mathematics

MDPI

Article

Multi-Partitions Subspace Clustering

Vincent Vandewalle [1,2]

[1] Biostatistics Department, Univ. Lille, CHU Lille, ULR 2694—METRICS: Évaluation des Technologies de Santé et des Pratiques MéDicales, F-59000 Lille, France; vincent.vandewalle@univ-lille.fr

[2] Inria Lille—Nord Europe, 59650 Villeneuve d'Ascq, France

Received: 28 February 2020; Accepted: 10 April 2020; Published: 15 April 2020

Abstract: In model based clustering, it is often supposed that only one clustering latent variable explains the heterogeneity of the whole dataset. However, in many cases several latent variables could explain the heterogeneity of the data at hand. Finding such class variables could result in a richer interpretation of the data. In the continuous data setting, a multi-partition model based clustering is proposed. It assumes the existence of several latent clustering variables, each one explaining the heterogeneity of the data with respect to some clustering subspace. It allows to simultaneously find the multi-partitions and the related subspaces. Parameters of the model are estimated through an EM algorithm relying on a probabilistic reinterpretation of the factorial discriminant analysis. A model choice strategy relying on the BIC criterion is proposed to select to number of subspaces and the number of clusters by subspace. The obtained results are thus several projections of the data, each one conveying its own clustering of the data. Model's behavior is illustrated on simulated and real data.

Keywords: clustering; mixture model; factorial discriminant analysis; EM algorithm

1. Introduction

In exploratory data analysis, the statistician often uses clustering and visualization in order to improve his knowledge on the data. In visualization he looks for some principal components explaining some major characteristics of the data. For example in principal component analysis (PCA) the goal is to find a linear combination of the variables explaining the major variability of the data. In cluster analysis, the goal is to find some clusters explaining the major heterogeneity of the data. In this article we suppose that the data can contain several clustering latent variables, that is, we are in the multiple partition setting, and we are simultaneously looking for clustering subspaces, that is, linear projections of the data each one related to some clustering latent variable thus the developed model is later called multi-partitions subspace clustering. A solution to perform multi-partition subspace clustering is to use a probabilistic model on the data such as a mixture model [1], it allows to perform the parameters estimation, and model selection such as the choice of the number of subspaces and the number of clusters per subspace using standard model choice criteria such as BIC [2]. Thus the main fields related to our work are model based subspace clustering and multi-partitions clustering.

In the model based subspace clustering framework, let first notice that PCA can be re-interpreted in a probabilistic way by considering a parsimonious version of a multivariate Gaussian distribution [3] and that the k-means algorithm can be re-interpreted as a particular parsimonious Gaussian mixture model estimated using a classification EM algorithm [4]. A re-interpretation of the probabilistic PCA has also been used in clustering by Bouveyron et al. [5] in order to cluster high-dimensional data. Although the proposed high dimensional mixture does not performs dimension reduction, it rather operates a class per class dimension reduction which does not allow to have a global model-based data visualization. Thus Bouveyron and Brunet [6] proposed the so called Fisher-EM algorithm which simultaneously performs clustering and dimension reduction. This is performed through a modified version of the EM algorithm [7] by including a Fisher step between the E and the M step. This approach allows the same projection to be applied to all data, but does not guarantee the increasing of the likelihood at each iteration of the algorithm.

In the context of multi-partitions clustering, Galimberti and Soffritti [8] assumed that the variables can be partitioned into several independent blocks, each one following a full-covariance Gaussian mixture model. The model selection was done by maximizing the BIC criterion by a forward/backward approach. Then, Galimberti et al. [9] generalized thier previous work by relaxing the assumption of block independence. The proposed extension takes into account three types of variables, classifying variables, redundant variables and non-classifying variables. In this context, the choice of the model is difficult because several roles have to be taken into account for each variable, which requires a lot of calculations, even for the reallocation of only one variable. Poon et al. [10] also took into account the multi-partition setting, called as facet determination in their article. The model considered is similar to that of Galimberti and Soffritti [8], but it also allows tree dependency between latent class variables, resulting in the Pouch Latent Tree Models (PLTM). Model selection is performed by a greed search to maximize the BIC criterion. The resulting model allows a broad understanding of the data, but the tree structure search makes estimation even more difficult as the number of variables increases. More recently, Marbac and Vandewalle [11] proposed a tractable muti-partition clustering algorithm not limited to continuous data; in the Gaussian setting it can be seen as particular case of Galimberti and Soffritti [8] where they assume a diagonal covariance matrix allowing a particularly efficient search of the partition of the variables in sub-vectors.

In this article we suppose that the data can contain *several* clustering latent variables, that is we are in the multiple partition setting. But contrary to Marbac and Vandewalle [11] where it is assumed that variables are divided into blocks each one related to some clustering of the data, we are looking for *clustering subspaces*, i.e., linear projections of the data each one related to some particular clustering latent variable thus replacing the combinatorial question of finding the partition of the variables in independent sub-vectors by the question of finding the coefficients of the linear combinations. The proposed approach can be related to the independent factor analysis [12] where the author deals with source separation, in our framework a source can be interpreted as some specific clustering subspace; however, their approach becomes intractable as the numbers of sources increases and does not allow to consider multivariate subspaces. Moreover, it is not invariant up to a rotation and rescaling of the data, where our proposed methodology is.

The organisation of the paper is the following, in Section 2 we present a reinterpretation of the factorial discriminant analysis as a search of discriminant components and of independent non-discriminant components. In Section 3, the multi-partitions subspace clustering model and the EM algorithm to estimate the parameters of the model will be presented. In Section 4, results on simulated and real data will show the interest of the method in practice. In Section 5, a conclusion and discussion of future extension of the paper will be made.

2. Probabilistic Interpretation of the Factorial Discriminant Analysis

2.1. Linear Discriminant Analysis (LDA)

It is supposed that n quantitative data in dimension d are available, the data number i will be denoted by $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{id})^{\mathrm{T}}$, where x_{ij} is the value of variable j of data i. The whole dataset will be denoted by $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$. Let assume that the data is clustered in K clusters, the class label of data i will be denoted by $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iK})^{\mathrm{T}}$, with z_{ik} equals to 1 if data i belongs to cluster k and 0 otherwise. Let also denote by $\mathbf{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)$ the partition of $\mathbf{x}$. In this section $\mathbf{z}$ is supposed to be known. For sake of simplicity the random variables and their realisations will be denoted in lower case, and p will be used as a generic notation to denote a probability distribution function (p.d.f.) which will be interpreted according to its arguments.

In the context of linear discriminant analysis [13], it is supposed that the distribution $\boldsymbol{x}_i$ given the cluster follows a d-variate Gaussian distribution with common covariance matrices:

$$\forall k \in \{1,\ldots,K\}, \;\; \boldsymbol{x}_i | z_{ik} = 1 \sim \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}),$$

with $\boldsymbol{\mu}_k$ the vector of means in cluster k and $\boldsymbol{\Sigma}$ the common class conditional covariance matrix. Let also denote by $\pi_k = p(z_{ik} = 1)$ the prior weights of each cluster.

The posterior cluster membership probabilities can be computed using the Bayes formula:

$$p(z_{ik} = 1 | \boldsymbol{x}_i) = \frac{\pi_k \phi_d(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})}{\sum_{k'=1}^{K} \pi_{k'} \phi_d(\boldsymbol{x}_i; \boldsymbol{\mu}_{k'}, \boldsymbol{\Sigma})}, \tag{1}$$

where $\phi_d(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ stands for the p.d.f. of the d-variable Gaussian distribution with expectation $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}$.

Let $\bar{\boldsymbol{x}}_k$ denote the class conditional mean in cluster k:

$$\bar{\boldsymbol{x}}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} \boldsymbol{x}_i,$$

with $n_k = \sum_{i=1}^{n} z_{ik}$ the number of data in cluster k, and by $\bar{\boldsymbol{x}}$ the unconditional mean. Let also denote by $\boldsymbol{W}$ the empirical intra-class covariance matrix:

$$\boldsymbol{W} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} z_{ik} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k)^{\mathrm{T}},$$

and by $\boldsymbol{B}$ the empirical between class covariance matrix:

$$\boldsymbol{B} = \frac{1}{n} \sum_{k=1}^{K} n_k (\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}})^{\mathrm{T}}.$$

If the data are supposed to be independent, the likelihood can simply be written as:

$$\begin{aligned} \ell(\pi_1,\ldots,\pi_K,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,\boldsymbol{\Sigma};\mathbf{x},\mathbf{z}) &= -\frac{n}{2}\log(\det(\boldsymbol{\Sigma})) - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|^2_{\boldsymbol{\Sigma}^{-1}} \\ &\quad + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\log(\pi_k) - \frac{n}{2}\log(2\pi). \end{aligned}$$

The maximum likelihood estimators of the parameters of π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ are $\hat{\pi}_k = \frac{n_k}{n}$, $\hat{\boldsymbol{\mu}}_k = \bar{\boldsymbol{x}}_k$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{W}$. A new data point can be then classified by plugin the estimated values of the parameters in Equation (1):

$$\hat{z}_i = \underset{k \in \{1,\ldots,K\}}{\operatorname{argmax}} \; \hat{\boldsymbol{\mu}}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{x}_i - \frac{1}{2} \boldsymbol{\mu}_k^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{\mu}_k + \log(\hat{\pi}_k),$$

the resulting classification boundary being linear in this case.

2.2. Factorial Discriminant Analysis (FDA)

Let us note that, from a descriptive viewpoint, one can be interested in dimension reduction in order to visualize the data. This could be done by using PCA, but from a classification perspective the component explaining the largest variability in the data are often not the same that the components providing the best separation between the clusters.

The goal of factorial discriminant analysis (FDA) is to find the component maximizing the variance explained by the cluster above the intra-class variance. The coefficients of the first discriminant component $v_1 \in \mathbb{R}^d$ is defined by

$$v_1 = \arg\max_{v \in \mathbb{R}^d} \frac{\boldsymbol{v}^{\mathrm{T}} \boldsymbol{B} \boldsymbol{v}}{\boldsymbol{v}^{\mathrm{T}} \boldsymbol{W} \boldsymbol{v}}.$$

It is well known that v_1 is the eigen vector associated with the highest eigen value λ_1 of $\boldsymbol{W}^{-1}\boldsymbol{B}$ [14]. The remaining discriminant components are obtained through the remaining eigen vectors of $\boldsymbol{W}^{-1}\boldsymbol{B}$. Let denote by $\lambda_1, \ldots, \lambda_{K-1}$ the eigen values of $\boldsymbol{W}^{-1}\boldsymbol{B}$ sorted in decreasing order and by $v_1, \ldots, v_{K-1}$ the associated eigen vectors. Moreover, if each component is constrained to have an intra-class variance equal to one (i.e., $v_k{}^{\mathrm{T}} \boldsymbol{W} v_k = 1, \forall k \in \{1, \ldots, K-1\}$), the classification obtained using the Mahalonobis distance can simply be obtained by using the Euclidean distance on the data projected on the discriminant components.

2.3. Equivalence between LDA and FDA

As proved in Campbell [15] and detailed in Trevor Hastie [16], the FDA can be interpreted in a probabilistic way as an LDA where the rank of $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ is constrained to be equal to p with $p \leq K-1$ under the common class covariance matrix assumption. This allows us to reparametrize the probabilistic model in the following way:

$$\boldsymbol{x}_i | z_{ik} = 1 \sim \mathcal{N}_d\left(\boldsymbol{A}\begin{pmatrix}\boldsymbol{\nu}_k \\ \boldsymbol{\gamma}\end{pmatrix}, \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}\right),$$

where $\boldsymbol{\nu}_k \in \mathbb{R}^p$, $\boldsymbol{\gamma} \in \mathbb{R}^{d-p}$ and $\boldsymbol{A} \in \mathcal{M}_{d,d}(\mathbb{R})$. Let notice that this new parametrization at this step is not unique but the model can easily be made identifiable by imposing some constraints on the parameters.

Let $\boldsymbol{y}_i \in \mathbb{R}^p$ and $\boldsymbol{u}_i \in \mathbb{R}^{d-p}$ two random variables, the new parametrization can be reinterpreted in the following generative framework:

- Draw $\boldsymbol{z}_i : \boldsymbol{z}_i \sim \mathcal{M}(1; \pi_1, \ldots, \pi_K)$ where $\mathcal{M}$ stands for the multinomial distribution
- Draw $\boldsymbol{y}_i | \boldsymbol{z}_i : \boldsymbol{y}_i | z_{ik} = 1 \sim \mathcal{N}_p(\boldsymbol{\nu}_k, \boldsymbol{I}_p)$
- Draw $\boldsymbol{u}_i : \boldsymbol{u}_i \sim \mathcal{N}_{d-p}(\boldsymbol{\gamma}, \boldsymbol{I}_{d-p})$
- Compute $\boldsymbol{x}_i$ based on $\boldsymbol{y}_i$ and $\boldsymbol{u}_i : \boldsymbol{x}_i = \boldsymbol{A}\begin{pmatrix}\boldsymbol{y}_i \\ \boldsymbol{u}_i\end{pmatrix}$.

Thus the p.d.f. of $\boldsymbol{x}_i$ can be factorized in the following way:

$$p(\boldsymbol{x}_i) = \frac{1}{|\boldsymbol{A}|} p(\boldsymbol{u}_i) p(\boldsymbol{y}_i).$$

From a graphical angle the model can be reinterpreted as in Figure 1, where $\boldsymbol{u}_i$ and $\boldsymbol{y}_i$ are latent random variables.

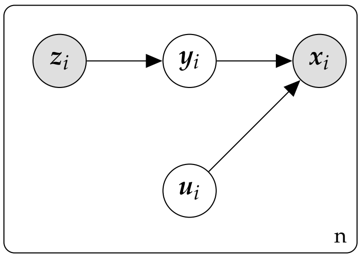

Figure 1. Bayesian dependency graph for factorial discriminant analysis.

In practice we are interested in finding $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$ from $\boldsymbol{x}_i$. We will denote by $\boldsymbol{V} \in \mathcal{M}_{p,d}(\mathbb{R})$ and $\boldsymbol{R} \in \mathcal{M}_{d-p,d}(\mathbb{R})$ the matrices which allow this computation: $\boldsymbol{y}_i = \boldsymbol{V}\boldsymbol{x}_i$, $\boldsymbol{u}_i = \boldsymbol{R}\boldsymbol{x}_i$. It is obvious that $\begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{R} \end{pmatrix} = \boldsymbol{A}^{-1}$, for the rest of the article only the parametrization in terms of $\boldsymbol{V}$ and $\boldsymbol{R}$ will be used.

The main interest of this parametrization is that

$$p(z_{ik} = 1|\boldsymbol{x}_i) = p(z_{ik} = 1|\boldsymbol{y}_i, \boldsymbol{u}_i) = p(z_{ik} = 1|\boldsymbol{y}_i) = p(z_{ik} = 1|\boldsymbol{V}\boldsymbol{x}_i).$$

It means that only $\boldsymbol{V}\boldsymbol{x}_i$ is required to compute the posterior class membership probabilities:

$$p(z_{ik} = 1|\boldsymbol{x}_i) = \frac{\pi_k \phi_p(\boldsymbol{V}\boldsymbol{x}_i; \boldsymbol{\nu}_k, \boldsymbol{I}_p)}{\sum_{k'=1}^{K} \pi_{k'} \phi_p(\boldsymbol{V}\boldsymbol{x}_i; \boldsymbol{\nu}_{k'}, \boldsymbol{I}_p)}. \tag{2}$$

Parameters are estimated by maximum likelihood, using the variable change formulae the likelihood can be written:

$$\begin{aligned} \ell(\pi_1, \ldots, \pi_K, \boldsymbol{V}, \boldsymbol{R}, \gamma, \boldsymbol{\nu}_1, ..., \boldsymbol{\nu}_K; \mathbf{x}, \mathbf{z}) &= n \log \left| \det \begin{pmatrix} \boldsymbol{V} \\ \boldsymbol{R} \end{pmatrix} \right| - \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \|\boldsymbol{V}\boldsymbol{x}_i - \boldsymbol{\nu}_k\|^2 \\ &+ \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_k) - \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{R}\boldsymbol{x}_i - \gamma\|^2 - \frac{n}{2} \log(2\pi). \end{aligned}$$

The first term is related to the variable change, the second term is related to the discriminant components and the third term is related to prior class membership probabilities and the fourth term is related to the non-discriminant components. In practice, this decomposition is the corner stone of the proposed approach since it separates the clustering part from the non-clustering part.

As stated in Campbell [15] the maximum likelihood estimator of $\boldsymbol{V}$ are the first p eigen vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$ of $\boldsymbol{W}^{-1}\boldsymbol{B}$ in rows:

$$\hat{\boldsymbol{V}} = \begin{pmatrix} \boldsymbol{v}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{v}_p^{\mathrm{T}} \end{pmatrix},$$

renormalized such that:

$$\hat{\boldsymbol{V}}\boldsymbol{W}\hat{\boldsymbol{V}}^{\mathrm{T}} = \boldsymbol{I}_p.$$

The maximum likelihood estimator of $\boldsymbol{R}$ is obtained such that :

$$\hat{\boldsymbol{R}}\boldsymbol{W}\hat{\boldsymbol{V}}^{\mathrm{T}} = 0$$

and that

$$\hat{\boldsymbol{R}}\boldsymbol{W}\hat{\boldsymbol{R}}^{\mathrm{T}} = \boldsymbol{I}_{d-p}, \tag{3}$$

with $\boldsymbol{D}_{d-p}$ the diagonal matrix of the $d-p$ last eigen-values of $\boldsymbol{W}^{-1}\boldsymbol{B}$ (with the $d-K$ last eigen-values which are null). $\hat{\boldsymbol{R}}^{\mathrm{T}}$ can simply be obtained by multiplying the last $d-p$ egien-vectors of $\boldsymbol{W}^{-1/2}\boldsymbol{B}\boldsymbol{W}^{-1/2}$ by $\boldsymbol{W}^{-1/2}$ then renormalize them such that Equation (3) is satisfied. Such renormalization makes the parameters V et R identifiable up to a sign.

Moreover $\boldsymbol{\nu}_k$ and γ are estimated by

$$\hat{\boldsymbol{\nu}}_k = \frac{1}{n_k} \sum_{i=1}^{n} z_{ik} \hat{\boldsymbol{V}}\boldsymbol{x}_i,$$

$$\hat{\gamma} = \frac{1}{n} \sum_{i=1}^{n} \hat{\boldsymbol{R}}\boldsymbol{x}_i.$$

As stated in Trevor Hastie [16] the link between the two parametrizations is as follows:

$$\hat{\boldsymbol{\mu}}_k = \boldsymbol{W}\hat{\boldsymbol{V}}^{\mathrm{T}}\hat{\boldsymbol{R}}(\bar{\boldsymbol{x}}_k - \bar{\boldsymbol{x}}) + \bar{\boldsymbol{x}},$$

and

$$\hat{\boldsymbol{\Sigma}} = \boldsymbol{W} + \boldsymbol{W}\hat{\boldsymbol{R}}^{\mathrm{T}}\hat{\boldsymbol{R}}\boldsymbol{B}\hat{\boldsymbol{R}}^{\mathrm{T}}\hat{\boldsymbol{R}}\boldsymbol{W}.$$

From this formula we can clearly see that the reduced rank constraint operates some regularization on the parameters estimation. Moreover, from a practical perspective the reparametrization Equation (2) is more efficient for computing the posterior class membership probabilities.

2.4. Application in the Clustering Setting

As in Trevor Hastie [16], the model can easily be used in the clustering setting using the EM algorithm to maximise the likelihood. Thus $\boldsymbol{W}$, $\boldsymbol{B}$, $\bar{\boldsymbol{x}}_k$ are recomputed at each iteration by using their version weighted by their posterior membership probabilities.

The EM algorithm is now presented. The algorithm is first initialized with some starting value of the parameters or of the partition. Then the E step and the M step are iterated until convergence. Let $^{(r)}$ denote the value of the parameters at iteration r, the E and the M steps are the followings:

- E step: compute the posterior class membership probabilities.

$$t_{ik}^{(r+1)} = \frac{\pi_k \phi_d(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}^{(r)})}{\sum_{k'=1}^{K} \pi_{k'} \phi_d(\boldsymbol{x}_i; \boldsymbol{\mu}_{k'}^{(r)}, \boldsymbol{\Sigma}^{(r)})} = \frac{\pi_k \phi_p(\boldsymbol{y}_i^{(r)}; \boldsymbol{\nu}_k^{(r)}, \boldsymbol{I}_p)}{\sum_{k'=1}^{K} \pi_{k'} \phi_p(\boldsymbol{y}_i^{(r)}; \boldsymbol{\nu}_{k'}^{(r)}, \boldsymbol{I}_p)},$$

  with $\boldsymbol{y}_i^{(r)} = \boldsymbol{V}^{(r)}\boldsymbol{x}_i$.

- M step: Compute

$$\bar{\boldsymbol{x}}_k^{(r+1)} = \frac{1}{n_k^{(r+1)}} \sum_{i=1}^{n} t_{ik}^{(r+1)} \boldsymbol{x}_i,$$

$$\boldsymbol{W}^{(r+1)} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik}^{(r+1)} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k^{(r+1)})(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_k^{(r+1)})^{\mathrm{T}},$$

$$\boldsymbol{B}^{(r+1)} = \frac{1}{n} \sum_{k=1}^{K} n_k^{(r+1)} (\bar{\boldsymbol{x}}_k^{(r+1)} - \bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_k^{(r+1)} - \bar{\boldsymbol{x}})^{\mathrm{T}}.$$

  Then deduce $\boldsymbol{V}^{(r+1)}$, $\boldsymbol{R}^{(r+1)}$, $\boldsymbol{\nu}_1^{(r+1)}, \ldots, \boldsymbol{\nu}_K^{(r+1)}$ and $\gamma^{(r+1)}$ as in the previous section using the eigenvalue decomposition of ${\boldsymbol{W}^{(r+1)}}^{-1}\boldsymbol{B}^{(r+1)}$.

As noticed in Trevor Hastie [16], this approach is not equivalent to performing a standard EM algorithm and then performing FDA at the end of the EM algorithm. FDA must be computed at each iteration of the EM algorithm since the posterior membership probabilities are only computed based on the p first clustering projections.

Let us also notice that the M step can be interpreted in a Fisher-M step since FDA is required. In this sense it can be interpreted as a particular version of the Fisher-EM algorithm of Bouveyron and Brunet [6]. Although the homoscedasticity could be seen as particularly constraining, it is the best framework for introducing our model in the next section, since it is easily interpretable and allows for efficient computation owing to the closed form of FDA. This limitation could be easily overcome by using rigorous extensions of the FDA in the heteroscedastic setting as in Kumar and Andreou [17]; however, the computation would be much more intensive, since in this case no closed form formula is available and an iterative algorithm would be required even for finding the best projection.

3. Multi-Partition Subspace Mixture Model

3.1. Presentation of the Model

Let us now suppose that instead of having only one class variable $\boldsymbol{z}_i$ for data i, we now have H class variables $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ with $K_1, \ldots, K_H$ modalities. It is assumed that $\boldsymbol{z}_i^1, \ldots, \boldsymbol{z}_i^H$ are independent, with $p(z_{ik}^h = 1)$ denoted by π_k^h. Let also denote by $\boldsymbol{y}_i^h$ the variables related to the clustering variable $\boldsymbol{z}_i^h$ such that:

$$\boldsymbol{y}_i^h | z_{ik}^h = 1 \sim \mathcal{N}_{p_h}(\boldsymbol{\nu}_k^h, I_{p_h})$$

and that we will denote by $p_\bullet = \sum_{h=1}^H p_h$.

Let us still denote by $\boldsymbol{u}_i$ the non clustering variables

$$\boldsymbol{u}_i \sim \mathcal{N}_{d-p_\bullet}(\gamma, I_{d-p_\bullet}).$$

Let us also define $\boldsymbol{x}_i$ by:

$$\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{y}_i^1 \\ \vdots \\ \boldsymbol{y}_i^H \\ \boldsymbol{u}_i \end{pmatrix}.$$

Thus,

$$p(\boldsymbol{x}) = \left| \det \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix} \right| p(\boldsymbol{u}) \prod_{h=1}^H p(\boldsymbol{y}^h).$$

The Figure 2 illustrates the model in the case of $H = 2$.

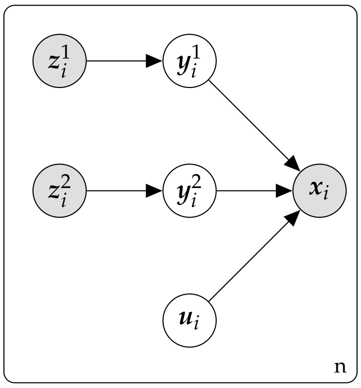

Figure 2. Adapted Bayesian dependency graph to the multi-partition setting, for $H = 2$ clustering variables.

Let us notice that this model allows us to visualize many clustering viewpoints in a low dimensional space since $\boldsymbol{x}_i$ can be summarized by $\boldsymbol{y}_i^1$, …, $\boldsymbol{y}_i^H$. For instance, one can suppose that $p_1 = \ldots = p_H = 1$. In this case each clustering variable can be visualized on one component. We will denote by $\boldsymbol{\theta} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_H, \boldsymbol{R}, \gamma, \boldsymbol{\nu}_1^1, ..., \boldsymbol{\nu}_{K_H}^H)$ the parameters of the model to be estimated.

3.2. Discussion about the Model

The Cartesian product of cluster spaces results in $\prod_{h=1}^H K_h$ clusters, which can be very large without needing many parameters. Thus the proposed model can be interpreted as being a very sparse Gaussian mixture model allowing the possibility to deal with a very large number of clusters, the resulting conditional means and covariances matrices are given in the following formulas:

$$\mathbb{E}(\boldsymbol{x}_i | z^1_{ik_1} = 1, z^2_{ik_1} = 1, \ldots, z^H_{ik_H} = 1) = \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\nu}^1_{k_1} \\ \vdots \\ \boldsymbol{\nu}^H_{k_H} \\ \gamma \end{pmatrix},$$

and

$$\mathbb{V}(\boldsymbol{x}_i | z^1_{ik_1} = 1, z^2_{ik_1} = 2, \ldots, z^H_{ik_H} = 1) = (\boldsymbol{V}_1^{\mathrm{T}} \boldsymbol{V}_1 + \cdots + \boldsymbol{V}_H^{\mathrm{T}} \boldsymbol{V}_H + \boldsymbol{R}^{\mathrm{T}} \boldsymbol{R})^{-1}.$$

Thus, the expectation of $\boldsymbol{x}_i$ given in all the clusters is a linear combination of the cluster specific means which can be referred to as a multiple-way MANOVA setting. On the one hand, as a particular homoscedastic Gaussian mixture, our model is more prone to model bias than free homoscedastic Gaussian mixture, and in the case when our model would be well-specified the homoscedastic Gaussian mixture would give a similar clustering for a large sample size (i.e., the same partitions with respect to the partition resulting from the product space of our multi-partitions model). On the other hand, our approach produces a factorised version of the partition space as well as the related clustering subspaces which is not a standard output of clustering methods, and it can deal with a large number of clusters in a sparse way which can be particularly useful for a moderated sample size. In practice, the choice between our model and an other mixture model can simply be performed through the BIC criterion.

In some sense our model can be linked with the mixture of factor analyzers [18]. In mixture of factor analyzers the model is of the type:

$$\boldsymbol{x}_i = \boldsymbol{A} \boldsymbol{y}_i + \boldsymbol{u}_i,$$

where $\boldsymbol{A}$ is a low rank matrix. But here we have chosen a model of the type

$$\boldsymbol{x}_i = \boldsymbol{A} \begin{pmatrix} \boldsymbol{y}_i \\ \boldsymbol{u}_i \end{pmatrix},$$

which allows us to deal with the noise in a different way. Actually, our model is invariant up to a bijective linear transformation of the data which is not the case for the mixtures of factor analyzers. On the other hand, our model can only deal with data with moderated dimension with respect to the number of statistical units; it assumes that the sources $\boldsymbol{y}_i$ can be recovered from the observed data $\boldsymbol{x}_i$.

3.3. Estimation of the Parameters of the Model in the Supervised Setting

The likelihood of the model can be written:

$$\begin{aligned} \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) \quad = \quad & n \log \left| \det \begin{pmatrix} \boldsymbol{V}_1 \\ \vdots \\ \boldsymbol{V}_H \\ \boldsymbol{R} \end{pmatrix} \right| - \sum_{i=1}^{n} \sum_{h=1}^{H} \sum_{k=1}^{K_h} z^h_{ik} \| \boldsymbol{V}_h^{\mathrm{T}} \boldsymbol{x}_i - \boldsymbol{\nu}^h_k \|^2 \\ & + \sum_{i=1}^{n} \sum_{h=1}^{H} \sum_{k=1}^{K_h} z^h_{ik} \log(\pi^h_k) - \sum_{i=1}^{n} \| \boldsymbol{R}^{\mathrm{T}} \boldsymbol{x}_i - \gamma \|^2 - \frac{n}{2} \log(2\pi). \end{aligned}$$

The likelihood cannot be maximized directly. However, in the case of $H = 1$, it reduces to the problem of Section 2. Let notice that if all the parameters are fixed except $\boldsymbol{V}_h$ and $\boldsymbol{R}$, $\boldsymbol{\nu}^h_k$ and γ, the optimisation can be easily performed by constraining $\boldsymbol{V}_h^{(r+1)}$ and $\boldsymbol{R}^{(r+1)}$ to be linear combinations of $\boldsymbol{V}_h^{(r)}$ and $\boldsymbol{R}^{(r)}$. Thus the likelihood will be optimized by using an alternate optimization algorithm.

Let $\boldsymbol{M} \in \mathcal{M}_{d-p_\bullet+p_h,d-p_\bullet+p_h}(\mathbb{R})$ the matrix which allow to compute $\boldsymbol{V}_h^{(r+1)}$ and $\boldsymbol{R}^{(r+1)}$ based on $\boldsymbol{V}_h^{(r)}$ and $\boldsymbol{R}^{(r)}$:

$$\begin{pmatrix} \boldsymbol{V}_h^{(r+1)} \\ \boldsymbol{R}^{(r+1)} \end{pmatrix} = \boldsymbol{M} \begin{pmatrix} \boldsymbol{V}_h^{(r)} \\ \boldsymbol{R}^{(r)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}_1 \\ \boldsymbol{M}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_h^{(r)} \\ \boldsymbol{R}^{(r)} \end{pmatrix},$$

where $\boldsymbol{M}_1$ is the sub-matrix containing the p_h first rows of $\boldsymbol{M}$ and $\boldsymbol{M}_2$ the matrix containing the last $d - p_\bullet$ rows of $\boldsymbol{M}$.

Thus, the increase of the likelihood when all the parameters are fixed except $\boldsymbol{V}_h$, $\boldsymbol{R}$, $\boldsymbol{\nu}_k^h$ and γ becomes:

$$\begin{aligned} C(\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{\nu}_k, \gamma) &= n \log|\det(\boldsymbol{M})| - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K_h} z_{ik}^h \left\| \boldsymbol{M}_1 \begin{pmatrix} \boldsymbol{V}_h^{(r)} \\ \boldsymbol{R}^{(r)} \end{pmatrix} \boldsymbol{x}_i - \boldsymbol{\nu}_k^h \right\|^2 \\ & -\frac{1}{2}\sum_{i=1}^{n} \left\| \boldsymbol{M}_2 \begin{pmatrix} \boldsymbol{V}_h^{(r)} \\ \boldsymbol{R}^{(r)\mathrm{T}} \end{pmatrix} \boldsymbol{x}_i - \gamma \right\|^2. \end{aligned}$$

By denoting

$$\begin{pmatrix} \boldsymbol{y}_i^{h(r)} \\ \boldsymbol{u}_i^{(r)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{V}_h^{(r)} \\ \boldsymbol{R}^{(r)} \end{pmatrix} \boldsymbol{x}_i,$$

we have to maximize:

$$\begin{aligned} C(\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{\nu}_k, \gamma) &= n \log|\det(\boldsymbol{M})| - \frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K_h} z_{ik}^h \left\| \boldsymbol{M}_1 \begin{pmatrix} \boldsymbol{y}_i^{h(r)} \\ \boldsymbol{u}_i^{(r)} \end{pmatrix} - \boldsymbol{\nu}_k^h \right\|^2 \\ & -\frac{1}{2}\sum_{i=1}^{n} \left\| \boldsymbol{M}_2 \begin{pmatrix} \boldsymbol{y}_i^{h(r)} \\ \boldsymbol{u}_i^{(r)} \end{pmatrix} - \gamma \right\|^2. \end{aligned}$$

Consequently, $\boldsymbol{M}$ and the others parameters can be obtained by applying a simple FDA on the data $(\boldsymbol{y}_i^{h(r)\mathrm{T}}, \boldsymbol{u}_i^{(r)\mathrm{T}})$. In order to optimise over all the parameters, we can loop over all the clustering dimensions.

Thus, in the case of mixed continuous and categorical data, this model can be used to visualize the clustering behavior of the categorical variables with respect to the quantitative ones.

3.4. Estimation of the Parameters of the Model in the Clustering Setting

Here our main goal is to consider the clustering setting, that is, when $z_i^1, \ldots, z_i^H$ are unknown. Consequently we will use an EM algorithm to "reconstitute the missing label" in order to maximize the likelihood. Thus the algorithm stays the same as in the supervised setting, except that the data at each iteration are now weighted by $t_{ik}^{h\,(r+1)}$ instead of z_{ik}^h.

The algorithm is the following:

- Until convergence, for $h \in \{1, \ldots, H\}$ iterates the following steps:
  - E step: compute
  $$t_{ik}^{h\,(r+1)} = \frac{\pi_k p(\boldsymbol{y}_i^{h(r)}; \boldsymbol{\nu}_k^{h(r)}, \boldsymbol{I}_p)}{\sum_{k'=1}^{K} \pi_{k'} p(\boldsymbol{y}_i^{h(r)}; \boldsymbol{\nu}_{k'}^{h\,(r)}, \boldsymbol{I}_p)}.$$
  - M step: compute $\pi_1^{h(r+1)}, \ldots, \pi_{K_h}^{h\ (r+1)}$, $\boldsymbol{V}_h^{(r+1)}$, $\boldsymbol{R}^{(r+1)}$, $\gamma^{(r+1)}$ and $\boldsymbol{\nu}_k^{h(r+1)}$ based on formulas given in the supervised setting.

3.5. Parsimonious Models and Model Choice

The proposed model needs the user to define the number of clustering subspaces H, the number of cluster in each clustering subspace $K_1, \ldots, K_H$, and the dimensionality $p_1, \ldots, p_H$ of each subspace. The constraints are that $H < d$, that $p_h \leq K_h - 1$ and $p_\bullet = p_1 + \cdots + p_H < d$. It is clear that the number of possible models can become very high. To limit the combinatorial aspect, one can impose $K_1 = \cdots = K_H = K$ and/or $p_1 = \cdots = p_h = p$. In practice the choice of $p = 1$ enforces to find clustering which could be visualized in one dimension, which can help the practitioner. Moreover, choosing $K = 2$ is the minimal requirement in order to investigate a clustering structure. However, if possible we recommend to explore the largest possible number of models and choosing the best one with the BIC. Let us define the following parsimonious models and their related number of parameters:

- $[K_h p_h]$ the general form of the proposed model, the index h will be removed if values are the same for each clustering subspace.

$$\sum_{h=1}^{H}(K_h - 1) + \sum_{h=1}^{H} \frac{p_h(2K_h + 2d - p_h + 1)}{2} + \frac{(d - p_\bullet)(d + p_\bullet + 3)}{2},$$

- $[Kp_h]$ where the number of clusters is the same for each subspace

$$H(K-1) + \sum_{h=1}^{H} \frac{p_h(2K + 2d - p_h + 1)}{2} + \frac{(d - p_\bullet)(d + p_\bullet + 3)}{2},$$

- $[K_h p]$ where the dimensionalities are the same for each subspaces

$$\sum_{h=1}^{H}(K_h - 1) + \sum_{h=1}^{H} \frac{p(2K_h + 2d - p + 1)}{2} + \frac{(d - Hp)(d + Hp + 3)}{2},$$

- $[K_h 1]$ where the dimensionalities are equals to one for each subspaces

$$\sum_{h=1}^{H}(K_h - 1) + \sum_{h=1}^{H}(K_h + d) + \frac{(d - H)(d + H + 3)}{2},$$

- $[Kp]$ where the number of clusters is the same for each subspace and the dimensionalities are the same for each subspace

$$H(K-1) + \frac{Hp(2K + 2d - p + 1)}{2} + \frac{(d - Hp)(d + Hp + 3)}{2},$$

- $[K1]$ where the number of clusters is the same for each subspace and the dimensionalities are equals to one for each subspaces

$$H(K-1) + H(K+d) + \frac{(d - H)(d + H + 3)}{2}.$$

For a given model m the BIC is computed as:

$$BIC(m) = \ell(\hat{\boldsymbol{\theta}}_m; \mathbf{x}) - \frac{\nu_m}{2} \log n,$$

where ν_m is the number of parameters of the model detailed above. Thus the model choice consists of choosing the model maximising the BIC. BIC enjoys good theoretical consistency properties, thus providing a guarantee to select the true model as the number of data increases. The ICL criterion [19] could also be used to enforce the choice of well separated clusters, since from a classification perspective BIC is known to over-estimate the number of clusters if model assumption are violated. Let us however notice that in practice the user could be mainly interested by a low value of H, since even $H = 2$ can provide him with new insights about his data, focusing on finding several clustering view points.

4. Experiments

4.1. Experiments on Simulated Data

We now present a tutorial example. Let us consider $n = 100$ data, with $H = 2$ clustering subspaces each one of dimension one ($p_1 = p_2 = 1$) and containing each of two clusters ($K_1 = K_2 = 2$). Let us draw y_i^1, the first clustering variable, according to a mixture of two Gaussian univariate distributions: $y_i^1 \sim 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(4,1)$ and draw independently y_i^2, the second clustering variable, according to the same mixture distribution: $y_i^2 \sim 0.5\mathcal{N}(0,1) + 0.5\mathcal{N}(4,1)$, then draw $\boldsymbol{u}_i$ the non-classifying components according to a multivariate Gaussian distribution in $\mathbb{R}^4$, $\boldsymbol{u}_i \sim \mathcal{N}_4(0, \boldsymbol{I}_4)$. Finally compute $\boldsymbol{x}_i$ based on the formula: $\boldsymbol{x}_i = A \begin{pmatrix} y_i^1 \\ y_i^2 \\ \boldsymbol{u}_i \end{pmatrix}$, where the 36 entries of the matrix A have been drawn according to independent $\mathcal{N}(0,1)$ for sake of simplicity.

Thus the proposed model, based on the observed data $\mathbf{x}$, aims at recovering the clustering variables $\mathbf{y}^1$ and $\mathbf{y}^2$ as well as the associated cluster variables $\mathbf{z}^1$ and $\mathbf{z}^2$. The initial data $\mathbf{x}$ are presented on Figure 3, we see that the clustering structure of the data is not apparent from these scatter plots. The underlying clustering variables are presented on Figure 4, where we see the separation of the colors on the first clustering variable Y_1 and separation of the shapes on the second clustering variable Y_2. Let us notice that such factorization gives a more synthetic view of the clustering than seeing these clusters as four clusters. Using standard dimension reduction techniques such as PCA does not succeed in recovering the clustering subspace see Figure 5. Performing a factorial discriminant analysis considering the four clusters in the supervised setting we get Figure 6, it finds good separation between the clusters, however we do not obtain the factorised interpretation of the clustering as the Cartesian product of two independent clusterings. Finally by performing the estimation of the model parameters in an unsupervised setting based on the data $\mathbf{x}$ we get Figure 7. We see that $\hat{Y}_1$ succeeds in recovering Y_1 up to a sign and that $\hat{Y}_2$ succeeds in recovering Y_2.

Moreover, supposing $p_1 = p_2 = 1$ we can choose K_1 and K_2 according to the BIC criterion. Values of the BIC criterion are presented in Table 1, where we show that the selected model is the true one with $K_1 = K_2 = 2$. The lower diagonal of the table is not presented for symmetry reasons, and we limited ourselves to $K_1, K_2 \in \{1, \ldots, 5\}$.

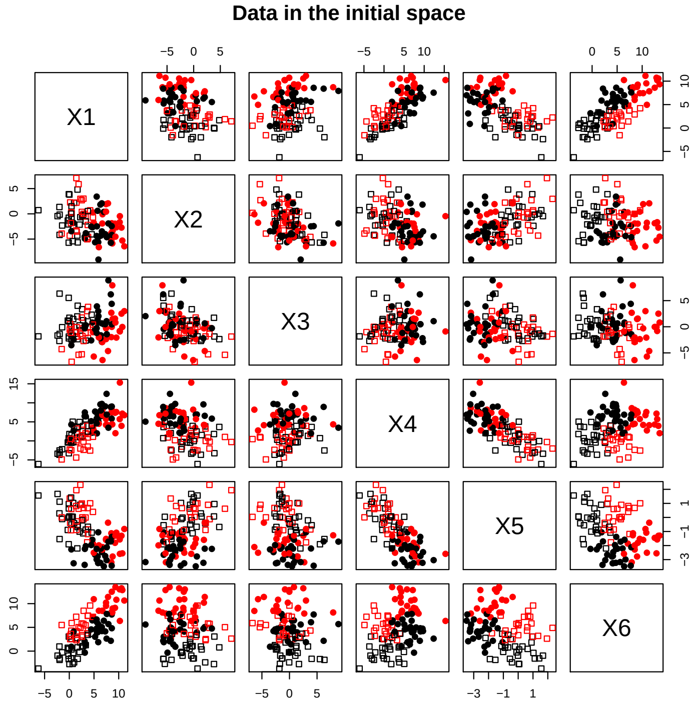

Figure 3. Scatter plots of the initial data on the illustrative example. The color depends on the first cluster variable, and the shape depends on the second cluster variable.

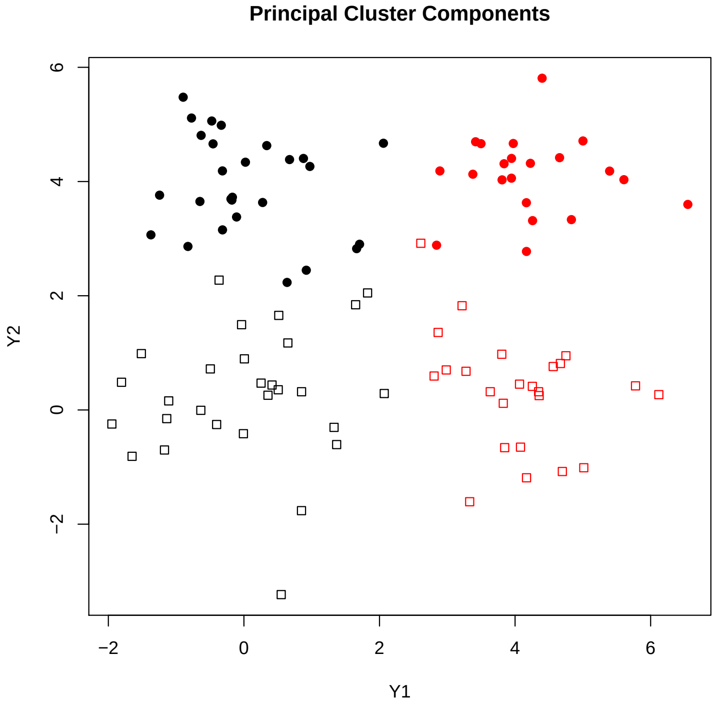

Figure 4. Scatter plot of the illustrative data on the two original clustering subspaces.

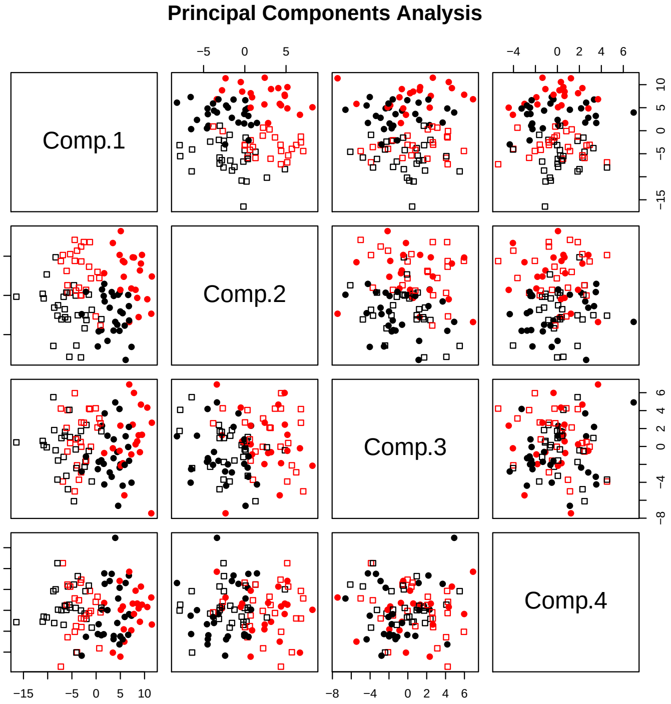

Figure 5. Scatter plot of the illustrative data on components one to four of the proimcipal component analysis (PCA).

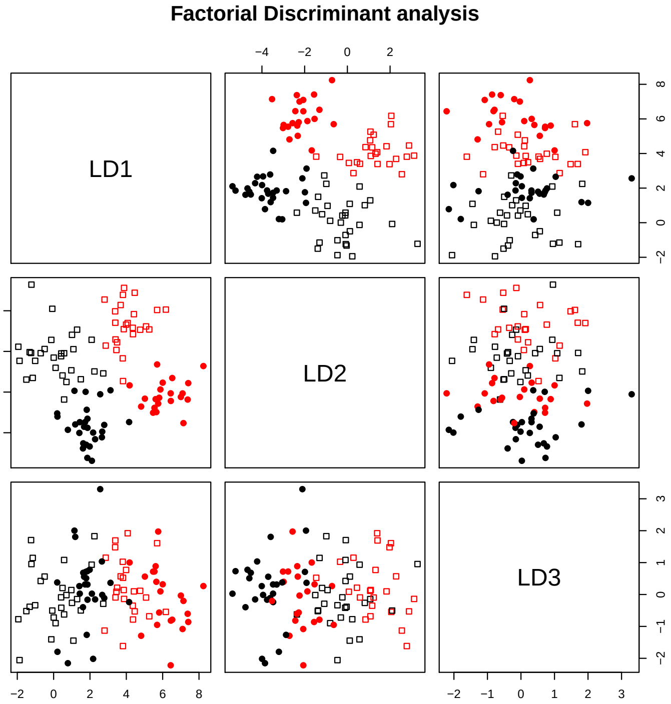

Figure 6. Scatter plot of the component of the factorial discriminant analysis for the illustrative example.

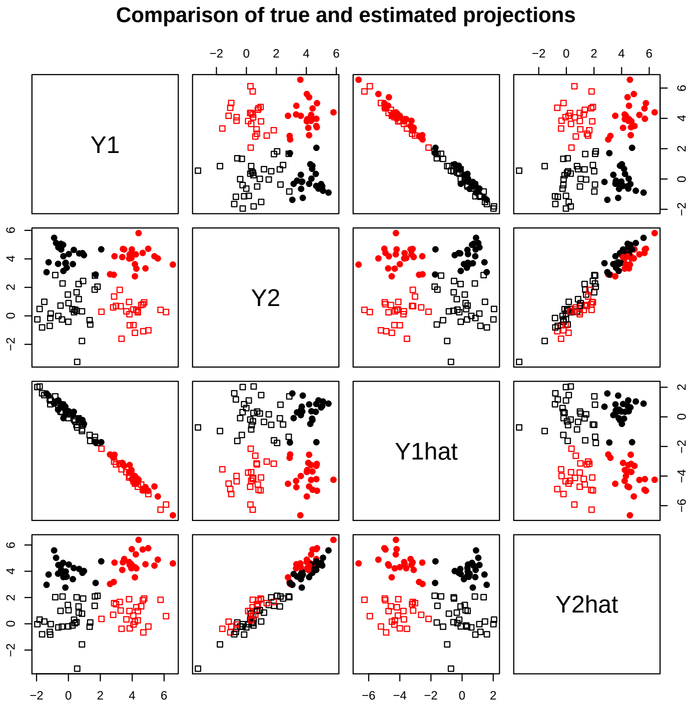

Figure 7. Comparison of the true and of the estimated clustering subspaces on the illustrative example, points are marked according to the estimated clusters.

Table 1. Value of the BIC criterion according to K_1 and K_2, for the choice of the number of clusters on the illustrative example, best value in bold.

$K_1 \setminus K_2$	1	2	3	4	5
1	−1318.40	−1301.15	−1305.20	−1305.47	−1310.08
2		**−1291.80**	−1293.22	−1292.28	−1307.07
3			−1296.70	−1303.68	−1310.95
4				−1306.09	−1320.29
5					−1319.22

4.2. Experiments on Real Data

Let us consider the crabs dataset [20]. It consists of 200 crabs morphological data, each crab has two categorical (cluster) attributes—the species, orange or blue, and the sex, male or female. The dataset is composed of 50 males orange, 50 males blue, 50 females orange, 50 females blue for which 5 numerical attributes have been measured: the frontal lobe size, the rear width, the carapace length, the carapace width and the body depth. We can see the PCA of the data in Figure 8. We see that component two separates males and females well, whereas component three separates orange and blue subspecies. However, we will see that by applying our model we obtained a better separation of the clusters.

Like in the tutorial example we will take $p_1 = p_2 = 1$ and $K_1, K_2 \in \{1, \ldots, 5\}$. The resulting BIC tabular is given in Table 2, it suggests the choice of $K_1 = 3$ and $K_2 = 4$. The resulting visualization of the clustering variables in given Figure 9. Let us notice that Y_2 is divided in four clusters however, however we only see three since two of them have the same mean. We can see that even if the numbers of clusters do not correspond, the first clustering subspace finds the subspecies, whereas the second clustering subspace finds the sex. We could also look at the solution provided by $K_1 = K_2 = 2$ on Figure 10, this one has a lower BIC but seems more natural for the problem at hand. We see that the obtained map is in fact quite similar the map obtained Figure 9; however, we notice that from a density approximation point of view we obtain a lower fit. In fact, if we look at the correlations between Y_1 Figure 9 and Y_2 Figure 10 we have a correlation of -0.97, and a similar correlation is obtained between Y_2 Figure 9 and Y_1 Figure 10. Thus, the produced subspace are finally quite similar.

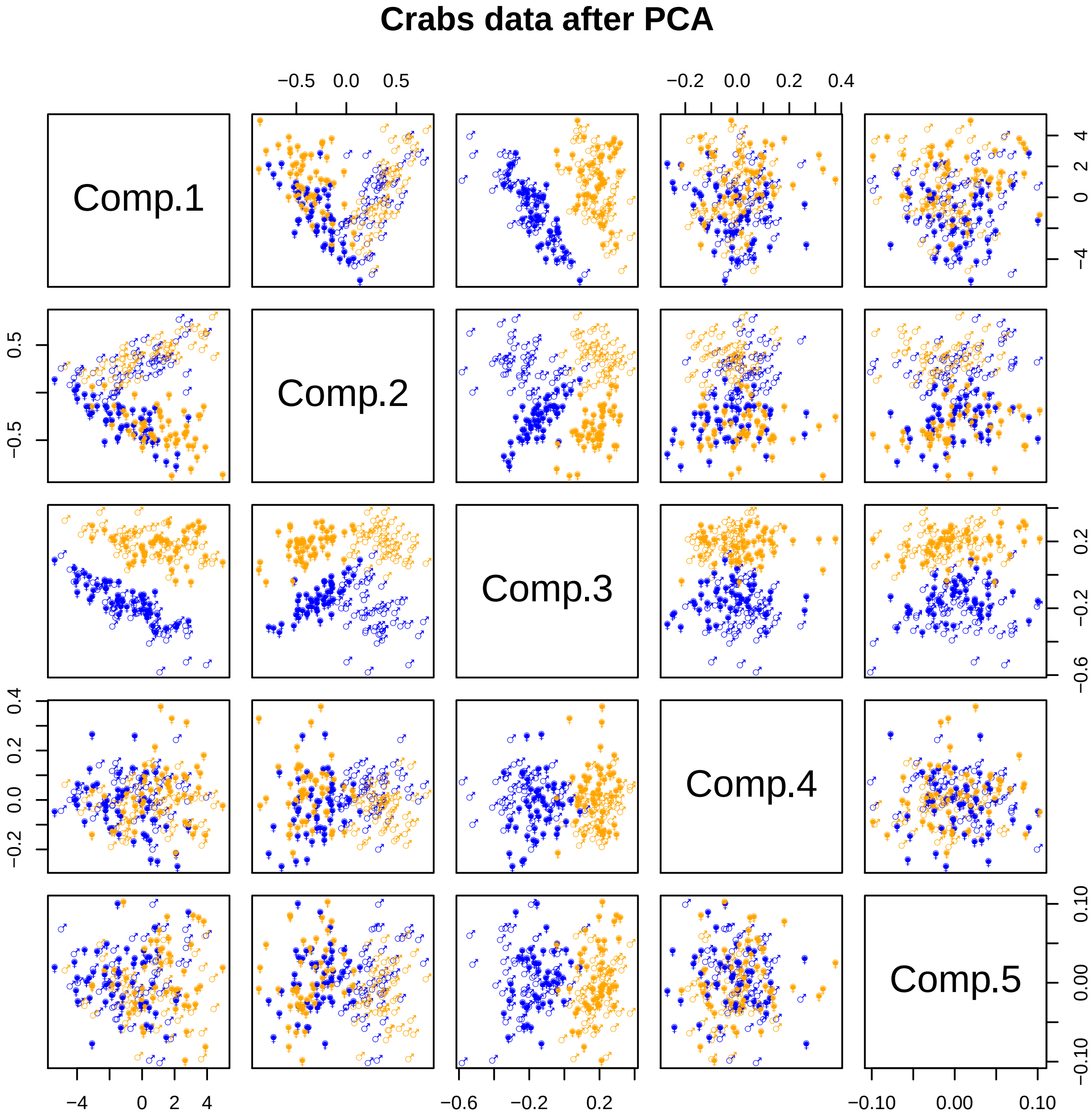

Figure 8. Scatter plots of the crabs data after PCA. Subspecies are represented according to their color, and sex is represented according to its symbol.

Table 2. Value of the BIC criterion according to K_1 and K_2, for the choice of the number of clusters on the crabs dataset, best value in bold.

$K_1 \setminus K_2$	1	2	3	4	5
1	−62.66	0.41	10.40	5.11	0.80
2		17.82	16.57	18.88	0.49
3			3.75	**22.52**	17.44
4				−26.65	−26.64
5					12.06

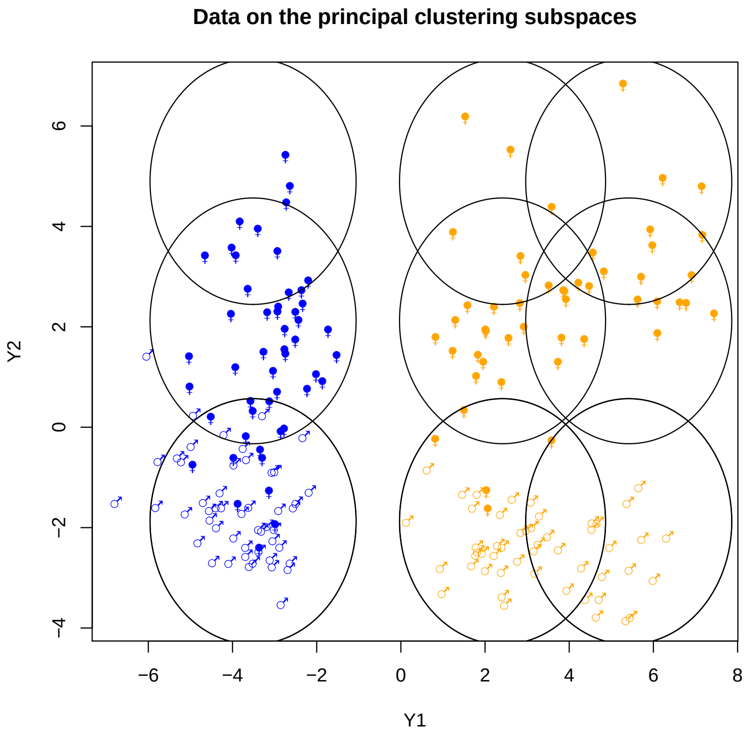

Figure 9. Scatter plots of the clustering subspace on the crabs data for $K_1 = 3$ and $K_2 = 4$, 95% isodensity is given for each component resulting of the Cartesian product.

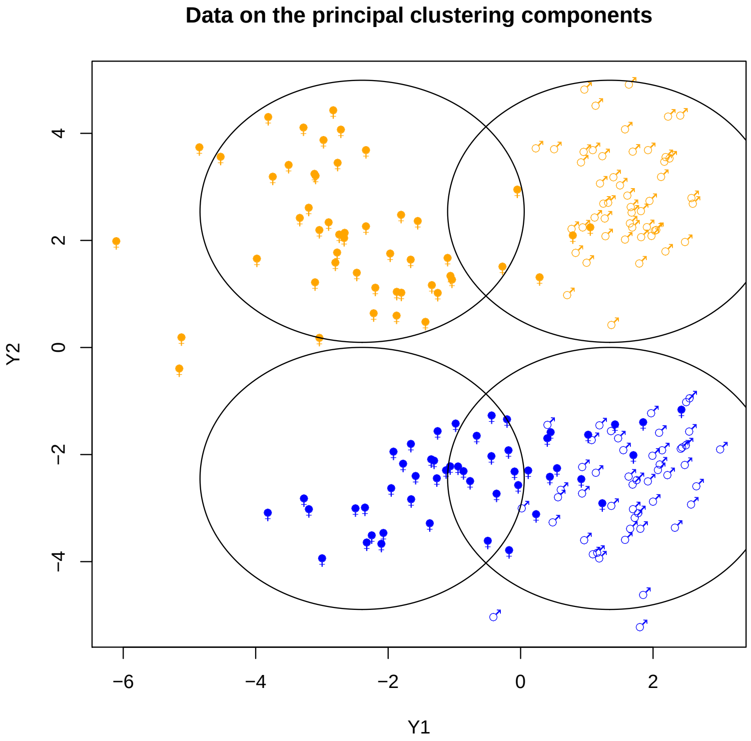

Figure 10. Scatter plots of the clustering subspace on the crabs data for $K_1 = K_2 = 2$, 95% isodensity is given for each component resulting of the Cartesian product.

5. Conclusions and Perspectives

We have proposed a model which allows us to combine visualization and clustering with many clustering viewpoints. Moreover, we have shown the possibility of performing model choice by using the BIC criterion. The proposed model can provide new information on the structure present in the data by trying to reinterpret the cluster as a result of the Cartesian product of several clustering variables.

The proposed model is limited to the homoscedatic setting, which could be seen as a limitation; however, from our point of view this is more robust than the heteroscedastic setting, which is known to be jeopardized by the degeneracy issue [21]. However, the extension of our work on the heteroscedastic setting can easily be performed from the modeling point of view; the main issue in this case would be the parameters estimation where an extension of the FDA to the heteroscedastic setting would be needed, as presented in Kumar and Andreou [17]. Another difficult issue is the choice of H, $K_1 \ldots, K_H$ and $p_1, \ldots, p_H$, which is very combinatorial. Here we have proposed an estimation strategy for all these tuning parameters being fixed, and then performed a selection of the best tuning according to BIC. However, in future work, a model selection strategy to perform the model selection through a modified version of the EM algorithm will also be investigated as in Green [22]; it would thus limit the combinatorial aspect of the global model search through EM-wise local model searches.

Funding: This research received no external funding.

Acknowledgments: I would like to thank Rohit BHAGWAT for his preliminary work on the experimental part of the topic during his internship. I also would like to thank anonymous reviewers for their comments which contributed to improve the quality of the article.

Conflicts of Interest: The author declares no conflict of interest.

References

1. McLachlan, G.; Peel, D. *Finite Mixture Models*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
2. Schwarz, G. Estimating the Dimension of a Model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
3. Tipping, M.E.; Bishop, C.M. Probabilistic principal component analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1999**, *61*, 611–622. [CrossRef]
4. Celeux, G.; Govaert, G. Gaussian parsimonious clustering models. *Pattern Recognit.* **1995**, *28*, 781–793. [CrossRef]
5. Bouveyron, C.; Girard, S.; Schmid, C. High-dimensional data clustering. *Comput. Stat. Data Anal.* **2007**, *52*, 502–519. [CrossRef]
6. Bouveyron, C.; Brunet, C. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Stat. Comput.* **2012**, *22*, 301–324. [CrossRef]
7. Dempster, A.; Laird, N.; Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Society. Ser. B Methodol.* **1977**, *39*, 1–38.
8. Galimberti, G.; Soffritti, G. Model-based methods to identify multiple cluster structures in a data set. *Comput. Stat. Data Anal.* **2007**, *52*, 520 – 536. [CrossRef]
9. Galimberti, G.; Manisi, A.; Soffritti, G. Modelling the role of variables in model-based cluster analysis. *Stat. Comput.* **2018**, *28*, 145–169. [CrossRef]
10. Poon, L.K.; Zhang, N.L.; Liu, T.; Liu, A.H. Model-based clustering of high-dimensional data: Variable selection versus facet determination. *Int. J. Approx. Reason.* **2013**, *54*, 196–215. [CrossRef]
11. Marbac, M.; Vandewalle, V. A tractable multi-partitions clustering. *Comput. Stat. Data Anal.* **2019**, *132*, 167–179. [CrossRef]
12. Attias, H. Independent factor analysis. *Neural Comput.* **1999**, *11*, 803–851. [CrossRef] [PubMed]
13. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
14. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: Berlin, Germany, 2001; Volume 1.
15. Campbell, N.A. Canonical variate analysis—A general model formulation. *Aust. J. Stat.* **1984**, *26*, 86–96. [CrossRef]
16. Hastie, T.; Tibshirani, R. Discriminant Analysis by Gaussian Mixtures. *J. R. Stat. Society. Ser. B Methodol.* **1996**, *58*, 155–176. [CrossRef]
17. Kumar, N.; Andreou, A.G. Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition. *Speech Commun.* **1998**, *26*, 283–297. [CrossRef]

18. Ghahramani, Z.; Hinton, G.E. *The EM Algorithm for Mixtures of Factor Analyzers*; Technical Report, Technical Report CRG-TR-96-1; University of Toronto: Toronto, ON, Canada, 1996.
19. Biernacki, C.; Celeux, G.; Govaert, G. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 719–725. [CrossRef]
20. Campbell, N.; Mahon, R. A multivariate study of variation in two species of rock crab of the genus Leptograpsus. *Aust. J. Zool.* **1974**, *22*, 417–425. [CrossRef]
21. Biernacki, C.; Chrétien, S. Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM. *Stat. Probab. Lett.* **2003**, *61*, 373–382. [CrossRef]
22. Green, P.J. On use of the EM for penalized likelihood estimation. *J. R. Stat. Soc. Ser. Methodol.* **1990**, *52*, 443–452.