On the estimation of the latent discriminative subspace in the Fisher-EM algorithm

Charles Bouveyron, Camille Brunet

Submission

Journal de la Société Française de Statistique

arXiv: math.PR/00000000

On the estimation of the latent discriminative subspace in the Fisher-EM algorithm

Titre: Sur l'estimation du sous-espace latent discriminant de l'algorithme Fisher-EM

Charles Bouveyron[1] and Camille Brunet[2]

Abstract: The Fisher-EM algorithm has been recently proposed in [2] for the simultaneous visualization and clustering of high-dimensional data. It is based on a discriminative latent mixture model which fits the data into a latent discriminative subspace with an intrinsic dimension lower than the dimension of the original space. The Fisher-EM algorithm includes an F-step which estimates the projection matrix whose columns span the discriminative latent space. This matrix is estimated *via* an optimization problem which is solved using a Gram-Schmidt procedure in the original algorithm. Unfortunately, this procedure suffers in some case from numerical instabilities which may result in a deterioration of the visualization quality or the clustering accuracy. Two alternatives for estimating the latent subspace are proposed to overcome this limitation. The optimization problem of the F-step is first recasted as a regression-type problem and then reformulated such that the solution can be approximated with a SVD. Experiments on simulated and real datasets show the improvement of the proposed alternatives for both the visualization and the clustering of data.

Résumé : L'algorithme Fisher-EM a été récemment proposé dans [2] pour simultanément visualiser et modéliser des données de grande dimension. Il se base sur un modèle de mélange latent discriminant qui modélise les données dans un sous-espace discriminant qui a une dimension intrinsèque plus petite que celle de l'espace des observations. L'algorithme Fisher-EM est composé d'une étape F qui estime la matrice de projection dont les colonnes engendrent le sous-espace latent discriminant. Cette matrice est estimée *via* un problème d'optimisation, lequel est résolu, dans l'algorithme original, par une procédure de Gram-Schmidt. Malheureusement, cette procédure souffre dans certains cas d'instabilités numériques qui peut engendrer une détérioration de la qualité de la visualisation ou de la classification automatique des données. Pour pallier cette limitation, nous proposons deux alternatives d'estimation du sous-espace latent. Le problème d'optimisation de l'étape F est réécrit comme un problème de régression puis reformulé de telle manière que la solution peut être approximée par une SVD. Des expériences sur des données simulées et réelles montre l'amélioration des alternatives proposées pour la visualisation et la classification automatique des données.

Keywords: clustering, Fisher-EM algorithm, regression problem, Fisher's criterion, discriminative latent subspace, dimension reduction
Mots-clés : classification automatique, algorithme Fisher-EM, problème de régression, critère de Fisher, sous-espace latent discriminant, réduction de dimension
AMS 2000 subject classifications: 35L05, 35L70

1. Introduction

Nowadays, the measured observations are very often high-dimensional and clustering such data remains a challenging problem. In particular, when considering the mixture model context, the corresponding clustering methods show a disappointing behavior in high-dimensional spaces.

[1] Université Paris 1-Panthéon-Sorbonne, 90 rue Tolbiac - 75013 PARIS - FRANCE.
E-mail: charles.bouveyron@univ-paris1.fr
[2] Université Paris X, 200 avenue de la République - 92001 NANTERRE - FRANCE.

Soumis au Journal de la Société Française de Statistique
File: essai.tex, compiled with jsfds, version : 2009/12/09
date: October 10, 2011

They suffer from the well-known *curse of dimensionality* [1] which is mainly due to the fact that model-based clustering methods are over-parametrized in high-dimensional spaces.

Hopefully, since the dimension of observed data is usually higher than their intrinsic dimension, it is theoretically possible to reduce the dimension of the original space without loosing any information. In the literature, a very common way to reduce the dimension is to use feature extraction methods such as principal component analysis (PCA) or feature selection methods. Alternatives to these methods are the subspace clustering methods [3, 13, 14, 15, 17] which avoid dimension reduction. These techniques have been proposed in the past few years to model the data of each group in low-dimensional subspaces. In a different approach, Raftery and Dean [19] and Maugis et al. [12] propose a method for feature selection in the model-based clustering context by recasting the variable selection problem as a model selection problem.

However, these approaches present certain limitations. For example, when the dimension reduction is operated before the clustering task, the discriminative information can be lost which is of course damageable for the classification task. In particular, Chang [4] showed that the principal components linked to the largest eigenvalues do not necessary contain the most relevant information about the group structure of the dataset. In the case of subspace clustering, even though these methods turned out to be very efficient in practice, they are usually not able to provide a global visualization of the clustered data since they model each group in a specific subspace. Finally, the main disadvantage in the works of [12, 19] remains in the estimation procedure which is too time-consuming in the case of high-dimensional data.

Recently, Bouveyron and Brunet [2] proposed a new statistical framework which aims to simultaneously cluster the data and produce a low-dimensional representation of the clustered data. To that end, the proposed model clusters the data into a common latent subspace which both best discriminates the groups according to the current fuzzy partition of the data and has an intrinsic dimension lower than the dimension of the observation space. Moreover, they propose an estimation procedure called the Fisher-EM algorithm. This algorithm is based on an EM procedure from which an additional step, named F-step, is introduced to estimate the projection matrix whose columns span the discriminative latent space. This matrix is estimated via an optimization problem which is solved using the concept of orthonormal discriminant vector developed by [6] through a Gram-Schmidt procedure. However, the set of column vectors built are not guaranteed to be optimal and such an approach remains numerically unstable because of the Gram-Schmidt process.

In order to improve the estimation procedure of the discriminative latent space, we propose in this paper two different alternatives. On the one hand, we reformulate the optimization problem originally based on an eigen-decomposition problem as a regression-type problem. To do so, our proposal is based on a result obtained in the supervised context by Qiao et al. [18]. In the other hand, we propose an approach based on the singular value decomposition (SVD) which best approximates the discriminative space while facilitating its estimation in practice.

This paper is organized as follows. Section 2 reviews the discriminative latent mixture model and its estimation procedure (the Fisher-EM algorithm). Section 3 presents the two proposed alternatives for estimating the discriminative latent mixture model. Then, numerical experiments are presented in Section 4 to highlight the improvements of the proposed alternatives. Some concluding remarks and ideas for further works are finally given in Section 5.

Soumis au Journal de la Société Française de Statistique
File: essai.tex, compiled with jsfds, version : 2009/12/09
date: October 10, 2011

2. The DLM model and the Fisher-EM algorithm

The discriminative latent mixture (DLM) model aims to both cluster the data at hand and reduce their dimensionality into a common latent subspace. Conversely to similar approaches such as [3, 14, 16, 17, 21] for example, this latent subspace is assumed to be discriminative, in the sense that it best discriminates K groups according to the current fuzzy partition of the data. Moreover, its intrinsic dimension is strictly bounded by the number of groups.

2.1. The DLM model

Let $\{y_1,\dots,y_n\} \in \mathbb{R}^p$ denote a dataset of n observations that one wants to cluster into K homogeneous groups, *i.e.* adjoin to each observation y_j a value $z_j \in \{1,\dots,K\}$ where $z_i = k$ indicates that the observation y_i belongs to the kth group. On the one hand, let us assume that $\{y_1,\dots,y_n\}$ are independent observed realizations of a random vector $Y \in \mathbb{R}^p$ and that $\{z_1,\dots,z_n\}$ are also independent realizations of a random vector $Z \in \{1,\dots,K\}$. On the other hand, let $\mathbb{E} \subset \mathbb{R}^p$ denote a 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 $\{x_1,\dots,x_n\} \in \mathbb{E}$ denote the actual data, described in the latent space $\mathbb{E}$ of dimension d, which are in addition presumed to be independent unobserved realizations of a random vector $X \in \mathbb{E}$. Finally, for each group, 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:

$$Y = UX + \varepsilon, \tag{1}$$

where U is a $p \times d$ orthogonal matrix common to K groups and satisfying $U^tU = \mathbf{I}_d$. The p-dimensional random vector ε stands for the noise term and conditionally to Z, ε is assumed to be distributed according to a centered Gaussian density function with covariance matrix Ψ_k ($\varepsilon_{|Z=k} \sim \mathcal{N}(0,\Psi_k)$). Besides, within the latent space, $X_{|Z=k}$ is $\mathcal{N}(\mu_k,\Sigma_k)$ where $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d\times d}$ are respectively the mean vector and the covariance matrix of the kth group. Given these distribution assumptions and according to equation (1), $Y_{|X,Z=k}$ is $\mathcal{N}(UX,\Psi_k)$ and its marginal distribution is therefore a mixture of Gaussians:

$$f(y) = \sum_{k=1}^{K} \pi_k \phi(y; m_k, S_k), \tag{2}$$

where π_k is the mixing proportion of the class k and $\phi(.)$ denotes a multivariate Gaussian density function parametrized by the mean vector $m_k = U\mu_k$ and the covariance matrix $S_k = U\Sigma_k U^t + \Psi_k$ of the kth group in the observation space. Furthermore, in the DLM model, a $p \times p$ matrix $W = [U,V]$ is defined, satisfying the condition $W^tW = WW^t = \mathbf{I}_p$, and the $(p-d) \times p$ matrix V is the orthogonal complement of U. Finally, the noise covariance matrix Ψ_k needs to satisfy the conditions $V\Psi_kV^t = \beta_k\mathbf{I}_{p-d}$ and $U\Psi_kU^t = \mathbf{0}_d$, such that $\Delta_k = W^tS_kW = \mathrm{diag}(\Sigma_k, \beta_k\mathbf{I}_{p-d})$. These last conditions imply that the discriminative subspace and the non discriminative one are orthogonal, which suggests in practice that all the relevant clustering information remains in the latent subspace. This model is referred to by $\mathrm{DLM}_{[\Sigma_k\beta_k]}$ in [2] and a graphical summary is given in Figure 1.

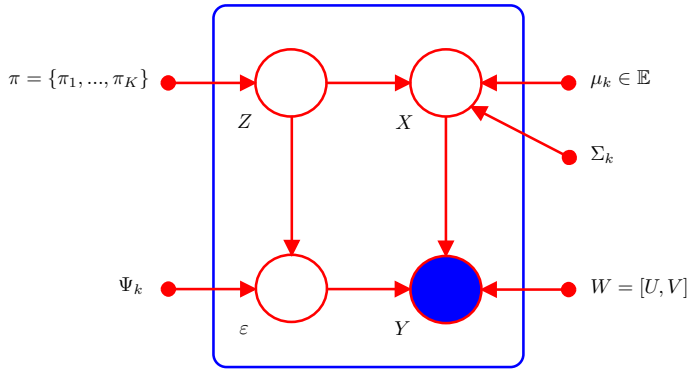


FIGURE 1. *Graphical summary of the* $DLM_{[\Sigma_k \beta_k]}$ *model.*

2.2. A family of parsimonious model

A family of parsimonious models can be obtained by constraining the parameters Σ_k or β_k to be common. For instance, the covariance matrices $\Sigma_1, \ldots, \Sigma_K$ in the latent space can be assumed to be common across groups and this submodel is referred to by $DLM_{[\Sigma \beta_k]}$. Similarly, in each group, Σ_k can be assumed to be diagonal, *i.e.* $\Sigma_k = \text{diag}(\alpha_{k1}, \ldots, \alpha_{kd})$. This submodel is referred to by $DLM_{[\alpha_{kj} \beta_k]}$. A constraint can also be applied in the parameter β_k by assuming it to be common to all classes ($\forall k$, $\beta_k = \beta$). 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. A detailed description of the 12 different DLM models can be found in [2]. Such a family yields very parsimonious models and allows, in the same time, to fit into various situations. In particular, the complexity of the $DLM_{[\Sigma_k \beta_k]}$ model mainly depends on the number of clusters K since the dimensionality of the discriminative subspace is such that $d \leq K - 1$. The complexity of the $DLM_{[\Sigma_k \beta_k]}$ grows linearly with p contrary to the traditional Gaussian models in which the complexity increases with p^2. As an illustration, if we consider the case with $p = 100$, $K = 4$ and $d = 3$, then the complexity of the $DLM_{[\Sigma_k \beta_k]}$ is $\gamma = 337$ which is drastically less than the number of parameters to estimate in the case of the Full-GMM ($\gamma = 20603$).

2.3. The Fisher-EM algorithm

An estimation procedure, called the Fisher-EM algorithm, is proposed in [2] in order to both estimate the discriminative space and the parameters of the mixture model. This algorithm is based on the EM algorithm from which an additional step is introduced, between the E and the M-step. This additional step, named F-step, aims to compute the projection matrix whose columns span the discriminative latent space. The Fisher-EM algorithm has therefore the following form, at iteration q:

The E-step This step computes the posterior probabilities $t_{ik}^{(q)}$ that the observations belong to the K groups, at iteration q, using the following update formula:

$$t_{ik}^{(q)} = \hat{\pi}_k^{(q-1)} \phi(y_i, \hat{\theta}_k^{(q-1)}) / \sum_{\ell=1}^{K} \hat{\pi}_\ell^{(q-1)} \phi(y_i, \hat{\theta}_\ell^{(q-1)}), \tag{3}$$

with $\hat{\theta}_k = \{\hat{\mu}_k, \hat{\Sigma}_k, \hat{\beta}_k, \hat{U}\}$.

The F-step This step estimates, conditionally to the posterior probabilities, the orientation matrix $U^{(q)}$ of the discriminative latent space by maximizing the Fisher's criterion [5, 7] under orthonormality constraints:

$$\begin{aligned} \hat{U}^{(q)} &= \max_{U^{(q)}} \quad \text{trace}\left((U^{(q)t} S U^{(q)})^{-1} U^{(q)t} S_B^{(q)} U^{(q)} \right), \\ &\quad \text{w.r.t.} \quad U^{(q)t} U^{(q)} = \mathbf{I}_d. \end{aligned} \tag{4}$$

where S stands for the covariance matrix and:

$$S_B = \frac{1}{n} \sum_{k=1}^{K} n_k^{(q)} (m_k^{(q)} - \bar{y})(m_k^{(q)} - \bar{y})^t, \tag{5}$$

denotes the soft within covariance matrix with $n_k^{(q)} = 1/n \sum_{i=1}^{n} t_{ik}$, $m_k^{(q)} = 1/n \sum_{i=1}^{n} t_{ik}^{(q)} y_i$ and $\bar{y} = 1/n \sum_{i=1}^{n} y_i$. This optimization problem is solved in [2] using the concept of orthonormal discriminant vector developed by [6] through a Gram-Schmidt procedure. Such a process enables to fit a discriminative and low-dimensional subspace while providing orthonormal discriminative axes conditionally to the current soft partition of the data. In addition, according to the optimization criterion defined in (4), the dimensionality of the discriminative space d is strictly bounded by the number of clusters K.

The M-step This final step estimates the parameters of the mixture model in the latent subspace by maximizing the conditional expectation of the complete log-likelihood:

$$Q(\theta) = \frac{\text{-}1}{2} \sum_{k=1}^{K} n_k^{(q)} \Big[\text{-}2\log(\pi_k) + \text{tr}(\Sigma_k^{-1} \hat{U}^{(q)t} C_k \hat{U}^{(q)}) + \log(|\Sigma_k|) + (p\text{-}d)\log(\beta_k) + \frac{\text{tr}(C_k)\text{-}\sum_{j=1}^{d} \hat{u}_j^{(q)t} C_k \hat{u}_j^{(q)}}{\beta_k} + \gamma \Big]. \tag{6}$$

where C_k is the empirical covariance matrix of the kth group, $\hat{u}_j^{(q)}$ is the jth column vector of $\hat{U}^{(q)}$, $n_k^{(q)} = \sum_{i=1}^{n} t_{ik}^{(q)}$ and $\gamma = p\log(2\pi)$ is a constant term. Hence, maximizing Q conditionally to $\hat{U}^{(q)}$ leads to the following update formula for the mixture parameters of the model $\text{DLM}_{[\Sigma_k \beta_k]}$:

$$\hat{\pi}_k^{(q)} = \frac{n_k^{(q)}}{n}, \tag{7}$$

$$\hat{\mu}_k^{(q)} = \frac{1}{n_k^{(q)}} \sum_{i=1}^{n} t_{ik}^{(q)} \hat{U}^{(q)t} y_i, \tag{8}$$

$$\hat{\Sigma}_k^{(q)} = \hat{U}^{(q)t} C_k \hat{U}^{(q)}, \tag{9}$$

$$\hat{\beta}_k^{(q)} = \frac{\text{tr}(C_k)\text{-}\sum_{j=1}^{d} \hat{u}_j^{(q)t} C_k \hat{u}_j^{(q)}}{p-d}. \tag{10}$$

Soumis au Journal de la Société Française de Statistique
File: essai.tex, compiled with jsfds, version : 2009/12/09
date: October 10, 2011

The Fisher-EM procedure iteratively updates the parameters until convergence of the likelihood. The convergence of the algorithm is guaranteed under certain conditions. Finally, since the latent subspace has a low dimension and is also common to all groups, the clustered data can be easily visualized.

3. Two alternatives for estimating the discriminative subspace

In [2], the estimation of the projection matrix U in the F-step is executed following the orthonormal discriminant vector (ODV) procedure. The main purpose of the ODV process is to build a set of column vectors which are both orthogonal and discriminative. However, the resulting set of vectors is not guaranteed to be optimal [9] and such an estimation remains numerically unstable because of the use of a Gram-Schmidt process. We therefore propose in this paper two different ways to efficiently solve the optimization problem (4).

3.1. Fisher's criterion as a regression criterion

In this first approach, we propose to reformulate the optimization problem as a regression-type problem in an unsupervised context by leaning on a result of Qiao *et al.* [18].

3.1.1. Related work in the supervised context

Fisher discriminant analysis [5, 7] is a supervised dimension reduction method looking for a linear transformation U which projects the observations in a discriminative and low dimensional subspace of dimension d. The $p \times d$ matrix U is chosen such as it 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 such that:

$$\hat{U} = \arg\max_{U} \quad \text{trace}\left((U^t S_W U)^{-1} U^t S_B U\right), \tag{11}$$

where $S_B = 1/n \sum_{k=1}^{K} n_k (m_k - \bar{y})(m_k - \bar{y})^t$ and m_k denotes the mean vector of the class k; $S_W = 1/n \sum_{k=1}^{K} n_k C_k$ where C_k and n_k stand respectively for the covariance matrix and the number of observations of the class k. The classical solution of this optimization problem is the eigenvectors associated to the d largest eigenvalues of the matrix $S_W^{-1} S_B$. Qiao *et al.* [18] have reformulated the optimization problem expressed in (11) as a ridge regression-type problem. The following theorem introduces such a reformulation as it was originally defined by Qiao *et al.*. Let us first consider the matrices H_W and H_B, defined by:

$$H_W = \frac{1}{n}\left[Y_1 - m_1 \mathbf{1}_{n_1}^t, \ldots, Y_K - m_K \mathbf{1}_{n_K}^t\right] \in \mathbb{R}^{p \times n}, \tag{12}$$

$$H_B = \frac{1}{n}\left[\sqrt{n_1}\,(m_1 - \bar{y}), \ldots, \sqrt{n_K}\,(m_K - \bar{y})\right] \in \mathbb{R}^{p \times K}, \tag{13}$$

such that $H_W H_W^t = S_W$ and $H_B H_B^t = S_B$. Then, Qiao *et al.* state the following theorem in the supervised classification context:

Theorem 1. *Consider the Cholesky decomposition of the within covariance matrix $S_W = R_W^t R_W$ where $R_W \in \mathbb{R}^{p\times p}$ is a upper triangular matrix. Let $H_B \in \mathbb{R}^{p\times K}$ be defined as in equation* (13). *Let $U_1,\ldots,U_d$ be d column vectors of dimension p and denote the eigenvectors linked to the $d \leq \min(p,K-1)$ largest values of the eigen decomposition of $S_W^{-1}S_B$. Let consider the $p\times d$ matrices $A = [\alpha_1,\ldots,\alpha_d]$ and $B = [\beta_1,\ldots,\beta_d]$. For $\rho > 0$, let $\hat{A}$ and $\hat{B}$ be the solutions of the following problem:*

$$\min_{A,V} \sum_{k=1}^{K} \left\| R_W^{-t} H_{B,k} - AB^t H_{B,k} \right\|_F^2 + \rho \sum_{j=1}^{d} \beta_j^t S_W \beta_j \text{ w.r.t. } A^tA = \mathbf{I}_d, \tag{14}$$

where $H_{B,k} = \sqrt{n_k/n}\,(m_k - \bar{y})$ is the kth column of H_B, $\|.\|_F$ stands for the Frobenius norm and ρ is a constant and positive term. Then:

$$\hat{A} = EP, \tag{15}$$

with P is an arbitrary $d\times d$ orthogonal matrix and E, respectively Λ, denotes the matrix containing the eigenvectors, respectively the eigenvalues, of $R_W^{-t}S_BR_W^{-1}$ satisfying $R_W^{-t}S_BR_W^{-1} = E\Lambda E^t$. The optimal loadings matrix $\hat{B}$ is therefore:

$$\hat{B} = R_W^{-1}E(\Lambda + \rho\mathbf{I})^{-1}\Lambda P,$$

which implies that the d column vectors of the fitted matrix $\hat{B}$ span the same linear subspace as the column vectors of U, solution of the eigen-decomposition problem.

3.1.2. Reformulation in the unsupervised context

We now propose to reformulate the eigen-decomposition problem associated with the estimation of the discriminative latent space in the F-step as a regression-type problem. To that end, we lean on the Qiao's result [18] defined previously and adapt their result to the unsupervised classification context. In their work, the matrices H_W and H_B are defined according to the class membership and this is not possible in our case since they are not observed. An additional problem occurs in our case: the DLM model assumes that the discriminative latent subspace has an orthonormal basis and this constraint is not taken into account in the Qiao's work.

Let us introduce the soft matrices $H_W^{(q)}$ and $H_B^{(q)}$ which are computed at each iteration q of the F-step and conditionally to the E-step:

Definition 1. *The soft matrices $H_W^{(q)} \in \mathbb{R}^{p\times n}$ and $H_B^{(q)} \in \mathbb{R}^{p\times K}$ are defined, conditionally to the posterior probabilities $t_{ik}^{(q)}$ computed in the E-step at iteration q, as follows:*

$$H_W^{(q)} = \frac{1}{n}\left[Y - \sum_{k=1}^{K} t_{1k}^{(q)} m_k^{(q)}, \ldots, Y - \sum_{k=1}^{K} t_{nk}^{(q)} m_k^{(q)} \right] \in \mathbb{R}^{p\times n} \tag{16}$$

$$H_B^{(q)} = \frac{1}{n}\left[\sqrt{n_1^{(q)}}(m_1^{(q)} - \bar{y}), \ldots, \sqrt{n_K^{(q)}}(m_K^{(q)} - \bar{y}) \right] \in \mathbb{R}^{p\times K}, \tag{17}$$

where $n_k^{(q)} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{(q)}$ and $m_k^{(q)} = \frac{1}{n}\sum_{i=1}^{n} t_{ik}^{(q)} y_i$ is the soft mean vector of the cluster k.

According to these definitions, the matrices satisfy the conditions:

$$H_W^{(q)} H_W^{(q)t} = S_W^{(q)} \quad \text{and} \quad H_B^{(q)} H_B^{(q)t} = S_B^{(q)}, \tag{18}$$

where $S_W^{(q)} = 1/n \sum_{k=1}^{K} n_k^{(q)} C_k$ stands for the soft within covariance matrix computed at iteration q and $S_B^{(q)}$ denotes the soft between covariance matrix defined in equation (5). According to the Qiao's theorem [18], the optimization problem (4) can be alternatively solved, at iteration q, by considering:

$$\begin{aligned} (\hat{A}^{(q)}, \hat{B}^{(q)}) \quad &= \quad \arg\min_{A^{(q)} B^{(q)}} \sum_{k=1}^{K} \left\| R_W^{(q)-t} H_{B,k}^{(q)} - A^{(q)} B^{(q)t} H_{B,k}^{(q)} \right\|_F^2 + \rho \sum_{j=1}^{d} \beta_j^{(q)t} S_W^{(q)} \beta_j^{(q)} \\ &\text{w.r.t.} \quad A^{(q)t} A^{(q)} = \mathbf{I}_d, \end{aligned} \tag{19}$$

where $S_W^{(q)} = R_W^{(q)t} R_W^{(q)}$ with $R_W^{(q)} \in \mathbb{R}^{p \times p}$ is an upper triangular matrix, $H_{B,k}^{(q)}$ is the kth column of the matrix $H_B^{(q)}$ defined from Equation (18), ρ is an hyper parameter to calibrate and finally, the norm $\|.\|_F$ denotes the Frobenius norm. By letting $\hat{B}^{(q)} = [\hat{\beta}_1^{(q)}, \ldots, \hat{\beta}_d^{(q)}]$ and according to the Qiao's results, the column vectors of the matrix $\hat{B} \in \mathbb{R}^{p \times d}$ span the same linear space as those of the projection matrix U.

However, the orthogonality constraint on the column vectors of the matrix U spanning the Fisher space is not guaranteed. To that end, we use a well-known result formulated in [8] which concerns the best approximation of a given matrix by an orthogonal matrix. In particular, it is stated that: *Obtaining the best approximation of a matrix $X \in \mathbb{R}^{d \times p}$ by an orthonormal matrix with the same dimensionality is equivalent to an orthogonal Procrustes problem:* $\min\{\|X - Q\|_F : Q^t Q = \mathbf{I}_p\}$, *then $Q = uv^t$ is the solution of such a problem where u and v are respectively the left and right singular vectors of the SVD of X*. In our case, this result becomes:

Proposition 1. *By considering $\hat{A}^{(q)}$ and $\hat{B}^{(q)}$ solutions of Problem* (19), *the best approximation of the projection matrix $U^{(q)}$ by an orthonormal matrix is solution of the following problem:*

$$\begin{aligned} \hat{U}^{(q)} \quad &= \quad \arg\min_{\mathscr{U}^{(q)}} \left\| \hat{B}^{(q)} - \mathscr{U}^{(q)} \right\|_F \\ &\textit{w.r.t.} \quad \mathscr{U}^{(q)t} \mathscr{U}^{(q)} = \mathbf{I_d}, \end{aligned}$$

where $\|.\|_F$ refers to the Frobenius norm. By considering the SVD of $\hat{B}^{(q)} = u'^{(q)} \Lambda'^{(q)} v'^{(q)t}$, then $\hat{U}^{(q)} = u'^{(q)} v'^{(q)t}$.

Proof. At iteration q, in the F-step and conditionally to the E-step, the following optimization problem is considered:

$$\begin{aligned} \left(\hat{A}^{(q)}, \hat{B}^{(q)}\right) \quad &= \quad \arg\min_{A^{(q)}, B^{(q)}} \sum_{k=1}^{K} \left\| \left(R_W^{(q)t}\right)^{-1} H_{B,k}^{(q)t} - A^{(q)} B^{(q)t} H_{B,k}^{(q)t} \right\| + \rho \sum_{j=1}^{d} \beta_j^{(q)t} S_W^{(q)} \beta_j^{(q)} \\ &\text{w.r.t.} \quad A^{(q)t} A^{(q)} = \mathbf{I}_d \end{aligned}$$

and is solved from the Qiao's theorem developed in Section 1 of [18]. Therefore, the column vectors of $\hat{B}^{(q)}$ span the same space as the solution of the eigendecomposition of $S_W^{(q)-1} S_B^{(q)}$ and

the estimation of $\hat{A}^{(q)}$ is obtained by equation (15). Moreover, as we search the best approximation of the matrix $\hat{B}^{(q)}$ to an orthogonal matrix, then the optimization problem is equivalent to the following one:

$$\begin{aligned} \hat{U}^{(q)} &= \arg\min_{\mathscr{U}^{(q)}} \left\| \hat{B}^{(q)} - \mathscr{U}^{(q)} \right\|_F \\ &\text{w.r.t.} \quad \mathscr{U}^{(q)t}\mathscr{U}^{(q)} = \mathbf{I_d}, \end{aligned}$$

where $\|.\|_F$ refers to the Frobenius norm. This problem is a nearest orthogonal Procrustes problem which can be solved by a singular value decomposition [8, 10]. The singular value decomposition of $\hat{B}^{(q)} = u'^{(q)}\Lambda'^{(q)}v'^{(q)t}$ allows to write $\hat{U}^{(q)} = u'^{(q)}v'^{(q)t}$. According to Qiao's theorem, since $\hat{B}^{(q)}$ spans the same subspace as those obtained by the standard Fisher's criterion and according to the nearest Procrustes problem, $\hat{U}^{(q)}$ is the orthogonal matrix which best approximates the projection matrix U whose column vectors span the orthogonal and discriminative latent subspace. □

3.1.3. Algorithm

From an algorithmic point of view, the optimization problem can be solved by alternatively optimizing over $B^{(q)}$ with $A^{(q)}$ fixed and over $A^{(q)}$ with $B^{(q)}$ fixed. This leads to the following algorithm:

Algorithm 1

1. At iteration q, compute the matrices $H_B^{(q)}$ and $H_W^{(q)}$ from Equations (17) and (16). Let $S_W^{(q)} = H_W^{(q)}H_W^{(q)t}$ and $S_B^{(q)} = H_B^{(q)}H_B^{(q)t}$.
2. Compute $R_W^{(q)}$ by using a Cholesky decomposition such that $R_W^{(q)}R_W^{(q)t} = H_W^{(q)}H_W^{(q)t}$.
3. Initialization:
   Let $B^{(q)} = Q$ the eigenvectors of $S^{-1}S_B^{(q)}$.
   Compute the SVD $R_W^{(q)-t}S_B^{(q)}B^{(q)} = u^{(q)}d^{(q)}v^{(q)t}$ and let $A^{(q)} = u^{(q)}v^{(q)t}$.
4. Solve d independent regression problems:

$$\hat{\beta}_j^{(q)} = \arg\min_{\beta_j} \left(\beta_j^{(q)}W^{(q)t}W^{(q)}\beta_j^{(q)t} - 2\tilde{Y}^{(q)t}W^{(q)}\beta_j^{(q)} \right),$$

   where $W^{(q)} = \begin{pmatrix} H_B^{(q)t} \\ \sqrt{\rho}R_W \end{pmatrix}$ and $\tilde{Y}^{(q)} = \begin{pmatrix} H_B^{(q)t}R_W^{(q)-1}\alpha_j^{(q)} \\ \mathbf{O}_p \end{pmatrix}$.
5. Let $\hat{B}^{(q)} = [\hat{\beta}_1, \ldots, \hat{\beta}_d]$. Compute the SVD $R_W^{(q)-t}S_B^{(q)}\hat{B}^{(q)} = u^{(q)}d^{(q)}v^{(q)t}$ and let $A^{(q)} = u^{(q)}v^{(q)t}$.
6. Repeat steps 3 and 4 several times until convergence.
7. Compute the SVD of $\hat{B}^{(q)} = u'^{(q)}\Lambda'^{(q)}v'^{(q)t}$ and let $\hat{U}^{(q)} = u'^{(q)}v'^{(q)t}$.

3.2. A modified Fisher criterion

In this second approach, we propose a modified Fisher's criterion which aims to efficiently approximate the discriminative latent subspace.

3.2.1. Optimization problem

Instead of considering the constrained Fisher's criterion (4) considered in the original algorithm, we look here for a $p \times d$ projection matrix U with orthogonal columns such as the associated latent subspace has a discrimination power as close as possible than the one of the whole observation space, *i.e.* such that the matrix $UU^tS^{-1}S_B^{(q)}$ best approximates the matrix $S^{-1}S_B^{(q)}$. This leads to consider the following optimization problem, in the F-step of the Fisher-EM algorithm, at iteration q:

$$\begin{aligned} \hat{U}^{(q)} &= \arg\min_U \left\| S^{-1}S_B^{(q)} - UU^tS^{-1}S_B^{(q)} \right\|_F \\ &\text{w.r.t.} \quad U^tU = \mathbf{I_d}, \end{aligned} \tag{20}$$

where U is a $p \times d$ orthogonal projection matrix, S stands for the covariance matrix of the input data, $S_B^{(q)}$ is the fuzzy between covariance matrix computed at iteration q and $\|.\|_F$ is the Frobenius norm. The solution of this new optimization problem is given by the following propostion:

Proposition 2. *At iteration q, the best approximation of the matrix $S^{-1}S_B^{(q)}$ onto an orthogonal subspace through a $p \times d$ projection matrix $(d < K-1)$ is the solution of the following optimization problem:*

$$\begin{aligned} \hat{U}^{(q)} &= \arg\max_U \quad \text{trace}\left(U^t(S^{-1}S_B^{(q)})(S^{-1}S_B^{(q)})^tU\right), \\ &\textit{w.r.t.} \quad U^tU = \mathbf{I}_d. \end{aligned} \tag{21}$$

and the columns of $\hat{U}$ are the d first left eigenvectors of the singular value decomposition of $S^{-1}S_B^{(q)}$.

Proof. In order to ease the reading of the proof, the index q is omitted. Let us first notice that:

$$\begin{aligned} \left\|S^{-1}S_B - UU^tS^{-1}S_B\right\|_F^2 &= \text{trace}\left((S^{-1}S_B - UU^tS^{-1}S_B)^t(S^{-1}S_B - UU^tS^{-1}S_B)\right), \\ &= -2\text{trace}((S^{-1}S_B)^tUU^tS^{-1}S_B) + \text{trace}((S^{-1}S_B)^tS^{-1}S_B) \\ &\quad +\text{trace}((S^{-1}S_B)^tUU^tUU^tS^{-1}S_B), \\ &= \left\|S^{-1}S_B\right\|_F^2 - \text{trace}((S^{-1}S_B)^tUU^tUU^tS^{-1}S_B), \\ &= \left\|S^{-1}S_B\right\|_F^2 - \left\|UU^tS^{-1}S_B\right\|_F^2. \end{aligned}$$

It implies that minimizing the quantity $\left\|S^{-1}S_B - UU^tS^{-1}S_B\right\|_F^2$ is equivalent to maximize $\left\|UU^tS^{-1}S_B\right\|_F^2$. Furthermore, since $U^tU = \mathbf{I}_d$, the following equalities hold:

$$\begin{aligned} \left\|UU^tS^{-1}S_B\right\|_F^2 &= \text{trace}\left((UU^tS^{-1}S_B)(UU^tS^{-1}S_B)^t\right) \\ &= \text{trace}\left(U^t(S^{-1}S_B)(S^{-1}S_B)^tU(U^tU)\right) \\ &= \text{trace}\left(U^t(S^{-1}S_B)(S^{-1}S_B)^tU\right). \end{aligned}$$

where $\|.\|_F$ denotes the Frobenius norm.

Let us now consider the SVD of the $n \times p$ matrix $S^{-1}S_B^{(q)} = u\Lambda v^t$ where u and v stands for respectively the left and right singular vectors of $S^{-1}S_B^{(q)}$ and Λ is a diagonal matrix containing its associated singular values. As the matrix $S_B^{(q)}$ has a rank d at most equal to $K-1 < p$, with K the number of clusters, then the matrix $S^{-1}S_B^{(q)}$ is also of rank $d = \text{rank}(S^{-1}S_B^{(q)})$ at most equal to $K-1 < p$. Consequently, only the d singular values of the matrix $S^{-1}S_B^{(q)}$ are non zeros, which enables us to write $S^{-1}S_B^{(q)} = u\Lambda_d v^t$, where $\Lambda_d = \text{diag}(\lambda_1, \ldots, \lambda_d, 0, \ldots, 0)$. Moreover, by letting $U = u_d$ the d first left eigenvectors of $S^{-1}S_B$, then:

$$\begin{aligned} \text{trace}\left(U^t(S^{-1}S_B)(S^{-1}S_B)^t U\right) &= \text{trace}\left(U^t(u\Lambda_d v^t)(u\Lambda_d v^t)^t U\right), \\ &= \text{trace}\left(U^t u\Lambda_d \Lambda_d^t u^t U\right), \\ &= \sum_{j=1}^{d} \lambda_j^2. \end{aligned}$$

Consequently, the $p \times d$ orthogonal matrix U such that $\left\|S^{-1}S_B - UU^tS^{-1}S_B\right\|_F^2$ is minimal is the matrix made of the d first left eigenvectors of $S^{-1}S_B$. □

Besides, let us consider $\tilde{U}$, the solution of the optimization problem (21). We can notice that, without loss of generality, in the case $S = \mathbf{I}_p$, the proposed modified Fisher's criterion becomes:

$$\begin{aligned} \left\|UU^tS^{-1}S_B\right\|_F^2 &= \text{trace}\left(U^tS_BS_B^tU\right), \\ &= \text{trace}\left(U^tS_B^2U\right). \end{aligned}$$

In this case and according to Proposition 2, $\tilde{U}$ stands for the right left eigenvectors of the SVD of S_B^2. Since S_B is symmetric and semi-definite positive, the matrix $\tilde{U}$ contains the eigenvectors associated with the d largest eigenvalues of S_B^2 and also to the ones of S_B. Therefore, in this case, $\tilde{U}$ is as well solution of the original optimization problem (4).

3.2.2. Algorithm

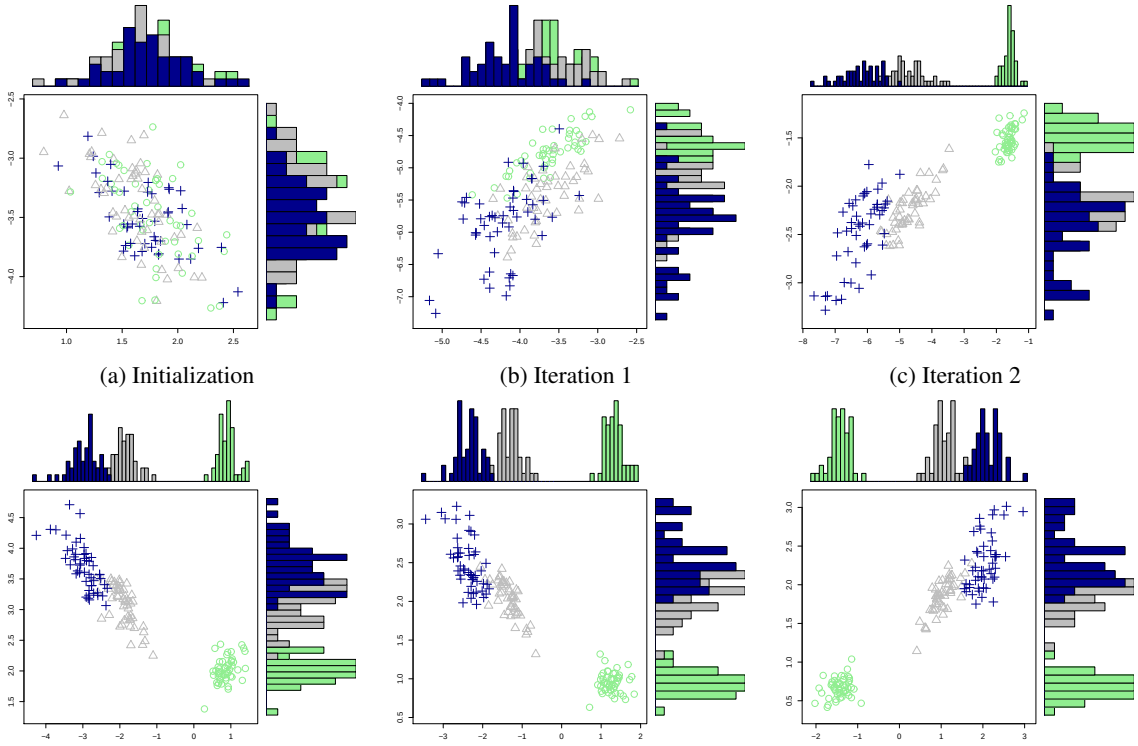
Algorithmically saying, one only needs to decompose by a singular value decomposition the matrix $S^{-1}S_B^{(q)}$ at iteration q. The projection matrix whose its columns span the discriminative latent subspace is fitted by the d first left singular vectors of $S^{-1}S_B^{(q)}$.

Algorithm 2

1. At iteration q, compute the matrix $S_B^{(q)}$ defined in Equation (5) and the covariance matrix S of the data.
2. Compute the singular value decomposition of $S^{-1}S_B^{(q)} = u^{(q)}\Delta_d^{(q)}v^{(q)t}$ with $d = \text{rank}(S^{-1}S_B^{(q)})$.
3. Let $\hat{U}_d^{(q)} = u_d^{(q)}$ where $u_d^{(q)}$ stands for the d first left eigenvectors of $u^{(q)}$.

4. Experiments

The following experiments allows to compare the Gram-Schmidt orthogonalization procedure originally used in [2] with the two alternatives proposed in this paper. For all the experiments, we



(a) Initialization (b) Iteration 1 (c) Iteration 2

(d) Iteration 3 (e) Iteration 4 (f) Iteration 5

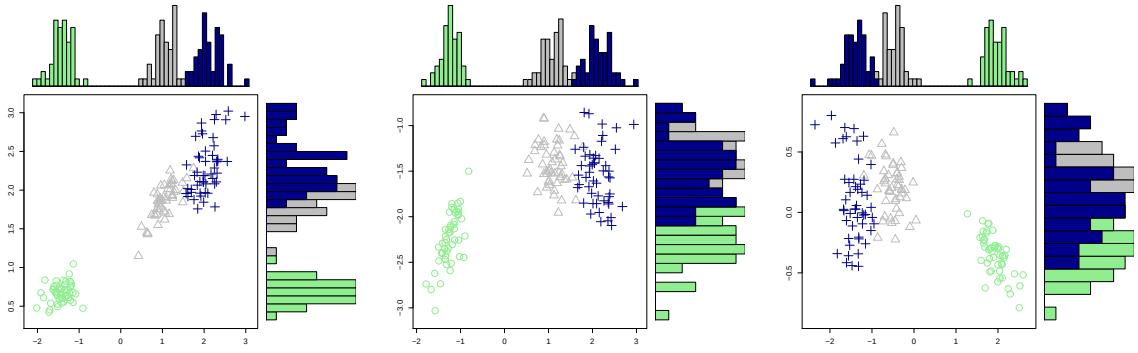
FIGURE 2. *Evolution of the fitted discriminative space according to the original Fisher-EM algorithm.*

refer to by FisherEM-GS the original procedure, by FisherEM-REG the algorithm associated with the regression criterion and finally by FisherEM-SVD the one corresponding to the approximation of the Fisher's criterion.

4.1. An introductory example: the Fisher's irises

This introductory example aims to highlight the main asset of the Fisher-EM algorithm based on the visualization of clustered data in a low and discriminative subspace. We first apply the original FisherEM algorithm to the iris dataset that Fisher used in [5] which is made of 3 groups corresponding to different species of iris (*setosa*, *versicolor* and *virginica*) among which the groups *versicolor* and *virginica* are difficult to discriminate. The dataset consists of 50 samples from each of 3 species and four features were measured from each sample. The four measurements are the length and the width of the sepal and the petal. As the FisherEM algorithm is an unsupervised procedure, the labels have been used only for performance evaluation and not for building the discriminative axes. The results have been obtained with a random initialization on the $\text{DLM}_{[\alpha_k\beta]}$ model where the number of classes has been fixed to 3. For this experiment, the clustering accuracy has reached 98% with FisherEM-GS.

Figure 2 shows, at each iteration, the estimated projection and the clustering of the the data with the Fisher-EM algorithm and, on each axis, the corresponding empirical group densities have



(a) FisherEM-GS (b) FisherEM-SVD (c) FisherEM-REG

FIGURE 3. *Projection of clustered Fisher's irises in the discriminative space fitted respectively by (a) FisherEM-GS, (b) FisherEM-SVD and (c) FisherEM-REG procedures.*

been drawn. As indicated by Panel (a), the initialization partition was randomly drawn. From the second iteration until the third one, the discrimination between 2 groups begins. From the fourth iteration (see Panel (d)), a structure of 3 different classes appears and the group densities begin to separate distinctly. In particular, the first Fisher's axis well-discriminates the 3 clusters. Finally, the last iterations enable to refine the estimations of the means and the covariance matrices of 3 clusters until convergence. Figure 3 illustrates the final partition obtained at convergence with the FisherEM-GS, FisherEM-SVD and FisherEM-REG procedures from the same random initialization on the $\text{DLM}_{[\alpha_k\beta]}$ model. As we can observe, the visualization of the clustered irises obtained by the 3 algorithms are relatively similar. As expected, the clustering accuracy of each method also remains very similar since the clustering accuracy has reached 98% for the FisherEM-GS procedure and 97.3% for the FisherEM-SVD and FisherEM-REG algorithms.

However, it appears that, for some datasets and in particular for high-dimensional data, the FisherEM-GS could fail in the visualization of clustered data while the clustering task remains performing. The experiment which will be presented in Paragraph 4.3 illustrates such a limitation and highlights the interest of the FisherEM-SVD and FisherEM-REG algorithms.

4.2. Comparison between the 3 F-steps

This second experiment aims to compare on simulations the three estimation procedures of the projection matrix in the F-step of the Fisher-EM algorithm. The three procedures will be compared on the basis of the produced latent subspace, the classification performance and the computing time.

On the one hand, we compare the produced latent subspaces and to that end, we consider the supervised context with $d < K-1$. For this comparison, 750 observations have been simulated following the $\text{DLM}_{[\Sigma\beta]}$ model with the parameters $\Sigma = 2\mathbf{I}_d$, $d = 8$ and $\beta = 15$. The difference between clusters happens to be entirely on the means vectors. The simulated dataset is made of 15 groups of 50 observations and each group is modeled by a Gaussian density in a 8-dimensional space completed by 7 orthogonal dimensions of Gaussian noise. The transformation matrix W has been randomly simulated such as $W^tW = WW^t = I_p$ and, for this experience, the dimension of

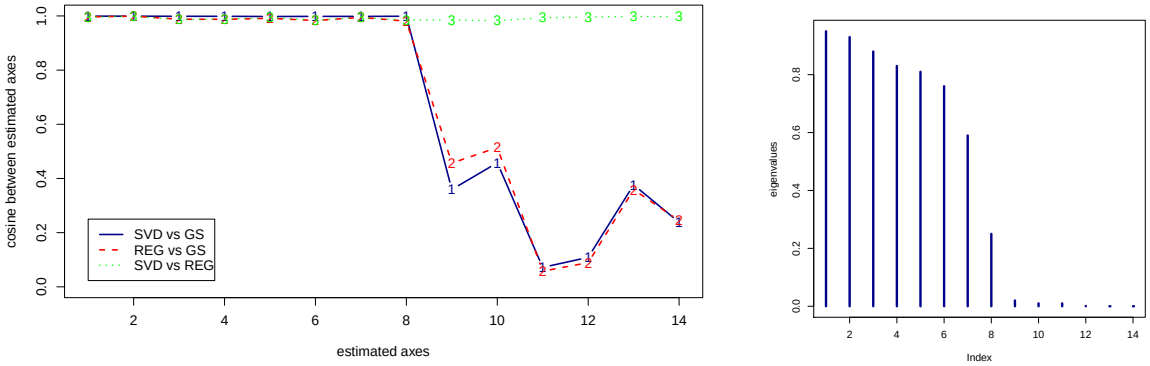
the observed space is fixed to 30. In order to ease the comparison, we consider in this experiment the supervised context. The true labels are used to initialize the Fisher-EM algorithm which is consequently iterated only once, i.e. only one F-step and M-step are considered before re-classifying the data with an E-step.

We first focus on the estimation of the discriminative latent subspace. Since the intrinsic dimension of the latent subspace is theoretically at most equal to $d = K - 1$, then the cosines of 14 potential discriminative axes have been computed. Figure 4.2 stands for cosine values computed between the discriminative axes estimated by the 3 procedures. Figure 4.2 illustrates the scree plot of the eigenvalues associated to the eigen-decomposition of the matrix $s^{-1}s_B$, where s and s_B stands for respectively the empirical covariance and the between covariance matrices in the fitted latent subspace. First of all, in Figure 4.2, it can be observed that the cosines computed on the 8 first are close to 1, whatever the procedures are, which implies that the 3 procedures seems to estimate the same discriminative subspace. Nevertheless, from the 8th axis, we can observe a gap between the axes estimated by the FisherEM-GS and those estimated by FisherEM-SVD or by FisherEM-REG. However, these last axes are no significance since they have no discriminative power. Indeed, since $\text{rank}(s_B) = 8$, we know the intrinsic dimension of the latent space is $d = 8$. This is confirmed by Figure 4.2 which shows that the discriminative power of the estimated axes is almost equal to 0 after the 8th dimension. Consequently, since the main difference between the 3 procedures remains in the axes which have no discriminative power, the 3 procedures used in the F-step can be considered as equivalent for estimating the latent subspace.

On the other hand, we compare the clustering accuracy and the computational of the 3 algorithms. To do this, we consider a traditional clustering situation from which the data are high-dimensional since the dimension of the input space is $p = 100$. For this simulation, 600 observations have been simulated following the $\text{DLM}_{[\alpha_{kj}\beta_k]}$ model and they are made of 3 balanced groups for which each group is modeled by a Gaussian density in a 2-dimensional space completed by orthogonal dimensions of Gaussian noise. The transformation matrix W has been randomly simulated such as $W^tW = WW^t = I_p$. The experimental process has been repeated 25 times for each dimension of the observed space in order to see both the average performances and the variances of the 3 algorithms. Regarding the classification performance, Figure 5 stands for the boxplots of clustering accuracy rate obtained on 25 trials for each procedure. The average correct classification rates of both the FisherEM-SVD (94.5%) and FisherEM-REG (92.2%) procedures are better compared to those obtained by the original algorithm (92.6%). In the same manner, one can also notice that FisherEM-SVD seems more stable than the nominal procedure whereas the FisherEM-REG is slightly less.

Finally, the elapsed real time for each procedure has been computed and Table 1 presents the computation time for the 3 procedures (FisherEM-GS, FisherEM-SVD, FisherEM-REG). We can observe that the F-step computed by SVD is faster than those obtained by the two other procedures. To summarize, the performances between these 3 procedures are comparable in terms of estimation of the latent subsapce and classification performance but the SVD procedure remains the quickest one.

Soumis au Journal de la Société Française de Statistique
File: essai.tex, compiled with jsfds, version : 2009/12/09
date: October 10, 2011



(a) Cosine between fitted axes. (b) Scree plot

FIGURE 4. *Evolution of the cosine between estimated axes according to the methodology used in the F-step : FisherEM-SVD, FisherEM-GS or FisherEM-REG procedures (a) and scree plot of the eigenvalues of the matrix $s^{-1}s_B$ in the latent space (b).*

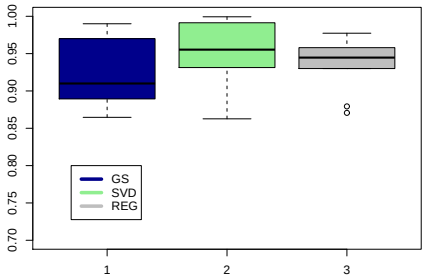


FIGURE 5. *Boxplots of correct classification rates obtained on 25 replications for the 3 different procedures for the F-step (FisherEM-GS, FisherEM-SVD and FisherEM-REG).*

Procedures:	Computational time
FisherEM-GS	111.3 ± 8.2
FisherEM-SVD	107.5 ± 11.7
FisherEM-REG	109.8 ± 8.8

TABLE 1. *Elapsed real time and CPU time computed for the 3 procedures of the F-step in the Fisher-EM algorithm.*

FIGURE 6. *Samples from the USPS358 dataset.*

4.3. The USPS358 dataset

In this last experiment, the Fisher-EM algorithm is executed on a high-dimensional real-world dataset. The data come from a sample of the USPS handwritten image data [11] collected by the Center of Excellence in Document Analysis and Recognition (CEDAR) at SUNY Buffalo. The overall dataset consists of 7291 digital numbers from $0, 1, 2, \ldots, 9$ scanned and stretched in a rectangular box 16×16 in a gray scale of 256 values from around 80 persons. In this experiment, only the classes which are difficult to discriminate are considered. Consequently, the studied dataset consists of 1756 records (rows) and 256 attributes (columns) divided in three classes: the numbers 3, 5 and 8. Figure 6 presents a sample from the USPS358. For this example, the $\text{DLM}_{[\alpha_{kj}\beta_k]}$ model is used, from which the Fisher-EM algorithm originally proposed by its authors is executed (FisherEM-GS). Figures 7 stand for the corresponding group means obtained from the group memberships estimated in the USPS358. Besides, the FisherEM-SVD and FisherEM-REG algorithms have also been executed from the same random initialization.

Figure 8 presents the projections of the USPS358 dataset into the latent discriminative subspace estimated by the 3 procedures. As previously, the empirical density of fitted clusters is in addition drawn on each axis. First of all, we can observe that the visualization of the group structure is really improved in the case of the FisherEM-SVD and FisherEM-REG algorithms compared to the original procedure. Indeed, the FisherEM-GS procedure can barely differentiate two different classes whereas the discrimination between the three groups is clear in the cases of the FisherEM-SVD and FisherEM-REG. The poor performance of FisherEM-GS is linked to the estimation of the second discriminative axis. Indeed, whereas the first axis enables to distinct the cluster 3 from the clusters 5 and 8 for the three procedures, the second axis discriminates the groups 5 and 8 from the group 3. It seems that only FisherEM-SVD and FisherEM-REG succeeded in estimating this second discriminative axis for this high-dimensional dataset. This difference is also illustrated in Figure 9 which stands for the loadings (in absolute value) of the 2 discriminative axes fitted respectively by the FisherEM-GS, FisherEM-SVD and FisherEM-REG algorithms.

Finally, even though the visualization of the clustered data is less satisfying for FisherEM-GS than for its two alternatives, the clustering accuracy remains nevertheless almost similar for FisherEM-GS (81.3%), FisherEM-SVD (82.2%) and FisherEM-REG (82.4%).

5. Conclusion

This work has presented two alternatives for estimating the projection matrix whose column vectors span the discriminative latent subspace of the DLM model. On the one hand, we have recasted the optimization problem originally defined in [2] as a ridge regression problem. On the other hand, we have proposed a modified Fisher's criterion in order to find the orthogonal discriminative space which best approximates the solution of the original optimization problem.

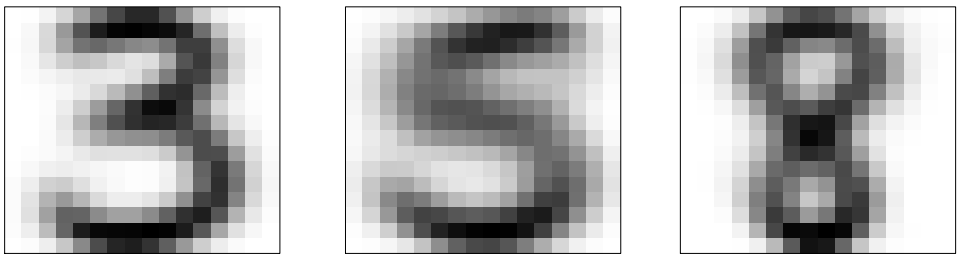
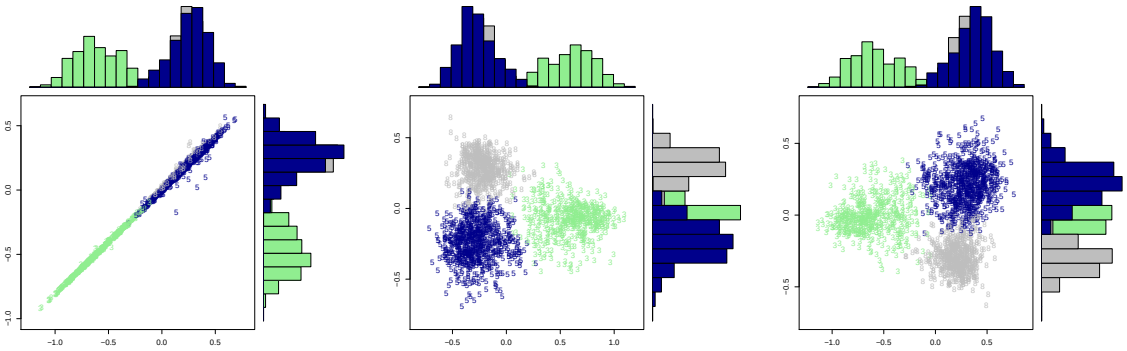
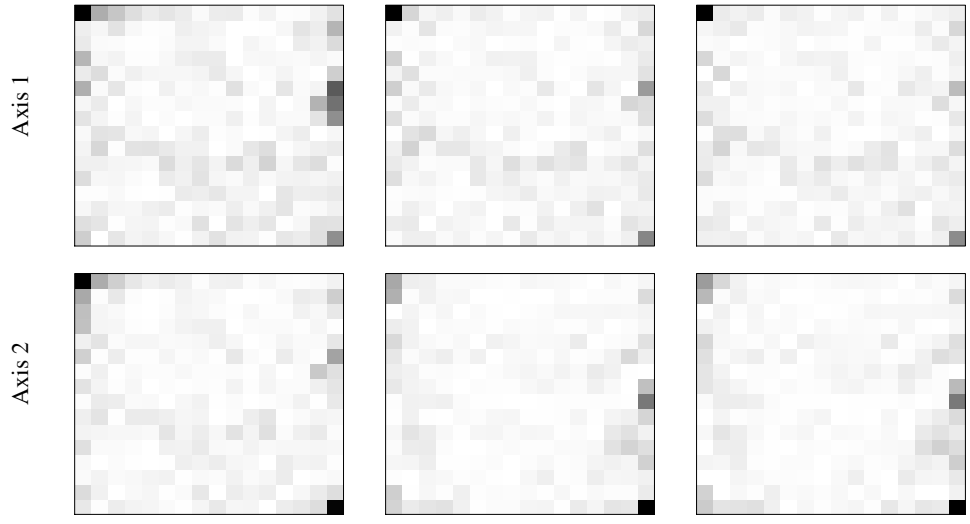


FIGURE 7. *Group means estimated by the original Fisher-EM algorithm (FisherEM-GS) which corresponds to* 81.3% *of miclassification rate.*



(a) FisherEM-GS (b) FisherEM-SVD (c) FisherEM-REG

FIGURE 8. *Projection of the usps358 in their discriminative latent space fitted respectively by (a) the FisherEM-GS, (b) the FisherEM-SVD and (c) the FisherEM-REG procedures.*



(a) FisherEM-GS (b) FisherEM-SVD (c) FisherEM-REG

FIGURE 9. *Absolute value of loadings of two discriminative axes estimated by (a) the FisherEM-GS, (b) the FisherEM-SVD and (c) the FisherEM-REG procedures.*

Soumis au Journal de la Société Française de Statistique
File: essai.tex, compiled with jsfds, version : 2009/12/09
date: October 10, 2011

Experiments on real-world datasets showed that the proposed alternatives enable to improve the visualization of clustered data particularly in the case of high-dimensional data while reducing the computing time.

The reformulation of the optimization problem into a regression-type one enables to extend this work in a sparse case. Indeed, the addition of an ℓ_1-penalty term into the regression problem could introduce sparsity into the loadings of the projection matrix and consequently could stress discriminative variables. According to a recent work of [20] based on a penalized matrix decomposition, we could also consider to penalize the modified Fisher's criterion in order to select discriminative variables.

References

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

[2] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, to appear:1–24, 2011.

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

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

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

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

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

[8] G. Golub and C. Van Loan. *Matrix Computations. Second ed.* The Johns Hopkins University Press, Baltimore, 1991.

[9] 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.

[10] N.J. Higham. *Matrix nearness problems and its applications*, chapter 1, pages 1–27. Oxford University Press, 1989.

[11] J. J. Hull. A database for handwritten text recognition research. *IEEE PAMI*, 16(5):550–554, 1994.

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

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

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

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

[16] A. Montanari and C. Viroli. Dimensionally reduced mixtures of regression models. *Electronic Proceedings of KNEMO, Knowledge Extraction and Modelling*, 2006.

[17] A. Montanari and C. Viroli. Heteroscedastic Factor Mixture Analysis. *Statistical Modeling: An International journal (forthcoming)*, (to appear), 2010.

[18] Z. Qiao, L. Zhou, and J.Z. Huang. Sparse linear discriminant analysis with applications to high dimensional low sample size data. *International Journal of Applied Mathematics*, 39(1), 2009.

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

[20] D.M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistic*, 10(3):515–534, 2009.

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

Soumis au Journal de la Société Française de Statistique
File: essai.tex, compiled with jsfds, version : 2009/12/09
date: October 10, 2011