui nécessite le choix d'un paramètre de réglage. Ensuite, on utilise un algorithme itératif pour propager les étiquettes des points étiquetés vers les points non étiquetés dans le graphe [60]. Cette propagation consiste principalement à utiliser la méthode de Jacobi pour la résolution de systèmes linéaires [51]. 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 afin d'obtenir un graphe moins dense, et par conséquent des algorithmes plus rapides. Si les paramètres de réglage sont bien qualibrés de bons résultats peuvent être obtenus [60].

2.6 Réduction de la dimension

Enfin les données non étiquetées peuvent servir à réduire la dimension des données. Cette réduction de la dimension peut être réalisée à partir de méthodes de type ACP [46],

MDS [56] ou ISOMAP [54]. L'enjeu est de représenter dans un sous-espace plus petit des données en grande dimension en perdant le moins d'information possible. En effet, de nombreuses méthodes de classification nécessitent que le nombre de variable soit petit comparativement au nombre de données. Dans de nombreux cas des méthodes comme l'ACP permettent une régularisation des solutions obtenues. La présence de nombreuses données non classées permet ainsi d'obtenir des bonnes projections des données dans des sous-espaces plus petits. Une fois la réduction de dimension réalisée, les méthodes d'apprentissage supervisé usuelles peuvent être utilisées sur l'espace de dimension réduite.

2.7 Modèles génératifs

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 classées dans le cas Gaussien homoscédastique [27, 45]. A notre connaissance l'approche la plus ancienne est celle d'Hosmer [33]. Dans ses travaux il s'agissait d'une population de flétans (grand poisson plat des mers froides), pour laquelle l'objectif était d'estimer la proportion de mâles et de femelles. Pour ces poissons la détermination du sexe n'étant possible qu'après dissection, des mesures biométriques de nombreux individus peuvent aider à déterminer la fraction de mâles et de femelles. Il avait à sa disposition de nombreuses données commerciales où seuls l'âge et la longueur étaient fournis. Tandis qu'une étude scientifique sur un nombre beaucoup plus petit de données fournissait aussi le sexe. Ainsi l'usage des données commerciales en plus des données de l'étude scientifiques permettent une estimation plus précise des paramètres. Dans ces travaux l'estimation est réalisée par maximum de vraisemblance grâce à un algorithme itératif, qui sera plus tard formalisé sous le nom d'algorithme EM [22]. Cet algorithme est bien adapté pour traiter les problèmes de données manquantes ce qui est le cas des données non-étiquetées pour lesquelles les étiquettes constituent les données manquantes. Les modèles considérés consistent principalement en des mélanges de distributions Gaussiennes et des mélanges de distributions multinomiales [18]. Cette approche connaît un renouveau depuis les années 1990, ceci notamment dans la communauté du Machine Learning. Elle sera détaillée dans la Section 3.

2.8 Entre estimation générative et prédictive

Une question qui revient souvent en analyse discriminante, ceci même dans le cadre supervisé, est celle du choix entre un modèle génératif de type analyse discriminante linéaire et un modèle prédictif de type régression logistique. D'un côté le modèle génératif fait des hypothèses de modélisation fortes qui si elles sont vérifiées permettent d'obtenir une erreur de classification plus faible que les modèles prédictifs. O'Neill a par exemple montré la supériorité de l'analyse discriminante linéaire face à la régression logistique si les données proviennent effectivement du modèle postulé (ici le modèle Gaussien homoscèdastique) [44]. Ceci peut se comprendre intuitivement car les modèles génératifs prennent aussi en compte l'information sur la distribution des covariables et estiment donc plus précisément les paramètres de la règle de classement. D'un autre côté si le modèle génératif postulé est faux, c'est le modèle prédictif qui produira la divergence de Kullback à la distribution conditionnelle la plus faible pour un nombre de données assez grand. Dans la pratique il est reconnu que les modèles génératifs produisent 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 [42].

Une approche récemment développée est la recherche d'un compromis entre estimation générative et prédictive [11]. Cette approche a ensuite été reformulée par Lasserre et al. [36] dans un cadre Bayesien permettant ainsi une interprétation plus naturelle. Leur approche se place profondément dans un cadre semi-supervisé. En théorie cette approche doit permettre de trouver de manière automatique le bon compromis entre estimation générative et prédictive. Cependant, en pratique le compromis choisi dépend en grande partie du choix des hyperparamètres. De plus cette approche double le nombre de paramètres à estimer et nécessite l'utilisation d'algorithmes de Newton numériquement instables en grande dimension.

2.9 Discussion sur les hypothèses du semi-supervisé

2.9.1 Hypothèses du semi-supervisé

En synthèse, la plupart des extensions des méthodes prédictives au semi-supervisé nécessitent des modifications plus ou moins ad-hoc, ainsi que le choix des paramètres de régularisation. D'autre part, les modèles génératifs permettent de prendre naturellement en compte l'information apportée par les données non étiquetées. L'approche compromis entre estimation prédictive et générative pose des difficultés techniques importantes bien qu'étant attrayante d'un point de vue théorique. Remarquons d'autre part que comme mentionné dans [17], l'approche semi-supervisée est susceptible de bien fonctionner quand les hypothèses suivantes sont vérifiées :

- 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 dimensionalité : Les données en grande dimension vivent dans des sous espaces de petite dimension.

2.9.2 Application aux modèles prédictifs

Les méthodes prédictives cherchent à introduire le type d'information de la partie précédente quand elles veulent s'adapter à la situation semi-supervisée. Par exemple pour la régularisation par l'entropie les paramètres sont estimés par maximisation de la vraisemblance conditionnelle pénalisée par l'entropie de classification, c'est à dire par le recouvrement des classes. Ceci force la recherche de séparation dans des zones de faible densité. Il en est de même pour les SVM transductifs. Les méthodes de réduction de la dimension supposent que les données en grande dimension vivent dans un sous espace de plus petite dimension. Ici on voit apparaître la difficulté qu'ont de nombreuses méthodes prédictives à utiliser les données non-étiquetées ; elles forcent certaines propriétés sur la distribution des covariables en lien avec les étiquettes, ceci sans pour autant accepter de la modéliser directement.

2.9.3 Application aux modèles génératifs

Dans le cadre des modèles génératifs un certain nombre des hypothèses précédentes est présent dès la construction du modèle. Si chaque classe est modélisée par une distribution gaussienne, on a bien l'hypothèse de cluster qui est faite d'office puisqu'on impose alors la correspondance entre une classe et un composant gaussien. D'autre part, dans la plupart des situations de classification il est souhaitable que les classes soient bien séparées pour obtenir une faible erreur de classement. Enfin des hypothèses sur la matrice de covariance peuvent être faites comme dans [39, 12] pour imposer à chaque classe de vivre dans un espace de dimension réduite. Ainsi à première vue les hypothèses sous lesquelles la classification semi-supervisée est susceptible de bien fonctionner sont les hypothèses pour lesquelles utiliser un modèle génératif fait sens. C'est à dire que conditionnellement à la classe les covariables ont effectivement une distribution spécifique. De plus l'intérêt de faire des hypothèses explicites sur la distribution des données et non pas des hypothèses implicites est que ces dernières peuvent être remises en cause lors d'une procédure de choix de modèle. Ainsi, si on dispose de modèles relativement parcimonieux épousant correctement la distribution des données, on peut améliorer l'estimation de la distribution jointe des données, et par suite obtenir une règle de classement plus précise.

D'autre part contrairement aux autres méthodes les modèles génératifs permettent par exemple de détecter l'apparition de nouvelles classes, de faire de la classification robuste ou bien de détecter des données abhérantes [41]. Précisons que ce point peut-être important en apprentissage actif, c'est à dire lorsque l'utilisateur a le choix des données à étiqueter. En effet, la présence de données non étiquetées dans certaines zones de l'espace peut conduire le praticien à étiqueter des données dans ces zones et par suite à éventuellement découvrir de nouvelles classes. Le lecteur intéressé pourra se référer à [4] pour la découverte de classes en classification de galaxies.

2.10 Synthèse

Nous avons dressé un panorama des principales approches utilisées pour l'incorporation des données non étiquetées dans le processus d'apprentissage de la règle de classement. Nous allons maintenant nous focaliser sur les modèles génératifs. Remarquons avant tout que l'utilisation des données non étiquetées ne permet pas toujours de réduire l'erreur de classement par rapport aux méthodes supervisées. Cependant, sous les hypothèses précédentes on peut espérer que cette réduction du taux d'erreur ait lieu. On trouve dans la littérature des exemples où les données non étiquetées dégradent la précision de la règle de classification [20]. Nous détaillerons les hypothèses sous lesquelles les données non classées permettent d'améliorer la précision de la règle de classement lorsqu'on utilise des modèles génératifs. Ainsi, il sera toujours important de vérifier que les données non étiquetées ne dégradent pas la règle de classement. Pour cela on pourra par exemple utiliser la validation croisée.

3 Mise en oeuvre des modèles génératifs

Dans cette section nous introduisons de manière plus précise le cadre probabiliste nécessaire à l'utilisation des modèles génératifs en classification semi-supervisée.

3.1 Hypothèses d'échantillonnage

Supposons que les données appartiennent à G classes différentes. Soit un couple de variables aléatoires $(\mathbf{X}_1, \mathbf{Z}_1)$ à valeurs dans $\mathcal{X} \times \mathcal{Z}$ où $\mathcal{X}$ est une espace mesurable ($\mathbb{R}^d$ par exemple) qui représente l'espace des covariables, et où $\mathcal{Z} = \{0,1\}^G$ est l'espace des classes. Supposons que n_ℓ données étiquetées et n_u données non étiquetées ont été observées et on note $n = n_\ell + n_u$. Les données étiquetées correspondent à $(\mathbf{x}_\ell, \mathbf{z}_\ell) = (\{(\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_{n_\ell}, \mathbf{z}_{n_\ell})\})$, et les données non étiquetées sont représentées par $\mathbf{x}_u = (\mathbf{x}_{n_\ell+1}, \ldots, \mathbf{x}_n)$, avec $\mathbf{z}_i \in \mathcal{Z}$, c'est-à-dire que $\mathbf{z}_i = (z_{i1}, \ldots, z_{iG})$ où $z_{ik} = 1$ si l'individu i appartient à la classe k et 0 sinon. On va supposer que les données non étiquetées proviennent de n_ℓ réalisations indépendantes et identiquement distribuées du couple $(\mathbf{X}_1, \mathbf{Z}_1)$ et que les données non étiquetées proviennent de n_u réalisations indépendantes et identiquement distribuées de $\mathbf{X}_1$. On peut se demander dans quel rapport évoluent n_ℓ et n_u quand n augmente. Il est naturel de supposer que chaque donnée est étiquetée indépendamment avec une probabilité β, ce qui implique que n_ℓ est la réalisation d'une variable aléatoire $N_\ell \sim \mathcal{B}(n, \beta)$. Ainsi quand n tend vers l'infini, $\frac{n_\ell}{n}$ tend vers β en probabilité.

Avant de passer aux questions de modélisation quelques remarques sur les hypothèses d'échantillonnage s'imposent. En effet dans le cadre des données manquantes, les hypothèses formulées précédemment impliquent que les données manquantes, en l'occurrence les étiquettes des données non étiquetées, sont manquantes complètement au hasard. C'est à dire, que le fait pour une étiquette de ne pas être observée ne dépend ni de sa valeur, ni de la valeur des covariables [50]. Cette hypothèse peut sembler relativement forte mais c'est l'hypothèse de travail faite explicitement ou non dans la majorité des travaux en classification supervisée et semi-supervisée. En effet, quand une règle de classement est apprise on souhaite ensuite l'appliquer à des données qui proviennent de la même distribution que celles qui ont servies à l'apprendre. Sans cette hypothèse l'estimation du taux d'erreur sur les données à venir par validation croisée sur les données d'apprentissage perd alors une grande partie de son sens.

Des approches existent pour prendre en compte d'éventuels biais mais elles compliquent grandement les choses [61, 24]. Toutefois notons que l'hypothèse que les données soient manquantes au hasard, c'est-à-dire que le fait pour une étiquette de ne pas être observée est indépendant de sa valeur conditionnellement aux covariables, est suffisante pour garantir la consistance de l'estimation par maximum de vraisemblance [50]. Dans le cas où les données classées et non-classées proviennent de distributions différentes mais qu'il existe des liens entre-elles des travaux montrent que la règle de classement apprise sur les données classées peut être transposée aux données non classées dans le cadre de l'analyse discriminante généralisée [7].

3.2 Définition du modèle

Passons maintenant au modèle paramétrique postulé sur la distribution des données. On suppose que le couple de variables aléatoires $(\mathbf{X}_1, \mathbf{Z}_1)$ admet une densité de probabilité (p.d.f.) f par rapport à une mesure sur $\mathcal{X} \times \mathcal{Z}$. On suppose que f appartient à une famille paramétrique paramétrèe par $\theta \in \Theta$, où Θ est l'espace des paramètres de dimension finie. Ainsi, on suppose $\exists \theta^* \in \Theta$ tel que $f(\mathbf{x}_1, \mathbf{z}_1) = f(\mathbf{x}_1, \mathbf{z}_1; \theta^*)\ \forall (\mathbf{x}_1, \mathbf{z}_1) \in \mathcal{X} \times \mathcal{Z}$. Plus précisément on effectue la décomposition suivante $f(\mathbf{x}_1, \mathbf{z}_1; \theta) = \prod_{k=1}^{G} (\pi_k f(\mathbf{x}_1; \theta_k))^{z_{1k}}$, où π_k représente la probabilité d'appartenir à la classe k ($pi_k > 0$, $\sum_{k=1}^{G} \pi_k = 1$) et

où θ_k représente les paramètres spécifiques à la distribution de la classe k. Ainsi $\theta = (\pi_1, \ldots, \pi_{G-1}, \theta_1, \ldots, \theta_G)$. Ces modèles sont appelés modèles génératifs puisqu'ils modélisent le processus de génération des données :

- $\mathbf{Z}_1 \sim \mathcal{M}(1, \pi_1^*, \ldots, \pi_G^*)$,
- $\mathbf{X}_1 | \mathbf{Z}_{1k} = 1$ a pour densité de probabilité $f(\cdot; \theta_k^*)$.

Il en résulte ainsi que la densité marginale pour $\mathbf{X}_1 = \mathbf{x}_1$ s'écrit sous la forme $f(\mathbf{x}_1; \theta_k) = \sum_k^G \pi_k f(\mathbf{x}_1; \theta_k)$ après marginalisation sur la variable $\mathbf{Z}_1$. On parle de loi mélange.

Si le paramètre θ^* est connu la règle de classement optimale pour un individu i consiste à le classer dans la classe $\hat{k} = \arg\max_k P(z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i; \theta^*)$ où $P(z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i; \theta^*) \propto \pi_k^* f(\mathbf{x}_i; \theta_k^*)$. L'enjeu est d'estimer θ^* de la manière la plus précise possible. Si les hypothèses sont correctes, alors une estimation précise de θ^* implique un règle de classement proche de la règle de classement optimale.

3.3 Estimation des paramètres du modèle

3.3.1 Motivation de l'estimation par maximum de vraisemblance

Compte tenu des hypothèses faites précédemment, la log-vraisemblance de θ s'écrit de la manière suivante :

$$\ell(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{G} z_{ik} \log(\pi_k f(\mathbf{x}_i; \theta_k)) + \sum_{i=n_\ell+1}^{n} \log \left(\sum_{k=1}^{G} \pi_k f(\mathbf{x}_i; \theta_k) \right) \quad (1)$$

Le premier terme représente la log-vraisemblance des données étiquetées, le second terme représente la log-vraisemblance des données non étiquetées. Ainsi, les données non étiquetées interviennent de façon naturelle dans la vraisemblance, et par suite dans l'estimation des paramètres. Sous certaines conditions de régularité sur le modèle l'estimateur du maximum de vraisemblance a de bonnes propriétés [58]. Ainsi les paramètres sont estimés par maximum de vraisemblance $\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u)$. L'intérêt théorique de l'utilisation des données non étiquetées dans l'estimation des paramètres est qu'elles permettent de réduire la variance de $\hat{\theta}$ comparativement à l'estimation supervisée. Sous l'hypothèse du bon modèle, cette réduction de variance conduit à une réduction de l'erreur moyenne du classifieur appris.

La maximisation de la vraisemblance ne peut pas être effectuée explicitement puisqu'il y a une somme dans la log-vraisemblance des données non étiquetées. Dans ce cadre comme mentionné plus haut il existe un algorithme très bien adapté à la maximisation de la vraisemblance qui s'appelle l'algorithme EM [22] et qu'on détaille dans la section suivante.

3.3.2 Algorithme EM

Historiquement l'approche itérative d'estimation des paramètres par maximum de vraisemblance a été utilisée dans un cadre semi-supervisé par Hosmer [33] dans le cadre gaussien homoscédastique avant même sa formulation plus générale sous le nom d'algorithme EM par Dempster et al. [22].

L'algorithme EM est un algorithme itératif qui consiste à itérer deux étapes :

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

– **Étape M (Maximisation) :** Calcul de $\theta^{(r+1)} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^{(r)})$.

La succession de ces deux étapes permet de faire croître la vraisemblance à chaque itération. Elle permet donc la convergence de l'algorithme vers une racine de la vraisemblance.

Ici les données manquantes sont les étiquettes des données non étiquetées $\mathbf{z}_u = (\mathbf{z}_{n_\ell+1}, \ldots, \mathbf{z}_n)$, et la vraisemblance complétée s'écrit alors :

$$\ell_c(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u, \mathbf{z}_u) = \sum_{i=1}^{n_\ell}\sum_{k=1}^{G} z_{ik}\log(\pi_k f(\mathbf{x}_i;\theta_k)) + \sum_{i=n_\ell+1}^{n}\sum_{k=1}^{G} z_{ik}\log(\pi_k f(\mathbf{x}_i;\theta_k)). \quad (2)$$

On prend l'espérance de la vraisemblance complétée conditionnellement au paramètre courant $\theta^{(r)}$ et aux données observées $\mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u$, ce qui donne en utilisant la linéarité de l'espérance et le fait que les données soient i.i.d. :

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n_\ell}\sum_{k=1}^{G} z_{ik}\log(\pi_k f(\mathbf{x}_i;\theta_k)) + \sum_{i=n_\ell+1}^{n}\sum_{k=1}^{G} E[z_{ik}|\mathbf{x}_i,\theta^{(r)}]\log(\pi_k f(\mathbf{x}_i;\theta_k)). \quad (3)$$

Enfin remarquons que $E[z_{ik}|\mathbf{x}_i,\theta^{(r)}] = P(z_{ik}=1|\mathbf{x}_i,\theta^{(r)})$ et nous avons d'après le théorème de Bayes $P(z_{ik}=1|\mathbf{x}_i,\theta^{(r)}) = \pi_k^{(r)} f(\mathbf{x}_i;\theta_k^{(r)})/(\sum_{l=1}^{G}\pi_l^{(r)} f(\mathbf{x}_i;\theta_l^{(r)}))$. Dans la suite on notera $t_{ik}^{(r+1)}$ cette quantité. Ainsi l'espérance de la log-vraisemblance complétée s'écrit :

$$Q(\theta|\theta^{(r)}) = \sum_{i=1}^{n_\ell}\sum_{k=1}^{G} z_{ik}\log(f(\mathbf{x}_i;\theta_k)) + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)}\log(f(\mathbf{x}_i;\theta_k)) + \sum_{k=1}^{G} n_k^{(r+1)}\log(\pi_k), \quad (4)$$

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 on est ramené à une expression qui ne comporte pas le log d'une somme. L'étape de M, consiste à maximiser $Q(\theta|\theta^{(r)})$ en θ, pour l'actualisation des proportions on obtient la formule suivante :

$$\pi_k^{(r+1)} = \frac{n_k^{(r+1)}}{n}. \quad (5)$$

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 situation supervisée.

Des variantes de l'algorithme EM existent. C'est par exemple le cas de l'algorithme CEM (Classification EM) [16] qui introduit une étape de classification entre l'étape E et l'étape M de l'algorithme, celle ci consiste à affecter chaque donnée de manière dure à la classe qui a la plus grande probabilité a posteriori. Précisons que cette variante ne maximise pas la vraisemblance mais la vraisemblance complétée. Cette version de l'algorithme EM a pour avantage de converger rapidement. Cependant dans le cas où les classes se chevauchent elle produit une estimation biaisée de paramètres. Dans le cadre semi-supervisé l'algorithme CEM peut être interprété comme de l'auto-apprentissage itéré. Ainsi pour des modèles génératifs, on voit que l'auto-apprentissage hérite des problèmes de biais présents dans l'algorithme CEM.

Comme pour la plupart des algorithmes itératifs, l'algorithme EM nécessite d'être initialisé. En semi-supervisé on dispose d'une bonne initialisation de l'algorithme qui consiste à l'initialiser à partir des paramètres estimés uniquement à partir des données

étiquetées. Toutefois des initialisations de type non supervisé peuvent être utilisées, ainsi que des chaînages d'algorithmes [8]. En effet, la vraisemblance comportant en général de nombreux maxima locaux, il n'y a priori pas de garantie de trouver le maximum global de la vraisemblance en initialisant l'algorithme à partir du paramètre estimé à partir des seules données étiquetées. Le paramètre retenu au final est celui qui produit le plus grande vraisemblance. Une autre question est le choix du moment où arrêter l'algorithme, pour ce faire on peut soit arrêter l'algorithme au bout d'un nombre d'itérations fixé, ou on peut l'arrêter quand la croissance de la vraisemblance entre deux étapes est inférieure à un certain seuil.

Une autre stratégie possible pour maximiser la vraisemblance est le recuit déterministe [48]. Cette dernière consiste à commencer à une température importante pour laquelle le problème à résoudre est convexe puis petit à petit faire décroître la température pour résoudre à nouveau le problème en partant de la solution précédente. Si la décroissance de la température est suffisamment lente alors la solution obtenue au final devrait être le maximum global. Remarquons que la décroissance lente de la température implique de nombreuses itérations. Un des problèmes rencontrés par le recuit déterministe est un problème d'inversion des classes. Pour éviter ce problème les classes doivent ensuite être réassignées. Si l'étape de réassignation est effectuée correctement, le recuit déterministe présente alors de meilleures performances comparées à l'initialisation à partir des données étiquetées [43]. Remarquons que dans le cadre de la classification de texte la convergence de EM est très rapide, typiquement une dizaine d'étapes sont suffisantes à la convergence de l'algorithme [43]. Ainsi, dans la plupart des situations réelles, il est préférable de relancer plusieurs fois l'algorithme à partir de solutions initiales tirées au hasard plutôt que d'utiliser le recuit déterministe.

3.3.3 Conditions sur les modèles utilisés

Avant d'aller plus loin dans le détail des modèles utilisés dans ce contexte il est nécessaire de détailler un peu plus les conditions généralement requises sur les modèles utilisés. Dans un cadre non supervisé, l'identifiabilité de la famille mélange à une permutation des classes près est généralement requise [37]. Cette condition est équivalente au fait que la famille paramétrée de densité est algébriquement libre sur $\mathbb{R}$. Ceci est notamment le cas pour les mélanges de distributions gaussiennes, exponentielles, de Poisson et de Cauchy. Cependant ceci n'est pas le cas pour les mélanges d'uniformes et de Bernoulli. Ces conditions permettent ensuite d'assurer la consistance de l'estimation des paramètres à une permutation des classes près. Dans le cadre semi-supervisé, cette permutation est éliminée puisque les données étiquetées induisent une asymétrie dans l'estimation des classes. Il est ici important de préciser que les mélanges de produits de distributions de Bernoulli ne sont pas identifiables [29] au sens précédent. Cependant, d'un point de vue pratique, ces modèles conservent une interprétation utile minimisant ainsi l'importance du problème d'identifiabilité [14]. En fait, ce problème est levé dans [2] en introduisant la notion d'identifiabilité générique, c'est à dire que le modèle est identifiable sauf pour un ensemble de points de l'espace des paramètres de mesure de Lebesgue nulle. Ainsi [2] donne une condition simple entre le nombre de classes et le nombre de variables binaires pour que le mélange de produits de distributions de bernouilli soit génériquement identifiable. Dans la section suivante nous détaillons l'utilisation des mélanges gaussiens ainsi que l'utilisation des produits de distributions multinomiales d'ordre 1, et des mélanges de

distributions multinomiales d'ordre supérieur à 1. Nous parlerons enfin de l'utilisation des mélanges à plusieurs composants par classe qui sont particulièrement utiles dans le cadre semi-supervisé.

3.4 Modèles utilisés

3.4.1 Modèle gaussien

Un modèle génératif très populaire quand $\mathcal{X} = \mathbb{R}^d$ est $\mathbf{X}_i|\mathbf{Z}_{ik} = 1 \sim \mathcal{N}(\mu_k, \Sigma_k)$ 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 :

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

Ainsi les formules d'actualisation pour les paramètres spécifiques du modèle sont :

$$\mu_k^{(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik}\mathbf{x}_i \sum_{i=n_\ell+1}^{n} t_{ik}\mathbf{x}_i}{n_k^{(r+1)}} \tag{6}$$

$$\Sigma_k^{(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik}(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)' + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)}(\mathbf{x}_i-\mu_k)(\mathbf{x}_i-\mu_k)'}{n_k^{(r+1)}}. \tag{7}$$

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 un cadre non supervisé, si aucune contrainte n'est imposée à la matrice de covariance, la vraisemblance est non bornée. Cependant [47], montre qu'il existe une solution consistante parmi les racines de la vraisemblance, et que si l'algorithme EM est initialisé avec une solution suffisamment proche, alors l'algorithme EM y conduit. Dans le cadre semi-supervisé, si on dispose d'au moins $d+1$ données classées dans chaque classe la vraisemblance reste bornée. Dans le cas où les matrices de covariance par classe sont supposée égales, on retrouve la célèbre analyse discriminante linéaire de Fisher [25].

Une reparamètrisation de la matrice de covariance permet de définir des modèles parcimonieux allant du cas homoscédastique au cas hétéroscédastique. En effet, la matrice de covariance peut être décomposée en valeurs singulières sous la forme $\Sigma_k = \lambda_k D_k A_k D_k'$ où $\lambda_k = |\Sigma_k|^{1/d}$ définit le volume de la classe k , D_k est la matrice orthogonale des vecteurs propres qui peut être interprétée en terme d'orientation de la classe k, A_k est la matrice diagonale des valeurs propres normalisées rangées par ordre décroissant qui peut être interprétée en terme de forme de la classe k. Ainsi en imposant à λ_k, D_k ou A_k d'être identiques pour chaque classe 14 modèles parcimonieux sont obtenus [5]. Leur estimation est elle aussi facile en pratique, même si elle nécessite éventuellement le recours à un algorithme itératif lors de l'étape M. Récemment reposant sur le même principe de décomposition en valeurs singulières mais en imposant aux plus petites valeurs propres d'être identiques, des modèles parcimonieux en grande dimension ont étés développés. Le fait d'imposer aux plus petites valeurs propres d'être identiques a un effet de régularisation

sur la matrice de covariance et permet ainsi son inversion même dans le cas où la dimension est supérieure au nombre de données observées [12]. Cette décomposition qui se place dans le cadre des facteurs analysants [39] avait déjà montré de bonnes performances en classification non supervisée de données génomiques de type biopuces où 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). Dans un cadre supervisé ces modèles ont montré des performances comparables aux SVM [12] qui est une méthode de référence dans ce contexte.

3.4.2 Produit de modèles multinomiaux d'ordre 1

Supposons maintenant que $\mathcal{X}$ soit un espace discret, c'est à dire que d variables discrètes sont observées et que chaque variable $j \in \{1, \ldots, d\}$ comporte m_j modalités. Ainsi l'espace probabilisé $\mathcal{X}$ est composé de $\prod_{j=1}^{d} m_j$ événements élémentaires. Dans la grande majorité des situations il est impossible de modéliser séparément chaque réalisation. Ainsi un modèle très couramment utilisé 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 [23]. Ce modèle bien que rudimentaire donne de bons résultats dans de nombreuses situations réelles [30]. Notons ici que l'indépendance des covariables conditionnellement à la classe n'implique pas l'indépendance des covariables. 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 $d \geq \lceil \log_m G \rceil + 1$ est vérifiée, alors le modèle est génériquement identifiable [2]. Ainsi, si le nombre de variables est suffisamment grand devant le nombre de classes, le modèle est génériquement identifiable. Soit $x_i^{jh} = 1$ si l'individu i présente la modalité h de la variable j et 0 sinon. On note $\alpha_k^{jh} = P(X_i^{jh} = 1 | Z_{ik} = 1)$, $\alpha_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$ 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 : $f(\mathbf{x}_i | \alpha_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}$. La log-vraisemblance s'écrit alors :

$$\ell(\theta; \mathbf{x}_\ell, \mathbf{z}_\ell, \mathbf{x}_u) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{G} z_{ik} [\log \pi_k + \sum_{j=1}^{d} \sum_{h=1}^{m_j} x_i^{jh} \log(\alpha_k^{jh})] + \sum_{i=n_\ell+1}^{n} \log(\sum_{k=1}^{G} \pi_k \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}). \tag{8}$$

Puis l'espérance de la vraisemblance complétée s'écrit :

$$Q(\theta | \theta^{(r)}) = \sum_{i=1}^{n_\ell} \sum_{k=1}^{G} \sum_{j=1}^{d} \sum_{h=1}^{m_j} z_{ik} x_i^{jh} \log(\alpha_k^{jh}) \tag{9}$$

$$+ \sum_{i=n_\ell+1}^{n} \sum_{k=1}^{G} \sum_{j=1}^{d} \sum_{h=1}^{m_j} t_{ik}^{(r+1)} x_i^{jh} \log(\alpha_k^{jh}) + \sum_{k=1}^{G} n_k^{(r+1)} \log \pi_k. \tag{10}$$

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

$$\alpha_k^{jh(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik} x_i^{jh} + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} x_i^{jh}}{n_k^{(r+1)}}. \tag{11}$$

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. Comme dans le cas continu, des versions parcimonieuses de ces modèles existent dans le cas où les covariables sont des variables binaires [15]. Ces modèles ont montré un intérêt en classification non supervisée où ils permettent de faire des liens avec des critères de classification non supervisée sur variables discrètes.

3.4.3 Modèle multinomial d'ordre quelconque

Un modèle assez similaire a été utilisé en classification de texte [43], où l'usage des données non étiquetées a permis d'améliorer les performances du classifieur appris. Il consiste à considérer un texte comme un sac de mots. Soit un dictionnaire de d mots $(w_1, \ldots, w_d)$. Soit le texte i de longueur ℓ_i et $\mathbf{x}_i = (x_i^1, \ldots, x_i^d)$ avec x_i^j qui est égal au nombre d'occurrences du mot j dans le texte i. On suppose que $\mathbf{X}_i | Z_{ik} = 1 \sim \mathcal{M}(\ell_i, \alpha_k^1, \ldots, \alpha_k^d)$. α_k^j représente la fréquence du mot w_j dans les textes appartenant à la classe k. Ce modèle peut être interprété comme le modèle précédent où chaque mot est la réalisation d'une variable multinomiale, et où on impose à chaque mot d'un type de texte de provenir de cette même distribution multinomiale. Ainsi, on retrouve des conditions d'identifiabilité générique assez similaires à celles du modèle précédent. Cette condition étant vérifiée lorsque la longueur des textes observés est suffisamment grande par rapport au nombre de classes, ceci est en généralement le cas en pratique. La formule d'actualisation est :

$$\alpha_k^{j\,(r+1)} = \frac{\sum_{i=1}^{n_\ell} z_{ik} x_i^j + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} x_i^j}{\sum_{i=1}^{n_\ell} z_{ik} \ell_i + \sum_{i=n_\ell+1}^{n} t_{ik}^{(r+1)} \ell_i}. \tag{12}$$

Cette formule est assez proche de celle rencontrée dans le cas du modèle d'indépendance conditionnelle. En effet 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. Ce modèle est bien sûr assez irréaliste dans le procédé d'écriture. Cependant, l'information sur la fréquence d'apparition des mots suffit dans de nombreux cas à classifier correctement les textes.

3.4.4 Modèles à plusieurs composants par classe

Dans certaines situations la modélisation d'une classe par un seul composant peut se montrer trop rigide. Une idée assez naturelle consiste à modéliser une classe par plusieurs composants. Cette approche se justifie d'autant plus par les bonnes propriétés d'approximation des mélanges de densité. 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 classées trop petit. Cependant dans le cas 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 dans [40]. Notons que cette extension à plusieurs composants par classe est possible pour les modèles gaussiens, les modèles multinomiaux et les modèles d'indépendance conditionnelle.

Deux hypothèses sont possibles :

– Soit les composants sont communs aux classes ; sachant le composant h la donnée appartient à la classe k avec une probabilité $\tau_{kh} \in [0, 1]$.
– Soit chaque classe est modélisée par des composants différents ; sachant le composant h la donnée appartient à la classe k avec une probabilité $\tau_{kh} \in \{0, 1\}$.

L'intérêt de la première approche est qu'elle nécessite simplement de fixer le nombre total de composants et d'estimer les τ_{kh}, tandis que la seconde nécessite de fixer le nombre de classes par composant et peut donc par suite nécessiter d'étudier un nombre de modèles relativement grand. Ce qui est fait dans [31] consiste à imposer aux nombres de composants par classe à être identiques pour éviter tout problème combinatoire. Dans le cadre supervisé [55] montre 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 une estimation de la densité marginale. Cependant dans le cadre semi-supervisé cela est moins évident puisqu'il s'agit justement en grande partie de bien estimer la densité marginale. 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 manquer de parcimonie par rapport au modèle à composants séparés. D'autre part 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. Cependant il peut tout de même rester présent compte-tenu de l'initialisation de l'algorithme et de la convergence vers des maxima locaux.

3.5 Synthèse

L'intérêt des modèles génératifs est qu'ils permettent une réutilisation quasi-immédiate dans le cadre semi-supervisé des modèles qui ont montré de bonnes performances dans le cadre supervisé ou dans le cadre non supervisé. Remarquons aussi que dans ce contexte de modélisation, le modèle peut être complexifié pour permettre l'apparition de nouvelles classes, l'existence d'une classe de bruit, ou encore la possibilité d'une supervision imparfaite. Ainsi ces modèles ne permettent pas seulement de prédire, mais aussi de comprendre.

Dans la section suivante on illustre l'utilisation de ces modèles sur des jeux de données réelles et simulées. On verra que les résultats peuvent grandement varier d'un modèle à un autre, ce qui mettra en avant l'importance du choix de modèle. Ce problème dépassant assez largement le cadre de cet article, on se limitera à décrire une procédure de validation croisée du taux d'erreur permettant d'effectuer ce choix.

4 Exemples d'utilisation

La mise en oeuvre de l'estimation semi-supervisée des paramètres est implémentée dans le logiciel MIXMOD [9] pour le modèle Gaussien et le produit de distributions multinomiales.

4.1 Exemples sur certains jeux de 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 supervisées. Ces jeux de données peuvent facilement être rendus semi-supervisés 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 et de pouvoir valider les résultats puisque les étiquettes des données non étiquetées sont en fait connues. Ici on compare les performances des approches supervisées et semi-supervisées dans les cas homo et hétéroscédastiques. Le dispositif expérimental est illustré Table 4.1. Pour chaque jeu de données, si un échantillon test est fourni, les données 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 classées et n_ℓ données classées 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. 1 – Dispositif expérimental.

Les résultats sont présentés Table 2, l'écart-type du taux d'erreur est obtenu à partir des 100 séparations données étiquetées données non étiquetées et est écrit entre parenthèse.

	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	6,89 (2,21)	8,86 (2,30)	11,36 (4,76)	6,47 (3,46)
Iris	2,72 (1,32)	2,05 (1,02)	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 – Taux d'erreur moyen dans différentes configurations.

4.2 Exemples sur les données du SSL book

4.2.1 Utilisation des modèles de classification en haute dimension

Dans le livre d'O.Chappelle sont fournis des jeux de données avec séparation données étiquetées données non étiquetées (`http://www.kyb.tuebingen.mpg.de/ssl-book/benchmarks.html`), ces derniers vont nous permettre d'illustrer l'utilisation des modèles gaussiens en grande dimension. On se limite à l'utilisation du modèle qui décompose la matrice de variance Σ_k en valeurs singulières et qui suppose que les p plus petites valeurs propres

de ces dernières sont identiques, ce qui correspond au modèle $[A_{ij}B_iQ_iD]$ de [12]. Pour les formules d'estimation des paramètres le lecteur intéressé pourra se référer à [12]. 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. Ces dernières sont utilisées pour comparer les performances des différents modèles utilisés. On va du modèle qui suppose que toutes les valeurs propres sont identiques 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 1, les boîtes à moustache [57] 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. Cela met en avant le phénomène de surapprentissage quand des modèles trop complexes sont utilisés.

Le second jeu de données intitulé g241n illustre une situation où une classe est modélisée par deux composants gaussiens ; ainsi le modèle postulé est faux. Figure 2 on remarque que les données non classées contribuent encore à réduire le taux d'erreur moyen quand au moins une valeur propre est supposée différente des autres. On remarque un phénomène intéressant qui est que ce n'est ni le modèle le plus complexe ni le modèle le plus simple qui produit les meilleurs résultats, mais qu'il y a un compromis entre la bonne approximation de la distribution des données par le modèle, et la variance dans l'estimation des paramètres.

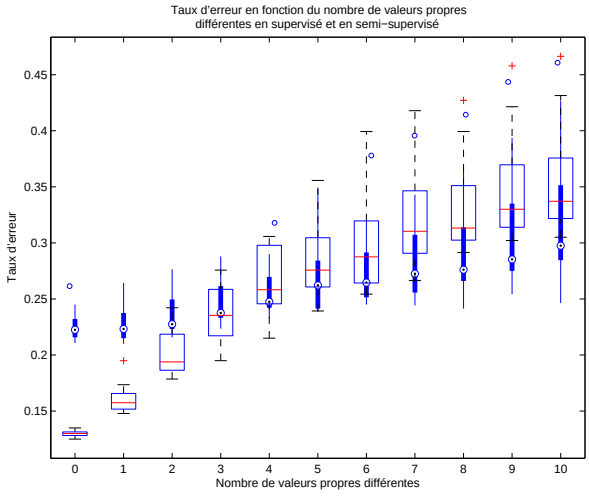

FIG. 1 – Jeu de données g241c (semi-supervisée : boîtes larges, supervisé : boîtes fines).

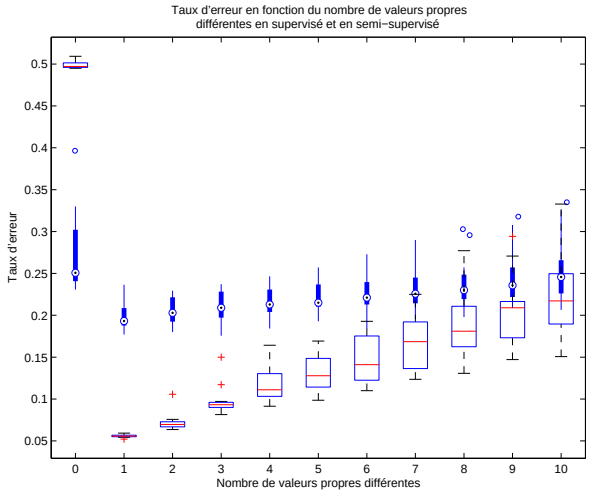

FIG. 2 – Jeu de données g241n (semi-supervisée : boîtes larges, supervisé : boîtes fines).

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

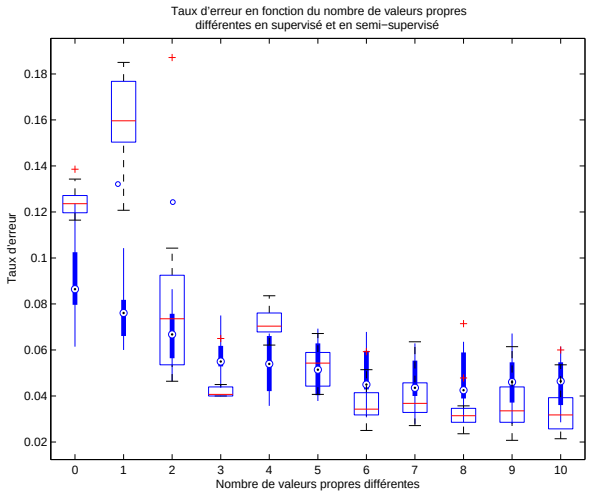

FIG. 3 – Jeu de données Digit1 (semi-supervisée : boîtes larges, supervisé : boîtes fines).

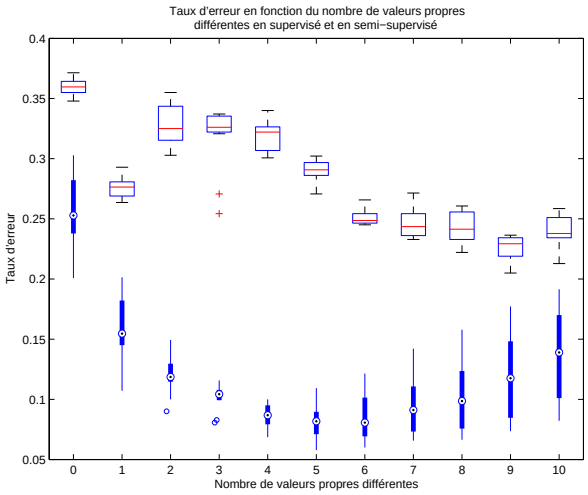

FIG. 4 – Jeu de données USPS (semi-supervisée : boîtes larges, supervisé : boîtes fines).

au maximum. On voit Figure 3 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 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 4. Cependant cette dégradation semble diminuer au fur et à mesure que des modèles plus complexes sont utilisés.

Ainsi ces exemples nous ont permis d'illustrer un certain nombre de situations rencontrées. On voit que les performances de la règle de classement peuvent varier grandement selon le modèle choisi ce qui justifie l'importance de la question du choix de modèle dans ce contexte.

4.2.2 Utilisation des mélanges à plusieurs composants par classe

Un autre type de modèle potentiellement 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. Cette hypothèse correspond 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 choisi sur les mêmes jeux de données que précédemment.

Pour le premier jeu de données, le vrai nombre de composants par classe est 1. C'est donc le modèle le plus simple qui produit les meilleures performances. On note encore une amélioration des performances dans le cadre semi-supervisé.

Pour le second jeu de données, les données sont issues d'un modèle à deux composants par classe, c'est donc 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 mauvaises, le semi-supervisé dégrade les performances du classifieur appris. On voit ici que le semi-supervisé tire un très bon parti des données non étiquetées lorsque le modèle postulé est correct.

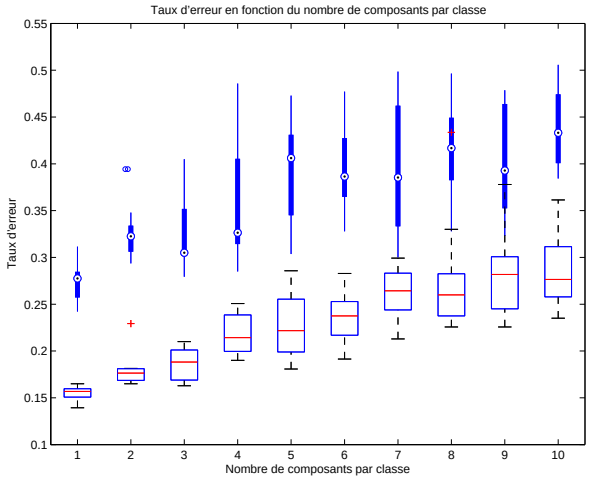

FIG. 5 – Jeu de données g241c (semi-supervisée : boîtes larges, supervisé : boîtes fines).

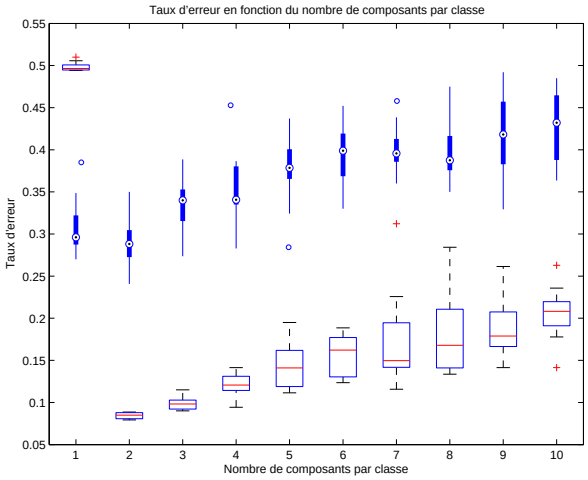

FIG. 6 – Jeu de données g241n (semi-supervisée : boîtes larges, supervisé : boîtes fines).

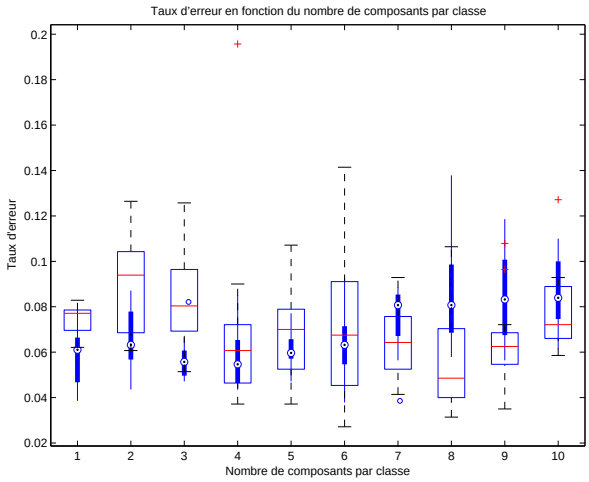

FIG. 7 – Jeu de données Digit1 (semi-supervisée : boîtes larges, supervisé : boîtes fines).

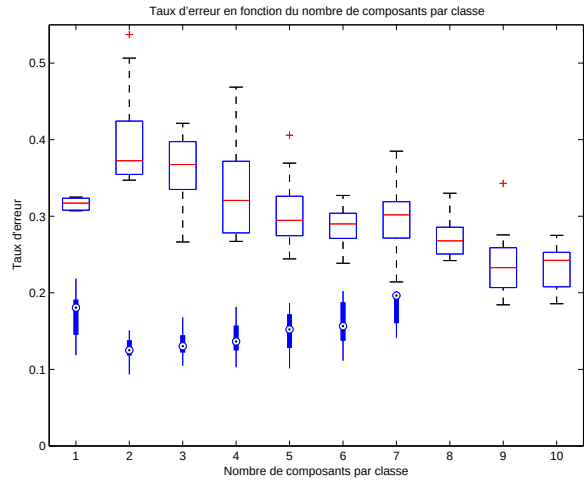

FIG. 8 – Jeu de données USPS (semi-supervisée : boîtes larges, supervisé : boîtes fines).

Pour le troisième jeu de données on voit que les résultats obtenus sont à peu près les mêmes dans les cadre supervisé et semi-supervisé et qu'ils ne varient pas trop selon le

nombre de composants choisis.

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

Remarquons aussi qu'on aurait pu combiner l'approche à plusieurs composants par classe et l'approche modèle en grande dimension, ce que nous n'avons pas fait ici pour rester simple. En effet, dans ce cas il faut 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, l'approche composants communs étant aussi possible.

4.3 Comparaison du semi-supervisé au supervisé dans le cas du bon modèle

Dans cette partie on explique l'intérêt du semi-supervisé sur des jeux de données simulées. Soit deux classes Gaussiennes en dimension 50, et en proportions égales. On a $\mathbf{X}|Z_1 = 1 \sim \mathcal{N}(0, I_{50})$ et $\mathbf{X}|Z_2 = 1 \sim \mathcal{N}(\mu, I_{50})$ avec $\mu_i = \frac{1}{i}$ pour i allant de 1 à 50. Soit $n_\ell = 100$ données classées et $n_u = 10000$ données non classées. On trace Figure 9 l'erreur moyenne en fonction du nombre de variables utilisées dans les cadres supervisé et semi-supervisé. Cette figure nous permet de remarquer deux intérêts de l'approche semi-supervisée. Le premier est que lorsqu'on considère le nombre de variables optimal dans le cadre supervisé (taux d'erreur moyen 29,36%), à même nombre de variable le semi-supervisé permet d'obtenir un taux d'erreur plus bas (taux d'erreur moyen 27,79%). Le second est que le semi-supervisé permet d'utiliser efficacement un plus grand nombre de variables et permet ainsi 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 usage plus efficace des modèles plus complexes.

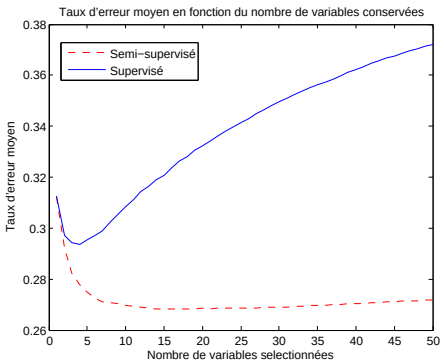

FIG. 9 – Taux d'erreur en fonction du nombre de variables utilisées

4.4 Choix de modèle par validation croisée

Le principal objet de cet article est l'utilisation des modèles génératifs pour la classification semi-supervisée. Cependant, les exemples précédents mettent clairement en évidence le fait que la question du choix d'un modèle est de première importance. On explique maintenant comment le choix de ce dernier peut se faire par validation croisée.

Comme on souhaite sélectionner le modèle qui produit l'erreur de classement la plus faible, il est naturel de sélectionner le modèle qui produit l'erreur de classement la plus faible estimée par validation croisée. Notons ici que la validation croisée nécessite des adaptations. En effet, pour estimer correctement l'erreur de classement il faut retirer la même fraction de données étiquetées et non-étiquetées en V fold validation croisée. Cette remarque est d'autant plus importante que V est petit. Remarquons aussi que si le nombre de données étiquetées est petit les performances de la validation croisée peuvent être relativement médiocres tout comme dans le cadre supervisé. Le principe de la V fold validation croisée est le suivant :

- Couper au hasard $\mathcal{D}_u = \mathbf{x}_u$ et $\mathcal{D}_\ell = (\mathbf{x}_\ell, \mathbf{z}_\ell)$ en V blocks de tailles à peu près identiques : $\mathcal{D}_\ell = \bigcup_{i=1}^{V} \{\mathcal{D}_\ell^{\{i\}}\}$, and $\mathcal{D}_u = \bigcup_{i=1}^{V} \{\mathcal{D}_u^{\{i\}}\}$
- **Pour** $i = 1$ **à** V
  - $\hat{e}_i = \frac{1}{\mathrm{card}(\mathcal{D}_\ell^{\{i\}})} \sum_{(\mathbf{x}_p, \mathbf{z}_p) \in \mathcal{D}_\ell^{\{i\}}} \mathbf{1}_{\{r(\mathbf{x}_p; \hat{\theta}^{\{-i\}}) \neq \mathbf{z}_p\}}$ avec $\hat{\theta}^{\{-i\}}$ l'estimateur du maximum de vraisemblance calculé en utilisant $\{\mathcal{D}_\ell, \mathcal{D}_u\} \backslash \{\mathcal{D}_\ell^{\{i\}}, \mathcal{D}_u^{\{i\}}\}$.
- **Fin**
- Calculer $\hat{e} = \frac{1}{V} \sum_{i=1}^{V} \hat{e}_i$.

Remarquons que cette approche est relativement coûteuse puisqu'elle nécessite de recourir V fois à l'algorithme EM, cet algorithme pouvant bien sur être initialisé à partir de $\hat{\theta}$. Ainsi comparativement à la situation supervisée où les paramètres peuvent parfois être recalculés de façon peu coûteuse à partir des anciens paramètres [10], ceci est impossible dans le cadre semi-supervisé puisque l'estimateur du maximum de vraisemblance n'est pas explicite.

5 Conclusion

Nous avons dressé un panorama des méthodes utilisées en semi-supervisé. Nous avons justifié l'intérêt des modèles génératifs dans ce contexte, et nous avons détaillé leur mise en oeuvre : hypothèses faites, intérêt théorique de l'utilisation des données non étiquetées, modèles pouvant être utilisés, formules d'actualisation pour ces modèles. Nous avons montré sur quelques exemples qu'ils pouvaient effectivement améliorer la précision de la règle de classement apprise. Ces expériences nous ont permis de mettre en évidence une question importante qui est celle du choix de modèle, cette question pouvant être résolue en utilisant une procédure de validation croisée.

Références

[1] A. Agrawala. Learning with a probabilistic teacher. *Information Theory, IEEE Transactions on*, 16(4) :373–379, 1970.

[2] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of latent class models with many observed variables. `http://www.citebase.org/abstract?id=oai:arXiv.org:0809.5032`, 2008.

[3] J. A. Anderson and S. C. Richardson. Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, 21(1) :71–78, 1979.

[4] D. Bazell and D. J. Miller. Class discovery in galaxy classification. *The Astrophysical Journal*, 618(2) :723–732, Jan. 2005.

[5] H. Bensmail and G. Celeux. Regularized discriminant analysis. *Journal of the American Statistical Association*, 91 :1743–1748, 1996.

[6] T. De Bie and N. Cristianini. Convex methods for transduction. In *Advances in Neural Information Processing Systems 16*, pages 73–80. MIT Press, 2004.

[7] C. Biernacki, F. Beninel, and V. Bretagnolle. A generalized discriminant rule when training population and test population differ on their descriptive parameters. *Biometrics*, 58(2) :387–397, 2002.

[8] C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate gaussian mixture models. *Computational Statistics and Data Analysis*, 41 :561–575, 2003.

[9] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the MIXMOD software. *Computational Statistics & Data Analysis*, 51(2) :587–600, November 2006.

[10] C. Biernacki and G Govaert. Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation*, 64 :49–71, 1999.

[11] G. Bouchard and B. Triggs. The tradeoff between generative and discriminative classifiers. In *IASC International Symposium on Computational Statistics (COMPSTAT)*, pages 721–728, Prague, August 2004.

[12] C. Bouveyron, S. Girard, and Cordelia Schmid. *Class-specific subspace discriminant analysis for high-dimensional data*, pages 139–150. Number 3940 in Lecture Notes in Computer Science. Springer Verlag, 2006.

[13] L. Breiman, J. Friedman, R.A. Olsen, and C. J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

[14] M. Á. Carreira-Perpiñn and S. Renals. Practical identifiability of finite mixtures of multivariate Bernoulli distributions. *Neural Computation*, 12(1) :141–152, 2000.

[15] G. Celeux and G. Govaert. Clustering criteria for discrete data and latent class models. *Journal of Classification*, 8 :157–17, 1991.

[16] G. Celeux and G. Govaert. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3) :315–332, 1992.

[17] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[18] D.B. Cooper and J.H. Freeman. On the asymptotic improvement in the out-come of supervised learning provided by additional nonsupervised learning. *Computers, IEEE Transactions on*, C-19(11) :1055–1063, Nov. 1970.

[19] A. Corduneanu and T. Jaakkola. Distributed information regularization on graphs. In *Neural Information Processing Systems*, 2004.

[20] F.G. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In *Fifteenth International Florida Artificial Intelligence Society Conference*, pages 327–331, 2002.

[21] B. V. Dasarathy. *Nearest neighbor (NN) norms : NN pattern classification techniques*. Los Alamitos : IEEE Computer Society Press, 1990, 1990.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1) :1–38, 1977.

[23] B. Everitt. *A Introduction to Latent Variable Models*. Chapman and Hall, 1984.

[24] W. Fan, I. Davidson, B. Zadrozny, and P. S. Yu. An improved categorization of classifier's sensitivity on sample selection bias. In *ICDM*, pages 605–608. IEEE Computer Society, 2005.

[25] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 :179–188, 1936.

[26] S. Fralick. Learning to recognize patterns without a teacher. *Information Theory, IEEE Transactions on*, 13(1) :57–64, 1967.

[27] S. Ganesalingam and G. J. McLachlan. The efficiency of a linear discriminant function based on unclassified initial samples. *Biometrika*, 65(3) :658–665, 1978.

[28] Y. Grandvalet and Y. Bengio. Entropy regularization. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-Supervised Learning*, pages 151–168. MIT Press, 2006.

[29] M. Gyllenberg, T. Koski, E. Reilink, and M. Verlann. Non-uniqueness in probabilistic numerical identification of bacteria. *Journal of Applied Probability*, 31(2) :542–548, 1994.

[30] David J. Hand and Keming Yu. Idiot's Bayes - not so stupid after all ? *International Statistical Review*, 69(3) :385–398, 2001.

[31] T. Hastie and R. Tibshirani. Discriminant analysis by Gaussian mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58 :155–176, 1996.

[32] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, 2001.

[33] D. W. Hosmer. 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, 1973.

[34] T. Joachims. *Making large-scale support vector machine learning practical*, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[35] T. Joachims. Transductive inference for text classification using support vector machines. In Ivan Bratko and Saso Dzeroski, editors, *Proceedings of ICML-99, 16th International Conference on Machine Learning*, pages 200–209, Bled, SL, 1999. Morgan Kaufmann Publishers, San Francisco, US.

[36] J.A. Lasserre, C.M. Bishop, and T.P. Minka. Principled hybrids of generative and discriminative models. *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, 1 :87–94, June 2006.

[37] G. Mclachlan and D. Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, October 2000.

[38] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. Wiley, New York, 1992.

[39] G. J. McLachlan, D. Peel, and R. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis*, 41 :379–388, 2003.

[40] D. J. Miller and H. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Proceedings in Neural Information Processing Systems Conference*, volume 9, pages 321–328, 1997.

[41] David J. Miller and John Browning. A mixture model and em-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets. *IEEE Transaction on Pattern Analysis Machine Intelligence*, 25(11) :1468–1483, 2003.

[42] A. Ng and M. Jordan. On discriminative vs. generative classifiers : A comparison of logistic regression and naive bayes. `http://citeseer.ist.psu.edu/542917.html`, 2002.

[43] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3) :103–134, 2000.

[44] T. O'Neill. 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, 1980.

[45] Terence O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73(364) :821–826, 1978.

[46] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6) :559–572, 1901.

[47] R. Redner and H. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2) :195–239, 1984.

[48] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. In *Proceedings of the IEEE*, pages 2210–2239, 1998.

[49] F. Rosenblat. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65 :386–408, 1958.

[50] D. B. Rubin. Inference and missing data. *Biometrika*, 63 :581–592, 1976.

[51] Y. Saad. *Iterative Methods for Sparse Linear Systems, Second Edition*. Society for Industrial and Applied Mathematics, April 2003.

[52] H. J. Scudder, III. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, IT-11 :363–371, 1965.

[53] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36 :111–147, 1974.

[54] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500) :2319–2323, December 2000.

[55] Michalis K. Titsias and Aristidis Likas. Mixture of experts classification using a hierarchical mixture model. *Neural Computation*, 14(9) :2221–2244, 2002.

[56] Warren Torgerson. Multidimensional scaling of similarity. *Psychometrika*, 30(4) :379–393, December 1965.

[57] J. W. Tukey. Exploratory data analysis, 1977.

[58] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, June 2000.

[59] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[60] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. S. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16*, volume 16, pages 321–328, 2004.

[61] H. Zou, J. Zhu, and T. Hastie. Automatic bayes carpentry using unlabeled data in semi-supervised classification. `http://www-stat.stanford.edu/~hastie/Papers/NIPS04/abc_nips04.pdf`, June 21 2004.