Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data

Julien Jacques, Charles Bouveyron, Stéphane Girard, Olivier Devos, Ludovic Duponchel, Cyril Ruckebusch

Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data

Julien Jacques[1], Charles Bouveyron[2], Stéphane Girard[3],
Olivier Devos[4], Ludovic Duponchel[4] and Cyril Ruckebusch[4*]

[1] Laboratoire Paul Painlevé, UMR CNRS 8524, Université Lille 1, Villeneuve d'Ascq, France
[2] Laboratoire SAMM, Université Paris I Panthéon-Sorbonne, Paris, France
[3] MISTIS, INRIA Rhône-Alpes and Laboratoire Jean Kuntzmann, Grenoble, France
[4] LASIR, UMR CNRS 8516, Université Lille 1, Villeneuve d'Ascq, France

Abstract

In this work, a family of generative Gaussian models designed for the supervised classification of high-dimensional data is presented as well as the associated classification method called High Dimensional Discriminant Analysis (HDDA). The features of these Gaussian models are: i) the representation of the input density model is smooth; ii) the data of each class are modeled in a specific subspace of low dimensionality; iii) each class may have its own covariance structure; iv) model regularization is coupled to the classification criterion to avoid data over-fitting. To illustrate the abilities of the method, HDDA is applied on complex high-dimensional multi-class classification problems in mid-infrared and near infrared spectroscopy and compared to state-of-the-art methods.

KEYWORDS: model-based classification, high-dimensional gaussian model, generative model, vibrational spectroscopy.

1 Introduction

Supervised classification, which aims at attributing unlabeled samples to known classes based on the knowledge of labeled learning samples, is of particular relevance in analytical spectroscopy. In this field, infrared spectroscopy is a traditional technique for the application of chemometric methods which deal with high-dimensional data. In general terms, classification methods can be divided into two categories: discriminative methods on the one hand and generative methods on the other. Discriminative methods consist in estimating directly the classification rule. Support Vector Machines (SVM) [8, chap. 12] is one of the most popular discriminative methods. Conversely, generative classification methods model the complex system which has generated the observed data and then, from the modeling, build the classification rule. Linear Discriminant Analysis (LDA) [8, chap. 4] is probably the most famous generative classification method. Let us note that, despite frequent mistakes in published articles, LDA is actually different from Fisher Linear Discriminant Analysis

*Corresponding author: cyril.ruckebusch@univ-lille1.fr

which first projects the data on the Fisher's axes before applying LDA on the projections. Unfortunately, conventional classification methods usually suffer from the "curse of dimensionality" [2] in high-dimensional spaces. In order to classify high-dimensional spectroscopic data, many of the conventional approaches incorporate Principal Component Analysis (PCA) [9] for dimension reduction: PCA is often applied to the full set of observations as a pre-processing step before applying a classical low-dimensional classification method in the reduced feature space. The PCA-DA method [10] is one of the approaches using such a strategy. PCA dimensionality reduction can also be incorporated in the classification model as in Soft Independent Modeling of Class Analogies (SIMCA), which is a powerful multivariate classification method [5, 22]. SIMCA carries out disjoint PCA analyses for each class, and classifies new data according to the distance to the class-PCA subspaces. The main criticisms are that this classification model does not explicitly consider between-class separation and that the interpretation of group differences is consequently difficult. Multivariate regression can also be used for discrimination as in Partial Least Squares-Discriminant Analysis (PLS-DA) [1, 10, 21]. This method was recently extended (OPLS-DA) for the discrimination between predictive and non-predictive data variation in order to improve interpretation ability [5]. There also exists non-linear discriminative methods such as Support Vector Machines [7, 16], as mentioned previously. SVM classifiers are powerful discriminative methods where dimensionality reduction can be performed but is not mandatory. Nevertheless, as for other non-linear methods, the tuning of the learning parameters is a critical step for SVM and the interpretability of the results is seriously lacking [7, 16].

An alternative and recent way for dealing with the problem of high-dimensional data classification is to model and classify the data in low-dimensional class specific subspaces. The Gaussian models for high-dimensional data and their associated classification method HDDA (High-Dimensional Discriminant Analysis), defined in [4] and under study in the present paper, allow to efficiently model and classify complex high-dimensional data in a parsimonious way since the model complexity is controlled by the intrinsic dimensions of the classes. In contrast to other generative methods which incorporate dimensionality reduction or variable selection for dealing with high-dimensional data [14, 20, 23], HDDA does not reduce the dimension while modeling the data of each class in specific low-dimensional subspace. Thus, no information loss due to data dimensionality reduction is to be deplored and all the available information is used to discriminate the classes. Furthermore, several submodels are defined by introducing constraints on the parameters in order to be able to model different types of data. The choice between these submodels can be done using classical model selection tools as cross-validation or penalized likelihood criteria [17]. An additional advantage is that HDDA models require the tuning of only one parameter, which contributes to the selection of the dimension of each class specific subspace (see Section 2.4). Finally, HDDA presents several numerical advantages compared to other generative classification methods: explicit formulation of the inverse covariance matrix and possibility of building the classifier when the number of learning observations is smaller than the dimension.

The paper is organised as follows. Section 2 introduces the Gaussian models for high-dimensional data, estimation of their parameter and their use in supervised classification. The data sets and experiments are detailed in Section 3. Section 4 presents the results obtained dealing with mid-infrared and near infrared data, in which HDDA is compared with other classification methods. Finally, some concluding remarks are proposed in Section 5.

2 Gaussian models for high-dimensional data classification

Supervised classification aims to associate a new observation $\mathbf{x}$ (a spectrum) with one of the k known classes through a learning set of labeled observations. We refer to [12] for more details on the general classification framework. In this context, a popular approach is the use of the Gaussian mixture model which assumes that each class can be represented by a Gaussian density. This approach assumes that the observations $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ are independent realizations of a random vector $\mathbf{X} \in \mathbb{R}^p$ with density:

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{i=1}^{k} \pi_i \phi(\mathbf{x}, \boldsymbol{\theta}_i), \tag{1}$$

where ϕ is the Gaussian density parametrized by $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ and π_i is the mixture proportion of the ith class. This model gives rise to the well-known Quadratic Discriminant Analysis (QDA) which unfortunately requires the estimation of a very large number of parameters (proportional to p^2). Hopefully, due to the *empty space* phenomenon [18], it can be assumed that high-dimensional data live in subspaces with a dimension lower than the one of the original space. We thus present hereafter Gaussian models modeling the data in low-dimensional and class specific subspaces.

2.1 The Gaussian model $[a_{ij}b_i\mathbf{Q}_i d_i]$

As in the classical Gaussian mixture model framework [12], we assume that class conditional densities are Gaussian $\mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with means $\boldsymbol{\mu}_i$ (the mean spectrum of the i-th class) and covariance matrices $\boldsymbol{\Sigma}_i$ (the covariance matrix between the different wavelengths), for $i = 1, ..., k$. Let $\boldsymbol{Q}_i$ be the orthogonal matrix with the eigenvectors of $\boldsymbol{\Sigma}_i$ as columns. The class conditional covariance matrix $\boldsymbol{\Delta}_i$ is therefore defined in the eigenspace of $\boldsymbol{\Sigma}_i$ by:

$$\boldsymbol{\Delta}_i = \boldsymbol{Q}_i^t \, \boldsymbol{\Sigma}_i \, \boldsymbol{Q}_i. \tag{2}$$

The matrix $\boldsymbol{\Delta}_i$ is thus a diagonal matrix which contains the eigenvalues of $\boldsymbol{\Sigma}_i$. It is further assumed that $\boldsymbol{\Delta}_i$ can be divided into two blocks:

$$\boldsymbol{\Delta}_i = \begin{pmatrix} \boxed{\begin{matrix} a_{i1} & & 0 \\ & \ddots & \\ 0 & & a_{id_i} \end{matrix}} & & \mathbf{0} \\ & \boxed{\begin{matrix} b_i & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & b_i \end{matrix}} & \\ \mathbf{0} & & \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} d_i \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} (p - d_i) \end{matrix} \tag{3}$$

with $a_{ij} > b_i$, $j = 1, ..., d_i$, and where $d_i \in \{1, \ldots, p-1\}$ is unknown. This Gaussian model will be denoted by $[a_{ij}b_i\mathbf{Q}_i d_i]$ in the sequel and Figure 1 summarizes these notations. Let us also remark that assuming $d_i = (p-1)$ for all $i = 1, ..., k$ leads to the classical Gaussian mixture model with full covariance matrices for each mixture component which yields in the supervised framework the well-known QDA. The class specific subspace $\mathbb{E}_i$ is defined as the affine space spanned by the d_i eigenvectors associated to the eigenvalues a_{ij} and such that $\boldsymbol{\mu}_i \in \mathbb{E}_i$. Similarly, the affine subspace $\mathbb{E}_i^{\perp}$ is the affine space spanned by the $p - d_i$

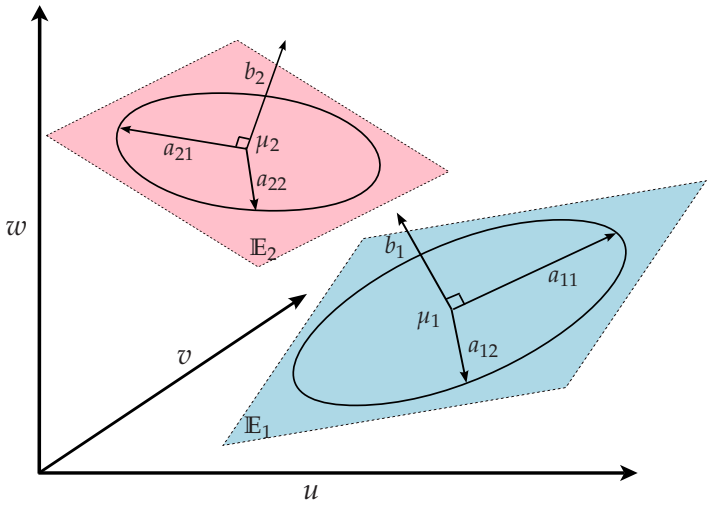

Figure 1: The parameters of model $[a_{ij}b_i\boldsymbol{Q}_i d_i]$ in the case of two classes.

eigenvectors associated with the smallest eigenvalues and such that $\boldsymbol{\mu}_i \in \mathbb{E}_i$. ($\mathbb{E}_i$ and $\mathbb{E}_i^{\perp}$ are illustrated on Figure 2).

In this subspace $\mathbb{E}_i^{\perp}$, the variance is modeled by the single parameter b_i. Let $P_i(\boldsymbol{x}) = \tilde{\boldsymbol{Q}}_i\tilde{\boldsymbol{Q}}_i{}^t(\boldsymbol{x}-\boldsymbol{\mu}_i)+\boldsymbol{\mu}_i$ and $P_i^{\perp}(\boldsymbol{x}) = \bar{\boldsymbol{Q}}_i\bar{\boldsymbol{Q}}_i^t(\boldsymbol{x}-\boldsymbol{\mu}_i)+\boldsymbol{\mu}_i$ be the projection of $\boldsymbol{x}$ on $\mathbb{E}_i$ and $\mathbb{E}_i^{\perp}$ respectively, where $\tilde{\boldsymbol{Q}}_i$ is made of the d_i first columns of $\boldsymbol{Q}_i$ supplemented by $(p-d_i)$ zero columns and $\bar{\boldsymbol{Q}}_i = (\boldsymbol{Q}_i - \tilde{\boldsymbol{Q}}_i)$. Thus, $\mathbb{E}_i$ is called the specific subspace of the ith group since most of the data live on or near this subspace. In addition, the dimension d_i of the subspace $\mathbb{E}_i$ can be considered as the intrinsic dimension of the ith group, *i.e.* the number of variables required to describe the main features of this group.

2.2 The submodels of $[a_{ij}b_i\boldsymbol{Q}_i d_i]$

By fixing some parameters to be common within or between classes, we obtain particular models which correspond to different regularizations. The family of models $[a_{ij}b_i\boldsymbol{Q}_i d_i]$ is divided into three categories: models with free orientations, common orientations and common covariance matrices. Common orientation models are however not discussed in the following due to their expensive computational cost. The different submodels considered in the following can be organized in a hierarchy according to their complexity, as illustrated by Figure 3. More details regarding model complexity are given in Table 2 of [4].

Models with free orientations They assume that the groups live in subspaces with different orientations, *i.e.* the matrices $\boldsymbol{Q}_i$ are specific to each group. Clearly, the general model $[a_{ij}b_i\boldsymbol{Q}_i d_i]$ belongs to this category. Fixing the dimensions d_i to be common between the classes yields the model $[a_{ij}b_i\boldsymbol{Q}_i d]$ which corresponds to the model proposed in [19] in the unsupervised classification framework. As a consequence, our approach encompasses the mixture of probabilistic principal component analyzers introduced in [19] and extended in [13]. In our model, d_i depends on the class and this permits the modeling of a dependence between the number of factors and the class whereas the model of [19] does not. Moreover, our approach can be combined with a "parsimonious models" strategy to further limit the

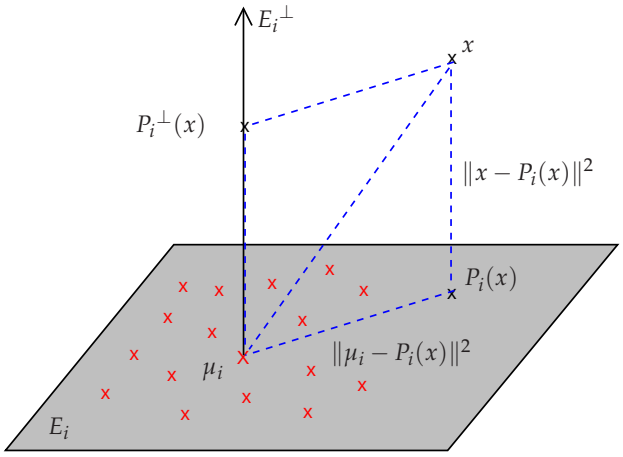

Figure 2: The subspaces $\mathbb{E}_i$ and $\mathbb{E}_i^\perp$ of the ith class.

number of parameters to estimate. It is indeed possible to add constraints on the different parameters to obtain more regularized models. Fixing the first d_i eigenvalues to be common within each class, we obtain the more restricted model $[a_i b_i \mathbf{Q}_i d_i]$. The model $[a_i b_i \mathbf{Q}_i d_i]$ often gives satisfying results, *i.e.* the assumption that each matrix $\mathbf{\Delta}_i$ contains only two different eigenvalues, a_i and b_i, seems to be an efficient way to regularize the estimation of $\mathbf{\Delta}_i$. Another type of regularization is to fix the parameters b_i to be common between the classes. This yields the model $[a_{ij} b \mathbf{Q}_i d_i]$ and $[a_i b \mathbf{Q}_i d_i]$ which assume that the variance outside the class specific subspaces is common. This can be viewed as modeling the noise in $\mathbb{E}_i^\perp$ by a single parameter b, which is natural when the data are obtained in a common acquisition process. This category of models also contains the models $[a b_i \mathbf{Q}_i d_i]$, $[a b \mathbf{Q}_i d_i]$ and all models with common d_i. The total number of models of this category considered in the present paper is thus equal to twelve.

Models with common covariance matrices This branch of the family only includes two models $[a_j b \mathbf{Q} d]$ and $[a b \mathbf{Q} d]$. Both models indeed assume that the classes have the same covariance matrix $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Delta}\mathbf{Q}^t$. Particularly, fixing $d = (p-1)$, the model $[a_j b \mathbf{Q} d]$ reduces to a Gaussian mixture model which yields in the supervised framework the well-known LDA. Note that if $d < (p-1)$, the model $[a_j b \mathbf{Q} d]$ can be viewed as a combination of a "dimension reduction" technique with a Gaussian model with common covariance matrices, but without losing information since the information carried by the smallest eigenvalues is not discarded.

2.3 Supervised classification: the HDDA method

The use of the models presented in the previous paragraphs gave birth to a method called high-dimensional discriminant analysis (HDDA [4]).

Parameter estimation In the context of supervised classification, the learning data are complete, *i.e.* a label z indicating the class belonging is available for each learning observation

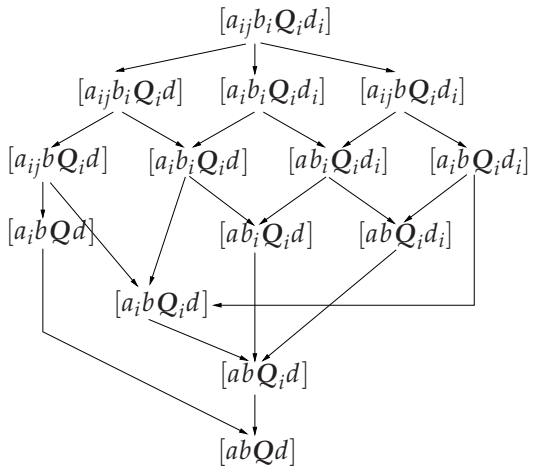

Figure 3: Hierarchy of the fourteen HDDA models, from the most complex (top) to the simplest (bottom).

$\boldsymbol{x}$. The estimation of model parameters is therefore direct through the maximum likelihood method and yields the following estimators. The mixture proportions and the means are respectively estimated by:

$$\hat{\pi}_i = \frac{n_i}{n},\ \hat{\mu}_i = \frac{1}{n_i}\sum_{j/z_j=i} \boldsymbol{x}_j,$$

where n_i is the number of observations in the ith class and z_j indicates the class number of observation $\boldsymbol{x}_j$. The estimation of the specific parameters of the models with free orientations is detailed hereafter. For the other models, details are given in [4]. For models with free orientations, the maximum likelihood estimators are closed-form:

- Orientation matrix $\boldsymbol{Q}_i$: the d_i first columns of $\boldsymbol{Q}_i$ are estimated by the eigenvectors associated with the d_i largest eigenvalues λ_{ij} of the empirical covariance matrix $\boldsymbol{W}_i$ of the ith class:

$$\boldsymbol{W}_i = \frac{1}{n_i}\sum_{j/z_j=i} (\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j - \hat{\boldsymbol{\mu}}_i)^t.$$

- Model $[a_{ij}b_i\boldsymbol{Q}_i d_i]$: the estimator of a_{ij} is $\hat{a}_{ij} = \lambda_{ij}$ and the estimator of b_i is the mean of the $(p - d_i)$ smallest eigenvalues of $\boldsymbol{W}_i$. It can be reformulated as follows:

$$\hat{b}_i = \frac{1}{(p - d_i)}\left(\mathrm{tr}(\boldsymbol{W}_i) - \sum_{j=1}^{d_i}\lambda_{ij}\right), \tag{4}$$

  where $\mathrm{tr}(\boldsymbol{W}_i)$ is the trace of matrix $\boldsymbol{W}_i$.

- Model $[a_{ij}b\boldsymbol{Q}_i d_i]$: the estimator of a_{ij} is $\hat{a}_{ij} = \lambda_{ij}$ and the estimator of b is:

$$\hat{b} = \frac{1}{(p - \delta)}\left(\mathrm{tr}(\boldsymbol{W}) - \sum_{i=1}^{k}\hat{\pi}_i\sum_{j=1}^{d_i}\lambda_{ij}\right), \tag{5}$$

  where $\delta = \sum_{i=1}^{k}\hat{\pi}_i d_i$ and $\boldsymbol{W} = \sum_{i=1}^{k}\hat{\pi}_i\boldsymbol{W}_i$ is the within covariance matrix.

- Model $[a_i b_i \boldsymbol{Q}_i d_i]$: the estimator of b_i is given by (4) and the estimator of a_i is :

$$\hat{a}_i = \frac{1}{d_i} \sum_{j=1}^{d_i} \lambda_{ij}. \tag{6}$$

- Model $[a b_i \boldsymbol{Q}_i d_i]$: the estimator of b_i is given by (4) and the estimator of a is :

$$\hat{a} = \frac{1}{\delta} \sum_{i=1}^{k} \hat{\pi}_i \sum_{j=1}^{d_i} \lambda_{ij}. \tag{7}$$

- Model $[a_i b \boldsymbol{Q}_i d_i]$: the estimators of a_i and b are respectively given by (6) and (5).
- Model $[a b \boldsymbol{Q}_i d_i]$: the estimators of a and b are respectively given by (7) and (5).

Classification of new observations As in the usual case, the classification of a new observation $\boldsymbol{x} \in \mathbb{R}^p$ can be done using the *maximum a posteriori* (MAP) rule which assigns the observation $\boldsymbol{x}$ to the class with the largest posterior probability. Therefore, the classification step mainly consists in computing $\mathbb{P}(Z = i|X = \boldsymbol{x})$ for each class $i = 1, ..., k$:

$$\mathbb{P}(Z = i|X = \boldsymbol{x}) = 1 \Big/ \sum_{\ell=1}^{k} \exp\left(\frac{1}{2}(K_i(\boldsymbol{x}) - K_\ell(\boldsymbol{x}))\right),$$

where the cost function $K_i(\boldsymbol{x}) = -2\log(\pi_i \phi(\boldsymbol{x}, \theta_i))$ has the following form in the case of the model $[a_i b_i \boldsymbol{Q}_i d_i]$:

$$K_i(\boldsymbol{x}) = \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x - P_i(x)\|^2 + \sum_{j=1}^{d_i} \log(a_{ij}) + (p - d_i)\log(b_i) - 2\log(\pi_i). \tag{8}$$

Let us note that $K_i(\boldsymbol{x})$ is mainly based on two distances (illustrated in Figure 2): the distance between the projection of $\boldsymbol{x}$ on $\mathbb{E}_i$ and the mean of the class and the distance between the observation and the subspace $\mathbb{E}_i$. This function favors the assignment of a new observation to the class for which it is close to the subspace and for which its projection on the class subspace is close to the mean of the class. The variance terms a_i and b_i balance the importance of both distances. For example, if the data are very noisy, *i.e.* b_i is large, it is natural to balance the distance $\|\boldsymbol{x} - P_i(\boldsymbol{x})\|^2$ by $1/b_i$ in order to take into account the large variance in $\mathbb{E}_i^{\perp}$. At this point, we can make a link with the SIMCA method. Indeed, SIMCA classifies a new data according to the distance $\|\boldsymbol{x} - P_i(\boldsymbol{x})\|^2$ whereas HDDA classifies it according to both distances.

2.4 Intrinsic dimension estimation

For HDDA, the intrinsic dimension of each subclass has to be estimated and this is a difficult problem with no unique technique to use. Our approach is based on the eigenvalues of the class conditional covariance matrix $\boldsymbol{\Sigma}_i$ of the ith class. The jth eigenvalue of $\boldsymbol{\Sigma}_i$ corresponds to the fraction of the full variance carried by the jth eigenvector of $\boldsymbol{\Sigma}_i$. The class specific dimension d_i, $i = 1, ..., k$ can be estimated through the scree-test of Cattell [6] which looks for

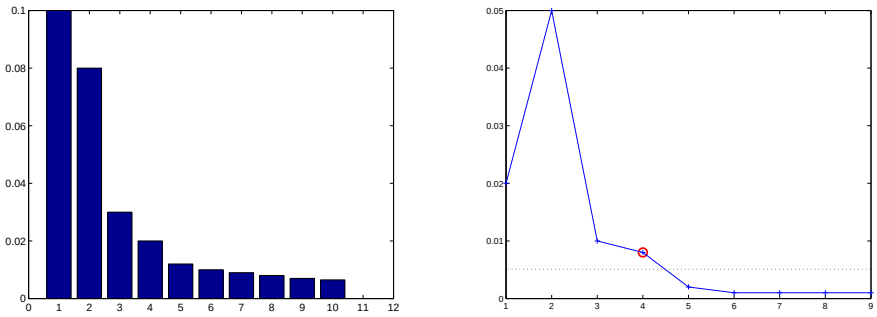

Figure 4: Estimation of the intrinsic dimension d_i using the scree-test of Cattell: plot of ordered eigenvalues of Σ_i (left) and differences between consecutive eigenvalues (right).

a break in the eigenvalue scree. The selected dimension is the one for which the subsequent eigenvalues differences are smaller than a threshold. Figure 4 illustrates this method: the graph on the right shows that the differences between eigenvalues after the fourth one are smaller than the threshold (dashed line). Thus, in this case, four dimensions will be chosen and this corresponds indeed to a break in the scree (left graph). The threshold can be chosen by either cross-validation on the learning set or using the Bayesian Information Criterion (BIC) [17]. In addition, this approach allows to estimate k parameters by choosing only the value of the threshold t. In the case of common intrinsic dimensions between the groups, the dimension d is directly determined using either cross-validation or BIC. Finally, in the specific case of the model $[a_i b_i \boldsymbol{Q}_i d_i]$, it has been recently demonstrated in [3] that it is possible to determine the intrinsic dimensions d_i by likelihood maximization. This allows an automatic and fast intrinsic dimension selection for this specific model.

2.5 Model selection

The previous paragraphs proposed a family of parsimonious Gaussian models ranging from the most complex to the simplest. In a real situation, the practician will have to choose one of the models for applying it to his data set. This choice can be done either based on the practician knowledge on the data set (common noise, ...) or using model selection tools. Among the existing model selection tools, we propose to use one of the two following tools depending on the experimental conditions. The first tool is cross-validation (CV) which approximates the actual classification performance by iteratively evaluating it on subsets of the learning sample. It is often a good way to select a model in the framework of supervised classification but this method has usually a high computational cost. Alternatively, the BIC criterion, only available for generative models, consists in selecting the best model according to the adequacy to the data penalized by the model complexity. The BIC criterion is formally defined by $\text{bic}(m) = -2\log(L(m)) + \nu(m)\log(n)$, where $\nu(m)$ is the number of parameters of the model, $L(m)$ is the maximum of the likelihood and n is the number of observations. The great interest of such a criterion is that it does not need any additional computation, since the likelihood is already computed during the estimation of the model parameters. We refer to [8, chap. 7] for a comparison of these tools.

2.6 Numerical considerations

The nature of the HDDA models implies several numerical advantages. First, it is important to remark that the parametrization of the Gaussian model proposed here provides an explicit expression of $\mathbf{\Sigma}_i^{-1}$ whereas other classical methods, such as QDA and LDA for instance, need to numerically invert empirical covariance matrices which usually fails for singularity reasons. To avoid this problem, it is a common practice to first reduce the data dimension with PCA but this implies a potential loss of information. In contrast, this problem does not arise with HDDA since the cost function K_i, equation (8), does not require to invert $\mathbf{\Sigma}_i$. Moreover, it appears in (8) that the cost function K_i does not use the projection on the subspace $\mathbb{E}_i^{\perp}$ and consequently does not require the computation of the last $(p - d_i)$ columns of the orientation matrix $\boldsymbol{Q}_i$. It has been shown in the previous paragraphs that the maximum likelihood estimators of these columns are the eigenvectors associated to the $(p - d_i)$ smallest eigenvalues of the empirical covariance matrix $\boldsymbol{W}_i$. Therefore, HDDA does not depend on these eigenvectors whose determination is numerically unstable. Thus, HDDA is robust with respect to ill-conditioning and singularity problems. In addition, it is also possible to use this feature to reduce computing time by using the Arnoldi method [11] which only provides the largest eigenvalues and the associated eigenvectors of an ill-conditioned matrix. During our experiments, we noticed a reduction by a factor 60 of the computing time on a 1024-dimensional data set compared to the classical approach. Furthermore, in the special case where the number of observations of a group n_i is smaller than the dimension p, our parametrization allows to use a linear algebra trick. Indeed, in this case, it is better from a numerical point of view to compute the eigenvectors of the $n_i \times n_i$ matrix $\mathbf{Y}_i\mathbf{Y}_i^t$ than those of the $p \times p$ matrix $\mathbf{Y}_i^t\mathbf{Y}_i$, where $\mathbf{Y}_i$ is the $n_i \times p$ matrix containing the mean-centered observations. In the case of data containing 13 observations in a 1024-dimensional space, it has been observed a reduction by a factor 500 of the computing time compared to the classical approach.

3 Datasets

In the present paper, the HDDA models are applied to the analysis of two different multiclass data sets. The first one is a 3-class problem where the observations are near infrared (NIR) spectra of different manufactured textile materials. The second example is a 4-class problem where the observations are mid-infrared (MIR) spectra of natural products. These two challenging situations are typical examples of the search for alternative analytical methods capable of rapid analysis and robust characterization or identification of raw materials. These two data sets were respectively explained and published in [7] and [16], and therefore only brief description is given thereafter. They represent realistic and challenging situations to evaluate the classification power of HDDA.

3.1 3-class NIR data set

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 samples were separated in a learning subset (130 samples) and a test subset (91 samples). 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

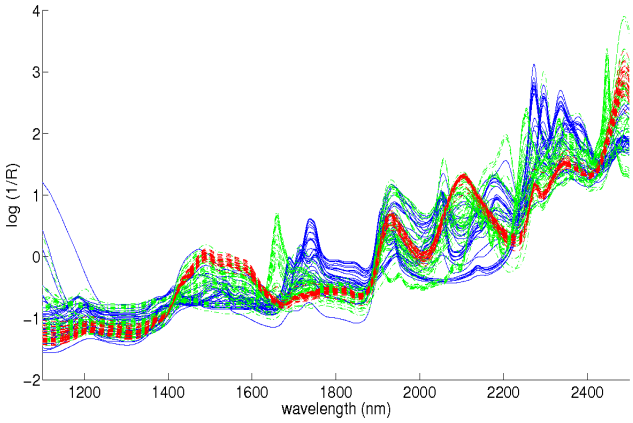

Figure 5: Spectra of the 3-class NIR learning set. Spectra are SNV pre-treated and coloured-coded according to class membership (blue solid line: class 1, green dashed line: class 2, red dash-dotted line: class 3).

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 5 shows the corresponding spectra.

3.2 4-class MIR data set

The second data set is composed by 258 MIR spectra of modified starches samples from different origins and of four different classes. This 4-class data set was proposed for a chemometric contest during the "Chimiométrie 2005" conference [15]. We also refer to [16]. Pierna *et al.* [15] have obtained very good classification results by using SVM. The data set studied in the current work is composed of a learning subset of 215 samples and of a test subset of 43 samples, among which 4 outliers were artificially introduced. The spectroscopic data were analyzed as provided for the contest, without pretreatment. The MIR spectra are depicted on Figure 6.

3.3 Software

The HDDA method is currently available through both stand-alone Matlab toolboxes and within the Mixmod software. For our study, we have used the Matlab toolboxes, available for download at *http://samm.univ-paris1.fr/-charles-bouveyron-*. Alternatively, the Mixmod software provides eight of the most useful models presented in this article (available for download at *http://www-math.univ-fcomte.fr/mixmod/*).

4 Results and discussion

4.1 3-class NIR data set

The results obtained for the fourteen HDDA models described in Section 2.2 are presented in Table 1. The choice of the model dimension (for models with fixed dimensions) and

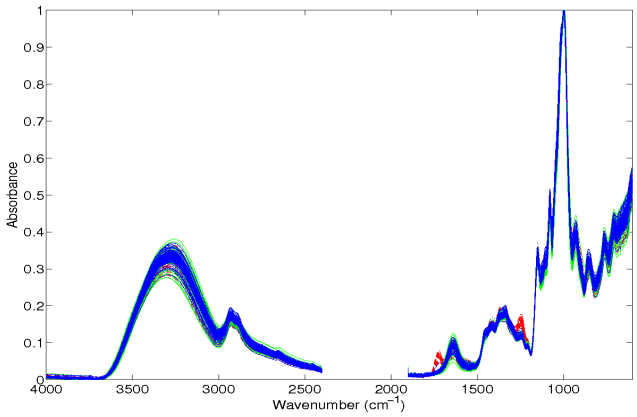

Figure 6: Spectra of the [16] 4-class MIR learning set. Spectra are coloured-coded according to class membership.

the choice of the threshold (for models with free dimensions) were done by a 5-fold cross-validation. For each model, the following results are presented: the correct classification rate on the learning subset estimated by 5-fold cross-validation (*learning CV-CCR*), the value of the BIC criterion, the correct classification rate on the test subset (*test CCR*) and the class specific subspace dimensions (d_i). Using cross-validation on the learning sample leads to select the three models with fixed dimensions and common variance outside the class specific subspaces: $[a_{ij}b\mathbf{Q}_id]$, $[a_ib\mathbf{Q}_id]$ and $[ab\mathbf{Q}_id]$, for which 92.3% correct classification rate was obtained. The dimensions d_i retained for these models were the same for each class: 16. It should be noticed the good agreement between the cross-validation and the BIC criterion, retaining the model $[a_{ij}b\mathbf{Q}_id]$, which is one of the models providing the best test CCR (96.7%). When compared to the results obtained with more classical chemometric methods, such as SIMCA (CCR 82.4%) and PLS-DA (CCR 87.7%), the results obtained with HDDA models show improved performances. The fact that these two methods showed poorer classification performance can be explained by the complexity of the problem where the three classes strongly overlap. The results provided for the SVM (91% test CCR) are the ones obtained by the authors on exactly the same data set (for details regarding SVM parameter optimization, we refer to [7]). Even if the HDDA models show better CCR performance on this data set, the purpose is here more to demonstrate the potential of the method for real life spectroscopy applications than to perform a formal comparison between the methods. Note also that HDDA is computationally faster than SVM: the learning step takes about 3 minutes[1] for SVM, less than 1 minute for PLS-DA and for HDDA (about 3 seconds per model). In addition to good classification performance, one of the most important features of HDDA models is that one may benefit from the interpretation of their estimated parameters. We thereafter focus on the three main points.

- First, as a result of generative modeling, each class is finally characterized by a mean spectrum and a covariance matrix. This latter expresses the dispersion of the spectra of the class around the mean spectrum. Figure 7 represents the mean spectra of the three classes obtained with the $[a_{ij}b\mathbf{Q}_id]$ model. This enables to point out which vari-

[1]Computing times are given for a 2.6 GHz bi-processor with 2Go RAM.

Table 1: 3-class NIR data set: Correct classification rates on the learning sample evaluated by 5-fold cross-validation (*CV-CCR on learning*), BIC value and correct classification rates on the test sample (*CCR on test*), and dimensions of the class spectific subspace for the fourteen HDDA models, SVM, SIMCA and PLS-DA.

model	learning CV-CCR	BIC	test CCR	d_i
$[a_{ij}b_i\boldsymbol{Q}_id_i]$	85.4%	-1422737	83.5%	$(5,7,6)$
$[a_{ij}b\boldsymbol{Q}_id_i]$	91.5%	-2133807	94.5%	$(16,15,14)$
$[a_ib_i\boldsymbol{Q}_id_i]$	85.4%	-1422097	83.5%	$(5,7,6)$
$[ab_i\boldsymbol{Q}_id_i]$	85.4%	-1421695	83.5%	$(5,7,6)$
$[a_ib\boldsymbol{Q}_id_i]$	91.5%	-2128701	94.5%	$(16,15,14)$
$[ab\boldsymbol{Q}_id_i]$	91.5%	-2127724	94.5%	$(16,15,14)$
$[a_{ij}b_i\boldsymbol{Q}_id]$	85.4%	-1422205	82.4%	$(8,8,8)$
$[a_{ij}b\boldsymbol{Q}_id]$	**92.3**%	**-2162226**	**96.7**%	$(16,16,16)$
$[a_ib_i\boldsymbol{Q}_id]$	85.4%	-1420973	82.4%	$(8,8,8)$
$[ab_i\boldsymbol{Q}_id]$	85.4%	-1420407	82.4%	$(8,8,8)$
$[a_ib\boldsymbol{Q}_id]$	**92.3**%	-2156407	**96.7**%	$(16,16,16)$
$[ab\boldsymbol{Q}_id]$	**92.3**%	-2155267	**96.7**%	$(16,16,16)$
$[a_jb\boldsymbol{Q}d]$	70.8%	-381399	73.6%	$(3,3,3)$
$[ab\boldsymbol{Q}d]$	70.8%	-381367	73.6%	$(3,3,3)$
SVM	88.5%	-	91.2%	-
SIMCA	-	-	82.4%	-
PLS-DA	87.7%	-	84.7%	-

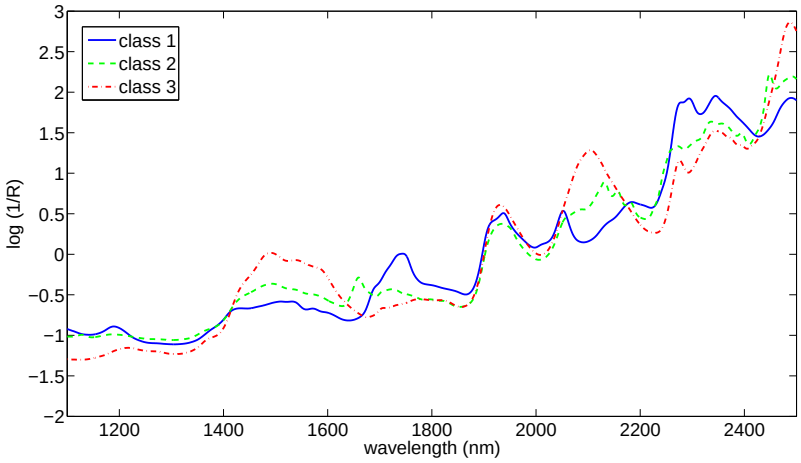

Figure 7: Mean spectra of the classes of NIR data set (blue solid line: class 1, green dashed line: class 2, red dash-dotted line: class 3).

ables or which characteristic features are directly responsible for the discrimination between the different classes. In addition, Figure 8 illustrates the variance which can be attributed to each of the three classes in the problem studied here. Very different situations are observed for Class 1 and Class 2 on the one hand and Class 3 on the other hand, for which variance is very small. In the example investigated here, this result can be explained by the fact that class 3 mainly contains pure samples in term of composition (e.g. pure cotton), whereas the two other classes mainly contain blends of different fibres (e.g. cotton/polyester). This is confirmed by the poor performances of the models with common covariance matrices $[a_j b \mathbf{Q} d]$ and $[ab\mathbf{Q}d]$. Gathered together, these results may lead to better interpretation in terms of the underlying chemical properties.

- Second, the competition between the parcimonious HDDA models can also be interpreted. Effectively, the three models $[a_{ij}b\mathbf{Q}_i d]$, $[a_i b\mathbf{Q}_i d]$ and $[ab\mathbf{Q}_i d]$ retained by cross-validation model selection share the property that the noise is assumed to be common to the three classes. Indeed, a unique parameter b is set to describe $\mathbb{E}_i^{\perp}$. In the current situation, this could be explained by the fact that all the spectra were acquired with same instrument and in the same experimental conditions. From a more general perspective, the parametrization of the models provides information that can be interpreted.

- Third, since HDDA leads to the parametrisation of the class specific subspaces, the correlation between the spectroscopic variables and the ones of each class specific subspace can be computed, as illustrated on Figure 9. These figures are equivalent to the correlation circle in PCA and they allow to determine which spectroscopic variables contribute to the discriminant subspaces of each class. For instance, comparing these three figures leads to suggest that two groups of wavelengths, respectively around 1800 nm and 2300 nm, contribute to the subspace specific to the first class. This information can be confirmed by Figures 5 and 7. Nevertheless, direct chemical interpretation can hardly be expected in applications concerning qualitative prediction of raw

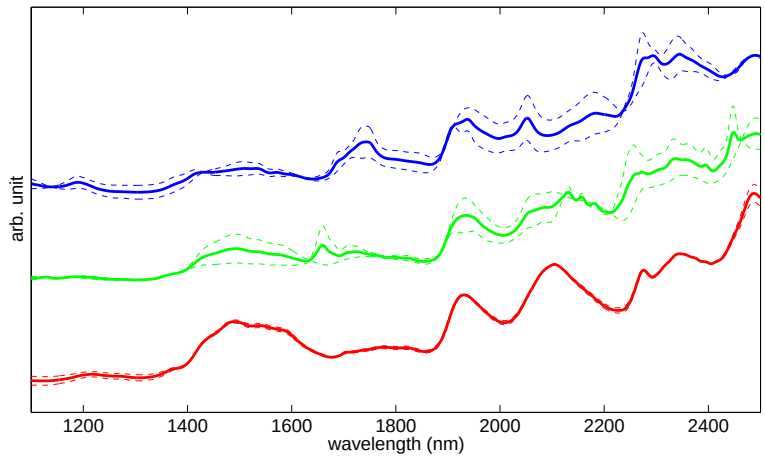

Figure 8: Mean spectra of the three classes of the NIR data set and confidence bounds expressing the variance of each class. (Class 1 to class 3 from the top to the bottom with an artificial shift).

sample properties from NIR spectroscopy.

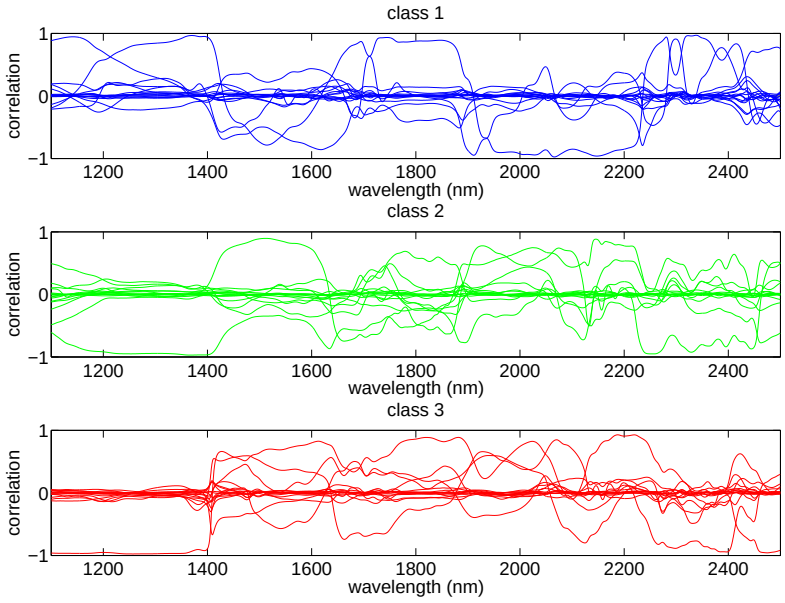

Figure 9: Correlation between wavelengths and class specific subspaces variables (one curve per variable), for class 1 (top) to class 3 (bottom).

Table 2: 4-class MIR data set: Correct classification rates on the learning sample evaluated by 5-fold cross-validation (*CV-CCR on learning*), BIC value and correct classification rates on the test sample (*CCR on test*), and dimensions of the class spectific subspace for the fourteen HDDA models, SVM, and PLS-DA. The two last lines corresponds to the results obtained by [15].

model	learning CV-CCR	BIC	test CCR	d_i
$[a_{ij}b_i\mathbf{Q}_i d_i]$	88.4%	-6560741	72.1%	$(9,9,8,10)$
$[a_{ij}b\mathbf{Q}_i d_i]$	91.2%	-6551380	81.4%	$(9,9,8,10)$
$[a_i b_i\mathbf{Q}_i d_i]$	88.4%	-6556468	72.1%	$(9,9,8,10)$
$[ab_i\mathbf{Q}_i d_i]$	88.4%	-6556244	72.1%	$(9,9,8,10)$
$[a_i b\mathbf{Q}_i d_i]$	91.2%	-6547107	81.4%	$(9,9,8,10)$
$[ab\mathbf{Q}_i d_i]$	91.2%	-6546883	81.4%	$(9,9,8,10)$
$[a_{ij}b_i\mathbf{Q}_i d]$	91.2%	-6145593	86.0%	$(5,5,5,5)$
$[a_{ij}b\mathbf{Q}_i d]$	**93.0**%	**-6599840**	**88.4**%	$(11,11,11,11)$
$[a_i b_i\mathbf{Q}_i d]$	91.2%	-6144271	86.0%	$(5,5,5,5)$
$[ab_i\mathbf{Q}_i d]$	91.2%	-6144186	86.0%	$(5,5,5,5)$
$[a_i b\mathbf{Q}_i d]$	**93.0**%	-6594102	**88.4**%	$(11,11,11,11)$
$[ab\mathbf{Q}_i d]$	**93.0**%	-6593899	**88.4**%	$(11,11,11,11)$
$[a_j b\mathbf{Q} d]$	89.8%	-5965748	79.1%	$(4,4,4)$
$[ab\mathbf{Q} d]$	89.8%	-5964841	79.1%	$(4,4,4)$
SVM	93.0%	-	88.4%	-
PLS-DA	93.0%	-	83.7%	-

4.2 4-class MIR data set

The results obtained for the multi-class MIR data set are presented in Table 2. The performances of the fourteen HDDA models are compared to the results obtained with SVM and PLS-DA. As previously, the correct classification rate on the learning sample estimated by a 5-fold cross-validation (*learning CV-CCR*), the value of the BIC criterion, the correct classification rate on the test sample (*test CCR*) and the class specific subspace dimensions (d_i) were computed. According to the values of the CV-CCR obtained on the learning subset, three models $[a_{ij}b\boldsymbol{Q}_i d]$, $[a_i b\boldsymbol{Q}_i d]$ and $[ab\boldsymbol{Q}_i d]$ are retained corresponding to a correct classification rate of 93%. Among those models, the model $[a_{ij}b\boldsymbol{Q}_i d]$ corresponds to the lowest BIC value, and which also indicates that the noise is common to the four classes. The test correct classification rate is satisfying at 88.4% and can be compared to the results we obtained with SVM classifier (88.4% test CCR) or PLS-DA (83.7% test CCR). It should be noticed that in [15], Pierna *et al.* obtained 93% and 86% test CCR for SVM and PLS-DA, respectively, but with SNV pretreatement. It should be noticed that their SVM results could be reproduced when setting the same SVM parameters. Nevertheless, we did not succeed in finding these parameters applying our optimization procedure [7]. Furthermore, applying SNV as spectra pretreatment did not significantly improved the result we obtained with HDDA and presented on Table 2.

5 Conclusion

When dealing with spectroscopic data, generative methods present several advantages in terms of modeling and understanding. However, classical generative methods (such as QDA or LDA) can not be directly applied to high-dimensional data and a dimension reduction step is then necessary before the classification step. The dimension reduction is traditionally done using PCA. On the other hand, discriminative methods, such as SVM, do not suffer from the data dimensionality but their results are usually difficult to interpret. This article has presented a generative discriminant analysis method, called HDDA, designed to the classification of high-dimensional data and applied to the classification of multi-class spectroscopic data sets. HDDA is a generative discriminant analysis method based on a family of parsimonious Gaussian models which allow HDDA to be both flexible and efficient. For this, HDDA models and classifies the data in class specific and low-dimensional subspaces without reducing the data dimensionality. Therefore, no information loss is to be deplored due to data dimensionality reduction. This latter point is a specificity of HDDA, when compared to other generative methods, and it allows practicians to avoid the invasive dimension reduction step with PCA. In addition, HDDA can deal with the frequent situation in chemometrics where the number of observations is smaller than the data dimensionality. Experimental results have highlighted that HDDA outperforms classical classification methods and performs as good as SVM. Furthermore, interpretability strongly distinguishes HDDA from other classification methods and, in that, HDDA may be a powerful tool for real life applications.

6 Aknowledgments

We acknowledge Dr P. Dardenne and Dr J. A. Fernandez Pierna from the Wallon Agricultural Research Centre (CRA-W) for providing the data of the 4-class MIR data set. We also thank the referees for comments which greatly improved this paper.

References

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

[2] Bellman R. *Dynamic programming*, Princeton University Press, 1957.

[3] Bouveyron C, Celeux G, Girard S. Intrinsic dimension estimation by maximum likelihood in probabilistic PCA. Université Paris 1 Panthéon-Sorbonne, Technical report 440372, 2010.

[4] Bouveyron C, Girard S, Schmid C. High Dimensional Discriminant Analysis. *Comm. Statist. Theory Methods* 2007; **36**(14) :2607–2623.

[5] Bulesjo M, Rantalainen M, Cloarec O, Nicholson JK, Holmes E, Trygg J. OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification. *J. Chemometrics* 2006; **20** :341–351.

[6] Cattell R. The scree test for the number of factors. *Multivariate Behav. Res.* 1966; **1**(2) :245–276.

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

[8] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning*, Springer-Verlag: New York, 2001.

[9] Jolliffe, IT. *Principal Component Analysis*, Springer, 2002.

[10] Kemsley EK. Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods. *Chemometr. Intell. Lab. Syst.* 1996; **33** :47–61.

[11] Lehoucq R, Sorensen D, Yang C. *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, SIAM Publications: Philadelphia, 1998.

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

[13] McLachlan G, Peel D, Bean R. Modelling high-dimensional data by mixtures of factor analyzers. *Comput. Statist. Data Anal.* 2003; **41** :379–388.

[14] Murphy TB, Dean N, Raftery AE. Variable Selection and Updating in Model-Based Discriminant Analysis for High Dimensional Data with Food Authenticity Applications. *Ann. Appl. Stat.* 2010; **4**(1) :219–223.

[15] Fernandez Pierna JA, Dardenne P. Chemometric contest at 'Chimiométrie 2005': A discrimination study. *Chemometr. Intell. Lab. Syst.* 2007; **86** :219–223.

[16] Fernandez Pierna JA, Volery P, Besson R, Baeten V, Dardenne P. Classification of modified starches by FTIR spectroscopy using support vector machines. *J. Agric. Food Chem.* 2005; **53**(17) :6581–6585.

[17] Schwarz G. Estimating the dimension of a model. *Ann. Statist.* 1978; **6** :461–464.

[18] Scott D, Thompson J. Probability density estimation in higher dimensions. In *Fifteenth Symposium in the Interface*, 1983; 173–179.

[19] Tipping M, Bishop C. Mixtures of probabilistic principal component analysers. *Neur. Comput.* 1999; **11**(2) :443–482.

[20] Toher D, Downey G, Murphy TB. A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. *Chemometr. Intell. Lab. Syst.* 2007; **89** :102–115.

[21] Vong R, Geladi P, Wold S, Esbensen K. Source contributions to ambient aerosol calculated by discriminant partial least square regression (PLS). *J. Chemometrics* 1988; **2** :281–296.

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

[23] Wu W, Mallet Y, Walczak B, Penninckx W, Massart DL, Heuerding S, Erni F. Comparison of regularized discriminant analysis, linear discriminant analysis and quadratic discriminant analysis applied to NIR data. *Anal. Chim. Acta* 1996; **329** :257–265.