# Probabilistic Fisher discriminant analysis: A robust and flexible alternative to Fisher discriminant analysis

Charles Bouveyron, Camille Brunet

# Probabilistic Fisher discriminant analysis: A robust and flexible alternative to Fisher discriminant analysis

Charles Bouveyron[1] & Camille Brunet[2]

[1] *Laboratoire SAMM, EA 4543, University Paris 1 Panthéon-Sorbonne*
*90 rue de Tolbiac – 75013 PARIS - FRANCE*

[2] *Equipe Modal'X, EA 3454, Université Paris X Ouest Nanterre*
*200 av. de la République, 92000 Nanterre, France*

**Abstract**

Fisher discriminant analysis (FDA) is a popular and powerful method for dimensionality reduction and classification. Unfortunately, the optimality of the dimension reduction provided by FDA is only proved in the homoscedastic case. In addition, FDA is known to have poor performances in the cases of label noise and sparse labeled data. To overcome these limitations, this work proposes a probabilistic framework for FDA which relaxes the homoscedastic assumption on the class covariance matrices and adds a term to explicitly model the non-discriminative information. This allows the proposed method to be robust to label noise and to be used in the semi-supervised context. Experiments on real-world datasets show that the proposed approach works at least as well as FDA in standard situations and outperforms it in the label noise and sparse label cases.

## 1. Introduction

Fisher discriminant analysis (FDA) [10, 13], also known as LDA by misnomer, is a commonly used method for linear dimensionality reduction in supervised classification. FDA aims to find a linear subspace that well separates the classes in which a linear classifier can be learned. In this paper, FDA will refer to the strategy which first finds a discriminative subspace and then classify the data in this subspace using linear discriminant analysis

(LDA) [29, Chap. 3]. FDA is a popular method, appreciated for its simplicity, which works very well in numerous cases. However, FDA does have some well-known limitations. In particular, FDA has not been originally defined in a probabilistic framework and its theoretical justification can be obtained only under the homoscedastic assumption on the distribution of the classes, *i.e.* each class has the same covariance matrix. Moreover, FDA produces correlated axes and its prediction performances are sensitive to label noise and sparse labeled data (semi-supervised context).

Unfortunately, label noise and sparse labeled data are nowadays frequent situations in application fields where the human supervision is either imprecise, difficult or expensive. For instance, in bio-medical applications, domain experts are asked to manually label a sample of learning data (MRI images, DNA micro-array, ...) which are then used for building a supervised classifier. In such cases, the cost of the supervision phase is usually high due to the difficulty of labeling complex data. Furthermore, an human error is always possible in such a difficult task and an error in the supervision phase could have big effects on the decision phase, particularly if the size of the learning sample is small. It is therefore very important to provide supervised classifiers robust enough to deal with data with uncertain labels and able to exploit the unlabeled observations of the data.

In this paper, we propose a supervised classification method, called probabilistic Fisher discriminant analysis (PFDA), based on a Gaussian parametrization of the data in a latent orthonormal discriminative subspace with a low intrinsic dimension. This probabilistic framework relaxes the homoscedastic assumption on the class covariance matrices and adds a term to explicitly model the non-discriminative information. This allows PFDA to be robust to label noise and to be used in the semi-supervised context. Numerical experiments show that PFDA improves predictive effectiveness in the label noise and semi-supervised contexts compared to FDA. As we know that the scientific literature is full of extensions of FDA, we do not claim that the proposed discriminant analysis method outperforms all existing works related to FDA in all situations. Nevertheless, the present work proposes a probabilistic, robust and flexible alternative to FDA which compares positively with

reference methods such as heteroscedasctic discriminant analysis (HDA) [26], regularized discriminant analysis (RDA) [12] and mixture discriminant analysis (MDA) [19]. PFDA may be therefore used by practitioners for their daily uses in place of FDA with the same advantages but without the label noise and sparse labeled data issues.

The paper is organized as follows. Section 2 first reviews the original discriminant analysis of Fisher and then presents its major probabilistic, robust and semi-supervised extensions. Section 3 introduces the discriminative latent mixture model and Section 4 discusses its inference in both the supervised and semi-supervised contexts. Experiments on real datasets presented in Section 5 illustrate the qualities of PFDA and compare it to state-of-the-art methods in various contexts. Finally, Section 6 gives some concluding remarks and directions for further work.

## 2. Related works

This section first recalls the nominal Fisher's discriminant analysis method and then briefly presents its major probabilistic, robust and semi-supervised extensions.

### *2.1. Fisher's discriminant analysis*

In his precursor work [10], Fisher poses the problem of the discrimination of three species of iris described by four measurements. The main goal of Fisher was to find a linear subspace that best separates the classes according to a criterion (see [9] for more details). For this, Fisher assumes that the dimensionality $p$ of the original space is greater than the number $K$ of classes. Fisher's discriminant analysis looks for a linear transformation matrix $U$ which allows to project the observations $\{y_1, ..., y_n\}$ in a discriminative and low dimensional subspace of dimension $d$. To this end, the $p \times d$ transformation matrix $U$ maximizes a criterion which is large when the between-class covariance matrix ($S_B$) is large and when the within-covariance matrix ($S_W$) is small. Since the rank of $S_B$ is at most equal to $K-1$, the dimension $d$ of the discriminative subspace is therefore at most equal to $K-1$ as well. Four different criteria can be found in the literature which satisfy such a constraint

(see [13] for a review). The criterion which is traditionally used is:

$$J(U) = \text{trace}((U^t S_W U)^{-1} U^t S_B U), \tag{2.1}$$

where $S_W = \frac{1}{n}\sum_{k=1}^{K}\sum_{y_i \in C_k}(y_i - m_k)(y_i - m_k)^t$ and $S_B = \frac{1}{n}\sum_{k=1}^{K} n_k(m_k - \bar{y})(m_k - \bar{y})^t$ are respectively the within and the between covariance matrices, $n_k$ is the number of observations in the $k$th class, $m_k = \frac{1}{n_k}\sum_{i \in C_k} y_i$ is the empirical mean of the observed column vector $y_i$ in the class $k$ and $\bar{y} = \frac{1}{n}\sum_{k=1}^{K} n_k m_k$ is the mean column vector of the observations. The maximization of criterion (2.1) is equivalent to the generalized eigenvalue problem [25] $\left(S_W^{-1} S_B - \lambda I_p\right) U = 0$ and the classical solution of this problem is the eigenvectors associated to the $d$ largest eigenvalues of the matrix $S_W^{-1} S_B$. Once the discriminative axes determined, linear discriminant analysis (LDA) is usually applied to classify the data into this subspace. The optimization of the Fisher criterion supposes the non-singularity of the matrix $S_W$ but it appears that the singularity of $S_W$ occurs frequently, particularly in the case of very high-dimensional space or in the case of under-sampled problems. In the literature, different solutions [12, 13, 18, 21, 23] are proposed to deal with such a problem in the supervised classification framework.

### *2.2. Probabilistic extensions of FDA*

Many authors have proposed ways to overcome the theoretical limitations of the original method. A first probabilistic framework has been proposed by Hastie *et al.* [19] by considering the different classes as a mixture of Gaussians with common covariance matrices. In 1998, Kumar *et al.* [26] have rewritten the Fisher's problem through a probabilistic framework which relaxes the homoscedastic constraint of FDA. More recently, Ioffe [22] has proposed a probabilistic approach for LDA. The same year, Yu *et al.* [40] have adapted the framework of probabilistic principal component analysis (PPCA), developed by Tipping *et al.* [37], in a supervised context and have found that the maximum likelihood of their approach is equivalent to the one of FDA in the homoscedastic context. Besides, Zhang *et al.* [42] have presented an extension of the Yu's work by considering the heteroscedastic case in a supervised and semi-supervised context which implies that the linear transformation is

different for each class.

*2.3. Dealing with the label noise problem*

Learning a supervised classifier from data with uncertain labels can be achieved using three main strategies: cleaning the data, using robust estimations of model parameters and finally modeling the label noise. Early approaches tried to clean the data by removing the misclassified instances using some kind of nearest neighbor algorithm [8, 15, 38]. Other works handle the noisy data using the C4.5 algorithm [24, 43], neural networks [41] or a saturation filter [14]. Hawkins *et al.* [20] identified as outliers the data subset whose deletion leads to the smallest value of the determinant of the within-class covariance matrix. Other researchers proposed not to remove any learning instance but to build instead supervised classifiers robust to label noise. Bashir *et al.* [2] and Croux *et al.* [7] focused on robust estimation of the model parameters in the mixture model context. Maximum likelihood estimators of the mixture model parameters are replaced by the corresponding S-estimators (see Rousseeuw and Leroy [34] for a general account on robust estimation) but the authors only observed a slight reduction of the average probability of misclassification. Boosting [33, 35] can also be used to limit the sensitivity of the built classifier to the label noise. Among all these solutions, the model proposed in [27] by Lawrence *et al.* has the advantage of explicitly including the label noise in the model with a sound theoretical foundation in the binary classification case. Denoting by $z$ and $\tilde{z}$ the actual and the observed class labels of an observation $y$, it is assumed that their joint distribution can be factorized as $p(y, z, \tilde{z}) = p(y|z)P(z|\tilde{z})P(\tilde{z})$. The class conditional densities $p(y|z)$ are modeled by Gaussian distributions while the probabilistic relationship $P(z|\tilde{z})$ between noisy and observed class labels is specified by a $2 \times 2$ probability table. An EM-like algorithm is introduced for building a kernel Fisher discriminant classifier on the basis of the above model. Finally, Bouveyron and Girard [5] proposed to relax the distribution assumption of Lawrence *et al.* by allowing each class density $p(y|z)$ to be modeled by a mixture of several Gaussians and confront the class information with an unsupervised modeling of the data for detecting

label errors.

### *2.4. FDA in the semi-supervised context*

The supervision cost of modern data often limits the number of labeled observations and, unfortunately, an error in the supervision phase could have particularly big effects on the classification phase when the size of the learning sample is small. In particular, supervised dimension reduction methods, such as FDA, tend to over-fit and therefore perform poorly in such situations. To avoid such a drawback, semi-supervised techniques propose to exploit additional unlabeled observations to improve the robustness of the classifier. For this, semi-supervised techniques [3, 31, 32] often rely on the mixture model and use the EM algorithm to infer the model from the partially labeled dataset. In the dimension reduction context, Sugiyama *et al.* [36] proposed to combine FDA with PCA for finding a subspace which preserves the global structure of unlabeled samples while discriminating as much as possible the known classes. Unfortunately, the effect of label noise on semi-supervised discriminant analysis has not been studied to our knowledge and one can think that label noise will have a significant effect in such a situation.

## 3. A probabilistic model for Fisher discriminant analysis

This section first introduces a probabilistic model, named the discriminative latent model (DLM), which fits the data in a latent orthonormal discriminative subspace with an intrinsic dimension lower than the dimension of the original space.

### *3.1. The probabilistic model*

Let us consider a complete training dataset $\{(y_1, z_1), ..., (y_n, z_n)\}$ where $z_i \in \{1, \ldots, K\}$ indicates the class label of the observation $y_i \in \mathbb{R}^p$. On the one hand, let us assume that $\{y_1, \ldots, y_n\}$ are independent observed realizations of a random vector $Y \in \mathbb{R}^p$ and that $\{z_1, \ldots, z_n\}$ are also independent realizations of a random variable $Z \in \{1, \ldots, K\}$. With these notations, we can define the prior probability of the $k$th class by $\pi_k = P(Z = k)$, for $k = 1, ..., K$. On the other hand, let $\mathbb{E} \subset \mathbb{R}^p$ denote a linear latent space

assumed to be the most discriminative subspace of dimension $d \leq K-1$ such that $\mathbf{0} \in \mathbb{E}$ and where $d$ is strictly lower than the dimension $p$ of the observed space. Moreover, let $\{x_1, \ldots, x_n\} \in \mathbb{E}$ denote the latent data which are in addition presumed to be independent unobserved realizations of a random vector $X \in \mathbb{E}$. Finally, for each class, the observed variable $Y \in \mathbb{R}^p$ and the latent variable $X \in \mathbb{E}$ are assumed to be linked through a linear transformation:

$$Y = UX + \varepsilon, \tag{3.1}$$

where $d < p$, $U$ is the $p \times d$ orthonormal matrix common to the $K$ class, such as $U^t U = I_d$, and $\varepsilon \in \mathbb{R}^p$, conditionally to $Z$, is a centered Gaussian noise term with covariance matrix $\Psi_k$, for $k = 1, ..., K$:

$$\varepsilon | Z = k \sim \mathcal{N}(\mathbf{0}, \Psi_k). \tag{3.2}$$

Following the classical framework of model-based clustering, each class is in addition assumed to be distributed according to a Gaussian density function within the latent space $\mathbb{E}$. Hence, the random vector $X \in \mathbb{E}$ has the following conditional density function:

$$X | Z = k \sim \mathcal{N}(\mu_k, \Sigma_k), \tag{3.3}$$

where $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d \times d}$ are respectively the mean and the covariance matrix of the $k$th class. Conditionally to $X$ and $Z$, the random vector $Y \in \mathbb{R}^d$ has therefore the following conditional distribution:

$$Y | X, Z = k \sim \mathcal{N}(UX, \Psi_k), \tag{3.4}$$

and its marginal class-conditional distribution is:

$$Y | Z = k \sim \mathcal{N}(m_k, S_k), \tag{3.5}$$

where:

$$\begin{aligned} m_k &= U\mu_k, \\ S_k &= U\Sigma_k U^t + \Psi_k, \end{aligned}$$

are respectively the mean and the covariance matrix of the $k$th class in the observation space. Let us also define $W = [U, V]$ a $p \times p$ matrix which satisfies $W^t W = W W^t = I_p$ and for which the $p \times (p-d)$ matrix $V$, is the orthonormal complement of $U$ defined above. We finally assume that the noise covariance matrix $\Psi_k$ satisfies the conditions $V \Psi_k V^t = \beta_k I_{d-p}$ and $U \Psi_k U^t = 0_d$, such that $\Delta_k = W^t S_k W$ has the following form:

$$\Delta_k = \begin{pmatrix} \boxed{\Sigma_k} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} \beta_k & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & \beta_k \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \end{matrix}}\right\} d \leq K-1 \\ \left.\vphantom{\begin{matrix} a \\ a \end{matrix}}\right\} (p-d) \end{matrix}$$

This model, called the discriminative latent model (DLM) and referred to by $\text{DLM}_{[\Sigma_k \beta_k]}$ in the sequel, is summarized by Figure 1. The $\text{DLM}_{[\Sigma_k \beta_k]}$ model is therefore parametrized by $\pi_k$, $\mu_k$, $U$, $\Sigma_k$ and $\beta_k$, for $k = 1, ..., K$ and $j = 1, ..., d$. On the one hand, $\pi_k$ and $\mu_k$ parametrize in a classical way the prior probability and the average latent position of the $k$th class respectively. On the other hand, $U$ defines the latent subspace $\mathbb{E}$ by parametrizing its orientation according to the basis of the original space. Finally, $\Sigma_k$ parametrize the variance of the $k$th class within the latent subspace $\mathbb{E}$ whereas $\beta_k$ parametrizes the variance of the class outside $\mathbb{E}$. With these notations and from a practical point of view, one can say that the discriminative information for the $k$th class is therefore modeled by $\Sigma_k$ and non discriminative information for this class is modeled by $\beta_k$.

### *3.2. Sub-models of the* $\text{DLM}_{[\Sigma_k \beta_k]}$ *model*

Starting with the $\text{DLM}_{[\Sigma_k \beta_k]}$ model presented in the previous paragraph, several sub-models can be generated by applying constraints on parameters of the matrix $\Delta_k$. For instance, the covariance matrices $\Sigma_1, \ldots, \Sigma_K$ in the latent space can be assumed to be common across the classes and this sub-model will be referred to by $\text{DLM}_{[\Sigma \beta_k]}$. Similarly, in each class, $\Sigma_k$ can be assumed to be diagonal, *i.e.* $\Sigma_k = \text{diag}(\alpha_{k1}, \ldots, \alpha_{kd})$. This sub-model will be referred to by $\text{DLM}_{[\alpha_{kj} \beta_k]}$. In the same manner, the $p-d$ last values of $\Delta_k$

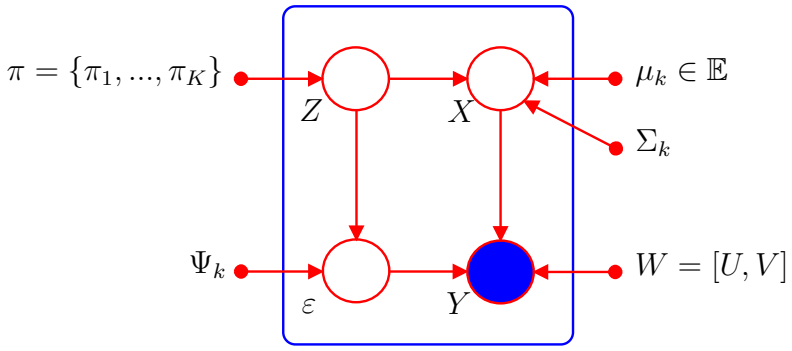



Figure 1: Graphical summary of the $\text{DLM}_{[\Sigma_k \beta_k]}$ model

can be assumed to be common for the $K$ classes, *i.e.* $\beta_k = \beta$, $\forall k = 1, ..., K$, meaning that the variance outside the discriminant subspace is common to all classes. This assumption can be viewed as modeling the non discriminative information with a unique parameter which seems natural for data obtained in a common acquisition process. Following the notation system introduces above, this sub-model will be referred to by $\text{DLM}_{[\alpha_{kj} \beta]}$. The variance within the latent subspace $\mathbb{E}$ can also be assumed to be isotropic for each class and the associated sub-model is $\text{DLM}_{[\alpha_k \beta_k]}$. In this case, the variance of the data is assumed to be isotropic both within $\mathbb{E}$ and outside $\mathbb{E}$. Similarly, it is possible to constrain the previous model to have the parameters $\beta_k$ common between classes and this gives rise to the model $\text{DLM}_{[\alpha_k \beta]}$. Finally, the variance within the subspace $\mathbb{E}$ can be assumed to be independent from the mixture component and this corresponds to the models $\text{DLM}_{[\alpha_j \beta_k]}$, $\text{DLM}_{[\alpha_j \beta]}$, $\text{DLM}_{[\alpha \beta_k]}$ and $\text{DLM}_{[\alpha \beta]}$. We therefore enumerate 12 different DLM models and an overview of them is proposed in Table 1. The table also gives the maximum number of free parameters to estimate (case of $d = K - 1$) according to $K$ and $p$ for the 12 DLM models and for some classical models. The Full-GMM model refers to the classical Gaussian model with full covariance matrices which yields the quadratic discriminant analysis (QDA) method. The Com-GMM model refers to the Gaussian model for which the covariance matrices are assumed to be equal to a common covariance matrix ($S_k = S$, $\forall k$) and this model

| Model | Nb. of parameters | $K = 4$ and $p = 100$ |
|---|---|---|
| $\text{DLM}_{[\Sigma_k \beta_k]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + K^2(K-1)/2 + K$ | 337 |
| $\text{DLM}_{[\Sigma_k \beta]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + K^2(K-1)/2 + 1$ | 334 |
| $\text{DLM}_{[\Sigma \beta_k]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + K(K-1)/2 + K$ | 319 |
| $\text{DLM}_{[\Sigma \beta]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + K(K-1)/2 + 1$ | 316 |
| $\text{DLM}_{[\alpha_{kj} \beta_k]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + K^2$ | 325 |
| $\text{DLM}_{[\alpha_{kj} \beta]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + K(K-1) + 1$ | 322 |
| $\text{DLM}_{[\alpha_k \beta_k]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + 2K$ | 317 |
| $\text{DLM}_{[\alpha_k \beta]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + K + 1$ | 314 |
| $\text{DLM}_{[\alpha_j \beta_k]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + (K-1) + K$ | 316 |
| $\text{DLM}_{[\alpha_j \beta]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + (K-1) + 1$ | 313 |
| $\text{DLM}_{[\alpha \beta_k]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + K + 1$ | 314 |
| $\text{DLM}_{[\alpha \beta]}$ | $(K-1) + K(K-1) + (K-1)(p - K/2) + 2$ | 311 |
| Full-GMM | $(K-1) + Kp + Kp(p+1)/2$ | 20603 |
| Com-GMM | $(K-1) + Kp + p(p+1)/2$ | 5453 |
| Diag-GMM | $(K-1) + Kp + Kp$ | 803 |
| Sphe-GMM | $(K-1) + Kp + K$ | 407 |

Table 1: Number of free parameters to estimate when $d = K - 1$ for the DLM models and some classical models. In particular, the Full-GMM model is the model of QDA and Com-GMM is the model of LDA (see text for details).

is the model of LDA. Diag-GMM refers to the Gaussian model for which $S_k = \text{diag}(s_{k1}^2, ..., s_{kp}^2)$ with $s_k^2 \in \mathbb{R}^p$ and Sphe-GMM refers to the Gaussian model for which $S_k = s_k^2 I_p$ with $s_k^2 \in \mathbb{R}$. In addition to the number of free parameters to estimate, Table 1 gives this number for specific values of $K$ and $p$ in the right column. The number of free parameters to estimate given in the central column can be decomposed in the number of parameters to estimate for the proportions $(K-1)$, for the means $(Kp)$ and for the covariance matrices (last terms). Among the classical models, the Full-GMM model is a highly parametrized model and requires the estimation of 20603 parameters when $K = 4$ and $p = 100$. Conversely, the Diag-GMM and Sphe-GMM model are very parsimonious models since they respectively require the estimation of only 803 and 407 parameters when $K = 4$ and $p = 100$. The Com-GMM model appears to have an intermediate complexity. Finally, the DLM models turn out to have low complexities whereas their modeling capacities are comparable to the one of the Full-GMM model.

### *3.3. Comparison with related models*

At this point, it is possible to highlight the main differences between the probabilistic model proposed in this work and the related models. Firstly, the DLM model differs from the FDA model on the fact that FDA only links the observed variable $Y$ with the latent variable $X$ through $U$ whereas the DLM model takes into account and model in addition the non discriminative information through the term $\varepsilon$. This specific feature of the proposed model implies that all the original variables (with different balancing terms however) are used for modeling the classes and classifying future observations. The DLM model also differs from the heteroscedasctic model of HDA, proposed by Kumar & Andreou [26], on two key points. Firstly, their model only relaxes the homoscedastic assumption on the covariances matrices within the latent space and not outside this subspace. Secondly, as in FDA, their approach does not keep all variables for the classification of new observations and retains only the $K-1$ dimensions assumed to carry all the discriminative information. Finally, although the parsimonious Gaussian model (HD-GMM) proposed by Bouveyron *et al.* [6] uses all variables to model and classify

high-dimensional data as the DLM model, this model however differs from our model in the fact that the HD-GMM model fits each class in a different latent subspace. Furthermore, the class-specific subspaces associated with the HD-GMM model are chosen such that the variance of the projected data is maximum whereas the DLM model chooses the latent subspace orientation such that it best discriminates the classes.

## 4. Parameter estimation and classification

This section presents parameter estimation for DLM parameters in both the supervised and semi-supervised cases. Classification of new observations through the MAP rule is discussed as well.

### *4.1. Parameter estimation in the supervised context*

Conversely to the probabilistic approaches reviewed in Section 2, the probabilistic model presented above is very general and there is no explicit solution for the likelihood maximization with respect to $U$. Therefore, we propose to estimate the linear transformation $U$ and the model parameters in two different steps.

*Estimation of the discriminative subspace.* Firstly, the estimate $\hat{U}$ of the latent subspace orientation $U$ is obtained through the optimization of the Fisher criterion with respect to the orthogonality of its column vectors,

$$\max_{U} \quad \text{tr}\left((U^t S_W U)^{-1} U^t S_B U\right) \quad \text{w.r.t.} \quad U^t U = \mathbf{I_d}, \tag{4.1}$$

where $S_W = \frac{1}{n}\sum_{k=1}^{K}\sum_{y_i \in C_k}(y_i - m_k)(y_i - m_k)^t$ and $S_B = \frac{1}{n}\sum_{k=1}^{K} n_k (m_k - \bar{y})(m_k - \bar{y})^t$ are respectively the within and the between covariance matrices, $m_k = \frac{1}{n_k}\sum_{i=1}^{n} \mathbf{1}_{\{z_i=k\}} y_i$, $n_k = \sum_{i=1}^{n} \mathbf{1}_{\{z_i=k\}}$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. This optimization problem can be solved using different ways (see [13, 21] for details) and the Gram-Schmidt procedure will be used in the experiments of Section 5.

*Estimation of model parameters.* Secondly, conditionally to the orientation matrix $\hat{U}$ estimated in the previous step, the estimation of model parameters is done by maximization of the likelihood. With the assumptions and

notations of the model $[\Sigma_k\beta_k]$, the log-likelihood for the learning data is:

$$\begin{aligned}\mathcal{L}(\theta) = -\frac{1}{2}\sum_{k=1}^{K}\Big[-2\log(\pi_k) + \operatorname{trace}(\Sigma_k^{-1}U^tC_kU) + \log(|\Sigma_k|) \\ + (p-d)\log(\beta_k) + \frac{1}{\beta_k}\Big(\operatorname{trace}(C_k) - \sum_{j=1}^{d}u_j^tC_ku_j\Big) + \gamma\Big].\end{aligned} \tag{4.2}$$

where $C_k$ is the empirical covariance matrix of the $k$th class, $u_j$ is the $j$th column vector of $U$ and $\gamma = p\log(2\pi)$ is a constant term. Given $U = \hat{U}$ and in conjunction with equation (4.1), the maximization of the log-likelihood (4.2) conduces to the following estimates in the case of the $\mathrm{DLM}_{[\Sigma_k\beta_k]}$ model:

- prior probabilities $\pi_k$ are estimated by $\hat{\pi}_k = \sum_{i=1}^{n}\mathbf{1}_{\{z_i=k\}}$,
- means $\mu_k$ are estimated by $\hat{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{n}\mathbf{1}_{\{z_i=k\}}\hat{U}^ty_i$,
- covariance matrices $\Sigma_k$ are estimated by $\hat{\Sigma}_k = \hat{U}^tC_k\hat{U}$,
- and variances $\beta_k$ are estimated by $\hat{\beta}_k = \frac{\operatorname{tr}(C_k) - \sum_{j=1}^{d}\hat{u}_j^tC_k\hat{u}_j}{p-d}$.

Proofs of these results can be deduced from the ones given in [4]. Finally, the intrinsic dimension $d$ of the discriminative latent subspace $\mathbb{E}$ is set to the rank of $S_B^{(q)}$ (see [13]).

*4.2. Parameter estimation in the semi-supervised context*

Let us consider now that $\{(y_i, z_i)\}_{i=1}^{n_\ell}$ where $n_\ell \leq n$ are the labeled data and there are $n - n_\ell$ unlabeled data referred to by $\{y_i\}_{i=n_\ell+1}^{n}$. The $n_\ell$ labeled observations are modeled by the probabilistic framework developed in Section 3 and the unlabeled data are modeled by a mixture model parametrized by $\pi_k$, the mixture proportion of the class $k$, and $\theta_k = (m_k, S_k)$, respectively its mean vector and its covariance matrix. Thus, the log-likelihood can be written as:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n_\ell}\sum_{k=1}^{K}\mathbf{1}_{\{z_i=k\}}\log\left(\pi_k\phi(y_i;\theta_k)\right) + \sum_{i=n_\ell+1}^{n}\log(\sum_{k=1}^{K}\pi_k\phi(y_i;\theta_k)) \tag{4.3}$$

In such a case, the direct maximization of $\mathcal{L}(\theta)$ is intractable and an iterative procedure has to be used. The Fisher-EM algorithm has been recently pro-

posed by [4] for iteratively maximizing $\mathcal{L}(\theta)$ in the case of the DLM models. For this, the Fisher-EM algorithm alternates 3 steps at iteration $q$:

*E-step.* This step computes the expectation of the complete log-likelihood conditionally to the current value of the parameter $\theta^{(q-1)}$. In practice, this step reduces to the computation for the unlabeled points of $t_{ik}^{(q)} = E[z_{ik}|y_i, \theta^{(q-1)}]$ where $z_i = k$ if $y_i$ comes from the $k$th component, $i = n_\ell, ..., n,$. Let us also recall that $t_{ik}^{(q)}$ is as well the posterior probability $P(Z = k|Y = y_i)$ that the observation $y_i$ belongs to the $k$th component of the mixture. For the labeled points, the value of $t_{ik}^{(q)}$ is set to $\mathbf{1}_{\{z_i=k\}}$ for $i = 1, ..., n_\ell$ and $k = 1, ..., K$.

*F-step.* This step aims to determinate, at iteration $q$, the discriminative latent subspace of dimension $d \leq K - 1$ in which the $K$ classes are best separated. Naturally, the estimation of this latent subspace has to be done conditionally to the current values of posterior probabilities $t_{ik}^{(q)}$ which indicates the current soft partition of the data. Estimating the discriminative latent subspace reduces to maximize the traditional criterion $J(U) = tr((U^t S_W U)^{-1} U^t S_B U)$. However, the traditional criterion $J(U)$ assumes that the data are complete (supervised classification framework). Unfortunately, in the present case, the matrices $S_B$ and $S_W$ have to be defined conditionally to the current soft partition for the unlabeled data. It is therefore necessary to introduce the soft between-covariance matrix $S_B^{(q)}$ and the soft within-covariance matrix $S_W^{(q)}$. The soft between-covariance matrix $S_B^{(q)}$ is defined conditionally to the posterior probabilities $t_{ik}^{(q)}$, obtained in the E step, as follows:

$$S_B^{(q)} = \frac{1}{n} \sum_{k=1}^{K} n_k^{(q)} (\hat{m}_k^{(q)} - \bar{y})(\hat{m}_k^{(q)} - \bar{y})^t, \tag{4.4}$$

where $n_k^{(q)} = \sum_{i=1}^{n} t_{ik}^{(q)}$, $\hat{m}_k^{(q)} = \frac{1}{n} \sum_{i=1}^{n} t_{ik}^{(q)} y_i$ is the soft mean of the $k$th class at iteration $q$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ is the empirical mean of the whole dataset. Since the relation $S = S_W^{(q)} + S_B^{(q)}$ holds in this context as well, it is preferable from a computational point of view to use the covariance matrix $S = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})^t$ of the whole dataset in the maximization problem instead of $S_W^{(q)}$ since $S$ remains fixed over the iterations. The F step of the

Fisher-EM therefore aims to solve the following optimization problem:

$$\begin{cases} \max\limits_{U} & \text{trace}\left((U^tSU)^{-1}U^tS_B^{(q)}U\right), \\ \text{w.r.t.} & u_j^tu_l = 0, \quad \forall j \neq l \in \{1,\ldots,d\}, \end{cases} \tag{4.5}$$

where $u_j$ is the $j$th column vector of $U$. The procedure then follows the concept of the orthonormal discriminant vector (ODV) method introduced by [11] in the supervised case and then extended by [16, 17, 28, 39], which sequentially selects the most discriminative features in maximizing the Fisher criterion subject to the orthogonality of features.

*M-step.* This third step estimates the model parameters by maximizing the conditional expectation of the complete likelihood and this conduces, at iteration $q$, to an estimation of the mixture proportions $\pi_k$ and the means $\mu_k$ for the $K$ components by their empirical counterparts:

$$\hat{\pi}_k^{(q)} = \frac{n_k^{(q)}}{n}, \qquad \hat{\mu}_k^{(q)} = \frac{1}{n_k}\sum_{i=1}^{n} t_{ik}^{(q)}\hat{U}^{(q)t}y_i, \tag{4.6}$$

with $n_k^{(q)} = \sum_{i=1}^{n} t_{ik}^{(q)}$. In the case of the $\text{DLM}_{[\Sigma_k\beta_k]}$ model, the remaining parameters are estimated by:

$$\hat{\Sigma}_k^{(q)} = \hat{U}^{(q)t}C_k^{(q)}\hat{U}^{(q)}, \tag{4.7}$$

and

$$\hat{\beta}_k^{(q)} = \frac{\text{trace}(C_k^{(q)}) - \sum_{j=1}^{d}\hat{u}_j^{(q)t}C_k^{(q)}\hat{u}_j^{(q)}}{p-d}, \tag{4.8}$$

where $C_k^{(q)} = \frac{1}{n_k^{(q)}}\sum_{i=1}^{n} t_{ik}^{(q)}(y_i - \hat{m}_k^{(q)})(y_i - \hat{m}_k^{(q)})^t$ and $\hat{m}_k^{(q)} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{(q)}y_i$. Parameter estimation for the other DLM models and proofs of these results can be found in [4].

### *4.3. Classification of new observations*

In the discriminant analysis framework, new observations are usually assigned to a class using the *maximum a posteriori* (MAP) rule which assigns a new observation $y \in \mathbb{R}^p$ to the class for which $y$ has the highest posterior probability $P(Z = k|Y = y)$. Therefore, the classification step mainly con-

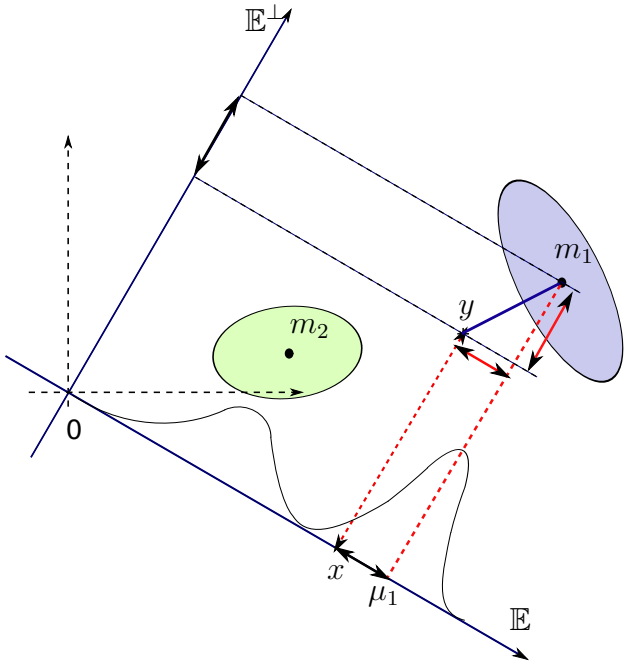



Figure 2: Two classes and their 1-dimensional discriminative subspace.

sists in calculating the posterior probability $P(Z = k|Y = y)$ for each class $k = 1, ..., K$. Maximizing the posterior probability over $k$ is equivalent to minimizing the classification function $\Gamma_k(y) = -2\log(\pi_k \phi(y; m_k, S_k)$ which is for our model equal to:

$$\begin{aligned}\Gamma_k(y) = \quad & \left\|UU^t(y-m_k)\right\|^2_{\vartheta_k} + \frac{1}{\beta_k}\left\|(y-m_k) - UU^t(y-m_k)\right\|^2 \\ & + \log(|\Sigma_k|) + (p-d)\ \log(\beta_k) - 2\log(\pi_k) + p\ \log(2\pi),\end{aligned} \tag{4.9}$$

where $\vartheta_k = [U, \mathbf{0}_{p-d}]\, \Delta_k^{-1}\, [U, \mathbf{0}_{p-d}]^t$ and $\|.\|_{\vartheta_k}$ is a norm on the latent space spanned by $[U, \mathbf{0}_{p-d}]$ such that $||y||^2_{\vartheta_k} = y^t \vartheta_k y$.

Besides its computational interest, the above formula provides as well a comprehensive interpretation of the classification function $\Gamma_k$ which mainly governs the computation of $P(Z = k|Y = y)$. Indeed, it appears that $\Gamma_k$ mainly depends on two distances: the distance between the projections on the discriminant subspace $\mathbb{E}$ of the observation $y_i$ and the mean $m_k$ on the one hand, and, the distance between the projections on the complementary subspace $\mathbb{E}^{\perp}$ of $y_i$ and $m_k$ on the other hand. Remark that the latter dis-

tance can be reformulated in order to avoid the use of the projection on $\mathbb{E}^{\perp}$. Indeed, as Figure 2 illustrates, this distance can be re-expressed according projections on $\mathbb{E}$. Therefore, the posterior probability $P(Z = k|Y = y)$ will be close to 1 if both the distances are small which seems quite natural. Obviously, these distances are also balanced by the variances in $\mathbb{E}$ and $\mathbb{E}^{\perp}$ and by the mixture proportions. Furthermore, the fact that the E step does not require the use of the projection on the complementary subspace $\mathbb{E}^{\perp}$ is, from a computational point of view, very important because it provides the stability of the algorithm and allows its use when $n < p$ (see [4] for details).

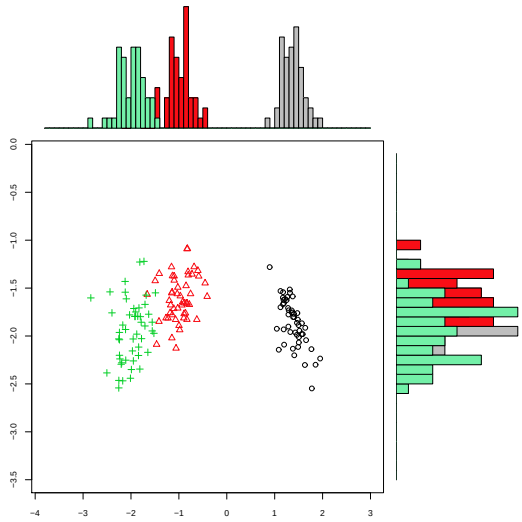
## 5. Experimental results

This section presents experiments on real-world datasets which aim to highlight the main features of the proposed probabilistic version of FDA and to show that PFDA can be considered as a robust and flexible alternative to FDA.

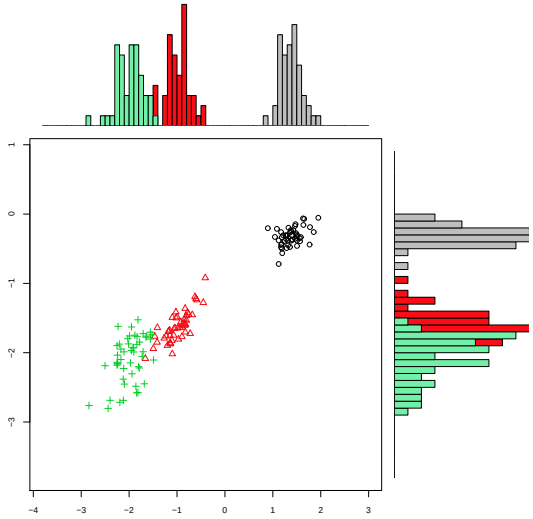
### *5.1. An introductory example: the Iris dataset*

It seemed to us natural to first apply PFDA to the Iris dataset that Sir R.A. Fisher used in [10] as an illustration for his discriminant analysis. This dataset, in fact collected by E. Anderson [1] in the Gaspé peninsula (Canada), is made of three classes corresponding to different species of iris (*setosa*, *versicolor* and *virginica*) among which the classes *versicolor* and *virginica* are difficult to discriminate (they are at least not linearly separable). The dataset consists of 50 samples from each of three species and four features were measured from each sample. The four measurements are the length and the width of the sepal and the petal. This dataset is used here as an introductory example because of the link with Fisher's work but also for its popularity in the classification community. For this introductory example, the $\text{DLM}_{[\alpha_k\beta]}$ model was used for PFDA and it is compared to FDA and orthonormalized FDA (OFDA) [17].

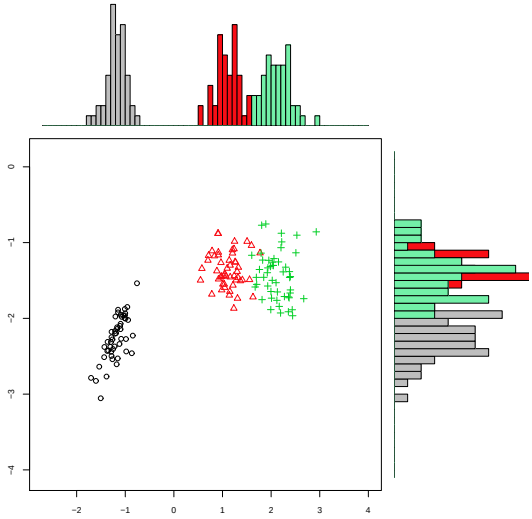
Figure 3 presents the projection of the Iris data into the latent discriminative subspaces respectively estimated by FDA, OFDA and PFDA. Unsurprisingly, all projections discriminate almost perfectly the data. One can



(a) FDA



(b) OFDA



(c) PFDA

Figure 3: Projection of the Iris data into the latent discriminative subspace estimated by FDA, OFDA and PFDA.

| | FDA | | OFDA | | PFDA | |
|---|---|---|---|---|---|---|
| | *axis* | | *axis* | | *axis* | |
| *variable* | 1 | 2 | 1 | 2 | 1 | 2 |
| sepal length | 0.208 | -0.006 | 0.208 | 0.152 | -0.203 | -0.062 |
| sepal width | 0.386 | -0.586 | 0.386 | -0.036 | -0.324 | -0.697 |
| petal length | -0.554 | 0.252 | -0.554 | -0.765 | 0.519 | 0.404 |
| petal width | -0.707 | -0.769 | -0.707 | 0.624 | 0.763 | -0.588 |

Table 2: Loadings associated with the discriminative axes estimated by FDA, OFDA and PFDA for the Iris data.

remark that OFDA provides however a slightly different projection compared to the one of FDA, due to its orthogonality constraint, and PFDA provides an intermediate projection between FDA and OFDA. Table 2 confirms this intuition. The first discriminative axis is overall estimated in the same manner by the three methods, but PFDA provides a closer estimation to the FDA estimation of the second axis than OFDA. Indeed, the cosine between the second discriminative axis estimated by PFDA and the one of FDA is 0.96 whereas it is -0.65 between OFDA and FDA. It is recalled that PFDA provides, as well as OFDA, discriminative axes which are orthogonal. Figure 3 presents the correct classification rates obtained by FDA, OFDA and PFDA for 25 bootstrap replications on the Iris data. It turns out that the three methods perform on average similarly even though PFDA provides sometimes better results than FDA and OFDA. As a partial conclusion, PFDA can be considered as a good alternative to FDA which produces in addition orthogonal discriminative axes.

*5.2. Comparison of PFDA and its sub-models with reference methods*

As described in Section 2, the family of probabilistic models of PFDA contains 12 models and this second experiment aims to compare their different performances. To do so, we chose 4 real-world datasets (Iris, Wine, Chiro and Ecoli) on the UCI Machine Learning repository (`http://archive.ics.uci.edu/ml/`) and we compared the prediction performances of PFDA for the 12 DLM models with the reference performances of FDA, OFDA, HDA [26] and RDA [12]. The Wine dataset is made of 178 Italian wines described by 13 variables and split up into 3 classes. The Chiro dataset con-

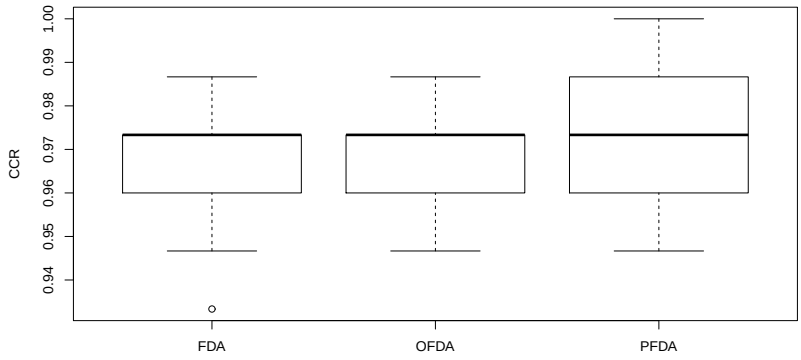



Figure 4: Correct classification rates obtained by FDA, OFDA and PFDA for 25 bootstrap replications on the Iris data.

tains 148 Chironomus larvae which are split up into 3 species and described by 17 morphometric attributes. This dataset is described in detailed in [30]. Finally, the Ecoli dataset is made of 272 observations of the E-coli bacteria which has recently received a lot of attention in the news due to the major epidemic in Germany. The 272 observations are described by 6 measures and are split up into 3 classes which correspond to different localization sites of the bacteria.

Table 3 provides the correct classification rates obtained by FDA, OFDA, HDA, RDA and PFDA for 25 bootstrap replications on the four considered datasets. For each bootstrap replication, the whole dataset was split into a learning set of 50% of the data and a test set with the remaining data. Correct classification rates are of course evaluated on a test dataset. On the one hand and from a global point of view, we can remark that the prediction performances of all methods are on average comparable. Nevertheless, OFDA seems penalized by its orthogonal constrain and performs less than the other studied methods. On the other hand, PFDA turns out to be once again a good alternative to FDA since it performs slightly better than FDA, HDA and RDA on these four datasets. The good performance of PFDA is certainly due to its flexibility. Indeed, the different probabilistic models of PFDA allows it to fit onto different situations. Beside, we can notice that the

| Method | Iris | Wine | Chiro | Ecoli |
|---|---|---|---|---|
| FDA | 97.5±1.4 | 97.5±1.4 | 98.1±1.1 | 94.1±2.0 |
| OFDA | 97.5±1.2 | 95.4±3.4 | 96.3±1.5 | 89.7±5.4 |
| HDA | 97.5±1.4 | 97.5±1.5 | 96.8±4.0 | 92.5±2.3 |
| RDA | 96.8±2.4 | 97.2±1.5 | 96.0±3.3 | 94.1±1.9 |
| PFDA $[\Sigma_k\beta_k]$ | 96.5±1.8 | 98.1±1.5 | 97.5±1.2 | 94.4±1.7 |
| PFDA $[\Sigma_k\beta]$ | 97.4±1.1 | 97.3±1.6 | 98.2±1.0 | 94.3±1.7 |
| PFDA $[\alpha_{kj}\beta_k]$ | 96.5±2.0 | 98.1±1.4 | 96.7±2.2 | 94.4±1.7 |
| PFDA $[\alpha_{kj}\beta]$ | 97.7±1.3 | 97.0±1.3 | 98.2±0.9 | 94.3±2.0 |
| PFDA $[\alpha_k\beta_k]$ | 96.7±2.2 | 98.5±1.4 | 96.2±2.6 | 94.5±1.7 |
| PFDA $[\alpha_k\beta]$ | 97.5±1.3 | 98.2±1.4 | 98.1±1.1 | 93.6±1.7 |
| PFDA $[\Sigma\beta_k]$ | 83.6±3.1 | 95.1±2.4 | 88.2±3.9 | 91.0±2.3 |
| PFDA $[\Sigma\beta]$ | 86.7±4.0 | 93.0±3.0 | 94.4±4.0 | 93.7±2.3 |
| PFDA $[\alpha_j\beta_k]$ | 86.0±3.1 | 95.0±2.3 | 87.9±4.2 | 91.2±2.2 |
| PFDA $[\alpha_j\beta]$ | 88.1±2.4 | 93.2±3.0 | 92.7±4.3 | 93.8±2.1 |
| PFDA $[\alpha\beta_k]$ | 86.9±3.7 | 95.0±2.5 | 85.8±3.9 | 91.3±2.2 |
| PFDA $[\alpha\beta_k]$ | 91.2±3.0 | 93.4±3.1 | 88.3±3.6 | 93.7±2.0 |

Table 3: Correct classification rates (in percentage) and standard deviations obtained by FDA, OFDA, HDA, RDA and PFDA for 25 bootstrap replications on real-world datasets (see text for details).

most efficient models for PFDA are models with intermediate complexities ($[\alpha_{kj}\beta]$ and $[\alpha_k\beta_k]$). Furthermore, the fact that the homoscedastic models of PFDA (bottom part of Table 3) perform less justify the necessity to propose heteroscedastic models in the context of discriminant analysis.

*5.3. Robustness to label noise: influence of the noise type*

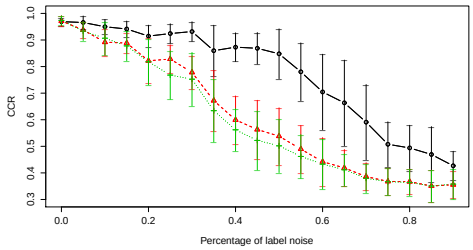
This third experiment aims to study the robustness of PFDA to different types of label noise. The FDA and OFDA methods are used as reference methods. The model used for PFDA was the model $[\Sigma_k\beta_k]$, which is the most general model of the DLM family. We have also tried other DLM models but we do not present their results here since their behaviors are similar to the presented one. The datasets used for this experimentation are the Iris and the USPS358 dataset. The USPS358 is a subset of the original USPS dataset (available at the UCI repository) which contains only observations of the digits 3, 5 and 8. It contains 1756 observations described by 256 measured variables which correspond to $16 \times 16$ gray scale images observed as vectors.

Since the aim of this experiment is to evaluate the robustness to label noise, let $\tau$ denote the percentage of false labels in the learning set. At each trial, the datasets are randomly divided in 2 balanced samples: a learning set of half the data in which a percentage $\tau$ of the data is mislabeled and a test set on which the prediction performances of the 3 methods are evaluated. This process was repeated 25 times for each value of $\tau$ in order to monitor both the average performances and their variances. Two kinds of label noise are considered. The first one corresponds to a scenario in which one class is overlapping the others. The second scenario corresponds to a random and equiprobable label noise.

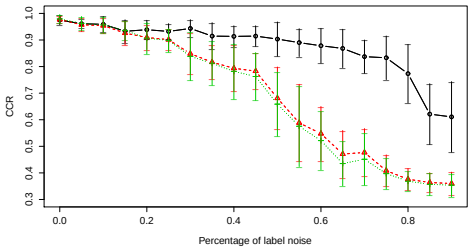
Figure 5 and 6 presents the evolution of correct classification rate computed on the test set for the studied methods according to $\tau$, respectively for the Iris and USPS358 datasets. First of all, it can be observed that the FDA and OFDA methods are sensitive to all types of label noise since their classification rates lower linearly with respect to $\tau$ in all considered cases. Conversely, PFDA turns out to be clearly more robust than FDA and OFDA in all the studied situations. On the Iris dataset, PFDA appears to be particularly robust in the overlapping situations whereas it is only slightly better than FDA in the other case. However, when dealing with high-dimensional data, PFDA outperforms clearly its challengers and shows a high robustness to all kinds of label noise. This robustness can be explained by the probabilistic model of PFDA which incorporates a term to model the *a priori* non discriminative information (which in fact carries discriminative information in the label noise case). This avoids to over-fit the embedding space on the labeled data and remains general enough to be robust to label noise conversely to FDA and OFDA.

### *5.4. Robustness to label noise: comparison with state-of-the-art methods*

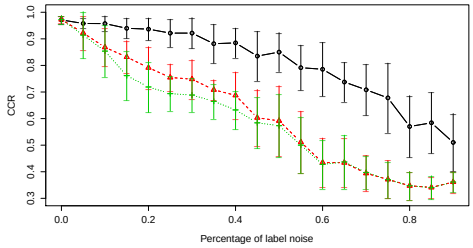
This experiment focuses now on the comparison of PFDA with other robust discriminant analysis methods. The methods used as reference methods are mixture discriminant analysis (MDA) [19], robust linear discriminant analysis (RLDA) [27] and robust mixture discriminant analysis (RMDA) [5]. In particular, RLDA and RMDA are very efficient and robust methods. Both



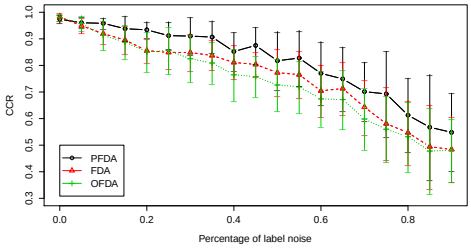

(a) Class 1 overlaps the others




(a) Class 2 overlaps the others

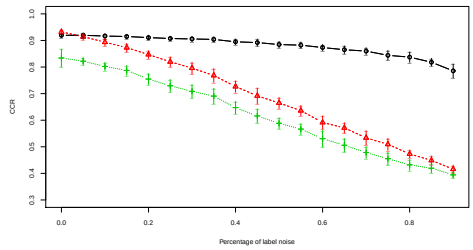



(c) Class 3 overlaps the others

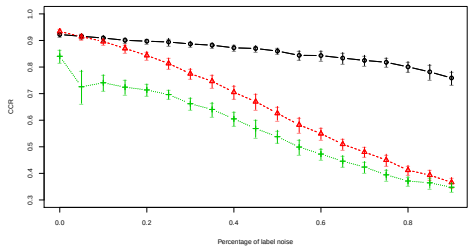



(d) equiprobable label noise

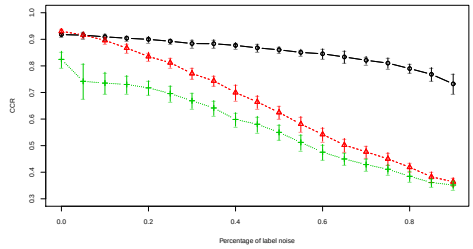
Figure 5: Effect of label noise in the learning dataset on the prediction effectiveness for FDA, OFDA and PFDA for the Iris dataset (3 classes, 4 dimensions). The prediction effectiveness is evaluated by the correct classification rate on the test set. Results are averaged on 25 bootstrap replications and vertical bars indicate the standard deviations.



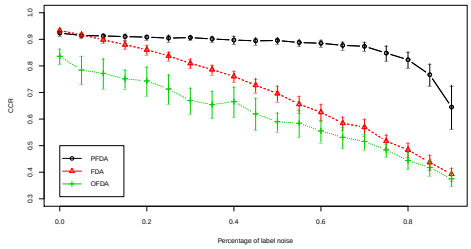

(a) Class 1 overlaps the others




(a) Class 2 overlaps the others




(c) Class 3 overlaps the others




(d) equiprobable label noise

Figure 6: Effect of label noise in the learning dataset on the prediction effectiveness for FDA, OFDA and PFDA on the test set for the USPS-358 dataset (3 classes, 256 dimensions). The prediction effectiveness is evaluated by the correct classification rate on the test set. Results are averaged on 25 bootstrap replications and vertical bars indicate the standard deviations.

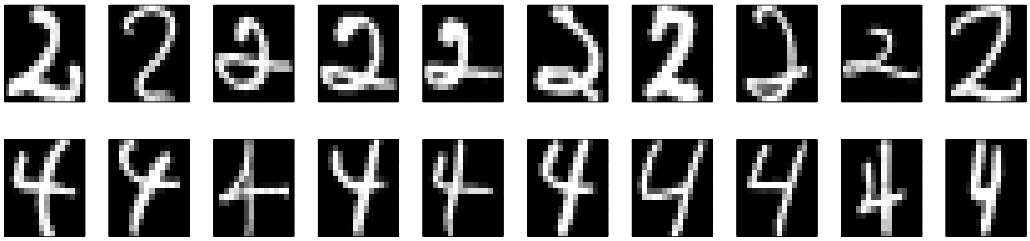


Figure 7: Samples from the USPS24 dataset.

were applied with success to object recognition in natural images. However, since RLDA is only able to consider binary classification cases, we restrict ourselves here to the classification of a datasets with only 2 classes. We chose to use the USPS24 dataset, considered in [5], which contains only observations of the digits 2 and 4. It contains 1383 observations described in a 256-dimensional space. Figure 7 presents a sample from the USPS24 dataset. For this comparison, we used the noise scenario which corresponds to a random and equiprobable label noise. This process was repeated 25 times for each value of $\tau$ in order to monitor both the average performances and their variances of the studied methods.

Figure 8 presents the evolution of correct classification rate computed on the test set for the studied methods according to $\tau$. As observed by [5], FDA, OFDA and MDA are very sensitive to label noise since their performances decrease linearly when $\tau$ increases. The RLDA method turns out to be significantly more robust than FDA, OFDA and MDA but its performance decreases quickly for contamination rate larger than 0.2. Finally, RMDA and PFDA appear to be very robust since they both provide very high correct classification rates for contamination rates up to 0.4. However, RMDA seems to be slightly less stable and reliable than PFDA due to its embedded EM algorithm. To summarize, PFDA can be considered as gathering the stability of RLDA and the robustness of RMDA while avoiding the drawbacks of these methods.

### *5.5. Robustness to label noise in the semi-supervised context*

This last experiment will focus on comparing on real-world datasets the efficiency of semi-supervised approaches with traditional supervised ones.

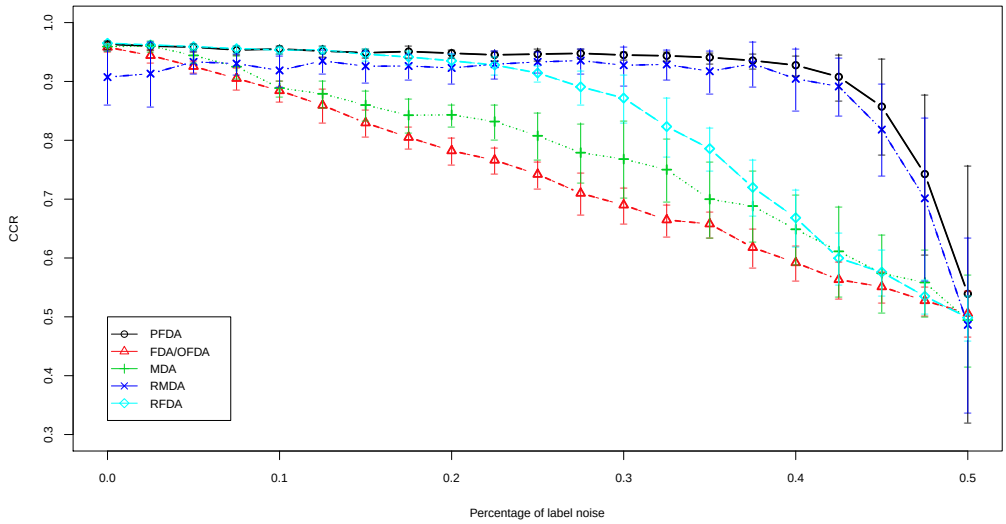



Figure 8: Comparison of PFDA with other robust discriminant analysis methods on the USPS24 dataset (2 classes, 256 dimensions). The prediction effectiveness is evaluated by the correct classification rate on the test set. Results are averaged on 25 Monte-Carlo replications and vertical bars indicate the standard deviations.

The used datasets are the same as in the previous sections and PFDA is compared here with FDA, OFDA and with a recent semi-supervised local approach, called SELF and proposed by [36]. This latter approach aims to find a discriminative subspace by considering both global and class structures. For this experiment, each dataset was randomly divided into a learning set and a test set containing 50% of the data each. In the learning set, a percentage $\gamma$ of data were randomly selected to constitute the known labeled data. Moreover, within these learning observations which are labeled, a percentage $\tau$ of the data is mislabeled according to the equiprobable label noise scenario. Therefore, the rate of correctly labeled observations in the learning set is $\gamma(1-\tau)$. Tables 4, 5 and 6 respectively present the correct classification rates, computed on the test sets, for a label noise rate equals to 0 (no label noise), 0.2 and 0.4 (strong label noise). In each case, three semi-supervised situations are considered: $\gamma = 0.2$ (almost unsupervised), 0.4 and 0.8 (almost supervised). This process was repeated 25 times for each value of $\gamma$ and $\tau$ in order to monitor both the average performances and variances of the studied

| Without label noise ($\tau = 0$) | | | | | |
|---|---|---|---|---|---|
| Sup. rate | Methods | Iris | Wine | Chiro | Ecoli |
| $\gamma = 0.2$ | PFDA $[\Sigma_k \beta_k]$ | 91.4±4.9 | 91.1±11.4 | 93.7±6.8 | 97.0± 1.9 |
| | PFDA $[\Sigma_k \beta]$ | 95.9±3.5 | 90.6±11.4 | 93.7±7.4 | 97.0±1.8 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 91.4±4.3 | 94.6±7.6 | 97.0±1.9 | 97.8±2.8 |
| | PFDA $[\alpha_{kj} \beta]$ | 95.9±2.4 | 93.9±7.5 | 96.5±2.9 | 93.9±5.7 |
| | SELF | 97.1±3.9 | 95.9±3.3 | 95.2±12.8 | 97.7±1.1 |
| | FDA | 95.6±3.1 | 80.4±10.1 | 87.5±5.4 | 87.3±7.4 |
| | OFDA | 95.6±2.1 | 77.8±10.2 | 86.8±5.6 | 92.1±4.8 |
| $\gamma = 0.4$ | PFDA $[\Sigma_k \beta_k]$ | 93.9±3.8 | 97.7±1.8 | 96.8±3.2 | 97.9±1.1 |
| | PFDA $[\Sigma_k \beta]$ | 96.2±2.7 | 96.8±2.2 | 96.7±4.4 | 98.1±1.2 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 93.8±3.7 | 97.8±1.2 | 97.8±1.4 | 98.3±1.4 |
| | PFDA $[\alpha_{kj} \beta]$ | 96.5±2.7 | 97.0±1.6 | 97.1±4.1 | 97.9±1.9 |
| | SELF | 98.1±0.7 | 97.6±1.5 | 100.0±0.0 | 98.0±1.3 |
| | FDA | 96.6±1.9 | 95.2±2.6 | 92.9±5.2 | 90.7±5.3 |
| | OFDA | 96.5±2.1 | 90.6±6.5 | 91.0±6.6 | 94.1±1.6 |
| $\gamma = 0.8$ | PFDA $[\Sigma_k \beta_k]$ | 95.5±1.8 | 97.9±1.3 | 98.0±1.7 | 98.4±0.7 |
| | PFDA $[\Sigma_k \beta]$ | 97.1±1.7 | 97.7±1.7 | 98.3±1.2 | 98.5±0.9 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 95.5±2.4 | 97.8±1.1 | 98.2±1.3 | 99.0±1.0 |
| | PFDA $[\alpha_{kj} \beta]$ | 97.1±1.4 | 97.5±1.8 | 98.3±1.2 | 98.4±2.2 |
| | SELF | 99.8±0.5 | 98.2±1.72 | 100.0±0.0 | 97.8±1.4 |
| | FDA | 97.3±1.6 | 97.1±1.5 | 97.6±1.4 | 93.7±2.2 |
| | OFDA | 97.2±1.1 | 95.9±3.7 | 95.6±2.1 | 94.7±1.5 |

Table 4: Prediction accuracies (in percentage) and standard deviations for different rates of labeled observations in the learning set (sup. rate) and without label noise.

methods.

On the one hand, one can notice that, as expected, the semi-supervised methods always outperform FDA and OFDA. This is due to the fact that the fully supervised methods estimate the embedding space only on the labeled data and thus over-fit it. Conversely, the semi-supervised methods take advantage of the unlabeled data in the discriminative subspace estimation, which enables them to be more effective. On the other hand, it appears that PFDA and SELF perform on average similarly when there is no label noise (Table 4). However, PFDA turns out to be more reliable than SELF when the labels of the learning observations are corrupted (Tables 5 and 6).

| | | 20% of label noise ($\tau = 0.2$) | | | |
|---|---|---|---|---|---|
| Sup. rate | Methods | Iris | Wine | Chiro | Ecoli |
| $\gamma = 0.2$ | PFDA $[\Sigma_k \beta_k]$ | 87.2±3.4 | 92.7±3.9 | 84.9±14.3 | 83.3±22.8 |
| | PFDA $[\Sigma_k \beta]$ | 77.0±14.6 | 90.1±2.1 | 84.4±11.5 | 93.2±4.0 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 96.2±3.3 | 76.1±3.2 | 89.0±7.3 | 97.6±1.4 |
| | PFDA $[\alpha_{kj} \beta]$ | 96.2±2.5 | 72.8±6.4 | 88.2±7.3 | 96.4±1.5 |
| | SELF | 95.2±9.7 | 92.7±5.3 | 94.9±13.5 | 97.5±1.8 |
| | FDA | 86.9±10.4 | 89.8±2.7 | 85.6±8.8 | 76.1±27.1 |
| | OFDA | 82.9±8.8 | 88.7±3.1 | 89.8±4.4 | 93.2±1.8 |
| $\gamma = 0.4$ | PFDA $[\Sigma_k \beta_k]$ | 85.8±3.4 | 94.0±2.8 | 88.0±9.3 | 84.8±8.5 |
| | PFDA $[\Sigma_k \beta]$ | 84.8±3.8 | 92.7±3.8 | 85.1±10.6 | 91.9±2.7 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 96.2±2.7 | 94.9±2.9 | 92.7±5.8 | 98.4±1.3 |
| | PFDA $[\alpha_{kj} \beta]$ | 96.5±2.2 | 92.7±4.8 | 90.0±6.5 | 98.2±1.5 |
| | SELF | 95.1±6.5 | 96.6±2.6 | 96.6±2.5 | 97.2±2.2 |
| | FDA | 81.8±4.2 | 87.6±6.5 | 88.0±3.8 | 82.9±7.4 |
| | OFDA | 84.5±5.1 | 89.2±1.8 | 88.5±4.3 | 94.5±0.4 |
| $\gamma = 0.8$ | PFDA $[\Sigma_k \beta_k]$ | 89.8±3.9 | 91.9±4.0 | 86.8±4.7 | 91.3±2.6 |
| | PFDA $[\Sigma_k \beta]$ | 86.1±4.3 | 90.7±3.1 | 87.4±6.0 | 91.0±3.2 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 98.1±1.1 | 97.1±2.1 | 97.9±1.4 | 99.0±0.4 |
| | PFDA $[\alpha_{kj} \beta]$ | 97.3±1.6 | 95.4±4.0 | 97.2±1.8 | 98.6±0.9 |
| | SELF | 91.9±5.9 | 97.0±1.7 | 97.1±1.8 | 96.8±1.1 |
| | FDA | 91.4±5.0 | 92.1±3.4 | 82.1±4.9 | 91.9±3.2 |
| | OFDA | 85.8±4.9 | 92.5±2.5 | 83.4±11.0 | 92.7±3.3 |

Table 5: Prediction accuracies (in percentage) and standard deviations for different rates of labeled observations in the learning set (sup. rate) and with 20% of label noise.

| | | 40% of label noise ($\tau = 0.4$) | | | |
|---|---|---|---|---|---|
| Sup. rate | Methods | Iris | Wine | Chiro | Ecoli |
| $\gamma = 0.2$ | PFDA $[\Sigma_k \beta_k]$ | 59.8±27.4 | 75.3±30.6 | 72.4±11.0 | 78.3±10.0 |
| | PFDA $[\Sigma_k \beta]$ | 58.6±27.6 | 77.9±24.3 | 71.0±12.2 | 85.2±10.3 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 94.4±4.7 | 78.8±9.0 | 77.8±15.9 | 97.7±2.1 |
| | PFDA $[\alpha_{kj} \beta]$ | 94.5±4.6 | 76.3±10.3 | 77.1±12.5 | 97.3±2.9 |
| | SELF | 91.0±10.3 | 91.7±9.1 | 97.2±6.9 | 96.9±1.9 |
| | FDA | 51.7±24.4 | 80.6±27.4 | 72.2±12.6 | 75.1±32.1 |
| | OFDA | 52.5±20.8 | 77.1±26.7 | 60.4±17.3 | 87.2±14.8 |
| $\gamma = 0.4$ | PFDA $[\Sigma_k \beta_k]$ | 56.0±20.8 | 86.2±10.9 | 68.4±12.7 | 78.0±2.7 |
| | PFDA $[\Sigma_k \beta]$ | 59.5±20.9 | 82.7±13.0 | 63.2±11.8 | 80.4±6.9 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 96.2±1.9 | 95.6±2.2 | 93.3±4.6 | 99.2±1.2 |
| | PFDA $[\alpha_{kj} \beta]$ | 96.1±2.0 | 90.7±6.1 | 91.2±5.5 | 99.1±1.5 |
| | SELF | 88.8±8.9 | 95.2±3.6 | 98.4±1.8 | 96.9±1.7 |
| | FDA | 56.0±22.9 | 89.1±5.8 | 66.7±11.0 | 86.4±8.6 |
| | OFDA | 56.0±22.1 | 89.1±3.6 | 56.2±12.2 | 86.0±9.5 |
| $\gamma = 0.8$ | PFDA $[\Sigma_k \beta_k]$ | 50.9±14.0 | 74.4±8.4 | 72.8±4.5 | 79.3±4.8 |
| | PFDA $[\Sigma_k \beta]$ | 54.1±12.6 | 74.2±8.0 | 68.1±5.5 | 75.0±5.0 |
| | PFDA $[\alpha_{kj} \beta_k]$ | 97.0±1.6 | 97.0±1.5 | 97.8±1.5 | 99.4±0.6 |
| | PFDA $[\alpha_{kj} \beta]$ | 97.5±1.1 | 94.2±3.4 | 96.4±2.6 | 99.2±0.5 |
| | SELF | 75.2±9.6 | 92.5±4.4 | 96.2±4.4 | 91.1±5.8 |
| | FDA | 54.4±16.8 | 78.7±10.6 | 70.8±4.8 | 81.6±3.5 |
| | OFDA | 54.4±18.8 | 78.4±11.0 | 64.2±5.3 | 79.4±1.0 |

Table 6: Prediction accuracies (in percentage) and standard deviations for different rates of labeled observations in the learning set (sup. rate) and with 40% of label noise.

## 6. Conclusion

This paper has presented a new probabilistic framework for FDA which relaxes the homoscedastic assumption on the class covariance matrices and adds a term to explicitly model the non-discriminative information. This allows PFDA to be robust to label noise and to be used in the semi-supervised context. Experiments on real-world datasets showed that the proposed PFDA method works at least as well as the traditional FDA method (even better in most cases) in standard situations and it clearly improves the modeling and the prediction when the dataset is subject to label noise and/or sparse labels. The practitioner may therefore replace without prejudice FDA by PFDA for its daily use.

Among the possible extensions of this work, it could be interesting to propose a unified estimation procedure for both the orientation matrix $U$ and the other model parameters. This should be at least possible in the isotropic case for which maximizing the Fisher's criterion is equivalent to maximizing the likelihood. Another interesting extension would be to introduce sparsity in the orientation matrix $U$ through a $\ell_1$ penalty in order to ease the interpretation of the discriminative axes.

## References


[1] E. Anderson. The irises of the Gaspé Peninsula. *Bulletin of the American Iris Society*, 59:2–5, 1935.

[2] S. Bashir and E. Carter. High breakdown mixture discriminant analysis. *Journal of Multivariate Analysis*, 93(1):102–111, 2005.

[3] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the Workshop on Computational Learning Theory*, pages 92–100, 1998.

[4] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, in press, 2011.

[5] C. Bouveyron and S. Girard. Robust supervised classification with mixture models: learning from data with uncertain labels. *Pattern Recognition*, 42:2649–2658, 2009.

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

[7] C. Croux and C. Dehon. Robust Linear Discriminant Analysis using S-estimators. *The Canadian Journal of Statistics*, 29:473–492, 2001.

[8] B. Dasarathy. Noising around the neighbourhood: a new system structure and classification rule for recognition in partially exposed environments. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2:67–71, 1980.

[9] R. Duda, P. Hart, and D. Stork. *Pattern classification*. John Wiley & Sons, 2000.

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

[11] D.H. Foley and J.W. Sammon. An optimal set of discriminant vectors. *IEEE Transactions on Computers*, 24:281–289, 1975.

[12] J.H. Friedman. Regularized discriminant analysis. *The journal of the American statistical association*, 84:165–175, 1989.

[13] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic. Press, San Diego, 1990.

[14] D. Gamberger, N. Lavrac, and C. Groselj. Experiments with noise filtering in a medical domain. In *16th International Conference on Machine Learning*, pages 143–151, USA, 1999.

[15] G. Gates. The reduced nearest neighbor rule. *IEEE Transactions on Information Theory*, 18(3):431–433, 1972.

[16] Y-F. Guo, S-J. Li, J-Y. Yang, T-T. Shu, and L-D. Wu. A generalized Foley-Sammon transform based on generalized fisher discriminant criterion and its application to face recognition. *Pattern Recognition letters*, 24:147–158, 2003.

[17] Y. Hamamoto, Y. Matsuura, T. Kanaoka, and S. Tomita. A note on the orthonormal discriminant vector method for feature extraction. *Pattern Recognition*, 24(7):681–684, 1991.

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

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

[20] D. Hawkins and G. McLachlan. High-breakdown linear discriminant analysis. *Journal of the American Statistical Association*, 92(437):136–143, 1997.

[21] P. Howland and H. Park. Generalizing discriminant analysis using the generalized singular decomposition. *IEEE transactions on pattern analysis and machine learning*, 26(8):995–1006.

[22] S. Ioffe. Probabilistic linear discriminant analysis. *Computer Vision ECCV 2006*, pages 531–542, 2006.

[23] Z. Jin, J.Y. Yang, Z.S. Hu, and Z. Lou. Face recognition based on the uncorrelated optimal discriminant vectors. *Pattern Recognition*, 10(34):2041–2047, 2001.

[24] G. John. Robust decision trees: Removing outliers from databases. In *First conference on Knowledge Discovery and Data Mining*, pages 174–179, 1995.

[25] W. Krzanowski. *Principles of Multivariate Analysis*. Oxford University Press, Oxford, 2003.

[26] N. Kumar and A.G. Andreou. Heteroscedastic discriminant analysis and reduced rank hmms for improved speech recognition. *Speech Communication*, 26(4):283–297, 1998.

[27] N. Lawrence and B. Schölkopf. Estimating a kernel Fisher discriminant in the presence of label noise. In *Proc. of 18th International Conference on Machine Learning*, pages 306–313. Morgan Kaufmann, San Francisco, CA, 2001.

[28] K. Liu, Y-Q. Cheng, and J-Y. Yang. A generalized optimal set of discriminant vectors. *Pattern Recognition*, 25(7):731–739, 1992.

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

[30] A. Montanari and C. Viroli. Heteroscedastic Factor Mixture Analysis. *Statistical Modeling: An International journal (forthcoming)*, 10(4):441–460, 2010.

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

[32] T. O'Neill. Normal discrimination with unclassified observations. *Journal of the American Statistical Association*, 73:821–826, 1978.

[33] J. Quinlan. Bagging, boosting and C4.5. In *13th National Conference on Artificial Intelligence*, pages 725–730, USA, 1996.

[34] P.J. Rousseeuw and A. Leroy. *Robust Regression and Outlier Detection*. Wiley, New York, 1987.

[35] R. Schapire. The strength of weak learnability. *Machine Learning*, 5:197–227, 1990.

[36] M. Sugiyama, T. Idé, S. Nakajima, and J. Sese. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. *Machine Learning*, 78:35–61, 2009.

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

[38] D. Wilson and T. Martinez. Instance pruning techniques. In *Fourteenth International Conference on Machine Learning*, pages 404–411, USA, 1997.

[39] J. Ye. Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005.

[40] S. Yu, K. Yu, V. Tresp, H.P. Kriegel, and M. Wu. Supervised probabilistic principal component analysis. In *Proc. of the 12th ACM SIGKDD*, pages 464–473, USA, 2006.

[41] X. Zeng and T. Martinez. A noise filtering method using neural networks. In *IEEE International Workshop on Soft Computing Techniques in Instrumentation, Measurement and Related Applications*, pages 26–31, 2003.

[42] Z. Zhang, G. Dai, and M.I. Jordan. A flexible and efficient algorithm for regularized fisher discriminant analysis, 2009.

[43] X. Zhu, X. Wu, and Q. Chen. Eliminating class noise in large datasets. In *20th ICML International Conference on Machine Learning*, pages 920–927, USA, 2003.