Probabilistic model-based discriminant analysis and clustering methods in chemometrics

Charles Bouveyron

Probabilistic model-based discriminant analysis and clustering methods in Chemometrics

Charles Bouveyron

Laboratoire MAP5, UMR CNRS 8145
Université Paris Descartes & Sorbonne Paris Cité
45 rue des Saints Pères, 75006 Paris, France

Abstract

In Chemometrics, the supervised and unsupervised classification of high-dimensional data has become a recurrent problem. Model-based techniques for discriminant analysis and clustering are popular tools which are renowned for their probabilistic foundations and their flexibility. However, classical model-based techniques show a disappointing behavior in high-dimensional spaces which up to now have been limited in their use within Chemometrics. The recent developments in model-based classification overcame these drawbacks and enabling the efficient classification of high-dimensional data, even in the "small n / large p" condition. This work presents a comprehensive review of these recent approaches, including regularization-based techniques, parsimonious modeling, subspace classification methods and classification methods based on variable selection. The use of these model-based methods is also illustrated on real-world classification problems in Chemometrics using R packages.

1 Introduction

In Chemometrics and many other scientific fields, recent technological developments have resulted in a dramatic increase in measurement capabilities. Nowadays, it is common to observe high-dimensional data (*i.e.* the number p of measured variables is large), mass of data (*i.e.* the number of observations n is large) or even data streams (*i.e.* the observations arrive over the time and $n \to \infty$). Due to the complex nature of these data, the requirement for statistical tools to analyze the data has also recently increased. Classification is one of those tools which allows one to model, understand and predict the studied phenomenon.

Classification is indeed an important and useful statistical tool in scientific fields where decisions have to be made. Depending on the availability of a learning data set, two main situations may occur: supervised classification (also known as discriminant analysis) and unsupervised classification (also known as clustering). Model-based techniques [30, 51] for clustering and classification are popular approaches renowned for their probabilistic foundations and their flexibility. One of the main advantages of these approaches is the fact that their models and results can be interpreted from both statistical and practical points of view.

Unfortunately, the behaviour of model-based methods may be disappointing in high-dimensional spaces. They suffer from the well-known curse of dimensionality [6] which is mainly due to the fact that model-based techniques are over-parametrized in high-dimensional spaces. Furthermore, in several applications, such as analytical spectroscopy, mass spectrometry or genomics, the number of available observations can be small compared to the number of variables. However and since the dimension of observed data is usually higher than their intrinsic dimension, it is theoretically possible to reduce the dimension of the original space without loosing any information. For this reason, dimension reduction methods are frequently used to reduce the dimension of the data before the clustering step. Feature extraction methods, such as principal component analysis (PCA), and feature selection methods are very popular. However, dimension reduction usually does not consider the classification task and provides a sub-optimal data representation for the classification step. Indeed, dimension reduction methods usually imply information loss which could have been discriminative.

To avoid the drawbacks of dimension reduction, several approaches have been proposed in the last decade which allow model-based methods to efficiently classify high-dimensional data. The earliest approaches included constrained models or regularization. More recently, subspace classification techniques and variable selection techniques have been proposed. Subspace classification techniques are based mostly on probabilistic versions of the factor analysis [68] model and allow data classification in low-dimensional subspaces without reducing the dimension. Conversely, variable selection techniques reduce the dimension of the data but select and retain the variables in respect to the classification objective. Both techniques turn out to be very efficient and their practical use will be discussed in this article.

This article is organized as follows. Section 2 briefly recalls the bases of model-based discriminant analysis and clustering. Solutions for the classification of high-dimensional data in the model-based context are presented in Section 3 and existing R packages and their practical uses are discussed in Section 4. Finally, Section 5 offers concluding remarks.

2 Model-based classification

Classification is a double-headed problem which includes supervised classification (also known as discriminant analysis) and unsupervised classification (also known as clustering). Discriminant analysis aims to build a classifier (or a decision rule) able to assign an observation y in an arbitrary space $\mathcal{Y}$ with an unknown class membership to one of K known classes $C_1, ..., C_K$. For building this supervised classifier, a learning data set $\{(y_1, z_1), ..., (y_n, z_n)\}$ is used, where the observation $y_i \in \mathcal{Y}$ and $z_i \in \{1, ..., K\}$ indicates the class belonging of the observation y_i. In a slightly different context, clustering aims to partition directly an incomplete dataset $\{y_1, ..., y_n\}$ into K homogeneous groups without any other information, *i.e.,* assign to each observation $y_i \in \mathcal{Y}$ its group membership $z_i \in \{1, ..., K\}$. In order to avoid any misunderstanding, the term "classification" is used in the following pages as a generic term which includes both the supervised case (discriminant analysis) and the unsupervised one (clustering). Notice that several intermediate situations exist, such as semi-supervised or weakly-supervised classifications [24], but they are not considered here.

2.1 Model-based discriminant analysis

As we said above, discriminant analysis aims to build from a complete data set $\{(y_1, z_1), ..., (y_n, z_n)\}$ a supervised classifier able to predict the belonging of a new and unlabeled observation y to one of the K known classes. From a theoretical point of view, this consists in learning a decision function δ:

$$\begin{array}{rcl} \delta : \mathcal{Y} & \mapsto & \{1, ..., K\} \\ y & \to & z. \end{array}$$

The decision rule which minimizes the probability of classification error is known as the maximum a posteriori (MAP) rule. The MAP rule assigns a new observation y to the class which has the largest posterior probability $P(Z = k|Y = y)$. The decision rule δ^* associated with the MAP rule has therefore the following form:

$$\delta^*(y) = \text{argmax}_k \, P(Z = k|Y = y).$$

At this point, it is possible to split the classification methods into two main families according to the way they compute this posterior probability. Discriminative methods, such that support vector machines [38], directly compute this posterior probability. Conversely, model-based methods model the joint probability $P(Y, Z)$ and deduce afterward, using the Bayes rule, the posterior probability $P(Z|Y)$. In the model-based context, discriminant analysis methods assumes that $\{y_1, ..., y_n\}$ are independent realizations of a random vector $Y \in \mathcal{Y}$ and that the class-conditional distribution of Y is parametric:

$$f(y|Z = k) = f_k(y; \theta_k),$$

with a prior probability $P(Z = k) = \pi_k$. In such a case, the MAP rule takes the following form:

$$\begin{array}{rcl} \delta^*(y) & = & \text{argmax}_k \, P(Z = k)p(Y = y|Z = k), \\ & = & \text{argmax}_k \, \pi_k f_k(y; \theta_k). \end{array}$$

This can be reformulated as:

$$\delta^*(y) = \text{argmin}_k \, \Gamma_k(y),$$

where $\Gamma_k(y) = -2\log(\pi_k f_k(y;\theta_k))$.

When the observed data are continuous, *i.e.* $\mathcal{Y} = \mathbb{R}^p$, the Gaussian distribution is often preferred among the possible parametric distributions for f_k and, in this case, the marginal distribution of Y is therefore a mixture of Gaussians:

$$g(y) = \sum_{k=1}^{K} \pi_k \phi(y;\mu_k,\Sigma_k),$$

where ϕ is the Gaussian density, π_k is the prior probability of the kth class (with the constraints $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$), μ_k is the mean of the kth class and Σ_k is its covariance matrix. This specific mixture model is commonly referred to as the Gaussian mixture model (GMM) among the literature. That stated, several recent works have focused on different distributions such as the skew normal, asymmetric Laplace or t-distributions. With the Gaussian assumption, the classification function Γ_k involved in the MAP rule can be rewritten as:

$$\Gamma_k(y) = -2\log|\Sigma_k| + (y-\mu_k)^t\Sigma_k^{-1}(y-\mu_k) - 2\log(\pi_k) + p\log(2\pi).$$

This method is known as quadratic discriminant analysis (QDA), and, under the additional assumption that $\Sigma_k = \Sigma$ for all $k \in \{1,\ldots,K\}$, it corresponds to linear discriminant analysis (LDA). In any case, the estimation of the model parameters is done by maximum likelihood and is usually straightforward in this context since the data are complete. A detailed overview on this topic can be found in [49].

2.2 Model-based clustering

Model-based clustering aims to partition a data set of n observations $\{y_1,\ldots,y_n\} \in \mathcal{Y}$ into K homogeneous groups, *i.e.* determine for each observation y_i the value of its unobserved label z_i such that $z_i = k$ if the observation y_i belongs to the kth cluster. To do so, model-based clustering [30, 51] considers the overall population as a mixture of the groups and each component of this mixture is modeled through its conditional probability distribution. In this context, the observations $\{y_1,\ldots,y_n\}$ are assumed to be independent realizations of a random vector $Y \in \mathcal{Y}$ whereas the unobserved labels $\{z_1,...,z_n\}$ are assumed to be independent realizations of a random variable $Z \in \{1,...,K\}$. The set of pairs $\{(y_i,z_i)\}_{i=1}^n$ is usually referred to as the complete data set. By denoting by g the probabilistic density function of Y, the finite mixture model is again:

$$g(y) = \sum_{k=1}^{K} \pi_k f_k(y;\theta_k), \tag{1}$$

where π_k , f_k and θ_k respectively represent the mixture proportion, the conditional density function and the parameter vector of the kth mixture component. As in the supervised context, the clusters are often modeled by the same parametric density function and the multivariate normal density is usually preferred. For a set of observations $y = \{y_1,...,y_n\}$, the log-likelihood of this mixture model is then:

$$\ell(\theta;y) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k f(y_i;\theta_k)\right). \tag{2}$$

However, the inference of this model cannot be directly done through the maximization of the likelihood since the group labels $\{z_1,...,z_n\}$ of the observations are unknown. Indeed, due to the exponential number of solutions to explore, the maximization of equation (2) is unfortunately intractable, even for limited numbers of observations and groups. The expectation-maximization (EM) algorithm, proposed by Dempster *et al.* [25], is certainly the most popular technique for inferring mixture models [10, 44, 46, 56, 59, 69]. It iteratively maximizes the likelihood through the maximization of the complete log-likelihood, which is defined by:

$$\ell_c(\theta;y,z) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log\left(\pi_k f(y_i;\theta_k)\right),$$

where $z_{ik} = 1$ if the ith observation belongs to the kth cluster and $z_{ik} = 0$ otherwise. We refer to [50] for a general description of the EM algorithm and we restrict ourselves here to the Gaussian case. In the case of the Gaussian mixture model, the EM algorithm has the following form, at iteration q:

E-step This step aims to compute the expectation of the complete log-likelihood conditionally to the current value of the parameter $\theta^{(q-1)}$. In practice, it reduces to the computation of $t_{ik}^{(q)} = E[z_i = k|y_i, \theta^{(q-1)}]$. Let us also recall that $t_{ik}^{(q)}$ is as well the posterior probability $P(z_i = k|y_i, \theta^{(q-1)})$ that the observation y_i belongs to the kth component of the mixture under the current model. Using Bayes' theorem, the posterior probabilities $t_{ik}^{(q)}$, $i = 1, ..., n$, $k = 1, ..., K$, can be expressed as follows:

$$t_{ik}^{(q)} = \frac{\pi_k^{(q-1)} \phi(y_i, \theta_k^{(q-1)})}{\sum_{l=1}^{K} \pi_l^{(q-1)} \phi(y_i|\theta_l^{(q-1)})}, \tag{3}$$

where $\pi_k^{(q-1)}$ and $\theta_k^{(q-1)} = \left\{\mu_k^{(q-1)}, \Sigma_k^{(q-1)}\right\}$ are the parameters of the kth mixture component estimated at the previous iteration.

M-step This step updates the model parameters by maximizing the conditional expectation of the complete log-likelihood. The maximization of $E\left[\ell_c\left(\theta; y, z\right)|\theta^{(q-1)}\right]$ conduces to an update of the mixture proportions π_k, the means μ_k and the covariance matrices Σ_k as follows, for $k = 1, ..., K$:

$$\hat{\pi}_k^{(q)} = \frac{n_k^{(q)}}{n}, \;\; \hat{\mu}_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} y_i, \;\; \hat{\Sigma}_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} \left(y_i - \hat{\mu}_k^{(q)}\right)\left(y_i - \hat{\mu}_k^{(q)}\right)^t,$$

where $n_k^{(q)} = \sum_{i=1}^{n} t_{ik}^{(q)}$.

The two steps of the EM algorithm are iteratively applied until a stopping criterion is satisfied. The stopping criterion may be simply $|\ell(\theta^{(q)}; y) - \ell(\theta^{(q-1)}; y)| < \varepsilon$ where ε is a positive value to provide. It would be also possible to use the Aitken's acceleration criterion [45] which estimates the asymptotic maximum of the likelihood and allows to detect in advance the algorithm convergence. Once the EM algorithm has converged, the partition $\{\hat{z}_1, \ldots, \hat{z}_K\}$ of the data can be deduced from the posterior probabilities $t_{ik} = P(Z = k|y_i, \hat{\theta})$ by using the *maximum* a posteriori (MAP) rule which assigns the observation y_i to the group with the highest posterior probability.

2.3 Classification results and model selection

Both in the supervised and unsupervised contexts, the learned model and the classification results are usually meaningful. Indeed, looking for instance at the estimated mean of each component is a good way to figure out what is the average behavior of each class or group. Another interesting results in both cases is the availability of the posterior probabilities that each observation belongs to the K components. This in particular allows to know the classification risk $P(Z \neq \hat{z}_i|Y = y_i)$ for each data point while allocating the observations to the components. This classification risk can be easily computed as follows:

$$P(Z \neq \hat{z}_i|Y = y_i) = 1 - P(Z = \hat{z}_i|Y = y_i),$$

where $\hat{z}_i = \text{argmax}_k \, P(Z = k|Y = y_i)$. This is obviously of particular interest in application fields where classification errors have important consequences, such as in medicine.

The use of model-based methods in classification also takes advantage of its probabilistic foundation in the selection of the appropriate model and the tuning of its hyper-parameters (usually all discrete parameters). Indeed, the choice of the most appropriate model and hyper-parameters for the data at hand is usually a tricky task which turns out to be natural here by considering those choices as model selection problems. The idea here is to consider, for instance, a Gaussian mixture model with $K = 2$ and another Gaussian mixture model with $K = 3$ as two different models among which one wants to choose. A naive strategy would choose the model associated with the highest likelihood but this fails because the two models are nested. The likelihood is indeed strictly increasing with K in such a case. The solution comes from the use of likelihood penalized criteria which add to the likelihood a penalty depending on the complexity of the model. Classical criteria for model selection include

the AIC [2], BIC [67] and ICL [9] criteria. The Bayesian Information Criterion (BIC) is certainly the most popular and consists in selecting the model which penalizes the likelihood by $\frac{\gamma(\mathcal{M})}{2}\log(n)$ where $\gamma(\mathcal{M})$ is the number of parameters in model $\mathcal{M}$ and n is the number of observations. On the other hand, the AIC criterion penalizes the log-likelihood by $\gamma(\mathcal{M})$ whereas the ICL criterion add the penalty $\sum_{i=1}^{n}\sum_{k=1}^{K} t_{ik}\log(t_{ik})$ to the one of the BIC criterion in order to favour well separated models. The value of $\gamma(\mathcal{M})$ is of course specific to the model selected by the practitioner. Although the interest of those criteria is most in the unsupervised case, they can also be efficient in the supervised case and circumvent the classical but time consuming cross-validation.

3 Model-based methods for high-dimensional classification

After having briefly reviewed the earliest approaches, this section focuses on recent model-based approaches for the supervised and unsupervised classification of high-dimensional data. Before moving forward, let us notice that we present in the following sections only the models and, unless a specific note, the inference of those models is done either by direct likelihood maximization in the supervised context or via the EM algorithm in the unsupervised case (see Section 2).

3.1 Early approaches

Early approaches dealing with the classification of high-dimensional data can be split into three families: dimension reduction methods, regularization methods and parsimonious methods.

Dimension reduction Approaches based on dimension reduction assume that the number p of measured variables is too large and that the data at hand live in a space of lower dimension, let us say $d < p$. Once the data are projected in a low-dimensional space, it is then possible to cluster or discriminate the projected observations to obtain a partition or a supervised classifier for the original data.

The most popular linear method used for dimension reduction is certainly principal component analysis (PCA). It was introduced by Pearson [61] who defines PCA as a linear projection that minimizes the average projection cost. Later, Hotelling [39] proposed another definition for PCA reducing the dimension of the data by keeping as much as possible the variation of the data set. In other words, this method aims to find an orthogonal projection of the data set in a low-dimensional linear subspace, such that the variance of the projected data is maximum. This leads to the classical result where the principal axes $\{u_1, ..., u_d\}$ are the eigenvectors associated with the largest eigenvalues of the empirical covariance matrix S of the data. Several decades later, Tipping and Bishop [70] proposed a probabilistic view of PCA by assuming that the observations are independent realizations of a random variable $Y \in \mathbb{R}^p$ which is linked to a latent variable $X \in \mathbb{R}^d$ through the linear relation:

$$Y = \Lambda^t X + \varepsilon.$$

It is further assumed that $X \sim \mathcal{N}(\mu, I_d)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$, such that the marginal distribution of Y is $\mathcal{N}(\Lambda\mu, \Lambda^t\Lambda + \sigma^2 I_p)$. The estimation of the parameters μ, Λ and b by maximum likelihood conduces in particular to estimate Λ by the eigenvectors associated with the largest eigenvalues of the empirical covariance matrix S of the data. Notice here that PCA is strongly related to factor analysis (FA) [68] which aims to both reduce the dimensionality of the space and to keep the observed covariance structure of the data. Many non-linear techniques have also been proposed such as Kernel PCA [66], non-linear PCA [36] and neural networks based techniques [42].

Despite the popularity of the dimension reduction approach, we would like to caution the reader that reducing the dimension without taking into consideration the classification goal may be dangerous. Such a dimension reduction may yield a loss of information which could have been useful for discriminating the classes or groups. In particular, when PCA is used for reducing the data dimensionality, only the components associated with the largest eigenvalues are kept. Such a practice is mathematically and practically disapproved by Chang [23] who shows that the first components do not necessary contain more discriminative information than the others. In addition, reducing the dimension of the data may not be a good idea since, as discussed in Section 3, it is easier to discriminate groups of points in high-dimensional spaces than in lower dimensional spaces, assuming that one can build a good classifier in high-dimensional spaces.

Regularization It is also possible to see the curse of dimensionality in classification as a numerical problem in the inversion of the covariance matrices Σ_k. From this point of view, a way to tackle the curse of dimensionality is to numerically regularize the estimates of the covariance matrices Σ_k before their inversion. As we will see, most of the regularization techniques have been proposed in the supervised classification framework, but they can be easily used for clustering as well. A simple way to regularize the estimation of Σ_k is to consider a ridge regularization which adds a positive quantity σ_k to the diagonal of the matrix:

$$\tilde{\Sigma}_k = \hat{\Sigma}_k + \sigma_k I_p.$$

Notice that this regularization is often implicitly used in statistical softwares, such as Matlab, for performing a linear discriminant analysis (LDA) where, for instance, the `lda` function spheres the data before analysis. A more general regularization has been proposed by Hastie *et al.* [35] which penalizes also correlations between the predictors. This regularization is used in particular in the supervised classification method penalized discriminant analysis (PDA). Friedman [31] also proposed, for his popular regularized discriminant analysis (RDA), the following regularization:

$$\hat{\Sigma}_k(\lambda, \gamma) = (1-\gamma)\hat{\Sigma}_k(\lambda) + \gamma \left(\frac{\mathrm{tr}(\hat{\Sigma}_k(\lambda))}{p} \right) I_p,$$

where :

$$\hat{\Sigma}_k(\lambda) = \frac{(1-\lambda)(n_k - 1)\hat{\Sigma}_k + \lambda(n-K)\hat{\Sigma}}{(1-\lambda)(n_k-1) + \lambda(n-K)}.$$

Thus, the parameter γ controls the ridge regularization whereas λ controls the contribution of the estimators $\hat{\Sigma}_k$ and $\hat{\Sigma}$, where $\hat{\Sigma}$ estimates the within covariance matrix. Finally, it is also possible to use the Moore–Penrose pseudo-inverse of $\hat{\Sigma}$ instead of the usual inverse $\hat{\Sigma}^{-1}$. The reader can also refer to [57] who provides a comprehensive overview of regularization techniques in discriminant analysis.

The solution based on regularization does not have the same drawbacks as dimension reduction and can be used with less fear. However, all regularization techniques require the tuning of a parameter which is difficult to tune in the unsupervised context, whereas this can be done easily in the supervised context using cross validation.

Constrained models A third way to look at the the curse of dimensionality in classification is to consider it as a problem of over-parameterized modeling. Indeed, the Gaussian model turns out to be highly parameterized which naturally yields inference problems in high-dimensional spaces. Consequently, the use of constrained models is another solution to avoid the curse of dimensionality in model-based classification.

A traditional way to reduce the number of free parameters of Gaussian models is to add constraints on the model through their parameters. Let us recall that the unconstrained Gaussian model (Full-GMM hereafter) is a highly parametrized model and requires the estimation of 20603 parameters when the number of components is $K = 4$ and the number of variables is $p = 100$. A first possible constraint for reducing the number of parameters to estimate is to constraint the K covariance matrices to be the same across all mixture components, *i.e.* $\Sigma_k = \Sigma$, $\forall k$. Notice that this model yields the famous linear discriminant analysis (LDA) [28] method in the supervised classification case. It will be referred hereafter to as the Com-GMM model.

In a similar spirit, Banfield & Raftery [4] and Celeux & Govaert [21] proposed, almost simultaneously, a parameterization of the Gaussian mixture model which yields a family of constrained models. To this end, they parametrize the covariance matrices from their eigenvalue decomposition:

$$\Sigma_k = \lambda_k D_k A_k D_k^t,$$

where D_k is the matrix of eigenvectors which determines the orientation of the cluster, A_k is a diagonal matrix proportional to the eigenvalues which explains its shape, and λ_k is a scalar which controls its volume. This model is referred to by the $[\lambda_k D_k A_k D_k^t]$ model in [21] and to by VVV in [4]. By constraining the parameters λ_k, D_k and A_k within and across the groups, 14 different parsimonious models can be enumerated. Among the 14 models, 4 models are highly parametrized as the Full-GMM model, 4 models have an intermediate level of parsimony as the Com-GMM model and, finally, 6 models

are very parsimonious. This reformulation of the covariance matrices can be viewed as a generalization of the constrained models, presented previously. For example, the Com-GMM model is equivalent to the model $[\lambda DAD^t]$. The reader can refer to [21] for further detail on these models.

The solution which introduces parsimony in the models is clearly a better solution in the context of model-based classification as it proposes a trade-off between the perfect modeling and what one can correctly estimate in practice. In the following sections, we show that recent solutions for high-dimensional classification are partially based on the idea of constrained modeling.

3.2 Model-based subspace classification

Conversely to previous solutions, subspace classification methods exploit the "empty space" phenomenon to ease the discrimination between groups of points. To do so, they model the data in low-dimensional subspaces and introduce some restrictions while keeping all dimensions. Subspace classification methods are mostly related to the factor analysis [64] model which assumes that the observation space is linked to a latent space through a linear relationship.

Mixture of factor analyzers (MFA) Mixture of factor analyzers (MFA) [33, 52] may be considered as the earliest subspace classification method which both classifies the data and reduces locally the dimensionality of each cluster. The MFA model differs from the FA model by the fact that it allows different local factor models, in different regions of the input space, whereas the standard FA assumes a common factor model. The MFA model was initially introduced by Ghahramani & Hinton [33], and then extended by McLachlan et al. [52]. We focus here only on the version of Baek & McLachlan which can be considered as the most general.

The MFA model is an extension of the factor analysis model to a mixture of K factor analyzers. Let $\{y_1, \ldots, y_n\}$ be independent observed realizations of a random vector $Y \in \mathbb{R}^p$. Let us also consider that Y can be expressed from an unobserved random vector $X \in \mathbb{R}^d$, named the factor and described in a lower dimensional space of dimension $d < p$. Moreover, the unobserved labels $\{z_1, \ldots, z_n\}$ are assumed to be independent unobserved realizations of a random vector $Z \in \{1, \ldots, K\}$ where $z_i = k$ indicates that y_i is generated by the kth factor analyzer. The relationship between these two spaces is finally assumed, conditionally to Z, to be linear:

$$Y_{|Z=k} = \Lambda_k X + \mu_k + \varepsilon, \tag{4}$$

where Λ_k is a $p \times d$ matrix and $\mu_k \in \mathbb{R}^p$ is the mean vector of the kth factor analyzer. Moreover $\varepsilon \in \mathbb{R}^p$ is assumed to be a centered Gaussian noise term with a diagonal covariance matrix Ψ, common to all factors:

$$\varepsilon \sim \mathcal{N}(0, \Psi_k). \tag{5}$$

As in the FA model, the factor $X \in \mathbb{R}^d$ is assumed to be distributed according to a Gaussian density function such as $X \sim \mathcal{N}(0, \mathbf{I}_d)$. This implies that the conditional distribution of Y is also Gaussian:

$$Y|X, Z = k \sim \mathcal{N}(\Lambda_k X + \mu_k, \Psi_k). \tag{6}$$

The marginal density of Y is thus a Gaussian mixture model such as $f(y) = \sum_{k=1}^{K} \pi_k \phi(y|\theta_k)$, where π_k stands for the mixture proportion and $\theta_k = \{\mu_k, \Lambda_k \Lambda_k^t + \Psi_k\}$. The complexity of the MFA model is $\gamma_{MFA} = (K-1) + Kp + Kd(p-(d-1)/2) + p$ and, by considering the practical case where $p = 100$, $K = 4$ and $d = 3$, then 1691 parameters have to be estimated for this MFA model. Extensions of this model include [3, 58, 77, 78].

Mixture of parsimonious Gaussian mixture models (PGMM) A general framework for the MFA model was also proposed by McNicholas & Murphy [54] which includes the previous works of Ghahramani and Hinton and of McLachlan et al. [52]. By considering the previous framework, defined by Equations (4) and (6), McNicholas & Murphy [55] proposed a family of models known as the expanded parsimonious Gaussian mixture model (EPGMM) family. They decline 12 EPGMM models by either constraining the terms of the covariance matrix to be equal or not, considering an isotropic variance for the noise term, or re-parametrizing the factor analysis covariance structure. According to this family of 12 models, the previous approaches developed by [3, 33, 52, 54, 71] then become sub-models of the

EPGMM approach. For example, by constraining only the noise variance to be isotropic on each class ($\Psi_k = \sigma_k^2 \mathbf{I}_p$), which by the way corresponds to the UUC and UUUC models, it produces the famous mixture of probabilistic PCA (Mixt-PPCA) of Tipping & Bishop [71]. In the same way, by considering the covariance structure of the UCU and UCCU models such that $\Psi_k = \Psi$ and Λ_k, then we obtain the mixture of factor analyzers model developed by Ghahramani & Hinton. The UUUU model is equivalent to the MFA model proposed by McLachlan et al. in [52]. Finally, by parameterizing the covariance structure as $\Psi_k = \omega_k \Delta_k$, where Δ_k is a diagonal matrix and $|\Delta_k| = 1$, McNicholas & Murphy proposed 4 additional models to their previous work [54]. In the case of the EPGMM models, McNicholas & Murphy [55] proposed in the clustering context to make use of the AECM algorithm for the model inference. It is worth noticing that the AECM could be used for inferring most of the MFA-based models. This model was also used in [53] for discriminant analysis and semi-supervised classification.

Mixture of high-dimensional Gaussian mixture models (HD-GMM) In a slightly different context, Bouveyron et al. [17, 18] proposed a family of 28 parsimonious and flexible Gaussian models to deal with high-dimensional data. Contrary to the previous approaches, this family of GMM was directly proposed in both supervised and unsupervised classification contexts. In order to ease the designation of this family, we propose to refer to these Gaussian models for high-dimensional data by the acronym HD-GMM. Bouveyron et al. [17] proposed to constraint the GMM model through the eigen-decomposition of the covariance matrix Σ_k of the kth group:

$$\Sigma_k = Q_k \Lambda_k Q_k^t,$$

where Q_k is a $p \times p$ orthogonal matrix which contains the eigenvectors of Σ_k and Λ_k is a $p \times p$ diagonal matrix containing the associated eigenvalues (sorted in decreasing order). The key idea of the work of Bouveyron et al. is to reparametrize the matrix Λ_k, such as Σ_k has only $d_k + 1$ different eigenvalues:

$$\Lambda_k = \mathrm{diag}\left(a_{k1}, \ldots, a_{kd_k}, b_k, \ldots, b_k\right),$$

where the d_k first values $a_{k1}, \ldots, a_{kd_k}$ parametrize the variance in the group-specific subspace and the $p - d_k$ last terms, the b_k's model the variance of the noise and $d_k < p$. With this parametrization, these parsimonious models assume that, conditionally to the groups, the noise variance of each cluster k is isotropic and is contained in a subspace which is orthogonal to the subspace of the kth group. Following the classical parsimony strategy, the authors proposed a family of parsimonious models from a very general model, the model $[a_{kj} b_k Q_k d_k]$, to very simple models.

Such an approach can be viewed in two different ways: on the one hand, these models enable to regularize the models in high-dimension. In particular, by constraining d_k such that $d_k = p - 1$ for $k = 1, ..., K$, the proposed approach can be viewed as a generalization of the works of [21, 30]. Indeed, the model $[a_{kj} b_k Q_k (p-1)]$ is equivalent to the Full-GMM model or the $[\lambda_k D_k A_k D_k]$ model in [21]. In the same manner, the model $[a_{kj} b_k Q (p-1)]$ is equivalent to the Diag-GMM and the $[a_j b Q (p-1)]$ is also the Com-Diag-GMM. On the other hand, this approach can also be viewed as an extension of the mixture of principal component analyzer (Mixt-PPCA) model [71] since it relaxes the equality assumption on d_k made in [71] and the model $[a_{kj} b_k Q_k d]$ is therefore equivalent to the Mixt-PPCA model. The estimation of the intrinsic dimensions d_k, $k = 1, ..., K$, relies on the scree test of Cattell [20] which looks for a break in the eigenvalue scree of the empirical covariance matrix of each group.

The discriminative latent mixture (DLM) models Recently, Bouveyron & Brunet [12] proposed a family of mixture models that fit the data into a common and discriminative subspace. This mixture model, called the discriminative latent mixture (DLM) model, differs from the FA-based models by the fact that the latent subspace is common to all groups and is assumed to be the most discriminative subspace of dimension d. This latter feature of the DLM model makes it significantly different from the other FA-based models. Generally speaking, the FA-based models choose the latent subspace(s) maximizing the projected variance whereas the DLM model chooses the latent subspace which maximizes the separation between the groups.

Nevertheless, let us start with a FA-like modeling. Let $Y \in \mathbb{R}^p$ be the observed random vector and let $Z \in \{1, \ldots, K\}$ be once again the unobserved random variable to predict. The DLM model then assumes that Y is linked to a latent random vector $X \in \mathbb{E}$ through a linear relationship of the form $Y = UX + \varepsilon$, where $\mathbb{E} \subset \mathbb{R}^p$ is assumed to be the most discriminative subspace of dimension $d \leq K - 1$

such that $\mathbf{0} \in \mathbb{E}$, $K < p$, U is a $p \times d$ orthonormal matrix common to the K groups and satisfying $U^t U = \mathbf{I}_d$, and $\varepsilon \sim \mathcal{N}(0, \Psi)$ models the non discriminative information. Within the latent space and conditionally to $Z = k$, X is assumed to be distributed as $X|Z = k \sim \mathcal{N}(\mu_k, \Sigma_k)$ where $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d \times d}$ are respectively the mean vector and the covariance matrix of the kth group. Given these distribution assumptions, the marginal distribution of Y is once again a mixture of Gaussians, *i.e.* $g(y) = \sum_{k=1}^{K} \pi_k \phi(y; m_k, S_k)$, where $m_k = U\mu_k$ and $S_k = U\Sigma_k U^t + \Psi$. Let $W = [U, V]$ be the $p \times p$ matrix such that $W^t W = WW^t = \mathbf{I}_p$ and V is an orthogonal complement of U. Finally, the noise covariance matrix Ψ is assumed to satisfy the conditions $V\Psi V^t = \beta \mathbf{I}_{p-d}$ and $U\Psi U^t = \mathbf{0}_d$, such that $\Delta_k = W^t S_k W$ is block-diagonal:

$$\Delta_k = \operatorname{diag}\left(\Sigma_k, \beta I_{p-d}\right)$$

These last conditions imply that the discriminative and the non-discriminative subspaces are orthogonal. This suggests in practice that all of the relevant classification information remains in the latent subspace.

This model is referred to by $\mathrm{DLM}_{[\Sigma_k \beta]}$ in [12]. Following the classical strategy, several other models can be obtained from the $\mathrm{DLM}_{[\Sigma_k \beta]}$ model by relaxing or adding constraints on model parameters. Firstly, it is possible to consider a more general case than the $\mathrm{DLM}_{[\Sigma_k \beta]}$ by relaxing the constraint on the variance term of the non discriminative information. Assuming that $\varepsilon|Z = k \sim \mathcal{N}(0, \Psi_k)$ yields the $\mathrm{DLM}_{[\Sigma_k \beta_k]}$ model which can be useful in some practical cases. From this extended model, 10 parsimonious models can be obtained by constraining the parameters Σ_k and β_k to be common between and within the groups. A model similar to the $\mathrm{DLM}_{[\alpha\beta]}$ model has been considered in [65].

Contrary to most of the MFA-based models, the inference of the DLM models in the unsupervised context cannot be directly done using the EM algorithm because of the specific features of its latent subspace. To overcome this problem, an estimation procedure, called the Fisher-EM algorithm, is proposed in [12] for estimating both the discriminative subspace and the parameters of the mixture model. The convergence properties of the Fisher-EM algorithm were also studied by the same authors in [14] from both the theoretical and the practical points of view. This modeling was also used in the context of supervised and semi-supervised classification and leads to the probabilistic Fisher discriminant analysis (pFDA) method [11]. Let us notice that pFDA turns out to be especially robust to label noise.

3.3 Variable selection for model-based classification

The literature regarding variable selection for supervised classification is abundant and is applied to model-based discriminant analysis methods. We refer to [34] for an overview of this literature. Conversely, variable selection for unsupervised model-based classification has received little consideration until the previous decade. Several recent works have been interested in simultaneously classifying the data and reducing their dimensionality by selecting relevant variables within the model-based classification context. A common assumption of these works is that the true underlying groups are assumed to differ only with respect to some of the original features.

Variable selection as a model selection problem The underlying idea in the works of Law *et al.* [43], Raftery and Dean [63] and Maugis et al. [47] is to find the variables relevant for the clustering task. The determination of the role of each variable is apprehended in [47, 63] as a model selection problem in the GMM context. In the Raftery & Dean approach, the authors define two different sets of variables: $\mathcal{S}$ which denotes the set of relevant variables and $\mathcal{S}_c$ which is the set containing the irrelevant variables. An interesting aspect of their approach is that they do not assume that the irrelevant variables are independent of the clustering variables conversely to Law *et al.* [43]. In particular, they define the irrelevant variables as those which are independent of the clustering but which remain dependent of the set of relevant variables according to a linear relationship. The models in competition are compared with the integrated log-likelihood via a BIC approximation. Thus, the selected model maximizes the following quantity:

$$\left(\hat{K}, \hat{m}, \hat{\mathcal{S}}\right) = \arg \max_{(K,m,r,\ell,V)} \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^{\mathcal{S}}|\mathrm{K}, \mathrm{m}) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{\mathcal{S}_c}|\mathbf{y}^{\mathcal{S}}) \right\}, \tag{7}$$

where K is the number of clusters and $m \in \mathcal{M}$ is a model which belongs to the family of parsimonious models available in the `mclust` [29] software. However, the dependence assumption which defines the irrelevant set of variables according to all the relevant ones remains debatable. Indeed, on the one

hand, considering only the case where the irrelevant variables are independent on both the clustering and the relevant partition, as it was considered in the work of Law et al. [43], seems to be unrealistic. On the other hand, considering that all the irrelevant variables depends on the relevant variables by a linear relationship seems to be a strong hypothesis which may be not valid in certain practical cases. To overcome these limitations, Maugis et al. [22, 47, 48] relax such restrictions and propose a more general variable role modelling. They define two subsets of variables: on the one hand, the relevant ones, which are grouped in $\mathcal{S}$ and, on the other hand, its complementary $\mathcal{S}_c$, which is formed by the irrelevant variables. Maugis et al. consider two types of behaviours among these irrelevant variables: a subset $\mathcal{U}$ of irrelevant variables which can be explained by a linear regression from a subset $\mathcal{R}$ of the clustering variables and a subset $\mathcal{W}$ of irrelevant variables which are totally independent of all relevant variables. Such a variable partition enables the consideration of both the approaches developed by Law et al. [43] and by Raftery & Dean [63].

Variable selection by likelihood penalization An other way to combine variable selection and clustering is to penalize the clustering criteria in order to yield sparsity in the features. This technique has been used, in particular, by penalizing the log-likelihood function to optimize. A general form for the penalized log-likelihood function is:

$$\mathcal{L}_{\mathsf{p}}(\theta) = \ell(\theta) - \mathsf{p}_\lambda(\theta) \tag{8}$$

where $\ell(\theta)$ stands for the log-likelihood function and $\mathsf{p}_\lambda(\theta)$ is the penalty function. In GMM context, Pan & Shen [60] proposed a penalized log-likelihood criterion by assuming a Gaussian mixture model with common diagonal covariance matrices, meaning that $\forall k \in \{1, \ldots, K\}$, $\Sigma_k = \Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_j^2, \ldots, \sigma_p^2)$ where $\sigma_j^2 \in \mathbb{R}$. The penalty function is focused on the means of K clusters $(m_{1k}, \ldots, m_{pk}, \forall k \in \{1, \ldots, K\})$ and has the following form:

$$\mathsf{p}_\lambda(\theta) = \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p} |m_{kj}|\,, \tag{9}$$

where m_{kj} denotes the mean of the jth variable in the component k and λ_1 an hyper-parameter which stands for the desired level of sparsity. Thus, since the observations are standardized, if the means of a variable j on each component are equal i.e. $m_{1j} = \ldots = m_{Kj} = 0$, then this variable is irrelevant and can be removed from the clustering variables. Therefore, a variable selection is realized when some m_{kj}'s can be shrunken toward 0. This situation occurs for an ℓ_1 penalty term large enough. In the same spirit, Wang & Zhou [72] proposed another penalty term based on ℓ_∞-norm which has the advantage to incorporate group information. Thus, this penalty tends to shrink all the m_{kj}'s toward 0 as soon as the jth variable is non informative. However, such a penalty tends to shrink the m_{kj}'s in the same magnitude and thus does no take into account the situation where a variable is different from 0 on only one component. Xie *et al.* [75] extended the model of Pan & Shen [60] by relaxing the equality constraint on the covariance matrices. Finally, Zhan et al. [79] recently proposed a penalization in the case of the GMM model where $S_k = S$, $\forall k \in \{1, \ldots K\}$:

$$\mathsf{p}_\lambda(\theta) = \lambda_1 \sum_{k=1}^{K} \sum_{j=1}^{p} |m_{kj}| + \lambda_2 \sum_{\ell=1}^{p} \sum_{j=1}^{p} |C_{j\ell}|\,,$$

where $\{C_{j\ell}\}_{j,\ell=1}^{p}$ are the elements of the inverse covariance matrix S^{-1}. In the same work, Zhan et al. also proposed an estimation procedure to deal with the $n < p$ case.

Variable selection by penalization of the loadings An alternative approach for selecting the relevant variables through penalization is to directly apply the lasso penalty on the loading matrix of a MFA-based model. This has been achieved in particular in [15, 32, 76]. In the case of the MFA model, Galimberti et al. [32] introduced an ℓ_1-penalty on the factor loadings in the log-likelihood function such as:

$$\mathsf{p}_\lambda(\theta) = \lambda_2 \sum_{\ell=1}^{d} \sum_{j=1}^{p} |b_{\ell j}| \tag{10}$$

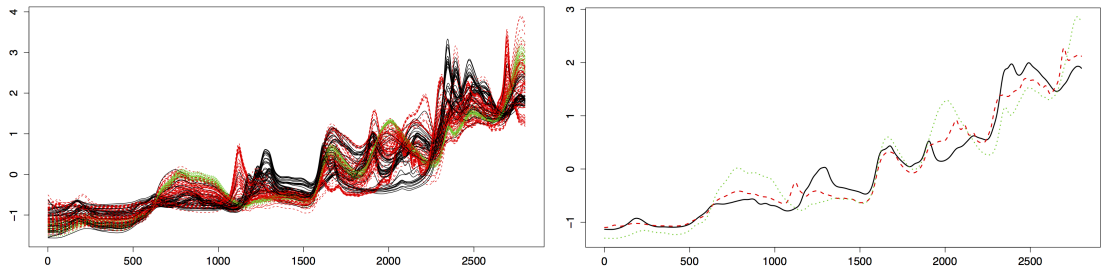


Figure 1: Spectra of the 3-class NIR data set (top) and mean of each class (bottom).

where $b_{\ell j}$ stands for the factor loadings. In a very recent work, Xie et al. [76] proposed a penalized MFA approach from the model introduced by Gharamani & Hinton where the covariance matrix of the noise term is diagonal and common to all factors. The penalty function has the following form:

$$\mathrm{p}_\lambda(\theta) = \lambda_1 \sum_{k=1}^{K}\sum_{j=1}^{p} |m_{kj}| + \lambda_2 \sum_{k=1}^{K}\sum_{j=1}^{p} b_{kj}^2, \tag{11}$$

where b_{kj} stands for the factor loading of the kth factor. As with the previous approaches, the first term based on the ℓ_1-norm is used to shrink the means m_{kj} to be exactly equal to 0 while the second term serves as a grouped variable penalty. Indeed, this last penalty aims to shrink the estimates of factor loadings b_{kj} which are close to 0 to be exactly equal to 0. Consequently, if a variable has a common mean equal to 0, a common variance on each factor across the clusters and is independent with all other cluster such as $b_{kj} = 0\ \forall k$, then this variable is irrelevant and does not contribute in the clustering task.

In the context of Fisher-EM, the direct penalization of the loading matrix U makes particularly sense since it is not estimated by likelihood maximization. The matrix U is indeed estimated in the F-step of the Fisher-EM algorithm by maximizing the Fisher criterion conditionally to the current partition of the data. To achieve the penalization of U, two solutions are proposed in [15]. The first solution is a two stage approach which first estimate U, at each iteration, with the F-step and then looks for its best sparse approximation as follows:

$$\min_U \left\| X^{(q)t} - Y^t U \right\|_F^2 + \lambda \sum_{j=1}^{d} \|u_j\|_1 ,$$

where u_j is the jth column vector of U, $X^{(q)} = \hat{U}^{(q)t} Y$ and $\|\ \|_F$ refers to the Frobenius norm. The solution of this penalized regression problem can be computed through the LARS algorithm [27]. The second solution directly recasts the maximization of the Fisher criterion as a regression problem and provides a sparse loading matrix by solving the lasso problem associated to this regression problem. However, solving this lasso problem is not direct in this case and requires the use of an iterative algorithm. Regarding the implementation details, it is proposed in [15] to initialize the sparseFEM algorithm with the result of the Fisher-EM algorithm and to determine the value of λ by model selection through a modified BIC criterion.

4 Some examples of applications to Chemometrics

This section now considers some specific R packages and data sets in order to illustrate the practical use of model-based classification techniques in Chemometrics.

4.1 Why it is important not to reduce the dimension before classification?

Before continuing, we would like to convince the reader that reducing the data dimension without taking into account the classification goal usually conduces to suboptimal results. Indeed, as we discussed

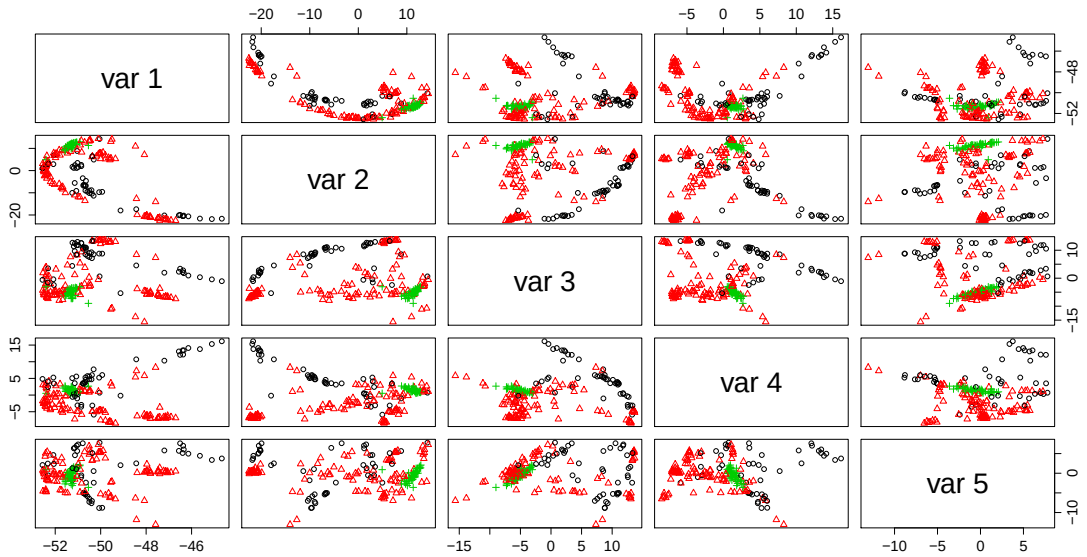


Figure 2: Scatter plot of the 3-class NIR data projected onto the 5 first principal axes.

earlier, the disconnection between the dimension reduction and the classification steps can lead to a loss of information which could have been discriminative.

We here focus on the problem of discriminant analysis and we consider the 3-class NIR data set presented in [26]. The 3-class NIR data set contains 221 NIR spectra of manufactured textiles of various compositions, the classification problem consisting in the determination of a physical property which can take three discrete values [7]. The NIR spectra were measured on a XDS rapid content analyzer instrument (FOSS) in reflectance mode in the range 1100-2500 nm at 0.5 nm apparent resolution (2800 data points per spectrum). Prior to model development, standard normal variate (SNV) was applied on the individual sample spectra as pretreatment. The SNV transformation consists of a centering and a reduction of each spectrum by its own standard deviation. Figure 1 shows the corresponding spectra and the class means.

In order to show that reducing the dimension of the data is usually not useful for the classification task, we consider a very simple approach. We computed the principal axes using PCA on the whole data set (using SVD since here $n \ll p$) and Figure 2 shows the scatter plot of the data projected onto the 5 first principal axes. Knowing the true classes, it appears that the two first principal axes are not the best axes for discriminating the classes. In particular, the projection onto the 4th and 8th principal axes seems a better subspace for classification than the two first ones. To verify this, we evaluated the classification performance of QDA and LDA on the two PCA subspaces (principal axes $1/2$ versus $4/8$) using a 25-fold cross-validation. Figure 3 presents the boxplots of correct classification rates. It clearly shows that the principal plane defined by the 4th and 8th axes allows to obtain better classification results than the first principal plane.

This practical example has therefore make the demonstration that dimension reduction is usually not well adapted to the classification task. Let us notice that supervised dimension reduction methods, such as PLS [74], usually provide better projections than PCA regarding the classification goal, even though they imply some information loss too. A efficient way to overcome such a problem is to use subspace classification methods which model and classify the data in class-specific low-dimensional subspaces without reducing the data dimensionality. Finally, it is worth noticing that the use of a model-based classification method which models each class with a specific mixture model, such as mixture discriminant analysis (MDA) [37], would be particularly useful on the data set used here.

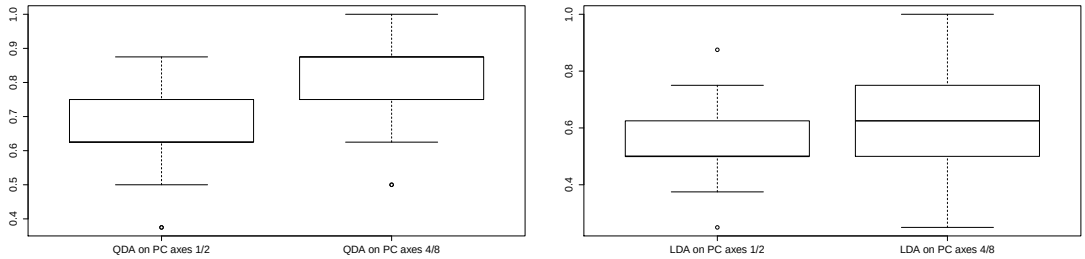


Figure 3: Boxplots of correct classification rates obtained by 25-fold cross validation for QDA (left) and LDA (right) on the PCA axes 1/2 and 4/8.

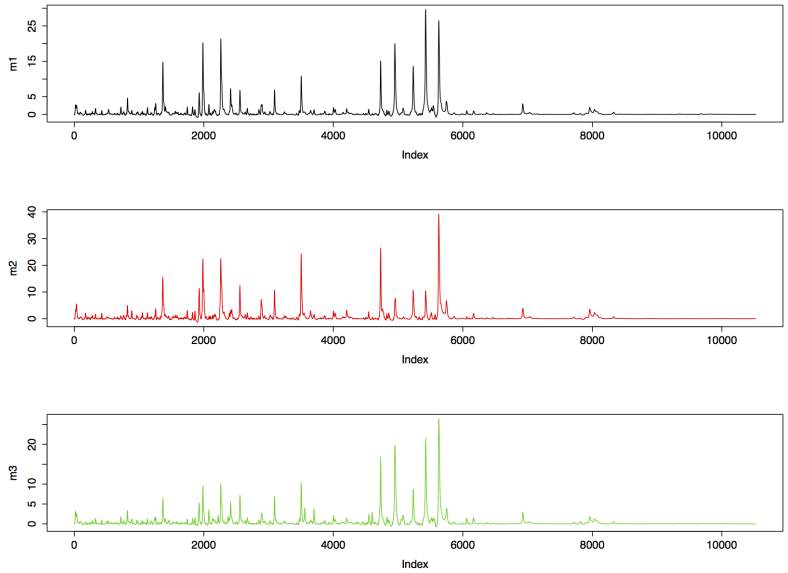


Figure 4: Mean spectra of the three classes for the prostate data set.

4.2 Discriminant analysis

The supervised classification context is considered here as well, but with the use of a recent model-based classification technique. We consider the HDDA method which is implemented in the `HDclassif` package [7] for R. HDDA is the supervised classification method associated with the HD-GMM model presented in Section 4.2 and it is a subspace classification method.

To illustrate the use of HDDA in Chemometrics, we consider the "prostate" data set which is available in the *ChemometricsWithR* package [62] for R. The data were presented in [1]. The data set consists in 327 spectra of blood samples measured on 10523 variables from 2000 to 20000 Da. The samples come from patients with prostate cancer, benign prostatic hypertrophy and normal controls. These spectra are already baseline corrected and normalized. Figure 4 presents the means of the three classes (prostate cancer, benign hypertrophy and control).

For this example, we therefore used the `hdda` function of the `HDclassif` package for R. The R code for loading the data and running the HDDA method on them is:

```
> library(ChemometricsWithR); data(prostate)
> library(HDclassif)
> res = hdda(prostate,prostate.type)
> res
HIGH DIMENSIONAL DISCRIMINANT ANALYSIS
```

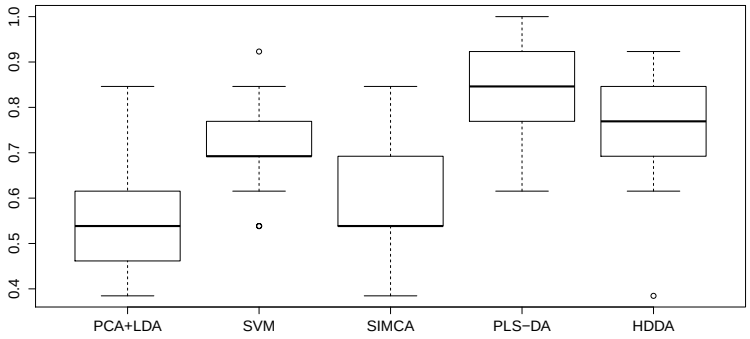


Figure 5: Boxplots of correct classification rates obtained by 25-fold double cross validation for PCA+LDA, SVM, SIMCA, PLS-DA and HDDA for the prostate data set.

```
MODEL: AKJBKQKDK
Prior probabilities of groups:
     bph control   pca
   0.239   0.248 0.514
Intrinsic dimensions of the classes:
       bph control pca
dim:   2       2   6
       Akj:
Class        a1    a2   a3   a4   a5   a6
  bph     20092  6516    .    .    .    .
  control 23257 10105    .    .    .    .
  pca     10860  8987 5717 3330 2683 1887
     bph control   pca
Bk: 1.56    1.42 0.749$
BIC:  -10726683
```

Let us briefly detail the obtained results. HDDA has used here the default model (referred to as `AKJBKQKDK`) and automatically selected the intrinsic dimensions of the classes. The prior probabilities of the classes were respectively 0.239, 0.248 and 0.514. The selected intrinsic dimensions were 2, 2 and 6. The following lines give the variances within and outside the class-specific subspaces.

We then compared HDDA to classical supervised methods used in Chemometrics: LDA on the projected data using PCA (referred to as PCA+LDA), SVM (with a Gaussian kernel), SIMCA [73] and PLS-DA [5]. The experiment was performed using a 25-fold double cross-validation setup (tuning parameters are set up by 5-fold cross-validation within each CV fold). The average selected dimensions for SIMCA, PLS-DA and HDDA where respectively 12, 22 and 77. The boxplots of the correct classification results for each method are presented by Figure 5. On the one hand, one can observe that learning a classifier on the projected data yields once again a less efficient classifier (PCA+LDA). On the other hand, it is particularly interesting to compare the results of HDDA and SIMCA. Indeed, both methods apply PCA on the data of each class but HDDA keeps all principal axes whereas SIMCA keeps only the first ones. It clearly appears here that reducing the dimension, even in a smart fashion, is worst than classifying the data on all dimensions. Finally, HDDA performs here better than SVM and slightly less than PLS-DA, which is in agreement with a recent study [16] on vibrational spectroscopy data.

Furthermore, the probabilistic model of HDDA provides, for each classified observations, the probability of classification error and usually allows meaningful interpretations. In particular, looking at the estimated specific subspaces for each class is usually helpful in understanding the data. Figure 6 shows the projection of the estimated class-specific subspaces by HDDA onto the principal components (PCA axes 2 and 3) for the prostate data set. One can see that here the projection of the data onto the principal axes is very messy. However, the visualizations of the two first axes of the class-specific

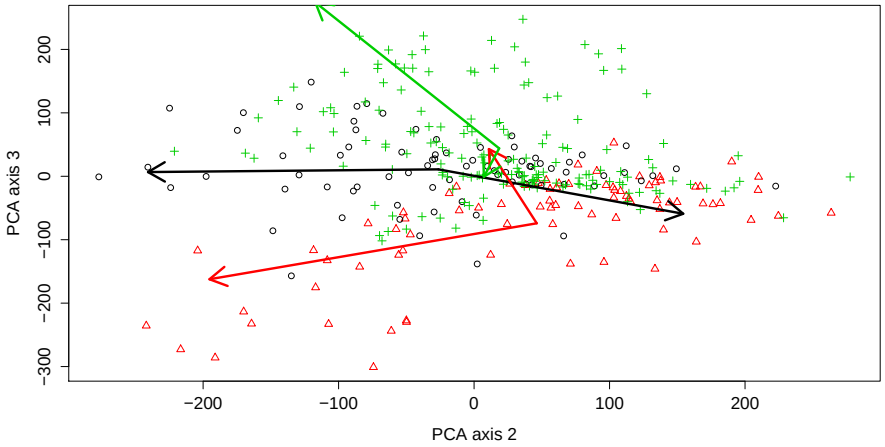


Figure 6: Projection of the estimated class-specific subspaces onto the principal components (PCA axes 2 and 3) for the prostate data set.

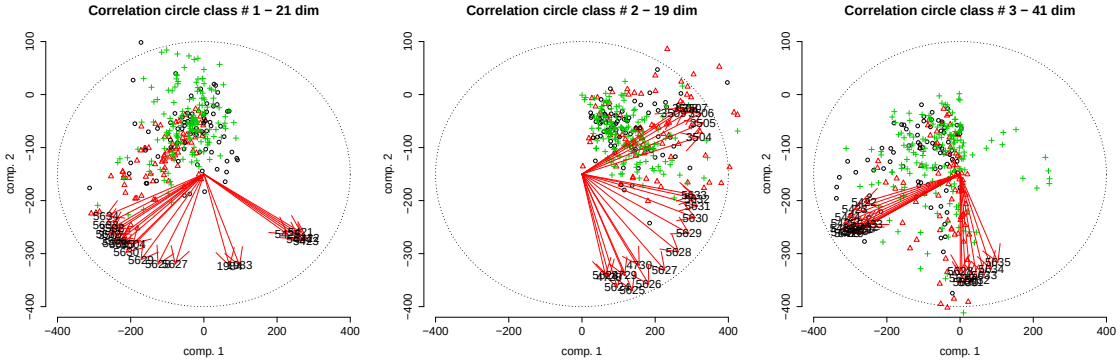


Figure 7: Correlation circles for the three classes of the prostate data set.

subspaces allows to have a better idea of the spaces where the data actually live. It is possible to have a deeper understanding of the data by looking at the correlation circles for the two first axes of the three class-specific subspaces. Figure 7 presents those correlation circles. On each biplot, the red arrows indicate the correlation between the original variables and the two first axes of the subspaces. Only the 25 original variables with the highest correlations are plotted. Those correlation circles should be read in view of the original spectra (at least the mean spectra). One can observe that the peak around the variable 5630 is important for describing the three classes. Classes 1 and 3 seem also characterized by the peak around the variable 5425 whereas class 2 is also characterized by the peak around the variable 3505. Class 1 has also a high correlation with the peak around variable 1943.

4.3 Clustering and variable selection

Here, we propose to use the Fisher-EM algorithm [12] and its sparse version [15] to segment hyperspectral images of the Martian surface. This problem is indeed by its very nature an unsupervised classification problem. Visible and near infrared imaging spectroscopy is a key remote sensing technique to study the system of the planets. Imaging spectrometers, which are onboard of an increasing number of satellites, provide high-dimensional hyperspectral images. In March 2004, the OMEGA instrument (Mars Express, ESA) [8] has collected 310 Gbytes of raw images. The OMEGA imaging spectrometer has mapped the Martian surface with a spatial resolution varying between 300 to 3000 meters depending on the spacecraft altitude. It acquired for each resolved pixel the spectrum from 0.36 to 5.2 µm

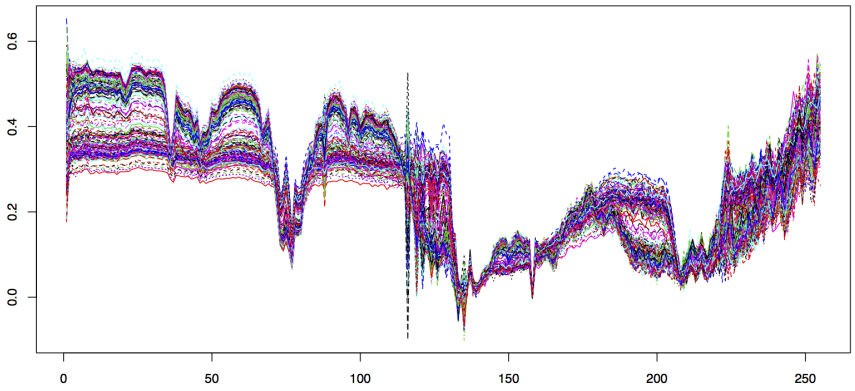


Figure 8: Some of the 38 400 measured spectra described on 256 wavelengths from 0.36 to 5.2 µm (see text for details).

in 256 contiguous spectral channels. OMEGA is designed to characterize the composition of surface materials, discriminating between various classes of silicates, hydrated minerals, oxides and carbonates, organic frosts and ices. For this experiment, a 300×128 image of the Martian surface is considered and a 256-dimensional spectral observation is therefore associated to each of the 38 400 pixels. Figure 8 shows some of the 38 400 measured spectra. According to the experts, there are $K = 5$ mineralogical classes to identify.

We first used the Fisher-EM algorithm to cluster the data set. We used the model $\mathrm{DLM}_{[\alpha_{kj}\beta]}$ with $K = 5$ and the intrinsic dimension d was fixed to $K - 1 = 4$. The R code for running `fem` function of the FisherEM package [13] for R on the Mars Express data is:

```
> # Data are stored in a dataframe named Y
> library(FisherEM)
> res = fem(Y,K=5,model='AkjB')
> str(res)
List of 15
$ model: chr "AkjB"
$ cls   : int [1:38400] 2 2 2 2 2 2 2 2 2 5 ...
$ P     : num [1:38400, 1:5] 0.00 1.09e-302 0.00 0.00 3.55e-303 ...
$ K     : int 5
$ p     : int 255
$ mean  : num [1:5, 1:4] -0.181 -0.144 -0.131 -0.165 -0.129 ...
$ my    : num [1:5, 1:255] 0.47 0.313 0.359 0.415 0.295 ...
$ prop  : num [1:5] 0.263 0.185 0.145 0.166 0.241
$ D     : num [1:5, 1:255, 1:255] 2.61e-05 4.49e-05 8.45e-05 7.39e-05 2.81e-05 ...
$ U     : num [1:255, 1:4] -0.00333 -0.00141 0.00297 -0.00466 0.07837 ...
$ aic   : num 28156951
$ bic   : num 28147159
$ icl   : num NaN
$ loglik: num [1:50] 28160856 28160398 28160059 28159827 28159684 ...
$ ll    : num 28159240
```

The object `res` contains several information which require some comments. First, `cls` and `P` contain respectively the partition into 5 groups of the data and the posterior probabilities that each observation belongs to the groups. The sub-object `prms` gathers all information about the learned mixture model. Chemometricians will be mostly interested in visualizing the group means which are stored in `prms$my`. The estimated group means are plotted on Figure 11 as well as other information that will be commented below. Another parameter which is useful from the practical point of view is the loading matrix `U`. This matrix contains the coordinates of the discriminative axes and allows therefore to project the original data onto the discriminative subspace for further analyses. Figure 9 presents the projection of the clustered data on the estimated discriminative axes with Fisher-EM.

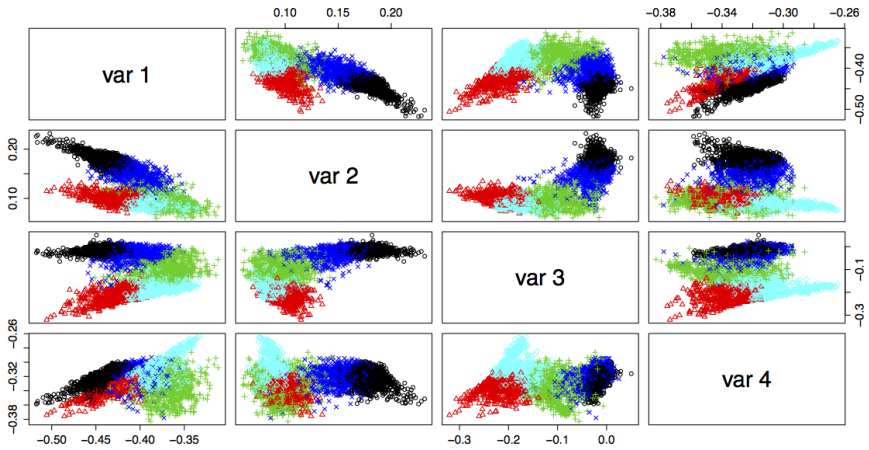
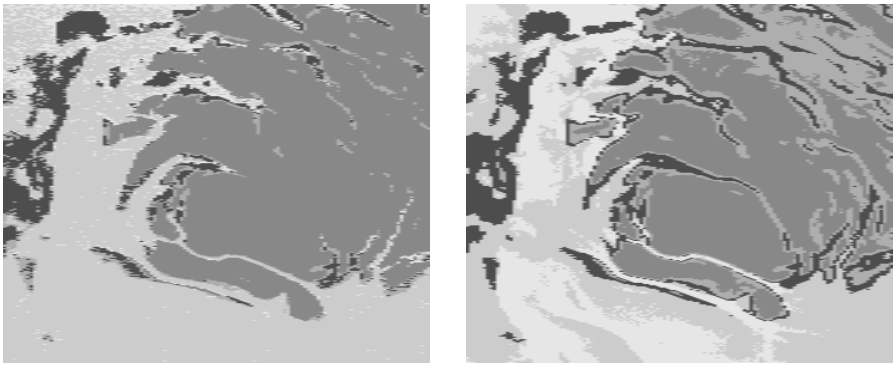


Figure 9: Projection of the clustered data on the estimated discriminative axes with Fisher-EM for the Mars data set.



Expert segmentation Fisher-EM segmentation

Figure 10: Segmentation of the hyperspectral image of the Martian surface using a physical model build by experts (left) and Fisher-EM (right).

Figure 10 presents, on the right panel, the segmentation into 5 mineralogical classes of the studied zone with the Fisher-EM algorithm. In comparison, the left panel of Figure 10 shows the segmentation obtained by experts of the domain using a physical model on the same data. It first appears that the two segmentations agree globally on the mineralogical nature of the surface of the studied zone (60.30% of global agreement). We recall that both segmentations do not exploit the spatial information. When looking at the top-right quarter of the image, we can notice that Fisher-EM seems to provide a precise segmentation of the fine "rivers" which can be seen on Figure 8. This point was particularly appreciated by the experts since their physical model was not able to segment those parts of the image.

The sparseFEM algorithm [15] was then applied to this dataset using a sparsity ratio equal to 0.1 (it refers to the ratio of the ℓ_1 norm of the coefficient vector relative to the norm at the full least square solution). The sparseFEM algorithm was initialized with the results of the Fisher-EM algorithm and the whole segmentation process took 18 hours on a 2.6 Ghz computer. The associated R code is:

```
> res2 = sfem(Y,K=5,model='AkjB',l1=0.1)
```

Figure 11 shows the mean spectra of the 5 groups formed by Fisher-EM and the selection of the discriminative wavelengths. SparseFEM selected 8 original variables (wavelengths) as discriminative variables, *i.e.* the rows associated to these variables were non-zero in the loading matrix U. Looking closely at the selection, we indeed notice that the first selected variable (from left to right) discriminates the blue group from the others. The second selected variable discriminates the red and green groups from the black, blue and light blue groups whereas the third selected variable allows to discriminate the

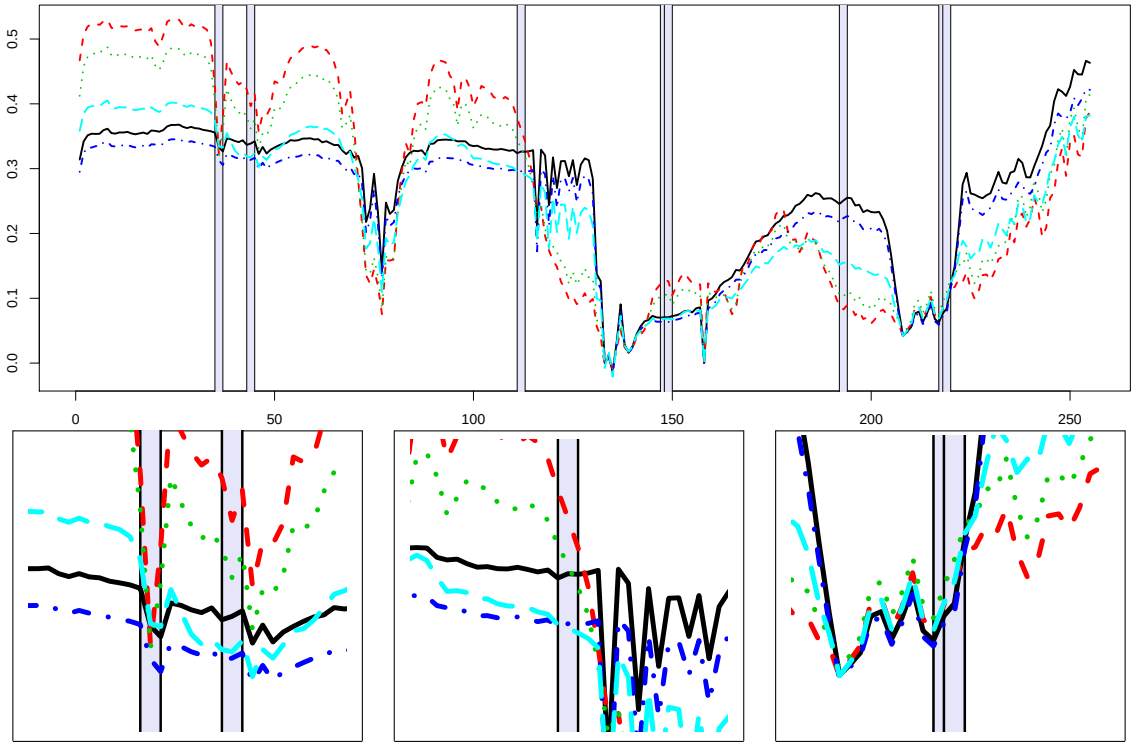


Figure 11: Mean spectra of the 5 groups formed by sparseFEM and selection of the discriminative wavelengths (indicated by gray rectangles). Some details are shown at the bottom to ease the visualization.

red, green and black groups from the blue and light blue groups. Similarly, the fourth and fifth selected variables discriminate the red and green groups from the black, blue and light blue groups whereas the sixth, seventh and eighth selected variable allows to discriminate the red, green and light blue groups from the blue and black groups. A possible interest of such a selection could be the measurement of only tens of wavelengths for future acquisitions instead of the 256 current ones for a result expected to be similar. This could reduce the acquisition time for each pixel from a few tens of seconds to less than one second.

5 Conclusion

This work presents a comprehensive overview of recent model-based methods for the supervised or unsupervised classification of data from Chemometrics. We emphasized the interest of using parsimonious models, subspace classification methods or variable selection methods designed for classification instead of preprocessing the data with dimension reduction. The few practical examples offered here may help the chemometricians applying recent model-based techniques to their own data. Let us finally notice that some recent works [19, 40, 41] have extended model-based classification methods to functional data. Such approaches are naturally of great interest for chemometricians.

References

[1] B.L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Y. Yasui, Z. Feng, and G.L. Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62(13):3609–3614, 2002.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[3] J. Baek, G.J. McLachlan, and L. Flack. Mixtures of Factor Analyzers with Common Factor Loadings: Applications to the Clustering and Visualisation of High-Dimensional Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–13, 2009.

[4] J. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

[5] M. Barker and W. Rayens. Partial least squares for discrimination. *J. Chemometrics*, 17:166–173, 2003.

[6] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[7] L. Bergé, C. Bouveyron, and S. Girard. HDclassif : an R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data. *Journal of Statistical Software*, 42(6):1–29, 2012.

[8] J.-P. Bibring and 42 co-authors. Mars Surface Diversity as Revealed by the OMEGA/Mars Express Observations. *Science*, 307(5715):1576–1581, 2005.

[9] C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725, 2001.

[10] C. Biernacki and J. Jacques. A generative model for rank data based on insertion sort algorithm. *Computational Statistics and Data Analysis*, 58(0):162 – 176, 2013.

[11] C. Bouveyron and C. Brunet. Probabilistic Fisher discriminant analysis: A robust and flexible alternative to Fisher discriminant analysis. *Neurocomputing*, 90(1):12–22, 2012.

[12] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, 22(1):301–324, 2012.

[13] C. Bouveyron and C. Brunet. The FisherEM package for the R software, 2012. `http://cran.r-project.org/web/packages/FisherEM/index.html`.

[14] C. Bouveyron and C. Brunet. Theoretical and practical considerations on the convergence properties of the Fisher-EM algorithm. *Journal of Multivariate Analysis*, 109:29–41, 2012.

[15] C. Bouveyron and C. Brunet. Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. *Computational Statistics*, page in press, 2013.

[16] C. Bouveyron, O. Devos, L. Duponchel, S. Girard, J. Jacques, and C. Ruckebusch. Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *Journal of Chemometrics*, 24(11-12):719–727, 2010.

[17] C. Bouveyron, S. Girard, and C. Schmid. High-Dimensional Data Clustering. *Computational Statistics and Data Analysis*, 52(1):502–519, 2007.

[18] C. Bouveyron, S. Girard, and C. Schmid. High Dimensional Discriminant Analysis. *Communications in Statistics : Theory and Methods*, 36(14):2607–2623, 2007.

[19] C. Bouveyron and J. Jacques. Model-based Clustering of Time Series in Group- specific Functional Subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300, 2011.

[20] R. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):145–276, 1966.

[21] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793, 1995.

[22] G. Celeux, M.-L. Martin-Magniette, C. Maugis, and A.E. Raftery. Letter to the editor. *Journal of the American Statistical Association*, 106(493), 2011.

[23] W.C. Chang. On using principal component before separating a mixture of two multivariate normal distributions. *Journal of the Royal Statistical Society, Series C*, 32(3):267–275, 1983.

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

[25] A. Dempster, N. Laird, and D. Robin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[26] O. Devos, C. Ruckebusch, A. Durand, L. Duponchel, and J-P. Huvenne. Support vector machines (svm) in near infrared (nir) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometr. Intell. Lab. Syst.*, 96:27–33, 2009.

[27] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, may 2004.

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

[29] C. Fraley and A.E. Raftery. MCLUST: Software for Model-Based Cluster Analysis. *Journal of Classification*, 16:297–306, 1999.

[30] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 2002.

[31] J.H. Friedman. Regularized discriminant analysis. *The Journal of the American Statistical Association*, 84:165–175, 1989.

[32] G. Galimberti, A. Montanari, and C. Viroli. Penalized factor mixture analysis for variable selection in clustered data. *Computational Statistics and Data Analysis*, 53(12):4301–4310, oct 2009.

[33] Z. Ghahramani and G.E. Hinton. The EM algorithm for factor analyzers. Technical report, University of Toronto, 1997.

[34] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[35] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23:73–102, 1995.

[36] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[37] T. Hastie and R. Tibshirani. Discriminant analysis by gaussian mixture. *Journal of the Royal Statistical Society*, 58(1):155–176, 1996.

[38] T. Hofmann, B. Schölkopf, and A. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008.

[39] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 1933.

[40] J. Jacques and C. Preda. Funclust: a curves clustering method using functional random variable density approximation. *Neurocomputing*, 112:164–171, 2013.

[41] J. Jacques and C. Preda. Model-based clustering of multivariate functional data. *Computational Statistics and Data Analysis*, in press, 2013. DOI 10.1016/j.csda.2012.12.004.

[42] T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, New York, 1995.

[43] M. Law, M. Figueiredo, and A. Jain. Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Trans. on PAMI*, 26(9):1154–1166, 2004.

[44] G. Lee and C. Scott. Em algorithms for multivariate gaussian mixture models with truncated and censored data. *Computational Statistics and Data Analysis*, 56(9):2816 – 2829, 2012.

[45] B.G. Lindsay. Mixture models: Theory, geometry and applications. In *NSF- CBMS Regional Conference Series in Probability and Statistics*, volume 5. Institute of Mathematical Statistics, 1995.

[46] I. Manolopoulou, T.B. Kepler, and D.M. Merl. Mixtures of gaussian wells: Theory, computation, and application. *Computational Statistics and Data Analysis*, 56(12):3809 – 3820, 2012.

[47] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection for Clustering with Gaussian Mixture Models. *Biometrics*, 65(3):701–709, 2009.

[48] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics and Data Analysis*, 53:3872–3882, 2009.

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

[50] G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley Interscience, New York, 1997.

[51] G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Interscience, New York, 2000.

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

[53] P.D. McNicholas. Model-based classification using latent gaussian mixture models. *Journal of Statistical Planning and Inference*, 1401(5):1175–1181, 2010.

[54] P.D. McNicholas and T.B. Murphy. Parsimonious Gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.

[55] P.D. McNicholas and T.B. Murphy. Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, 26(21):2705–2712, 2010.

[56] V. Melnykov and I. Melnykov. Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics and Data Analysis*, 56(6):1381 – 1395, 2012.

[57] A. Mkhadri, G. Celeux, and A. Nasrollah. Regularization in discriminant analysis: a survey. *Computational Statistics and Data Analysis*, 23:403–423, 1997.

[58] A. Montanari and C. Viroli. Heteroscedastic factor mixture analysis. *Statistical Modelling*, 10(4):441–460, 2010.

[59] A. O'Hagan, T.B. Murphy, and I.C. Gormley. Computational aspects of fitting mixture models via the expectation-maximization algorithm. *Computational Statistics and Data Analysis*, 56(12):3843 – 3864, 2012.

[60] W. Pan and X. Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8:1145–1164, 2007.

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

[62] R. Wehrens Pinheiro. *Chemometrics With R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer, Heidelberg, 2012.

[63] A.E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[64] D. Rubin and D. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[65] G. Sanguinetti. Dimensionality reduction of clustered datasets. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 30(3):1–29, 2008.

[66] B. Schölkopf, A. Smola, and K. Müller. Non linear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[67] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[68] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[69] P.M. Steiner and M. Hudec. Classification of large data sets with mixture models via sufficient em. *Computational Statistics and Data Analysis*, 51(11):5416 – 5428, 2007.

[70] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. Technical Report NCRG-97-010, Neural Computing Research Group, Aston University, 1997.

[71] M.E. Tipping and C.M. Bishop. Mixtures of Probabilistic Principal Component Analysers. *Neural Computation*, 11(2):443–482, 1999.

[72] S. Wang and J. Zhou. Variable selection for model-based high dimensional clustering and its application to microarray data. *Biometrics*, 64:440–448, 2008.

[73] S. Wold. Pattern recognition by means of disjoint principal component models. *Patt. Recogn.*, 8:127–139, 1976.

[74] S. Wold, M. Sjöström, and L. Eriksson. PLS regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

[75] B. Xie, W. Pan, and X. Shen. Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electrical Journal of Statistics*, 2:168–212, 2008.

[76] B. Xie, W. Pan, and X. Shen. Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data. *Bioinformatics*, 26(4):501–508, 2010.

[77] R. Yoshida, T. Higuchi, and S. Imoto. A mixed factor model for dimension reduction and extraction of a group structure in gene expression data. *IEEE Computational Systems Bioinformatics Conference*, 8:161–172, 2004.

[78] R. Yoshida, T. Higuchi, S. Imoto, and S. Miyano. Array cluster: an analytic tool for clustering, data visualization and model finder on gene expression profiles. *Bioinformatics*, 22:1538–1539, 2006.

[79] Z. Zhang, G. Dai, and M.I. Jordan. A flexible and efficient algorithm for regularized fisher discriminant analysis. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 632–647, 2009.