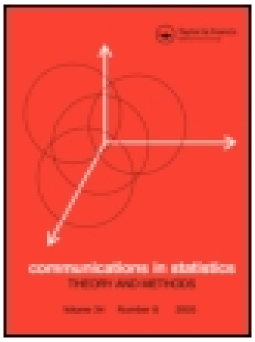

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: http://www.tandfonline.com/loi/lsta20

High-Dimensional Discriminant Analysis

Charles Bouveyron , Stéphane Girard & Cordelia Schmid

To cite this article: Charles Bouveyron , Stéphane Girard & Cordelia Schmid (2007) High-Dimensional Discriminant Analysis, Communications in Statistics - Theory and Methods, 36:14, 2607-2623, DOI: 10.1080/03610920701271095

To link to this article: http://dx.doi.org/10.1080/03610920701271095

Published online: 13 Oct 2007.

Communications in Statistics—Theory and Methods, 36: 2607–2623, 2007
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920701271095

Taylor & Francis
Taylor & Francis Group

High-Dimensional Discriminant Analysis

CHARLES BOUVEYRON[1], STÉPHANE GIRARD[1],
AND CORDELIA SCHMID[2]

[1]LMC-IMAG, Université Grenoble 1, Grenoble, France
[2]INRIA Rhône-Alpes, Saint-Ismier, France

We propose a new discriminant analysis method for high-dimensional data, called High-Dimensional Discriminant Analysis (HDDA). Our approach is based on the assumption that high-dimensional data live in different subspaces with low dimensionality. We therefore propose a new parameterization of the Gaussian model which combines the ideas of dimension reduction and constraints on the model. This parameterization takes into account the specific subspace and the intrinsic dimension of each class to limit the number of parameters to estimate. In addition, it is possible to make additional assumptions on the model to further limit the number of parameters. Our experiments on artificial and real datasets highlight that HDDA is more efficient than classical methods in high-dimensional spaces and with small learning datasets.

Keywords Class-specific subspaces; Discriminant analysis; High-dimensional data; Regularization.

Mathematics Subject Classification 62H25; 62H30

1. Introduction

Many scientific domains need to analyze data which are increasingly complex. For example, medical research, financial analysis, and computer vision provide high-dimensional data. Classifying such data is a challenging problem since the performance of classification methods suffers from the *curse of dimensionality*, first introduced by Bellman (1957), i.e., both classification accuracy and efficiency decrease rapidly in high dimensions. We therefore propose a new parameterization of the Gaussian model to classify high-dimensional data. This parameterization takes into account the specific subspace and the intrinsic dimension of each class to limit the number of parameters to estimate. In order to further limit the number of parameters, it is possible to make additional assumptions on the model and this gives rise to several particular models. We can, for example, assume that the

Address correspondence to Charles Bouveyron, SAMOS, Centre Pierre Mendès France, 90 rue de Tolbiac, 75634 Paris Cedex 13, France; E-mail: charles.bouveyron@univ-paris1.fr

classes are spherical in their subspaces or fix some parameters to be common between classes. A regularized discriminant analysis method for high-dimensional data is derived based on these models. This method is called High-Dimensional Discriminant analysis (HDDA). The article is organized as follows. Section 2 presents the discrimination problem and existing methods to regularize discriminant analysis in high-dimensional spaces. Section 3 introduces the theoretical framework of HDDA and Sec. 4 is devoted to the inference aspects. Our method is then compared to classical methods on artificial and real datasets in Sec. 5.

2. Discriminant Analysis Framework

In this section, we describe the general framework of the discrimination problem and present existing approaches of discriminant analysis in high-dimensional spaces.

2.1. *Discrimination Problem*

The goal of discriminant analysis is to assign an observation $x \in \mathbb{R}^p$ with unknown class membership to one of k classes $C_1, \ldots, C_k$ known *a priori*. For learning, we have a dataset $A = \{(x_1, c_1), \ldots, (x_n, c_n)/x_j \in \mathbb{R}^p \text{ and } c_j \in \{1, \ldots, k\}\}$, where the vector x_j contains p components of explanatory variables and c_j indicates the index of the class of x_j. The optimal decision rule, called *Bayes decision rule*, assigns the observation x to the class C_{i^*} which has the *maximum a posteriori* probability. This is equivalent to minimize a cost function $K_i(x)$, i.e., $i^* = \text{argmin}_{i=1,\ldots,k} K_i(x)$, with $K_i(x) = -2\log(\pi_i f_i(x))$, where π_i is the *a priori* probability of class C_i and $f_i(x)$ denotes the class conditional density of x, $\forall i = 1, \ldots, k$. For instance, Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA) rely on the assumption that $f_i(x)$ is a Gaussian density. An overview on this topic can be found in the book of McLachlan (1992). In this article, we focus on discriminant analysis methods based on Gaussian mixture models.

2.2. *Dimension Reduction and Parsimonious Models*

The majority of existing methods shows a disappointing behavior when the size of the training dataset is too small compared to the number of parameters to estimate which grows with the square of the dimension. On one hand, the direct application of standard discrimination method to high-dimensional data fails because of the singularity of the covariance matrices. Krzanowski et al. (1995) present a method to augment the covariance matrix such that it retains its major characteristics and becomes non singular. On the other hand, it is necessary to reduce the number of parameters to avoid overfitting. This is possible by either reducing the dimension of the data or by using a parsimonious model with additional assumptions on the model.

- **Dimension Reduction**. Many methods use global dimension reduction techniques to overcome problems due to high dimensionality. A widely used solution is to reduce the dimensionality of the data before using a classical discriminant analysis method. Dimension reduction can be done using Principal Components Analysis (PCA) or a variable selection technique, see, respectively, Jolliffe (1986) and Guyon and Elisseeff (2003) for further details. It is also possible to reduce the data dimension for classification purposes by using Fisher Discriminant Analysis (FDA) which projects the data on the $(k-1)$ discriminant axes and then classifies the projected data. The dimension reduction is often advantageous in terms of performance but loses information which could be discriminant due to the fact that most approaches are global and not designed for classification. Flury et al. (1997) proposed a discrimination method which uses dimension reduction for the purpose of classification by assuming that all differences between two classes occur in a low-dimensional subspace.

- **Parsimonious Models**. Another solution is to use a model which requires the estimation of fewer parameters. The parsimonious models most often used assume a Gaussian model with a common covariance matrix for all classes (used in LDA), i.e., $\forall i,\ \Sigma_i = \Sigma$, or diagonal covariance matrices, i.e., $\Sigma_i = \operatorname{diag}(\sigma_{i1}, \ldots, \sigma_{ip})$. Other approaches propose new parameterizations of the Gaussian model in order to find different parsimonious models. For example, the method proposed by Friedman (1989), called Regularized Discriminant Analysis (RDA), uses two regularization parameters to design an intermediate classifier between QDA and LDA. The Eigenvalue Decomposition Discriminant Analysis (EDDA), proposed by Bensmail and Celeux (1996), is based on a re-parameterization of the covariance matrices of the classes in their eigenspace. A survey on discriminant analysis regularization can be found in Mkhadri et al. (1997).

3. High-Dimensional Discriminant Analysis

The above-mentioned methods do not always allow to solve efficiently the problem of high dimensionality because the data usually contain clusters which are hidden in different subspaces of the original feature space. The *empty space* phenomenon, first noticed by Scott and Thompson (1983), allows us assume that high-dimensional data live in low-dimensional subspaces. We will therefore propose in this section a new parameterization of the Gaussian model which combines a local linear subspaces approach and a parsimonious model.

3.1. *Definitions and Assumptions*

Similar to classical discriminant analysis, we assume that class conditional densities are Gaussian $\mathcal{N}(\mu_i, \Sigma_i)$, $\forall i = 1, \ldots, k$. Let Q_i be the orthogonal matrix of the eigenvectors of Σ_i, then $\Delta_i = Q_i^t\, \Sigma_i\, Q_i$ is a diagonal matrix containing the eigenvalues of Σ_i. We further assume that Δ_i has the following form:

$$\Delta_i = \left(\begin{array}{cc} \boxed{\begin{matrix} a_{i1} & & 0 \\ & \ddots & \\ 0 & & a_{id_i} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_i & & 0 \\ & \ddots & \\ 0 & & b_i \end{matrix}} \end{array}\right) \begin{matrix} \left.\vphantom{\begin{matrix} a \\ \ddots \\ a \end{matrix}}\right\} d_i \\ \left.\vphantom{\begin{matrix} a \\ \ddots \\ a \end{matrix}}\right\} (p - d_i) \end{matrix}$$

where $a_{ij} \geq b_i$, for $j = 1, \ldots, d_i$ and $d_i < p$. The class-specific subspace $\mathbb{E}_i$ is generated by the d_i first eigenvectors corresponding to the eigenvalues a_{ij} with

$\mu_i \in \mathbb{E}_i$. Outside this subspace, the variance is modeled by the single parameter b_i. In addition, we respectively define the projection operators of x on $\mathbb{E}_i$ and $\mathbb{E}_i^{\perp}$:

$$P_i(x) = \widetilde{Q}_i \widetilde{Q}_i^t (x - \mu_i) + \mu_i, \tag{1}$$

$$P_i^{\perp}(x) = \overline{Q}_i \overline{Q}_i^t (x - \mu_i) + \mu_i \tag{2}$$

where $\widetilde{Q}_i$ is made of the d_i first columns of Q_i supplemented by zeros and $\overline{Q}_i = Q_i - \widetilde{Q}_i$. Figure 1 summarizes these notations. This model will be referred in the following by $[a_{ij} b_i Q_i d_i]$.

3.2. *Decision Rule*

Deriving the Bayes decision rule with the model $[a_{ij} b_i Q_i d_i]$ described in the previous section yields the decision rule of High-Dimensional Discriminant Analysis (HDDA).

Theorem 3.1. *Bayes decision rule yields the decision rule of HDDA which classifies x as the class C_{i^*} such that $i^* = \operatorname{argmin}_{i=1,\ldots,k}\{K_i(x)\}$ where K_i is defined by:*

$$K_i(x) = \|\mu_i - P_i(x)\|^2_{\mathcal{A}_i} + \frac{1}{b_i}\|x - P_i(x)\|^2 + \sum_{j=1}^{d_i} \log(a_{ij}) + (p - d_i)\log(b_i) - 2\log(\pi_i), \tag{3}$$

where $\|\cdot\|_{\mathcal{A}_i}$ is a norm on $\mathbb{E}_i$ such that $\|x\|^2_{\mathcal{A}_i} = x^t \mathcal{A}_i x$ with $\mathcal{A}_i = \widetilde{Q}_i \Delta_i^{-1} \widetilde{Q}_i^t$.

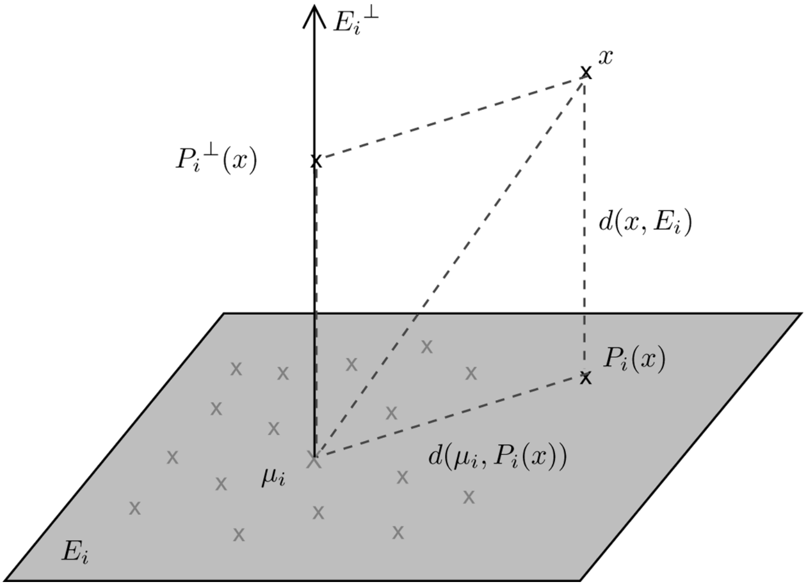

Figure 1. The subspaces $\mathbb{E}_i$ and $\mathbb{E}_i^{\perp}$ of the class C_i.

We can observe that this new decision rule is mainly based on two distances: the distance between the observation and the subspace $\mathbb{E}_i$, and the distance between the projection of x on $\mathbb{E}_i$ and the mean of the class. It also depends on the variances a_{ij} and b_i and on the *a priori* probability π_i. This rule is easily understood because it is natural to assign 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 variances a_{ij} 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 $\|x - P_i(x)\|^2$ by $1/b_i$ in order to take into account the large variance in $\mathbb{E}_i^{\perp}$. This rule allows a straightforward interpretation of the classification results, whereas other methods, such as Support Vectors Machine or logistic regression (see Hastie et al., 2001), provide results which are difficult to understand. Note that the decision rule of HDDA does not use the projection on $\mathbb{E}_i^{\perp}$ and thus requires only the estimation of the d_i first columns of Q_i. This reduces significantly the number of parameters to estimate. For example, if we consider 100-dimensional data, made of 4 classes and with common intrinsic dimensions d_i equal to 10, HDDA estimates only 4,231 parameters whereas QDA estimates 20,603 parameters. In addition, the fact of not using the projection $\mathbb{E}_i^{\perp}$ prevents numerical problems due to the singularity of covariance matrices. Finally, the *a posteriori* probability $\mathbb{P}(x \in C_i \mid x)$, which measures the probability that x belongs to C_i and allows to identify dubiously classified points, can be written as follows:

$$\mathbb{P}(x \in C_i|x) = 1/\sum_{\ell=1}^{k} \exp\left(\frac{1}{2}(K_i(x) - K_\ell(x))\right).$$

3.3. *A Family of Models Designed for High-Dimensionality*

By fixing some of the HDDA parameters to be common between classes, we obtain 28 different models (including the model $[a_{ij}b_iQ_id_i]$) which correspond to different types of regularization, some of them having geometrically interpretable decision rule. For instance, if we fix the d_i first eigenvalues a_{ij} to be common within each class, we obtain the more restricted model $[a_ib_iQ_id_i]$. In the following, "free Q_i" means that Q_i is specific to the class C_i and "common Q_i" means that for each $i = 1, \ldots, k$, $Q_i = Q$ and consequently the class orientations are the same. We split this family of models designed for high dimensionality into three categories: models with free orientations, models with common orientations, and models with common covariance matrices. Several models with common orientations require a complex iterative estimation based on the FG algorithm of Flury and Gautschi (1986) and therefore they will not be considered here. Table 1 summarizes the properties of these models. The second column of this table gives the number of parameters to estimate. The third column provides the asymptotic order of the number of parameters to estimate (with the assumption $k \ll d \ll p$). The fourth column gives this number for the particular case $k = 4$, $p = 100$, and $\forall i$, $d_i = 10$. The last column indicates whether the estimators are closed form or not. These values are also given for QDA, LDA, and the model $[\lambda_k B_k]$ of EDDA. We can observe that all HDDA models require the estimation of a number of parameters lower than both QDA and LDA. In addition, some particular cases of HDDA correspond to classical discriminant analysis. For example, if $d_i = (p-1)$, for $i = 1, \ldots, k$, then HDDA reduces to QDA. Moreover, if $a_{ij} = a_j$, $b_i = b$ and $Q_i = Q$,

Table 1

Properties of the HDDA models: $\rho = kp + k - 1$ is the number of parameters required for the estimation of means and proportions, $\bar{\tau} = \sum_{i=1}^{k} d_i[p - (d_i + 1)/2]$ and $\tau = d[p - (d + 1)/2]$ are the number of parameters required for the estimation of $\widetilde{Q}_i$ and $\widetilde{Q}$, and $D = \sum_{i=1}^{k} d_i$. For asymptotic orders, we assume that $k \ll d \ll p$. CF means that the ML estimates are closed form. IP means that the ML estimation needs an iterative procedure. FG means that the ML estimation requires the iterative FG algorithm

Model	Number of parameters	Asymptotic order	Nb of prm for $k = 4$, $p = 100$ and $d = 10$	ML estimation
$[a_{ij}b_iQ_id_i]$	$\rho + \bar{\tau} + 2k + D$	kpd	4231	CF
$[a_{ij}bQ_id_i]$	$\rho + \bar{\tau} + k + D + 1$	kpd	4228	CF
$[a_ib_iQ_id_i]$	$\rho + \bar{\tau} - 3k$	kpd	4195	CF
$[ab_iQ_id_i]$	$\rho + \bar{\tau} + 2k + 1$	kpd	4192	CF
$[a_ibQ_id_i]$	$\rho + \bar{\tau} + 2k + 1$	kpd	4192	CF
$[abQ_id_i]$	$\rho + \bar{\tau} + k + 2$	kpd	4189	CF
$[a_{ij}b_iQ_id]$	$\rho + k(\tau + d + 1) + 1$	kpd	4228	CF
$[a_jb_iQ_id]$	$\rho + k(\tau + 1) + d + 1$	kpd	4198	CF
$[a_{ij}bQ_id]$	$\rho + k(\tau + d) + 2$	kpd	4225	CF
$[a_jbQ_id]$	$\rho + k\tau + d + 2$	kpd	4195	CF
$[a_ib_iQ_id]$	$\rho + k(\tau + 2) + 1$	kpd	4192	CF
$[ab_iQ_id]$	$\rho + k(\tau + 1) + 2$	kpd	4189	CF
$[a_ibQ_id]$	$\rho + k(\tau + 1) + 2$	kpd	4189	CF
$[abQ_id]$	$\rho + k\tau + 3$	kpd	4186	CF
$[a_{ij}b_iQd_i]$	$\rho + \tau + D + 2k$	pd	1396	FG
$[a_{ij}bQd_i]$	$\rho + \tau + D + k + 1$	pd	1393	FG
$[a_ib_iQd_i]$	$\rho + \tau - 3k$	pd	1360	FG
$[a_ibQd_i]$	$\rho + \tau + 2k + 1$	pd	1357	FG
$[ab_iQd_i]$	$\rho + \tau + 2k + 1$	pd	1357	FG
$[abQd_i]$	$\rho + \tau + k + 2$	pd	1354	FG
$[a_{ij}b_iQd]$	$\rho + \tau + kd + k + 1$	pd	1393	FG
$[a_jb_iQd]$	$\rho + \tau + k + d + 1$	pd	1363	FG
$[a_{ij}bQd]$	$\rho + \tau + kd + 2$	pd	1390	FG
$[a_ib_iQd]$	$\rho + \tau + 2k + 1$	pd	1357	IP
$[ab_iQd]$	$\rho + \tau + k + 2$	pd	1354	IP
$[a_ibQd]$	$\rho + \tau + k + 2$	pd	1354	IP
$[a_jbQd]$	$\rho + \tau + d + 2$	pd	1360	CF
$[abQd]$	$\rho + \tau + 3$	pd	1351	CF
QDA	$\rho + kp(p + 1)/2$	$kp^2/2$	20603	CF
LDA	$\rho + p(p + 1)/2$	$p^2/2$	5453	CF
EDDA $[\lambda_kB_k]$	$\rho + kp$	kp	803	CF

for $i = 1, \ldots, k$, then HDDA reduces to LDA. Furthermore, the regularized method proposed by Flury et al. (1997) is equivalent to HDDA with the model $[a_{ij}bQd]$ and an additional assumption on the means.

4. Parameters Estimation

The parameters of HDDA are estimated using the maximum likelihood (ML) estimation technique based on the learning dataset A. In the following, parameters π_i, μ_i, and Σ_i of the class C_i are estimated by their empirical counterparts:

$$\hat{\pi}_i = \frac{n_i}{n}, \quad \hat{\mu}_i = \frac{1}{n_i}\sum_{x_j \in C_i} x_j, \quad \widehat{\Sigma}_i = \frac{1}{n_i}\sum_{x_j \in C_i}(x_j - \hat{\mu}_i)^t(x_j - \hat{\mu}_i),$$

where $n_i = card(C_i)$. We also introduce the following notations: $\xi = \sum_{i=1}^{k} \hat{\pi}_i d_i$ is the average dimension of the class-specific subspaces, $\widehat{W} = \sum_{i=1}^{k} \hat{\pi}_i \widehat{\Sigma}_i$ is the empirical estimate of the within-covariance matrix, λ_{ij} is the jth largest eigenvalue of $\widehat{\Sigma}_i$ and λ_j is the jth largest eigenvalue of $\widehat{W}$. We first present the estimators of HDDA parameters and then those of intrinsic dimensions. Proofs of following results are given in the Appendix.

4.1. *Models with Free Orientations*

The following three propositions provide closed form estimators for the model parameters.

Proposition 4.1. *The d_i first columns of Q_i are estimated by the eigenvectors associated with the d_i largest eigenvalues λ_{ij} of $\widehat{\Sigma}_i$.*

Proposition 4.2. *Model $[a_{ij}b_iQ_id_i]$: 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 $\widehat{\Sigma}_i$ and can be written as follows:*

$$\hat{b}_i = \frac{1}{(p - d_i)}\left(\operatorname{tr}(\widehat{\Sigma}_i) - \sum_{j=1}^{d_i} \lambda_{ij}\right). \tag{4}$$

Model $[a_{ij}bQ_id_i]$: The estimator of a_{ij} is $\hat{a}_{ij} = \lambda_{ij}$ and the estimator of b is:

$$\hat{b} = \frac{1}{(p - \xi)}\left(\operatorname{tr}(\widehat{W}) - \sum_{i=1}^{k} \hat{\pi}_i \sum_{j=1}^{d_i} \lambda_{ij}\right). \tag{5}$$

Model $[a_ib_iQ_id_i]$: 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 $[ab_iQ_id_i]$: The estimator of b_i is given by (4) and the estimator of a is:

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

Model $[a_ibQ_id_i]$: The estimators of a_i and b are, respectively, given by (6) and (5).
Model $[abQ_id_i]$: The estimators of a and b are, respectively, given by (7) and (5).

Proposition 4.3. *The estimators of the models with common dimensions d_i can be obtained from those of Proposition* 4.2 *by replacing the values d_i by d, for each $i = 1, \dots, k$. In this case, Eqs.* (5) *and* (7) *can be simplified as*:

$$\hat{a} = \frac{1}{d}\sum_{j=1}^{d}\lambda_j, \tag{8}$$

$$\hat{b} = \frac{1}{(p-d)}\left(\operatorname{tr}(\widehat{W}) - \sum_{j=1}^{d}\lambda_j\right). \tag{9}$$

Model $[a_j b_i Q_i d]$: The estimator of a_j is $\hat{a}_j = \lambda_j$ and the estimator of b_i is given by (4).
Model $[a_j b Q_i d]$: The estimator of a_j is $\hat{a}_j = \lambda_j$ and the estimator of b is given by (9).

4.2. *Models with Common Orientations*

Here, we assume that the orientations and the dimensions are common between classes. The following propositions both give rise to an iterative scheme for estimating the parameters.

Proposition 4.4. *Given a_i and b_i, the d first columns of Q are estimated by the eigenvectors associated to the d largest eigenvalues of the matrix M defined by*:

$$M(a_1, \dots, a_k, b_1, \dots, b_k) = \sum_{i=1}^{k} n_i\left(\frac{1}{b_i} - \frac{1}{a_i}\right)\widehat{\Sigma}_i.$$

Proposition 4.5. *Model $[a_i b_i Q d]$: Given Q, the estimator of a_i and b_i are*:

$$\hat{a}_i(Q) = \frac{1}{d}\sum_{j=1}^{d} q_j^t \widehat{\Sigma}_i q_j, \tag{10}$$

$$\hat{b}_i(Q) = \frac{1}{(p-d)}\left(\operatorname{tr}(\widehat{\Sigma}_i) - \sum_{j=1}^{d} q_j^t \widehat{\Sigma}_i q_j\right). \tag{11}$$

Model $[a_i b Q_i d_i]$: Given Q, the estimator of a_i is given by (10) *and the estimator of b is*:

$$\hat{b}(Q) = \frac{1}{(p-d)}\left(\operatorname{tr}(\widehat{W}) - \sum_{j=1}^{d} q_j^t \widehat{W} q_j\right). \tag{12}$$

Model $[a b_i Q d]$: Given Q, the estimator of b_i is defined by (11) *and the estimator of a is*:

$$\hat{a}(Q) = \frac{1}{d}\sum_{j=1}^{d} q_j^t \widehat{W} q_j. \tag{13}$$

Model $[a_i b Q d]$: Given Q, the estimators of a_i and b are, respectively, given by (10) *and* (12).

The estimators of these models require an iterative estimation procedure. For example, it is possible to use the following iterative procedure for the model $[a_i b_i Qd]$:

- Initialization: The d first columns of $Q^{(0)}$ are the eigenvectors associated with the d largest eigenvalues of $\widehat{W}$.
- Until convergence: $a_i^{(\ell)} = \hat{a}_i(Q^{(\ell-1)})$, $b_i^{(\ell)} = \hat{b}_i(Q^{(\ell-1)})$, and the d first columns of $Q^{(\ell)}$ are the eigenvectors associated to the d largest eigenvalues of $M(a_i^{(\ell)}, b_i^{(\ell)})$, for $i = 1, \dots, k$.

4.3. *Models with Common Covariance Matrices*

For this category of models, the parameters can be estimated in closed form.

Proposition 4.6. *The d first columns of the matrix Q are the eigenvectors associated to the d largest eigenvalues of $\widehat{W}$.*

Proposition 4.7. *Model $[a_j bQd]$: The estimator of a_j is $\hat{a}_j = \lambda_j$ and the estimator of b is given by* (9).

Model $[abQd]$: The estimator of a and b are, respectively, given by (8) *and* (9).

4.4. *Estimation of the Intrinsic Dimension*

The estimation of the dataset intrinsic dimension is a difficult problem which does not have an explicit solution. If the dimensions d_i are common between classes, i.e., $d_i = d$ for $i = 1, \dots, k$, we can determine the dimension d by cross-validation, i.e., by maximizing the correct classification rate on the learning dataset. Otherwise, we use an approach based on the eigenvalues of the class conditional covariance matrix $\widehat{\Sigma}_i$. The jth eigenvalue of $\widehat{\Sigma}_i$ corresponds to the fraction of the full variance carried by the jth eigenvector of $\widehat{\Sigma}_i$. We therefore propose to estimate dimensions d_i with the empirical method "scree-test" of Cattell (1966). For each class, the selected dimension d_i is the one for which the difference between the two subsequent eigenvalues is smaller than a given threshold t. The threshold t is found by cross-validation on the learning dataset. We also compared the scree-test of Cattell to the probabilistic criterion BIC proposed by Schwarz (1978) and we obtained similar choices of dimension.

5. Numerical Results

In this section, we present results for artificial and real datasets illustrating the main features of HDDA. They show the influence of the dimensionality and of the size of the learning dataset on the behavior of classification methods. We refer to our previous work, Bouveyron et al. (2005), for an application of HDDA to the recognition of object classes in natural images. In following experiments, HDDA will be compared to four classical methods: QDA, LDA, EDDA, and PCA + LDA. For EDDA, we used the model $[\lambda_k B_k]$ which is recommended by the authors. PCA + LDA reduces the dimension to 15 with PCA and then applies standard LDA. A numerical regularization was necessary in order to inverse the covariance matrices in the methods QDA, LDA, and EDDA so that they are able to work with data of dimension larger than 50.

5.1. *Simulation Study*

For this experiment, we simulated three Gaussian densities in $\mathbb{R}^p$, $p = 15, \ldots, 100$, according to the model $[a_i b_i Q_i d_i]$ with the following parameters: $\{d_1, d_2, d_3\} = \{2, 5, 10\}$, $\{\pi_1, \pi_2, \pi_3\} = \{0.4, 0.3, 0.3\}$, $\{a_1, a_2, a_3\} = \{150, 75, 50\}$, $\{b_1, b_2, b_3\} = \{10, 10, 10\}$, and with close means and random Q_i. The learning and the test datasets are respectively made of 250 and 1,000 points. The performance of methods is measured by the average classification rate computed on 50 replications. Firstly, Fig. 2 shows that data dimensionality does not influence the performance of HDDA and that it is very close to the performance of the Bayes decision rule (computed with the true densities). In addition, HDDA provides a classification rate similar to QDA in low dimensions. QDA is known to be very sensitive to the data dimensionality and, indeed, gives bad results in high-dimensional spaces. LDA is more robust to dimensionality, but cannot correctly find the class specific subspaces and therefore provides poor classification results. LDA is also penalized by the data dimensionality for dimensions larger than 60. The dimension reduction allows PCA+LDA to have a constant performance according to the data dimensionality but does not allow to improve over LDA results. Finally, the model $[\lambda_k B_k]$ of EDDA does not suffer from the curse of the dimensionality, but provides results which are worse than HDDA. This is certainly due to the fact that the model $[\lambda_k B_k]$ of EDDA is too parsimonious. To summarize, HDDA is not sensitive to the dimensionality and works very well in low- and in high-dimensional spaces. HDDA seems to have the right number of degrees of freedom since it outperforms methods requiring a higher number of parameters (QDA, LDA) and a method requiring a smaller number of parameters (model $[\lambda_k B_k]$ of EDDA).

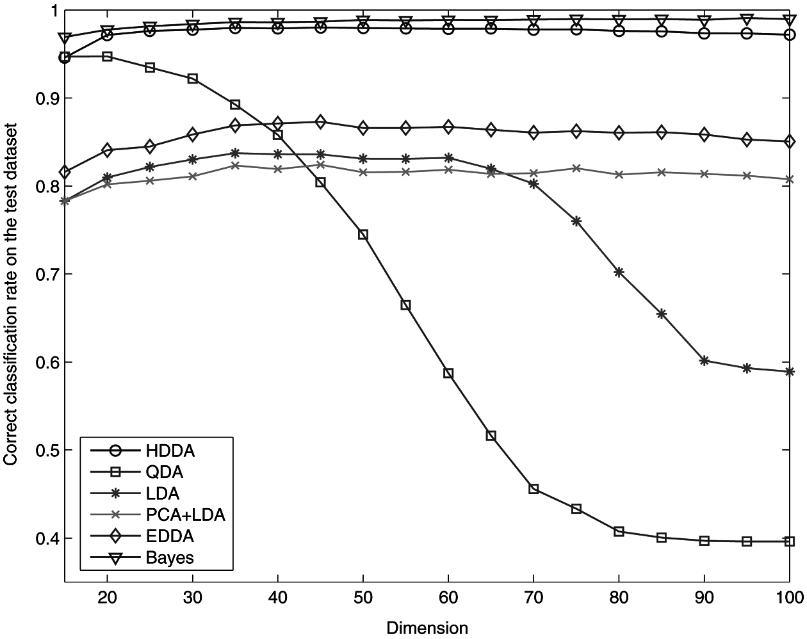

Figure 2. Influence of the dimensionality on classification results obtained with HDDA and classical methods on an artificial dataset.

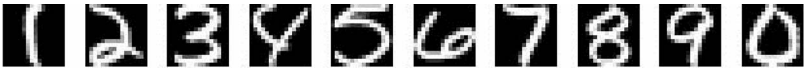

Figure 3. Some examples of the USPS dataset used for the OCR experiment.

5.2. *Real Data Study*

Figure 4 presents results for optical character recognition (OCR) on the USPS dataset available at www.kernel-machines.org. We dispose of 7,291 images for learning and 2,007 images for testing. The data are divided into 10 classes and each digit is a 16 × 16 grey level image, represented as a 256-dimensional vector. Figure 3 show some examples of the USPS dataset. In order to show the influence of the size of the learning dataset on classification results, we successively used an increasing part of the learning set to classify the test dataset. The performance of methods is measured by the average classification rate computed on 50 replications. Figure 4 highlights that HDDA works very well compared to the other methods when the size of the learning dataset is small. We can also observe that the dimension reduction step allows PCA + LDA to improve recognition results and to work with small learning datasets. Table 2 presents the classification results obtained with some HDDA models on the USPS dataset. These experiments illustrate that HDDA provides very satisfying performances in high-dimensional space and with small learning datasets. In addition, among all HDDA models, the models with common b_i seem particularly efficient.

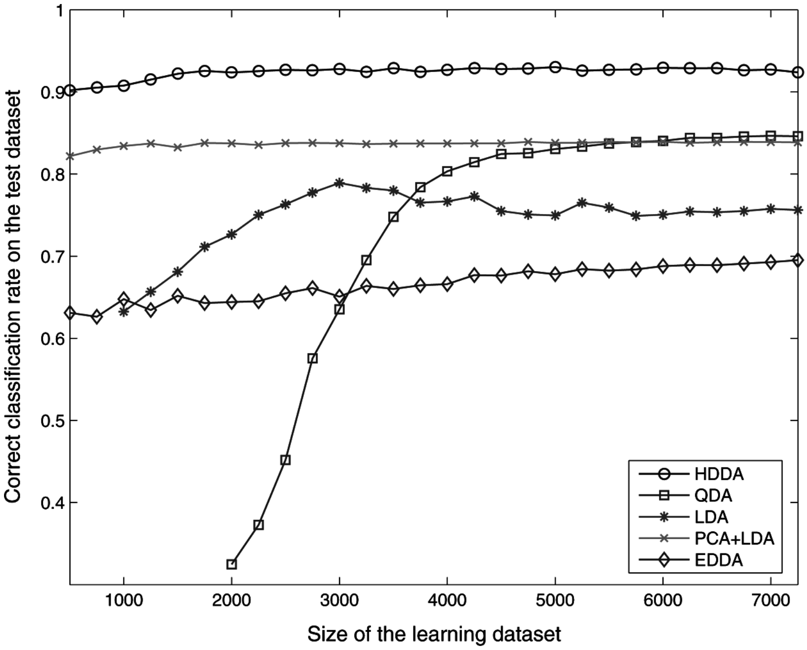

Figure 4. Influence of the size of the learning dataset on classification results obtained with HDDA and classical methods on a real dataset (USPS digits).

Table 2
Classification results for some HDDC models on the USPS dataset

Model	Classification rate
$[a_{ij}b_iQ_id_i]$	92.63
$[a_{ij}bQ_id_i]$	93.67
$[a_ib_iQ_id_i]$	92.78
$[a_ibQ_id_i]$	**93.72**
$[abQ_id_i]$	93.17
$[a_{ij}b_iQ_id]$	92.83
$[a_{ij}bQ_id]$	**94.77**
$[a_ib_iQ_id]$	92.83
$[a_ibQ_id]$	94.57
$[abQ_id]$	94.52

6. Conclusion

In this article we introduce a new parameterization of the Gaussian model appropriate for classifying high-dimensional data in a supervised framework. Based on this model, we derive a regularized discriminant analysis technique, called High-Dimensional Discriminant Analysis. The decision rule of HDDA is easy to interpret since it takes into account the distance to the class mean and the distance to the subspace. Experimental results confirm that HDDA works very well in high-dimensional spaces and with small learning datasets. A natural extension of this work is to use the same Gaussian model in the context of unsupervised classification.

A. Appendix

A.1. The Decision Rule

Proof of Theorem 3.1. We derive the Bayes decision rule for the Gaussian model presented in section. Writing f_i with the new class conditional covariance matrix Δ_i, we obtain:

$$-2\log(f_i(x)) = (x-\mu_i)^t(Q_i\Delta_iQ_i^t)^{-1}(x-\mu_i) + \log(\det\Delta_i) + p\log(2\pi).$$

Given the structure of Δ_i and using the relations $Q_i = \widetilde{Q}_i + \overline{Q}_i$, $\widetilde{Q}_i\widetilde{Q}_i^t\widetilde{Q}_i = \widetilde{Q}_i$ and $\overline{Q}_i\overline{Q}_i^t\overline{Q}_i = \overline{Q}_i$, we obtain:

$$-2\log(f_i(x)) = \|\widetilde{Q}_i^t\widetilde{Q}_i(x-\mu_i)\|^2_{\mathcal{A}_i} + \frac{1}{b_i}\|\overline{Q}_i^t\overline{Q}_i(x-\mu_i)\|^2 + \log(\det\Delta_i) + p\log(2\pi),$$

where $\| \cdot \|_{\mathscr{A}_i}$ is a norm on $\mathbb{E}_i$ such that $\|x\|^2_{\mathscr{A}_i} = x^t \mathscr{A}_i x$ and with $\mathscr{A}_i = \widetilde{Q}_i \Delta_i^{-1} \widetilde{Q}_i^t$. Using the definitions of P_i and $P_i^\perp$ and in view of Fig. 1, we obtain:

$$-2\log(f_i(x)) = \|\mu_i - P_i(x)\|^2_{\mathscr{A}_i} + \frac{1}{b_i}\|x - P_i(x)\|^2 + \log(\det \Delta_i) + p\log(2\pi).$$

The relation $\log(\det \Delta_i) = \sum_{j=1}^{d_i} \log(a_{ij}) + (p - d_i)\log(b_i)$ concludes the proof. □

A.2. Parameters Estimation

Lemma A.1. *First of all, we introduce the following useful formulation of the log-likelihood*:

$$-2\log(L) = \sum_{i=1}^{k} n_i \sum_{j=1}^{p} \left(\log(\delta_{ij}) + \frac{1}{\delta_{ij}} q_{ij}^t \widehat{\Sigma}_i q_{ij} \right) + c^{st}, \tag{14}$$

where δ_{ij} is the jth diagonal coefficient of Δ_i and q_{ij} is the jth column of Q_i.

We refer to Flury (1984) for a demonstration of this result.

Proof of Proposition 4.1. The log-likelihood is to be maximized under the constraint $q_{ij}^t q_{ij} = 1$, which is equivalent to finding a saddle point of the Lagrange function:

$$\mathscr{L} = -2\log(L) - \sum_{j=1}^{p} \theta_{ij}(q_{ij}^t q_{ij} - 1),$$

where θ_{ij} are the Lagrange multipliers. Using the expression (14) of the log-likelihood, the gradient of $\mathscr{L}$ with respect to q_{ij} is:

$$\nabla_{q_{ij}} \mathscr{L} = 2\frac{n_i}{\delta_{ij}} \widehat{\Sigma}_i q_{ij} - 2\theta_{ij} q_{ij},$$

and by multiplying this quantity on the left by q_{ij}^t, we obtain:

$$q_{ij}^t \nabla_{q_{ij}} \mathscr{L} = 0 \Leftrightarrow \theta_{ij} = \frac{n_i}{\delta_{ij}} q_{ij}^t \widehat{\Sigma}_i q_{ij}.$$

Consequently, $\widehat{\Sigma}_i q_{ij} = \frac{\theta_{ij}\delta_{ij}}{n_i} q_{ij}$ and thus q_{ij} is the eigenvector of $\widehat{\Sigma}_i$ associated with the eigenvalue $\lambda_{ij} = \frac{\theta_{ij}\delta_{ij}}{n_i} = q_{ij}^t \widehat{\Sigma}_i q_{ij}$. As the vectors q_{ij} are eigenvectors of the symmetric matrix $\widehat{\Sigma}_i$, this implies that $q_{ij}^t q_{i\ell} = 0$ if $j \neq \ell$. The log-likelihood can therefore be re-written as follows:

$$-2\log(L) = \sum_{i=1}^{k} n_i \left(\sum_{j=1}^{d_i} \left(\log(a_{ij}) + \frac{\lambda_{ij}}{a_{ij}} \right) + \sum_{j=d_i+1}^{p} \left(\log(b_i) + \frac{\lambda_{ij}}{b_i} \right) \right) + c^{st},$$

and, using the relation $\sum_{j=d_i+1}^{p} \lambda_{ij} = \mathrm{tr}(\widehat{\Sigma}_i) - \sum_{j=1}^{d_i} \lambda_{ij}$, we obtain:

$$-2\log(L) = \sum_{i=1}^{k} n_i \left(\sum_{j=1}^{d_i} \log(a_{ij}) + (p - d_i)\log(b_i) + \frac{\mathrm{tr}(\widehat{\Sigma}_i)}{b_i} + \sum_{j=1}^{d_i} \left(\frac{1}{a_{ij}} - \frac{1}{b_i} \right) \lambda_{ij} \right) + c^{st}. \tag{15}$$

Thus, minimizing $-2\log(L)$ with respect to λ_{ij} is equivalent to minimizing the quantity $\sum_{i=1}^{k} n_i \sum_{j=1}^{d_i} (\frac{1}{a_{ij}} - \frac{1}{b_i})\lambda_{ij}$. Since $(\frac{1}{a_{ij}} - \frac{1}{b_i}) < 0$, $\forall j = 1, \ldots, d_i$, λ_{ij} must therefore be as larger as possible. Thus, the column vector q_{ij}, $\forall j = 1, \ldots, d_i$, is estimated by the eigenvector associated to the jth largest eigenvalue of $\widehat{\Sigma}_i$. □

Proof of Proposition 4.2. Model $[a_{ij}b_iQ_id_i]$: starting from Eq. (15), the partial derivative of $-2\log(L)$ with respect to a_{ij} and b_i are:

$$-2\frac{\partial \log(L)}{\partial a_{ij}} = n_i \left(\frac{1}{a_{ij}} - \frac{\lambda_{ij}}{a_{ij}^2} \right) \quad \text{and} \quad -2\frac{\partial \log(L)}{\partial b_i} = \frac{n_i(p - d_i)}{b_i} - \frac{n_i}{b_i^2} \left(\mathrm{tr}(\widehat{\Sigma}_i) - \sum_{j=1}^{d_i} \lambda_{ij} \right).$$

The condition $\frac{\partial \log(L)}{\partial a_{ij}} = 0$ implies that $\hat{a}_{ij} = \lambda_{ij}$ and the condition $\frac{\partial \log(L)}{\partial b_i} = 0$ implies (4).

Model $[a_{ij}bQ_id_i]$: The partial derivative of $-2\log(L)$ with respect to b is:

$$-2\frac{\partial \log(L)}{\partial b} = \frac{n(p - \xi)}{b} - \frac{1}{b^2} \sum_{i=1}^{k} n_i \left(\mathrm{tr}(\widehat{\Sigma}_i) - \sum_{j=1}^{d_i} \lambda_{ij} \right),$$

and the condition $\frac{\partial \log(L)}{\partial b} = 0$ proves (5).

Model $[a_ib_iQ_id_i]$: From (15), the partial derivative of $-2\log(L)$ with respect to a_i is:

$$-2\frac{\partial \log(L)}{\partial a_i} = \frac{n_i d_i}{a_i} - \frac{n_i}{a_i^2} \sum_{j=1}^{d_i} \lambda_{ij},$$

and the condition $\frac{\partial \log(L)}{\partial a_i} = 0$ implies (6).

Model $[ab_iQ_id_i]$: The partial derivative of $-2\log(L)$ with respect to a is:

$$-2\frac{\partial \log(L)}{\partial a} = \frac{n\xi}{a} - \frac{1}{a^2} \sum_{i=1}^{k} n_i \sum_{j=1}^{d_i} \lambda_{ij},$$

and the condition $\frac{\partial \log(L)}{\partial a} = 0$ gives (7). □

Proof of Proposition 4.3. Model $[a_jb_iQ_id]$: the partial derivative of $-2\log(L)$ with respect to a_j is:

$$-2\frac{\partial \log(L)}{\partial a_j} = \frac{n}{a_j} - \frac{1}{a_j^2} \sum_{i=1}^{k} n_i \lambda_{ij}.$$

The condition $\frac{\partial \log(L)}{\partial a_j} = 0$ and the relation $\sum_{i=1}^{k} n_i \lambda_{ij} = n\lambda_j$ imply that $\hat{a}_j = \lambda_j$. □

Proof of Proposition 4.4. Starting from the likelihood expression of Lemma A.1, we can write:

$$\begin{aligned}-2\log(L) &= \sum_{i=1}^{k} n_i \sum_{j=1}^{d}\left(\log(a_i) + \frac{1}{a_i} q_j^t\widehat{\Sigma}_i q_j\right) + \sum_{i=1}^{k} n_i \sum_{j=d+1}^{p}\left(\log(b_i) + \frac{1}{b_i} q_j^t\widehat{\Sigma}_i q_j\right) + c^{st},\\ &= \sum_{i=1}^{k} n_i(d\log(a_i) + (p-d)\log(b_i)) + \sum_{j=1}^{d} q_j^tAq_j + \sum_{j=d+1}^{p} q_j^tBq_j + c^{st},\end{aligned}$$

where $A = \sum_{i=1}^{k}\frac{n_i}{a_i}\widehat{\Sigma}_i$ and $B = \sum_{i=1}^{k}\frac{n_i}{b_i}\widehat{\Sigma}_i$. Note that $\sum_{j=d+1}^{p} q_j^tBq_j$ can be written using the trace of B: $\sum_{j=d+1}^{p} q_j^tBq_j = \mathrm{tr}(B) - \sum_{j=1}^{d} q_j^tBq_j$. This yields:

$$-2\log(L) = \sum_{i=1}^{k} n_i(d\log(a_i) + (p-d)\log(b_i)) - \sum_{j=1}^{d} q_j^t(B-A)q_j + \mathrm{tr}(B) + c^{st}. \quad (16)$$

Consequently, the gradient of $\mathcal{L} = -2\log(L) - \sum_{j=1}^{p}\theta_j(q_j^tq_j - 1)$ with respect to q_j is:

$$\nabla_{q_j}\mathcal{L} = -2(B-A)q_j - 2\theta_j q_j,$$

where θ_j is the jth Lagrange multiplier. The relation $\nabla_{q_j}\mathcal{L} = 0$ is equivalent to $(B-A)q_j = -\theta_j q_j$ which means that q_j is eigenvector of the matrix $(B-A)$. In order to minimize the quantity $-2\log(L)$, the d first columns of Q must be the eigenvectors associated with the d largest eigenvalues of $(B-A)$. □

Proof of Proposition 4.5. Model $[a_ib_iQd]$: Starting from Eq. (16), the partial derivatives of $-2\log(L)$ with respect to a_i and b_i are:

$$-2\frac{\partial\log(L)}{\partial a_i} = \frac{n_id}{a_i} - \frac{n_i}{a_i^2}\sum_{j=1}^{d} q_j^t\widehat{\Sigma}_i q_j \quad \text{and}$$

$$-2\frac{\partial\log(L)}{\partial b_i} = \frac{n_i(p-d)}{b_i} - \frac{n_i}{b_i^2}\left(\mathrm{tr}(\widehat{\Sigma}_i) - \sum_{j=1}^{d} q_j^t\widehat{\Sigma}_i q_j\right).$$

The condition $\frac{\partial\log(L)}{\partial a_i} = 0$ gives (10) and $\frac{\partial\log(L)}{\partial b_i} = 0$ gives (11).

Model $[a_ibQd]$: The partial derivative of $-2\log(L)$ with respect to b is:

$$-2\frac{\partial\log(L)}{\partial b} = \frac{n(p-d)}{b} - \frac{n}{b^2}\left(\mathrm{tr}(\widehat{W}) - \sum_{j=1}^{d} q_j^t\widehat{W}q_j\right),$$

and the condition $\frac{\partial\log(L)}{\partial b} = 0$ implies (12).

Model $[ab_iQd]$: The partial derivative of $-2\log(L)$ with respect to a is:

$$-2\frac{\partial\log(L)}{\partial a} = \frac{nd}{a} - \frac{n}{a^2}\sum_{j=1}^{d} q_j^t\widehat{W}q_j,$$

and the condition $\frac{\partial\log(L)}{\partial a} = 0$ implies (13).

□

Proof of Proposition 4.6. The log-likelihood can be written as follows:

$$-2\log(L) = n\left(\sum_{j=1}^{d}\log(a_j) + (p-d)\log(b) + \frac{\operatorname{tr}(\widehat{W})}{b} + \sum_{j=1}^{d}\left(\frac{1}{a_j} - \frac{1}{b}\right)q_j^t\widehat{W}q_j\right) + c^{st}.$$

The gradient of $\mathscr{L} = -2\log(L) - \sum_{j=1}^{p}\theta_j(q_j^tq_j - 1)$ with respect to q_j is:

$$\nabla_{q_j}\mathscr{L} = 2n\left(\frac{1}{a_j} - \frac{1}{b}\right)\widehat{W}q_j - 2\theta_jq_j,$$

where θ_j is the jth Lagrange multiplier. The relation $\nabla_{q_j}\mathscr{L} = 0$ implies that q_j is eigenvector of $\widehat{W}$. In order to minimize $-2\log(L)$, the first columns of Q must be the eigenvectors associated to the d largest eigenvalues of $\widehat{W}$. □

Proof of Proposition 4.7. Model $[a_jbQd]$: The partial derivatives of $-2\log(L)$ with respect to a_j and b are:

$$-2\frac{\partial\log(L)}{\partial a_j} = \frac{n}{a_j} - \frac{n}{a_j^2}q_j^t\widehat{W}q_j \quad \text{and} \quad -2\frac{\partial\log(L)}{\partial b} = \frac{n(p-d)}{b} - \frac{n}{b^2}\sum_{j=d+1}^{p}q_j^t\widehat{W}q_j.$$

The condition $\frac{\partial\log(L)}{\partial a_i} = 0$ implies that $\hat{a}_j = \lambda_j$. The combination of the condition $\frac{\partial\log(L)}{\partial b} = 0$ with the relation $\sum_{j=d+1}^{p}\lambda_j = \operatorname{tr}(\widehat{W}) - \sum_{j=1}^{d}\lambda_j$ gives the estimator of b.

Model $[abQd]$: The partial derivatives of $-2\log(L)$ with respect to a is:

$$-2\frac{\partial\log(L)}{\partial a} = \frac{nd}{a} - \frac{n}{a^2}\sum_{j=1}^{d}q_j^t\widehat{W}q_j,$$

and the condition $\frac{\partial\log(L)}{\partial a} = 0$ implies that $\hat{a} = \frac{1}{d}\sum_{j=1}^{d}\lambda_j$ and concludes the proof. □

Acknowledgment

This work was supported by the French department of research under the program "*ACI Masse de données*" (MoViStaR project). The authors wish to thank two referees for their helpful comments on the first draft of this article.

References

Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Bensmail, H., Celeux, G. (1996). Regularized Gaussian discriminant analysis through eigenvalue decomposition. *J. Amer. Statist. Assoc.* 91:1743–1748.

Bouveyron, C., Girard, S., Schmid, C. (2005). High dimensional discriminant analysis. Int. Conf. Appl. Stochastic Mod. Data Anal. Brest, France, pp. 526–534.

Cattell, R. (1966). The scree test for the number of factors. *Multivariate Behav. Res.* 1:140–161.

Flury, B. (1984). Common principal components in k groups. *J. Amer. Statist. Assoc.* 79:892–897.

Flury, B., Gautschi, W. (1986). An algorithm for simultaneous orthogonal transformation of several positive definite symmetric matrices to nearly diagonal form. *SIAM J. Scientific Statist. Comput.* 7:169–184.

Flury, L., Boukai, B., Flury, B. (1997). The discrimination subspace model. *J. Amer. Statist. Assoc.* 92:758–766.

Friedman, J. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.* 84:165–175.

Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Machine Learn. Res.* 3:1157–1182.

Hastie, T., Tibshirani, R., Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer Verlag.

Jolliffe, I. (1986). *Principal Component Analysis*. New York: Springer-Verlag.

Krzanowski, W., Jonathan, P., McCarthy, W., Thomas, M. (1995). Discriminant analysis with singular covariance matrices. *Appl. Statist.* 44:101–105.

McLachlan, G. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: John Wiley & Sons.

Mkhadri, A., Celeux, G., Nasrollah, A. (1997). Regularization in discriminant analysis: an overview. *Computat. Statist. Data Anal.* 23:403–423.

Schwarz, G. (1978). Estimating the dimension of a model. *Annal. Stat.* 6:461–464.

Scott, D., Thompson, J. (1983). Probability density estimation in higher dimensions. *Fifteenth Symp. Interface*. Elsevier Science Publishers, pp. 173–179.