
A BAYESIAN FISHER-EM ALGORITHM FOR DISCRIMINATIVE GAUSSIAN SUBSPACE CLUSTERING

Nicolas Jouvin

Université Paris 1 Panthéon-Sorbonne, SAMM, France
FP2M, CNRS FR 2036, Paris, France

Charles Bouveyron

Université Côte d’Azur, Inria, CNRS, Laboratoire J.A. Dieudonné
Maasai research team, Nice, France

Pierre Latouche

Université de Paris, MAP5, CNRS,
FP2M, CNRS FR 2036, Paris, France

December 9, 2020

ABSTRACT

High-dimensional data clustering has become and remains a challenging task for modern statistics and machine learning, with a wide range of applications. We consider in this work the powerful discriminative latent mixture model, and we extend it to the Bayesian framework. Modeling data as a mixture of Gaussians in a low-dimensional discriminative subspace, a Gaussian prior distribution is introduced over the latent group means and a family of twelve submodels are derived considering different covariance structures. Model inference is done with a variational EM algorithm, while the discriminative subspace is estimated via a Fisher-step maximizing an unsupervised Fisher criterion. An empirical Bayes procedure is proposed for the estimation of the prior hyper-parameters, and an integrated classification likelihood criterion is derived for selecting both the number of clusters and the submodel. The performances of the resulting Bayesian Fisher-EM algorithm are investigated in two thorough simulated scenarios, regarding both dimensionality as well as noise and assessing its superiority with respect to state-of-the-art Gaussian subspace clustering models. In addition to standard real data benchmarks, an application to single image denoising is proposed, displaying relevant results. This work comes with a reference implementation for the **R** software in the **FisherEM** package¹ accompanying the paper.

Keywords Mixture model · High dimensionality · Dimensionality reduction · Linear Discriminant Analysis

1 Introduction

Clustering has become an important part of contemporary statistics and machine learning, with applications ranging from DNA microarray analysis in biology (Ghosh and Chinnaiyan 2002) to text analysis (Aggarwal and Zhai 2012) and image processing (Coleman and Andrews 1979; Jégou et al. 2010). However, the high dimensionality characterizing modern datasets causes geometrical and statistical issues, often subsumed under the term of *curse of dimensionality* (Bellman 1957; Giraud 2014), and affecting traditional approaches. In this paper, we focus on model-based clustering with Gaussian mixture models, and we propose an extension to the Bayesian framework of the work of Bouveyron and Brunet (2012a) along with a novel algorithm for simultaneously clustering and visualizing high-dimensional continuous data.

¹Available on CRAN, see <https://github.com/nicolasJouvin/FisherEM> for additional information.



Figure 1: Chang (1983) data set with $n = 300$ observations projected on the 1st, 2nd and 15th principal components respectively. Colors and shapes indicate the true cluster membership. We see that the last principal component contains important discriminative information in terms of clustering, while the second principal component is not suited for the task.

1.1 High-dimensional Gaussian clustering

Finite mixture models constitute one of the most popular approaches to model-based clustering (McLachlan and Peel 2004). Let us consider n continuous observations $\{\mathbf{y}_i\}$ in dimension p , summarized in a data matrix $\mathbf{Y} \in \mathbb{R}^{n \times p}$, that we want to cluster into K groups. The Gaussian mixture model (GMM, Bouveyron et al. 2019) posits the following distribution:

$$p(\mathbf{Y} \mid \boldsymbol{\pi}, \mathbf{m}, \mathbf{S}) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{m}_k, \mathbf{S}_k), \quad (1)$$

where $\boldsymbol{\pi}$ denotes the mixture proportions and $(\mathbf{m}_k, \mathbf{S}_k)$ respectively corresponds to the mean and covariance matrix of the k -th component. Clustering arises from this probabilistic formulation with the introduction of an unobserved random variable z_i following a multinomial distribution with parameter $\boldsymbol{\pi}$, and characterizing the cluster assignment of observation \mathbf{y}_i . The corresponding partition $\mathbf{Z} = \{z_i\}$ is then considered as the set of discrete latent variables which are to be estimated, along with the model parameters, traditionally by an Expectation-Maximization algorithm (EM, Dempster et al. 1977). Here, the curse of dimensionality comes from the covariance matrices \mathbf{S}_k , which involve a number of parameters growing with the squared of the dimension. In such context, the number of observations required to fit high-dimensional data may be very large and computationally impractical.

On the one hand, some approaches rely on unsupervised dimension reduction such as principal component analysis or factor analysis to project the data prior to model fitting (Ghosh and Chinnaiyan 2002). However, such transformations do not take into account the clustering task at hand and might induce a loss of relevant discriminative information, in addition to losing the principled approach to model-based clustering. In order to demonstrate the importance of a relevant transformation in a clustering context, Chang (1983) exhibited a 2-components simulation setting in dimension $p = 15$ where the groups are best discriminated on the space defined by the first and the last components, as represented in Figure 1. Another example of this phenomenon is also given in McLachlan and Peel (2004, sec. 8.2) in dimension $p = 5$. Model-free heuristics have also been proposed for subspace clustering, seeking for regions of high-density within the observed space. The CLIQUE algorithm (Agrawal et al. 1998) is a popular instance of such methods, and the building blocks for many others. We refer to Parsons et al. (2004) for a comprehensive review on this subject.

On the other hand, a wealth of literature has focused on developing parsimonious models based on Equation (1), consisting of constrained covariance matrices \mathbf{S}_k . Banfield and Raftery (1993) and Celeux and Govaert (1995) proposed

to perform a spectral decomposition of the latter, and derived a family of fourteen submodels by considering different constraints on the eigenvectors and eigenvalues (see *e.g.* Scrucca et al. 2016). Although efficient and flexible, this method still requires the estimation of full-rank $p \times p$ matrices in the observation space for complex models, which can be impractical. Other types of restrictions were considered by Ghahramani and Hinton (1996) and Tipping and Bishop (1999), with low-rank factorizations $\mathbf{S}_k = \mathbf{U}_k \mathbf{U}_k^\top + \mathbf{\Psi}_k$ where \mathbf{U}_k is a $p \times d$ matrix. Based on a factor analysis formulation, these models have a geometric interpretation: integrating model-based clustering and probabilistic linear dimension reduction, they seek to cluster the data in K low-dimensional subspaces \mathbf{U}_k of dimension d . This formulation was further refined to impose a common subspace \mathbf{U} across cluster, restricting the number of parameter and allowing a common visualization of the data points (Yoshida et al. 2004; Baek et al. 2009; McNicholas and Murphy 2008; Montanari and Viroli 2010). At the intersection of the spectral and geometric approaches, Bouveyron et al. (2007) proposed a family of 28 submodels, separating between signal and noisy directions. These models are often referred to as Gaussian subspace clustering, and maximum likelihood inference is always preferred, usually via an EM algorithm. We refer the reader to Bouveyron and Brunet-Saumard (2014b) for a thorough review of model-based high-dimensional clustering.

However, without further clustering information, the estimated latent subspace may be biased toward density estimation, preserving the variance of the observed data as much as possible, rather than clustering and explicit separation of the groups. These objectives are not always aligned as Figure 1 suggests, and in order to circumvent this issue, several works introduced the notion of a discriminative subspace.

1.2 Discriminative subspace: from classification to clustering

In the supervised framework, where the partition \mathbf{Z} is observed, the tension between signal *representation* and signal *classification* is well known and analogous to the distinction between density estimation and clustering. Fukunaga (1990, chap. 10) discusses this in detail, introducing the notion of class separability along with four different possible criteria to measure it. The idea is to find a linear subspace \mathbf{U} in which the group means are well separated while the within-class variance is small. The most popular criterion for such a task is a generalization of Fisher’s Linear Discriminant Analysis (LDA, Fisher 1936; Duda et al. 2000, chap. 4):

$$F(\mathbf{U}) = \text{Tr} \left[(\mathbf{U}^\top \mathbf{S}_W \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{S}_B \mathbf{U} \right], \quad (2)$$

where $\mathbf{S}_W = (1/n) \sum_k \sum_i z_{ik} (\mathbf{y}_i - \mathbf{m}_k)(\mathbf{y}_i - \mathbf{m}_k)^\top$ is the within-class covariance matrix of the data, while $\mathbf{S}_B = (1/n) \sum_k n_k (\mathbf{m}_k - \bar{\mathbf{y}})(\mathbf{m}_k - \bar{\mathbf{y}})^\top$ is the between-class covariance, with $n_k = \sum_i z_{ik}$ and $\mathbf{m}_k = (1/n_k) \sum_i z_{ik} \mathbf{y}_i$. This criterion computes the trace of the ratio between the within-class and between-class covariance matrices in the latent space, and its maximization with respect to \mathbf{U} translates the goal of a discriminative subspace. Without orthogonality constraints, the maximization of Equation (2) happens to be equivalent to a generalized eigenvalue problem $\mathbf{S}_B \mathbf{u} = \lambda \mathbf{S}_W \mathbf{u}$ which can be solved efficiently, even when \mathbf{S}_W is singular (Ye 2005). Moreover, since the rank of the matrix \mathbf{S}_B is at most $K - 1$, there is only $d \leq K - 1$ dimensions of interest.

In the unsupervised context, which is of interest in this paper, \mathbf{Z} is unknown, and the scatter matrices \mathbf{S}_W and \mathbf{S}_B cannot be formed. Still, building on these ideas, some works have been proposed to adapt the criterion. In the goal of feature selection for clustering, Dy and Brodley (2004) compared maximum likelihood approaches with the maximization criterion of Equation (2), highlighting the interest of both in different contexts. Feature selection can be cast in the framework of dimension reduction where \mathbf{U} is forced to be a $(0, 1)$ -matrix with non-zero index in each columns indicating the subset of d selected variables (Nie et al. 2008). For clustering applications, Torre and Kanade (2006) proposed the discriminative cluster analysis, combining k -means and linear discriminant analysis. In a visualization approach, Scrucca (2010) proposed to project the data in a subspace minimizing the Fisher criterion, using the partition given by the model of Equation (1). It relies on a modified version of \mathbf{S}_B taking into account variability between within-class covariance, and demonstrates good visualization power. However, this method is post-inference and still requires to fit a GMM in the observation space which is prohibitive in real high-dimensional scenarios. Finally, Bouveyron and Brunet (2012a) proposed the discriminative latent mixture (DLM) model, which treats the estimation of the latent subspace as a separate problem from maximum likelihood estimation. The proposed model is closed to the mixture of common factor analyzers (MCFA, Baek et al. 2009), although inference is different and done via the Fisher-EM algorithm which mixes the EM strategy with a specific Fisher-step. In the latter, the current posterior membership probabilities are used to compute the scatter matrices \mathbf{S}_W and \mathbf{S}_B , and \mathbf{U} is supposed to be discriminant, maximizing Equation (2) with orthogonality constraints.

Albeit not directly related to the high-dimensional setting, the idea of incorporating clustering information is popular in the context of mixture modeling. It dates back to the CEM algorithm (Celeux and Govaert 1992) aiming at maximizing the classification likelihood for inference. It is also present in the integrated classification likelihood (ICL, Biernacki et al. 2000) criterion for selecting the number of clusters. This trade-off between clustering and density estimation also

led to the proposition of differentiating between the notion of mixture component and cluster in Gaussian mixtures (Baudry et al. 2010), the former being a combination of the latter.

1.3 Contribution and organization of the paper

This paper introduces a new algorithm for the clustering of high-dimensional data with a constrained Gaussian mixture model. In Section 2, we introduce a Bayesian formulation of the DLM model putting a prior distribution on the mean in the latent space, with a hyper-parameter λ controlling the between-class variance. Following Bouveyron and Brunet (2012a), we derive a family of submodels with constraints on the latent covariance matrices, and discuss its links to existing methods. Then, the posterior distribution now being intractable, Section 3 introduces a variational extension of the Fisher-EM algorithm for simultaneous clustering and dimension reduction. In Section 4, two carefully designed scenarios, controlling for dimension and noise, demonstrate the superiority of the corresponding Bayesian Fisher-EM over state-of-art model-based subspace clustering models. Traditional benchmarks of the litterature are also used to assess the performances on real data. Finally, Section 5 proposes an application to the problem of image denoising, demonstrating the interest of the proposed methodology compared to state-of-the-art approaches and paving way for future improvements.

2 The Bayesian discriminative latent mixture

2.1 Discriminative latent mixture

Bouveyron and Brunet (2012a) proposed the following generative model, relying on the idea that $K - 1$ properly chosen dimensions are sufficient to discriminate between K classes. It is based on a Factor-Analysis-like formulation where the latent scores $\mathbf{x}_i \in \mathbb{R}^d$ are low-dimensional representation of the observations \mathbf{y}_i , and are assumed to follow a Gaussian mixture model in the subspace:

$$\begin{aligned} \mathbf{z}_i &\sim \mathcal{M}_K(1, \boldsymbol{\pi}), \\ \mathbf{x}_i \mid \{z_{ik} = 1\} &\sim \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \\ \mathbf{y}_i &= \mathbf{U}\mathbf{x}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \mid \{z_{ik} = 1\} \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Psi}_k). \end{aligned} \quad (\text{DLM})$$

Here, the matrix \mathbf{U} is constrained to be column-orthonormal, $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_d$, and forms the basis of a low-dimensional subspace of dimension $d \leq \min(K - 1, p)$, which is called the *discriminative subspace*. When the latent variables are integrated out, the marginal distribution is a constrained GMM:

$$\mathbf{y}_i \sim \sum_{k=1}^K \pi_k \mathcal{N}_p(\mathbf{U}\boldsymbol{\mu}_k, \mathbf{U}\boldsymbol{\Sigma}_k\mathbf{U}^\top + \boldsymbol{\Psi}_k), \quad \mathbf{m}_k = \mathbf{U}\boldsymbol{\mu}_k, \quad \mathbf{S}_k = \mathbf{U}\boldsymbol{\Sigma}_k\mathbf{U} + \boldsymbol{\Psi}_k.$$

This model relates to the MCFA model (Baek et al. 2009), except that the latent dimension is constrained to be at most $K - 1$, and \mathbf{U} is considered to be discriminative in the sense of Fisher's criterion. Moreover, the noise matrices $\boldsymbol{\Psi}_k$ are not constrained to be common across clusters nor diagonal anymore, but rather to be isotropic in the orthogonal of the subspace. Formally, let us define $\mathbf{D} = [\mathbf{U}, \mathbf{V}]$ where $\mathbf{V} \in \mathbb{R}^{p \times (p-d)}$ is the orthogonal complement of \mathbf{U} in \mathbb{R}^p . Then, $\boldsymbol{\Psi}_k$ is assumed to respect:

$$\begin{aligned} \mathbf{V}^\top \boldsymbol{\Psi}_k \mathbf{V} &= \beta_k \mathbf{I}_{p-d}, \\ \mathbf{U}^\top \boldsymbol{\Psi}_k \mathbf{U} &= \mathbf{0}_{d \times d} \text{ and } \mathbf{V}^\top \boldsymbol{\Psi}_k \mathbf{U} = \mathbf{0}_{(p-d) \times d}. \end{aligned}$$

These constraints amount to say that the covariance matrix \mathbf{S}_k is block diagonal after being rotated by \mathbf{D} . In other terms, writing $\boldsymbol{\Delta}_k = \mathbf{D}^\top \mathbf{S}_k \mathbf{D}$, the DLM model assumes that:

$$\boldsymbol{\Delta}_k = \text{diag}(\boldsymbol{\Sigma}_k, \beta_k \mathbf{I}_{p-d}) = \left(\begin{array}{cc} \boxed{\boldsymbol{\Sigma}_k} & \mathbf{0} \\ \mathbf{0} & \boxed{\beta_k \mathbf{I}_{p-d}} \end{array} \right) \left. \begin{array}{l} \left. \vphantom{\begin{array}{cc} \boxed{\boldsymbol{\Sigma}_k} & \mathbf{0} \\ \mathbf{0} & \boxed{\beta_k \mathbf{I}_{p-d}} \end{array}} \right\} d \leq K - 1 \\ \left. \vphantom{\begin{array}{cc} \boxed{\boldsymbol{\Sigma}_k} & \mathbf{0} \\ \mathbf{0} & \boxed{\beta_k \mathbf{I}_{p-d}} \end{array}} \right\} (p - d) \end{array} \right\}.$$

This hypothesis implies that the discriminative subspace contains the relevant clustering information, while the noise variance lies in the orthogonal directions. Denoting $\boldsymbol{\mu} = (\boldsymbol{\mu}_k)$ and $\boldsymbol{\beta} = (\beta_k)$, the model parameters are then denoted

as $\vartheta = (\pi, \mu, \Sigma, U)$, and the Fisher-EM proposed by Bouveyron and Brunet (2012a) decomposes in two steps. First, an EM-step is used to maximize the observed-data log-likelihood with respect to $(\pi, \mu, \Sigma, \beta)$. The resulting posterior probabilities $\tau_{ik} = p(z_{ik} = 1 \mid \mathbf{y}_i, \vartheta)$ are then used to replace z_{ik} in order to compute the *soft* within and between-class covariance matrices \mathbf{S}_W and \mathbf{S}_B . The Fisher-step chooses the discriminative space U as the one maximizing the Fisher criterion in Equation (2) with orthonormality constraints. These two steps are iterated over until convergence of the likelihood or a maximum number of iteration is reached. However, while efficient, this algorithm displays some instabilities due to poor conditioning of the scatter matrices, and can get stuck in poor local maxima in terms of clustering as demonstrated in Section 4.

2.2 A Bayesian formulation and the family of submodels

We propose a Bayesian extension of the DLM model where a Gaussian prior distribution is put on μ_k as in the standard Bayesian Gaussian mixture model:

$$\begin{aligned} \mu &= (\mu_k)_k, \quad \mu_k \stackrel{i.i.d.}{\sim} \mathcal{N}_d(\nu, \lambda \mathbf{I}_d), \\ z_i &\sim \mathcal{M}_K(1, \pi), \\ \mathbf{x}_i \mid \{z_{ik} = 1\} &\sim \mathcal{N}_d(\mu_k, \Sigma_k), \\ \mathbf{y}_i &= \mathbf{U} \mathbf{x}_i + \epsilon_i, \quad \epsilon_i \mid \{z_{ik} = 1\} \sim \mathcal{N}_p(\mathbf{0}_p, \Psi_k). \end{aligned} \tag{BDLM}$$

Here, λ is a hyper-parameter controlling the spreading of the μ_k 's in the latent space. The rest of the model and assumptions is unchanged and we refer to this Bayesian version as $\text{BDLM}_{[\Sigma_k \beta_k]}$, which is represented as a graphical model in Figure 2. The set of parameters is then $\vartheta = (\pi, \Sigma, U, \beta)$ of dimension

$$\gamma = K - 1 + K \frac{d(d+1)}{2} + pd - \frac{d(d+1)}{2} + K,$$

and Section 3 discusses inference and clustering.

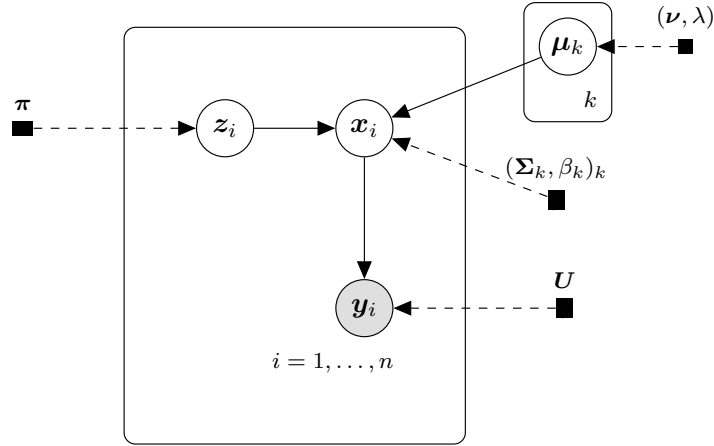


Figure 2: Graphical model representation of the Bayesian discriminative latent mixture model.

Considering specific constraints on the matrix Δ_k , we can derive a family of submodels for the BDLM as in the original DLM. Akin to the spectral constraints of Banfield and Raftery (1993), we can assume a combination of hypotheses on the structure of the latent space covariance Σ_k and on the noise covariance Ψ_k , for a total of 12 models. First homoscedasticity constraints of the type $\Sigma_k = \Sigma$ or $\Psi_k = \Psi$ may be considered, denoted as $\text{BDLM}_{[\Sigma \beta_k]}$, $\text{BDLM}_{[\Sigma_k \beta]}$, and $\text{BDLM}_{[\Sigma \beta]}$ where the subscript k denotes heteroscedasticity of the concerned parameter. Moreover, the covariance Σ_k can be further assumed to be diagonal $\Sigma_k = \text{diag}(\alpha_{k1}^2, \dots, \alpha_{kd}^2)$ leaving 4 possibilities depending on the homoscedasticity hypothesis denoted as $\text{BDLM}_{[\alpha_{kh} \beta_k]}$, $\text{BDLM}_{[\alpha_h \beta_k]}$, $\text{BDLM}_{[\alpha_{kh} \beta]}$ and $\text{BDLM}_{[\alpha_h \beta]}$. Finally, the latent covariance may be considered isotropic $\Sigma_k = \alpha_k^2 \mathbf{I}_d$, and the corresponding 4 submodels are noted as $\text{BDLM}_{[\alpha_k \beta_k]}$, $\text{BDLM}_{[\alpha \beta_k]}$, $\text{BDLM}_{[\alpha_k \beta]}$ and $\text{BDLM}_{[\alpha \beta]}$. A comprehensive summary of these submodels, along with their number of free parameters is given in Table 1.

Table 1: The BDLM family of submodels with their associated number of free parameters, along with related model-based subspace clustering models. Here, $\omega = K - 1 + pd - \frac{d(d+1)}{2}$ and the dimension of the latent space is fixed to $d = K - 1$ in the last examples. A comparison with other Gaussian subspace clustering models in the same setting is available in Bouveyron and Brunet-Saumard (2014b, Table 3 to 6)

Model	Number of free parameters γ	$p = 100, K = 4$
Full-GMM	$K - 1 + Kp + K \frac{p(p+1)}{2}$	20603
Sphe-GMM	$K - 1 + 2Kp$	803
BDLM $_{[\Sigma_k \beta_k]}$	$\omega + K \frac{d(d+1)}{2} + K$	325
BDLM $_{[\Sigma_k \beta]}$	$\omega + K \frac{d(d+1)}{2} + 1$	322
BDLM $_{[\Sigma \beta_k]}$	$\omega + \frac{d(d+1)}{2} + K$	307
BDLM $_{[\Sigma \beta]}$	$\omega + \frac{d(d+1)}{2} + 1$	304
BDLM $_{[\alpha_{kh} \beta_k]}$	$\omega + Kd + K$	313
BDLM $_{[\alpha_{kh} \beta]}$	$\omega + Kd + 1$	310
BDLM $_{[\alpha_h \beta_k]}$	$\omega + d + K$	304
BDLM $_{[\alpha_h \beta]}$	$\omega + d + 1$	301
BDLM $_{[\alpha_k \beta_k]}$	$\omega + K + K$	305
BDLM $_{[\alpha_k \beta]}$	$\omega + K + 1$	302
BDLM $_{[\alpha \beta_k]}$	$\omega + 1 + K$	302
BDLM $_{[\alpha \beta]}$	$\omega + 1 + 1$	299

2.3 Link with Gaussian subspace clustering models

As an extension of the DLM model, the BDLM model inherits its connections to the other model-based subspace clustering models. Indeed, they share the same *global* linear Gaussian model formulations, with a common loading matrix \mathbf{U} . The mixture of common factor analyzer (MCFA, Baek et al. 2009) model is the closest to the DLM model although the assumption on the nature of the subspace and the noise matrix Ψ_k are different. This model also shares deep ties with the heteroscedastic factor mixture analysis (HFMA, Montanari and Viroli 2010) where the latent scores again follow a mixture of Gaussian distributions and a common loading matrix. However, the linear transformation \mathbf{U} is not constrained to be orthonormal here, and other identifiability constraints are put on μ_k and Σ_k such that the scores are standardized in the latent space. The PGMM family of McNicholas and Murphy (2008) also contains 4 constrained models CUU, CCU, CUC and CCC, with common loadings across clusters $\mathbf{S}_k = \mathbf{U}\mathbf{U}^\top + \Psi_k$. Note that, in these models the matrix \mathbf{U} is not orthonormal and captures all the variance. As a consequence, the mixture is no longer in the latent space but in the observation one, which does not allow to put further constraints on Σ_k . Finally, Bouveyron et al. (2007) proposed a family of 28 constrained Gaussian mixtures for high-dimensional data, with the decomposition $\mathbf{S}_k = \mathbf{U}_k \text{diag}(\Lambda_k, \beta_k \mathbf{I}_{p-d_k}) \mathbf{U}_k^\top$ where $\Lambda_k = \text{diag}(\lambda_{k1}, \dots, \lambda_{kd_k})$. The submodels sharing the orientation matrices $\mathbf{U}_k = \mathbf{U}$ with same dimension $d_k = d$ are close to some BDLM submodels where Σ_k is diagonal. However, once again, the subspace are not assumed to be discriminant and are estimate through maximum likelihood estimation.

Related to our approach, a fully Bayesian extension of the MCFA model was proposed in Wei and Li (2013), putting a Dirichlet prior on the mixture proportions π , a standard Gaussian on each column of \mathbf{U} and a Gaussian-inverse-Wishart prior on (μ, Σ) . However, the marginal distribution of x_i in this model is now a mixture of Student t-distribution which differs from the mixture of Gaussian in our model. In addition, the factor loading matrix is no longer assumed to be column-orthonormal, and as in MCFA the subspace spanned by \mathbf{U} is not considered discriminant. The authors rely on a variational Bayes EM algorithm to approximate the posterior distribution of the parameters.

3 Clustering with the Bayesian Fisher-EM algorithm

In the following, we propose a clustering algorithm based on the joint maximization of the Fisher criterion and the observed-data likelihood. Contrary to the DLM model, the latter is intractable, and so is the posterior distribution of the latent variables $(\mathbf{Z}, \boldsymbol{\mu})$ given the data and the parameters. Thus, we propose to use variational inference, relying on a mean-field approximation of the posterior and the maximization of a lower bound of the log-likelihood (Jaakkola and Jordan 2000). First, we give the optimal form of the mean-field approximation in the variational E-step (VE-step). Then, we derive the expression of the variational lower bound as well as the M-step updates maximizing the latter with respect to the mixture parameters $(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Finally, following Bouveyron and Brunet (2012a), we propose to choose \mathbf{U} as the best current discriminative subspace maximizing the Fisher criterion at each iteration. Thus, the proposed clustering algorithm for BDLM is named Bayesian Fisher-EM (BFEM) and alternates between 3 steps:

- The VE-step which finds an approximation of the posterior $p(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{Y}, \boldsymbol{\vartheta})$ in the mean-field family,
- The M-step where the latent space mixture parameters are estimated by maximizing the variational lower bound,
- The F-step where \mathbf{U} is chosen to maximize the current variational Fisher criterion.

3.1 Variational approximation

Similarly to the Bayesian formulation of standard GMM, the observed-data likelihood is no longer tractable. Indeed, the latter is written as:

$$p(\mathbf{Y} \mid \boldsymbol{\vartheta}) = \int_{\boldsymbol{\mu}} p(\boldsymbol{\mu}) \prod_{i=1}^n \sum_{k=1}^K \pi_k \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{U} \boldsymbol{\mu}_k, \mathbf{S}_k) d\boldsymbol{\mu}. \quad (3)$$

Unfortunately, each $\boldsymbol{\mu}_k$ now appears in all n factors of the integrand. Thus, the integral does not reduce to products and sums of d -dimensional trivial integrals over $\boldsymbol{\mu}_k$. Another way of seeing the difficulty is to swap the integrals over \mathbf{Z} and $\boldsymbol{\mu}$, leaving:

$$p(\mathbf{Y} \mid \boldsymbol{\vartheta}) = \sum_{\mathbf{Z}} p(\mathbf{Z}) \prod_{k=1}^K \int_{\boldsymbol{\mu}_k} p(\boldsymbol{\mu}_k) \prod_{i: z_{ik}=1} \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{U} \boldsymbol{\mu}_k, \mathbf{S}_k) d\boldsymbol{\mu}_k.$$

Now each integral may be computed thanks to Gaussian conjugacy. However, there are K^n possible configurations to sum over, which is not computationally feasible.

Besides, the posterior $p(\mathbf{Z}, \boldsymbol{\mu} \mid \mathbf{Y})$ is not tractable either, and is only known up to its normalizing constant in Equation (3), which prevents from calculating its moments. This fact is well known in Bayesian treatment of mixtures, and leads to either MCMC algorithms or approximate inference (Fruhwirth-Schnatter et al. 2019, section 5.2). Here, we rely on the latter and introducing a *variational* distribution $q(\boldsymbol{\mu}, \mathbf{Z})$, the classical identity (see e.g. Blei et al. 2017, Equation (14)) holds for any q :

$$\log p(\mathbf{Y} \mid \boldsymbol{\vartheta}) = \mathcal{J}(q, \boldsymbol{\vartheta}) + \text{KL}(q \parallel p(\cdot \mid \mathbf{Y}, \boldsymbol{\vartheta})) \geq \mathcal{J}(q, \boldsymbol{\vartheta}), \quad (4)$$

with

$$\mathcal{J}(q, \boldsymbol{\vartheta}) = \mathbb{E}_q [\log p(\mathbf{Y}, \boldsymbol{\mu}, \mathbf{Z} \mid \boldsymbol{\vartheta})] - \mathbb{E}_q [\log q(\boldsymbol{\mu}, \mathbf{Z})]. \quad (5)$$

Here, KL denotes the Kullback-Leibler divergence between the variational distribution q and the posterior $p(\cdot \mid \mathbf{Y}, \boldsymbol{\vartheta})$:

$$\text{KL}(q \parallel p(\cdot \mid \mathbf{Y}, \boldsymbol{\vartheta})) = - \int_{\boldsymbol{\mu}} \sum_{\mathbf{Z}} q(\boldsymbol{\mu}, \mathbf{Z}) \log \frac{p(\boldsymbol{\mu}, \mathbf{Z} \mid \mathbf{Y}, \boldsymbol{\vartheta})}{q(\boldsymbol{\mu}, \mathbf{Z})} d\boldsymbol{\mu}.$$

The latter being non-negative, Equation (5) is a lower bound of the observed-data log-likelihood in the left-hand side of Equation (4). Moreover, maximizing the bound with respect to q is equivalent to minimizing the KL divergence of the latter to the posterior, the bound being tight when the two distributions are equal. However, the posterior being intractable, the KL minimization problem is restricted to a certain family of distributions, hence the *approximation* terminology. Here, we propose to rely on the classical mean-field variational family which posits a fully factorized distribution:

$$q(\boldsymbol{\mu}, \mathbf{Z}) = \prod_{k=1}^K q(\boldsymbol{\mu}_k) \prod_{i=1}^n q(\mathbf{z}_i). \quad (6)$$

The main interest of the mean-field approximation is that it comes with the coordinate ascent variational inference (CAVI, Bishop 2006; Blei et al. 2017) algorithm. The latter consists in sequentially optimizing the lower bound with respect to each individual distribution $q(\mathbf{z}_i)$ and $q(\boldsymbol{\mu}_k)$ while keeping the other fixed, and to cycle over until convergence is reached like a fixed point algorithm.

The following propositions give the optimal form of the CAVI updates for the BDLM model. Notice that, while no specific functional forms were assumed on q , the optimal distributions $q^*(\mathbf{z}_i)$ and $q^*(\boldsymbol{\mu}_k)$ happen to be multinomial and Gaussian as it is often the case with mean-field variational inference (Blei et al. 2017). In the VE-step, the updates of Equations (7) and (8) are cycled over until a local maximum of $\mathcal{J}(q)$ is reached.

Proposition 1 (Proof in Appendix A.1). *The coordinate update for the variational distribution $q(\mathbf{z}_i)$ is*

$$q^*(\mathbf{z}_i) = \mathcal{M}_K(\mathbf{z}_i \mid 1, \boldsymbol{\tau}_i), \quad (7)$$

with $\forall i, k$,

$$\tau_{ik} \propto \pi_k \exp \left\{ \mathbb{E}_{q^*(\boldsymbol{\mu}_k)} [\log \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{U} \boldsymbol{\mu}_k, \mathbf{S}_k)] \right\}.$$

Proposition 2 (Proof in Appendix A.2). *The coordinate update for the variational distribution $q(\boldsymbol{\mu}_k)$ is*

$$q^*(\boldsymbol{\mu}_k) = \mathcal{N}_d(\boldsymbol{\mu}_k \mid \tilde{\boldsymbol{\mu}}_k, \tilde{\mathbf{M}}_k), \quad (8)$$

with $\forall k$,

$$\begin{aligned} \tilde{n}_k &= \sum_i \tau_{ik}, \\ \tilde{\mathbf{M}}_k &= \left(\lambda^{-1} \mathbf{I}_d + \tilde{n}_k \boldsymbol{\Sigma}_k^{-1} \right)^{-1}, \\ \tilde{\boldsymbol{\mu}}_k &= \boldsymbol{\nu} + \tilde{\mathbf{M}}_k \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{U}^\top \left(\sum_i \tau_{ik} \mathbf{y}_i \right) - \tilde{n}_k \boldsymbol{\nu} \right). \end{aligned}$$

Finally, note that the expression of τ_{ik} in Equation (7) involves an expectation in the observation space that may be reworked in order to avoid inverting the $p \times p$ matrix \mathbf{S}_k . This is done by taking advantage of the specific block structure of the latter, hence only relying on the inverse of the $d \times d$ matrix $\boldsymbol{\Sigma}_k$:

$$\begin{aligned} \tau_{ik} &\propto \pi_k \exp \left\{ \mathbb{E}_{\boldsymbol{\mu}_k} [\log \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{U} \boldsymbol{\mu}_k, \mathbf{S}_k)] \right\}, \\ &\propto \pi_k \exp \left\{ -\frac{1}{2} \left(p \log(2\pi) + \log |\mathbf{S}_k| + (\mathbf{y}_i - \mathbf{U} \tilde{\boldsymbol{\mu}}_k)^\top \mathbf{U} \boldsymbol{\Sigma}_k^{-1} \mathbf{U}^\top (\mathbf{y}_i - \mathbf{U} \tilde{\boldsymbol{\mu}}_k) \right. \right. \\ &\quad \left. \left. + \frac{1}{\beta_k} (\|\mathbf{y}_i\|_2 - \|\mathbf{U}^\top \mathbf{y}_i\|_2) + \text{Tr} [\tilde{\mathbf{M}}_k \boldsymbol{\Sigma}_k^{-1}] \right) \right\}, \end{aligned} \quad (9)$$

and

$$\log |\mathbf{S}_k| = \log |\boldsymbol{\Sigma}_k| + (p - d) \log(\beta_k).$$

3.2 The M-step

In the M-step, the bound of Equation (5) is maximized with respect to the latent space mixture parameters $(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Note that \mathbf{U} is treated as a fixed, distinct parameter here, which will be dealt with in the next section. The following proposition gives the form of the lower bound as a function of $\boldsymbol{\vartheta}$.

Proposition 3 (Proof in Appendix A.3). *In model BDLM $_{[\boldsymbol{\Sigma}_k \beta_k]}$, the variational lower bound as a function of $\boldsymbol{\vartheta}$ may be written as:*

$$\begin{aligned} \mathcal{J}(\boldsymbol{\vartheta}) &= \text{const} - \frac{1}{2} \sum_{k=1}^K \tilde{n}_k \left\{ -2 \log(\pi_k) + \log |\boldsymbol{\Sigma}_k| + (p - d) \log(\beta_k) \right. \\ &\quad \left. + \text{Tr} [\boldsymbol{\Sigma}_k^{-1} \mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U}] + \frac{1}{\beta_k} \left(\text{Tr} [\hat{\mathbf{C}}_k] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U}] \right) \right\}, \end{aligned} \quad (10)$$

where

$$\hat{\mathbf{C}}_k = \frac{1}{\tilde{n}_k} \sum_{i=1}^n \tau_{ik} (\mathbf{y}_i - \mathbf{U} \tilde{\boldsymbol{\mu}}_k) (\mathbf{y}_i - \mathbf{U} \tilde{\boldsymbol{\mu}}_k)^\top + \mathbf{U} \tilde{\mathbf{M}}_k \mathbf{U}^\top.$$

At iteration (t) , in the M-step, the mixture proportions are estimated classically as in other mixture models:

$$\hat{\pi}_k^{(t)} = \frac{\tilde{n}_k^{(t)}}{n}. \quad (11)$$

The remaining parameters (Σ_k, β_k) depend on the chosen submodel and the following proposition details the estimates corresponding to each of the 12 submodels.

Proposition 4 (Proof in Appendix A.4). *The M-step estimates for Σ_k and β_k at iteration (t) are:*

- *Model* BDLM $_{[\Sigma_k \beta_k]}$:

$$\hat{\Sigma}_k^{(t)} = \mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U}, \quad \hat{\beta}_k^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}_k^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U}]}{p - d}. \quad (12)$$

- *Model* BDLM $_{[\Sigma_k \beta]}$:

$$\hat{\Sigma}_k^{(t)} = \mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U}, \quad \hat{\beta}^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U}]}{p - d}. \quad (13)$$

- *Model* BDLM $_{[\Sigma \beta_k]}$:

$$\hat{\Sigma}^{(t)} = \mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U}, \quad \hat{\beta}_k^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}_k^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U}]}{p - d}. \quad (14)$$

- *Model* BDLM $_{[\Sigma \beta]}$:

$$\hat{\Sigma}^{(t)} = \mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U}, \quad \hat{\beta}^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U}]}{p - d}. \quad (15)$$

- *Model* BDLM $_{[\alpha_{kh} \beta_k]}$:

$$\hat{\alpha}_{kh}^{(t)} = \mathbf{u}_h^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{u}_h, \quad \hat{\beta}_k^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}_k^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U}]}{p - d}. \quad (16)$$

- *Model* BDLM $_{[\alpha_{kh} \beta]}$:

$$\hat{\alpha}_{kh}^{(t)} = \mathbf{u}_h^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{u}_h, \quad \hat{\beta}^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U}]}{p - d}. \quad (17)$$

- *Model* BDLM $_{[\alpha_h \beta_k]}$:

$$\hat{\alpha}_h^{(t)} = \mathbf{u}_h^\top \hat{\mathbf{C}}^{(t)} \mathbf{u}_h, \quad \hat{\beta}_k^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}_k^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U}]}{p - d}. \quad (18)$$

- *Model* BDLM $_{[\alpha_h \beta]}$:

$$\hat{\alpha}_h^{(t)} = \mathbf{u}_h^\top \hat{\mathbf{C}}^{(t)} \mathbf{u}_h, \quad \hat{\beta}^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U}]}{p - d}. \quad (19)$$

- *Model* BDLM $_{[\alpha_k \beta_k]}$:

$$\hat{\alpha}_k^{(t)} = \frac{1}{d} \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U}], \quad \hat{\beta}_k^{(t)} = \frac{\text{Tr} [\hat{\mathbf{C}}_k^{(t)}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U}]}{p - d}. \quad (20)$$

- *Model* BDL $M_{[\alpha_k\beta]}$:

$$\hat{\alpha}_k^{(t)} = \frac{1}{d} \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U} \right], \quad \hat{\beta}^{(t)} = \frac{\text{Tr} \left[\hat{\mathbf{C}}^{(t)} \right] - \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U} \right]}{p-d}. \quad (21)$$

- *Model* BDL $M_{[\alpha\beta_k]}$:

$$\hat{\alpha}^{(t)} = \frac{1}{d} \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U} \right], \quad \hat{\beta}_k^{(t)} = \frac{\text{Tr} \left[\hat{\mathbf{C}}_k^{(t)} \right] - \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}_k^{(t)} \mathbf{U} \right]}{p-d}. \quad (22)$$

- *Model* BDL $M_{[\alpha\beta]}$:

$$\hat{\alpha}^{(t)} = \frac{1}{d} \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U} \right], \quad \hat{\beta}^{(t)} = \frac{\text{Tr} \left[\hat{\mathbf{C}}^{(t)} \right] - \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}^{(t)} \mathbf{U} \right]}{p-d}. \quad (23)$$

Here, \mathbf{u}_h denotes the h -th column of \mathbf{U} which is computed in the F -step at iteration (t) and:

$$\hat{\mathbf{C}}_k^{(t)} = \frac{1}{\tilde{n}_k^{(t)}} \sum_{i=1}^n \tau_{ik}^{(t)} (\mathbf{y}_i - \mathbf{U} \tilde{\boldsymbol{\mu}}_k^{(t)}) (\mathbf{y}_i - \mathbf{U} \tilde{\boldsymbol{\mu}}_k^{(t)})^\top + \mathbf{U} \tilde{\mathbf{M}}_k^{(t)} \mathbf{U}^\top, \quad (24)$$

$$\hat{\mathbf{C}}^{(t)} = \frac{1}{n} \sum_{k=1}^K \tilde{n}_k^{(t)} \hat{\mathbf{C}}_k^{(t)}. \quad (25)$$

3.3 The Fisher step

As explained above, the subspace \mathbf{U} is supposed to be discriminative in the sense of the Fisher criterion. The partition \mathbf{Z} being unknown, the scatter matrices in Equation (2) cannot be formed. Following Bouveyron and Brunet (2012a) we propose to replace them by the soft within and between-class scatter matrices:

$$\begin{aligned} \tilde{\mathbf{m}}_k^{(t)} &= \frac{1}{\tilde{n}_k^{(t)}} \sum_{i=1}^n \tau_{ik}^{(t)} \mathbf{y}_i, \\ \tilde{\mathbf{S}}_W^{(t+1)} &= \frac{1}{n} \sum_{k=1}^K \frac{1}{\tilde{n}_k^{(t)}} \sum_{i=1}^n \tau_{ik}^{(t)} \left(\mathbf{y}_i - \tilde{\mathbf{m}}_k^{(t)} \right) \left(\mathbf{y}_i - \tilde{\mathbf{m}}_k^{(t)} \right)^\top, \\ \tilde{\mathbf{S}}_B^{(t+1)} &= \frac{1}{n} \sum_{k=1}^K \tilde{n}_k^{(t)} \left(\tilde{\mathbf{m}}_k^{(t)} - \bar{\mathbf{y}} \right) \left(\tilde{\mathbf{m}}_k^{(t)} - \bar{\mathbf{y}} \right)^\top. \end{aligned}$$

Note that these matrices only involve the variational distribution of \mathbf{Z} , although the latter also depends on $q^{(t)}(\boldsymbol{\mu})$ through the fixed point algorithm of the VE-step. Moreover, we recover the classical identity of linear discriminant analysis $\mathbf{S}_T = \mathbf{S}_W^{(t)} + \mathbf{S}_B^{(t)}$ at any iteration (t) , where $\mathbf{S}_T = (1/n) \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})^\top$ is the sample covariance matrix which does not depend on the clustering and is constant throughout the algorithm.

Then, \mathbf{U} is supposed to maximize the following criterion:

$$\mathbf{U}^{(t)} = \arg \max_{\mathbf{U}^\top \mathbf{U} = \mathbf{I}_d} F(\mathbf{U}) = \text{Tr} \left[(\mathbf{U}^\top \mathbf{S}_T \mathbf{U})^{-1} \mathbf{U}^\top \tilde{\mathbf{S}}_B^{(t)} \mathbf{U} \right]. \quad (26)$$

This criterion is slightly different from the one in Equation (2) since $\tilde{\mathbf{S}}_W^{(t)}$ has been replaced by \mathbf{S}_T . Working this this criterion is justified using the identity above since the problems of minimizing $\text{Tr}[\mathbf{U}^\top \tilde{\mathbf{S}}_W \mathbf{U}]$ or $\text{Tr}[\mathbf{U}^\top \mathbf{S}_T \mathbf{U}]$ are the same (Fukunaga 1990, chap. 10). It is often used in practice (Ye 2005), and computationally efficient in this case since \mathbf{S}_T and its inverse need to be computed only once at the beginning of the algorithm.

Without the orthonormality constraints, the problem in Equation (26) is directly solved by taking the leading d eigenvectors of the generalized eigenvalue problem $\tilde{\mathbf{S}}_B^{(t)} \mathbf{u}_h = \gamma_h \mathbf{S}_T \mathbf{u}_h$ (Ghojogh et al. 2019). If \mathbf{S}_T is invertible, this can be done by computing the d leading eigenvectors of $\mathbf{S}_T^{-1} \tilde{\mathbf{S}}_B^{(t)}$. However, since $\mathbf{S}_T^{-1} \tilde{\mathbf{S}}_B$ is not necessarily symmetric,

the solution is not orthonormal with respect to the regular scalar product, but rather verifies $U^\top S_T U = I_d$. Unfortunately, there is no direct solution for the problem of Equation (26) with the constraint $U^\top U = I_d$. In the supervised context, algorithms have been derived to solve this problem which is called orthogonal LDA (OLDA). Foley and Sammon (1975) proposed an iterative algorithm to successively find $\mathbf{u}_1, \dots, \mathbf{u}_d$ in the 2-class problem. It was later generalized for arbitrary values of K by Okada and Tomita (1985), and coined the orthonormal discriminant vectors (ODV) by Hamamoto et al. (1991). Note that simultaneous algorithms also exist to optimize with respect to U , based on successive eigen and QR-decompositions of carefully designed matrices (see. Ye 2005; Lu et al. 2016).

Relying on the ODV method, Bouveyron and Brunet (2012a) proposed an iterative algorithm starting from \mathbf{u}_1 , the leading eigenvector of $S_T^{-1} \tilde{S}_B^{(t)}$, and greedily maximizing the criterion by computing the r -th direction as the solution of the unconstrained problem in the orthogonal of the current subspace $\mathcal{B}_{r-1} = \text{vect}(\mathbf{u}_1, \dots, \mathbf{u}_{r-1})$. An orthogonal basis $V_r = (\mathbf{v}_r, \dots, \mathbf{v}_p)$ of \mathcal{B}_{r-1}^\perp can be found by the Gram-Schmidt procedure:

$$\mathbf{v}_l = \alpha_l (\mathbf{I}_p - \sum_{l'=1}^{l-1} \mathbf{v}_{l'} \mathbf{v}_{l'}^\top) \psi_l, \quad \forall l = r, \dots, p.$$

where $\mathbf{v}_l = \mathbf{u}_l$ for $l = 1, \dots, r-1$, α_l is a normalization constant such that $\|\mathbf{v}_l\|_2 = 1$ and ψ_l are linearly independent vectors of $\mathbf{u}_1, \dots, \mathbf{u}_{r-1}$. Then, the matrix $P_r = (\mathbf{v}_r, \dots, \mathbf{v}_p)$ is used to project the scatter matrix in the orthogonal subspace \mathcal{B}_{r-1}^\perp :

$$\begin{aligned} S_{Tr} &= P_r^\top S_T P_r, \\ \tilde{S}_{Br}^{(t)} &= P_r^\top \tilde{S}_B^{(t)} P_r. \end{aligned}$$

Finally, the leading eigenvector \mathbf{a}_r of the generalized eigenvalue problem $\tilde{S}_{Br}^{(t)} \mathbf{a}_r = \gamma_r S_{Tr} \mathbf{a}_r$ is computed, and the r -th discriminant vector is chosen as

$$\mathbf{u}_r = \frac{P_r \mathbf{a}_r}{\|\mathbf{a}_r\|_2}. \quad (27)$$

Thus, \mathbf{u}_r meets the constraints $\mathbf{u}_r^\top \mathbf{u}_h = 0, \forall h < r$. This iterative procedure is repeated until $r = d$ discriminant vectors are found.

3.4 Estimation of the hyper-parameters ν and λ

The hyper-parameters (λ, ν) may be set by the user and kept fixed during the whole procedure. For instance, when the data is centered, $\bar{\mathbf{y}} = 0$, then $\bar{\mathbf{x}} = U^\top \bar{\mathbf{y}} = 0$ thus ν could be set to $\mathbf{0}_d$. However, λ controls the variance of μ_k and setting it by hand can lead to poor performances. On the one hand, a too small value would not allow the space to be discriminant. On the other hand, when $\lambda \rightarrow +\infty$, the prior becomes non-informative. A quick asymptotic analysis of the variational distribution $q(\mu_k)$ of Proposition 2 confirms this as:

$$\tilde{\mathbf{M}}_k \xrightarrow{\lambda \rightarrow +\infty} \frac{1}{\tilde{n}_k} \Sigma_k, \quad \tilde{\mu}_k \xrightarrow{\lambda \rightarrow +\infty} \Sigma_k \Sigma_k^{-1} \frac{1}{\tilde{n}_k} \sum_i \tau_{ik} U^\top \mathbf{y}_i = \hat{\mu}_k^{DLM}.$$

Thus, the variational posterior mean becomes the maximum-likelihood estimate of μ_k in the frequentist formulation of BDLM. Under the hypothesis that $\tilde{n}_k \rightarrow +\infty$ as $n \rightarrow +\infty$, the variational approximation of the posterior becomes a Dirac mass at $\hat{\mu}_k^{DLM}$. This fact is somewhat similar to the well-known behavior of the posterior in Bayesian formulations of ridge regression when the prior becomes vague (Bishop 2006, p. 153).

Here, we propose a parametric empirical Bayes approach (Morris 1983), using the variational bound as a surrogate for the type-II likelihood as it is commonly done in other well-known hierarchical Bayesian models (Blei et al. 2003; Airolidi et al. 2008). The following proposition gives the form of the empirical Bayes estimates $(\hat{\nu}, \hat{\lambda})$ maximizing $\mathcal{J}(\nu, \lambda) \leq \log p(\mathbf{Y} | \vartheta, \nu, \lambda)$.

Proposition 5 (Proof in Appendix A.5). *The following updates maximize the variational lower bound with respect to (ν, λ) :*

$$\hat{\nu} = \frac{\sum_{k=1}^K \tilde{\mu}_k}{K}, \quad (28)$$

$$\hat{\lambda} = \frac{\sum_{k=1}^K \|\tilde{\mu}_k - \hat{\nu}\|_2^2 + \text{Tr} [\tilde{\mathbf{M}}_k]}{dK} \quad (29)$$

3.5 Stopping criterion and properties

Starting from a subspace $\mathbf{U}^{(0)}$, the BFEM algorithm iterates over the VE-step, M-step and F-step updates, in this order. The algorithm is described in Algorithm 1 and this section discusses initialization, convergence and useful properties of the algorithm. Let us begin by discussing the link to the original FEM algorithm.

Link to the original Fisher-EM The proposed algorithm is largely inspired by the original FEM algorithm. However, note that the M-step updates of Bouveyron and Brunet (2012a) use:

$$\tilde{\mathbf{C}}_k = \frac{1}{\tilde{n}_k} \sum_{i=1}^n \tau_{ik} (\mathbf{y}_i - \tilde{\mathbf{m}}_k) (\mathbf{y}_i - \tilde{\mathbf{m}}_k)^\top, \quad \text{with: } \tilde{\mathbf{m}}_k = \frac{1}{\tilde{n}_k} \sum_{i=1}^n \tau_{ik} \mathbf{y}_i. \quad (30)$$

The latter does not exactly correspond to the matrix $\hat{\mathbf{C}}_k$ of Propositions 3 and 4, since it uses $\tilde{\mathbf{m}}_k$ instead of $\hat{\mathbf{m}}_k = \mathbf{U} \hat{\boldsymbol{\mu}}_k$. In particular, this has the consequence that the matrix $\tilde{\mathbf{C}}_k$ does not directly depend on \mathbf{U} , whereas it does in $\hat{\mathbf{C}}_k$. Therefore, our algorithm computes the true optimal updates in the M-step, while the FEM algorithm relies on the approximation $\tilde{\mathbf{C}}_k \approx \hat{\mathbf{C}}_k$.

Convergence and stopping criterion Since the F-step does not maximize the variational bound with respect to \mathbf{U} , the latter is no longer monotonically increasing. This is also the case for the original FEM algorithm, hence we propose to rely on the same stopping criterion: Aitken’s accelerated criterion (McLachlan and Krishnan 2007, p. 145). The latter was introduced as an acceleration method for EM when the sequence of likelihood is linearly converging. Here, we replace the likelihood sequence with the variational bound $\{\mathcal{J}^{(t)}\}_t$. Then, define the Aitken accelerated estimate of \mathcal{J}^* is defined for $t \geq 2$ as:

$$l_A^{(t+1)} = \mathcal{J}^{(t)} + \frac{1}{1 - c^{(t+1)}} \left(\mathcal{J}^{(t+1)} - \mathcal{J}^{(t)} \right), \quad \text{with: } c^{(t+1)} = \frac{\mathcal{J}^{(t+1)} - \mathcal{J}^{(t)}}{\mathcal{J}^{(t)} - \mathcal{J}^{(t-1)}}. \quad (31)$$

The stopping criterion is defined as $|l_A^{(t+1)} - l_A^{(t)}| < \epsilon$, where ϵ is a user-defined tolerance parameter. Since there is no guarantee that the sequence $\{\mathcal{J}^{(t)}\}_t$ is increasing here, a maximum number of iterations is also provided by the user as an alternative stopping criterion, which is always done in standard implementation of VEM algorithms anyway.

Another possible stopping condition is the absolute change of the Fisher criterion of Equation (26) between two successive F-step: $|\mathbf{F}(\mathbf{U}^{(t+1)}) - \mathbf{F}(\mathbf{U}^{(t)})| / |\mathbf{F}(\mathbf{U}^{(t)})|$. The latter was shown to have good performance for clustering applications (Bouveyron and Brunet 2012b).

Initialization The BFEM algorithm needs an initial variational distribution $q^{(0)}$ defined by its starting variational parameters $(\boldsymbol{\tau}, \tilde{\boldsymbol{\mu}}, \tilde{\mathbf{M}})$, an initial set of parameters $(\boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\pi})$ and an initial subspace $\mathbf{U}^{(0)}$. We recommend initializing by setting $\boldsymbol{\tau}^{(0)} = \mathbf{Z}^{(0)}$, a partition obtained by any suitable clustering algorithm, *e.g.* random or k -means partitions. Then, the matrix $\tilde{\mathbf{S}}_B^{(0)}$ can be formed to solve the problem in Equation (26), giving an initial $\mathbf{U}^{(0)}$. Next, the initial parameters $(\boldsymbol{\Sigma}, \boldsymbol{\beta}, \boldsymbol{\pi})$ are obtained by using the frequentist M-step of Bouveyron and Brunet (2012a) with $\mathbf{U}^{(0)}$. The remaining variational parameters $(\tilde{\boldsymbol{\mu}}_k, \tilde{\mathbf{M}}_k)$ can then be set using Proposition 2 with $\boldsymbol{\tau}^{(0)}$, $\mathbf{U}^{(0)}$ and $\boldsymbol{\vartheta}^{(0)}$. As for the hyper-parameters, we initialize $\boldsymbol{\nu}$ as $\boldsymbol{\nu}^{(0)} = (1/n) \sum_{i=1}^n \mathbf{U}^{(0)\top} \mathbf{y}_i$, and set $\lambda^{(0)} = 10^3$ as a vague prior for the first iteration, which is refined by empirical Bayes estimation throughout the algorithm.

Naturally, as in every algorithm with non-convex objective, the procedure can fall into poor local maxima of the bound. Thus, we recommend several restarts with different initializations. In the experiments of Section 4, we try several k -means initializations and take the one achieving the greatest variational lower bound.

Computational complexity The ODV method necessitates $d(p-1)$ Gram-Schmidt operations overall, and d generalized eigenvalue problems to solve, although only the first leading eigenvector needs to be found which can be done efficiently (Ge et al. 2016). Since $d \leq K-1$ is supposed to be small compared to p , this is not too computationally expensive. Moreover, this has to be compared to the computational cost of maximizing the lower bound of Proposition 3 with respect to \mathbf{U} as in a traditional M-step. Indeed, there is no closed-form solution for this problem and relying on gradient descent can rapidly become cumbersome since it necessitates relying on the steepest descent in the Stiefel manifold $St(p, d) = \{\mathbf{U} \in \mathbb{R}^{p \times d}, \mathbf{U}^\top \mathbf{U} = \mathbf{I}_d\}$.

Still, Section 3.7 introduces an alternative Fisher criterion which only necessitates performing one singular value decomposition at each step (t) . The empirical computing time of the BFEM algorithm and competing methods are shown in Figure 8 on page 19 of the experimental settings of Section 4.

Algorithm 1: Pseudo code of the BFEM algorithm

Data: \mathbf{Y}
Result: A clustering \mathbf{Z} and a discriminative subspace \mathbf{U}
Input: $K, \mathbf{Z}^{(0)}, \epsilon_{VE}, \epsilon_M, T_M, T_{VE}$, F-procedure, $\lambda^{(0)}$

// Initialization
Set $\boldsymbol{\tau} \leftarrow \mathbf{Z}^{(0)}$
Compute $\mathbf{S}_T, \tilde{\mathbf{S}}_B^{(0)}$ and subspace \mathbf{U} with F-procedure
Compute initial parameters $(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ with frequentist M-step
Compute variational parameters $(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{M}})$ with Proposition 2
Set $\boldsymbol{\nu} \leftarrow \frac{1}{n} \sum_{i=1}^n \mathbf{U}^\top \mathbf{y}_i$ and $\lambda \leftarrow \lambda^{(0)}$
Set $L[0] \leftarrow \mathcal{J}(\boldsymbol{\vartheta}, q, \boldsymbol{\nu}, \lambda)$

// Optimization
for $t \leftarrow 1$ **to** T_M **do**
 // F-step
 Compute $\tilde{\mathbf{S}}_B^{(t)}$
 Update \mathbf{U} with F-procedure
 // VE-step (fixed point algorithm)
 for $v \leftarrow 1$ **to** T_{VE} **do**
 Set temp $\leftarrow \mathcal{J}(q)$
 Update $\boldsymbol{\tau}$ using Proposition 1
 Update $(\tilde{\boldsymbol{\mu}}, \tilde{\mathbf{M}})$ using Proposition 2
 if $|(\mathcal{J}(q) - \text{temp}) / \mathcal{J}(q)| < \epsilon_{VE}$ **then Break;**
 end
 // M-step
 Update $(\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ with Equation (11) and Proposition 4
 // Empirical Bayes
 Update $(\boldsymbol{\nu}, \lambda)$ with Proposition 5
 // Compute the variational lower bound
 $L[t] \leftarrow \mathcal{J}(\boldsymbol{\vartheta}, q, \boldsymbol{\nu}, \lambda)$
 if $t \geq 2$ **then**
 $c \leftarrow \frac{L[t] - L[t-1]}{L[t-1] - L[t-2]}$
 aitken[t] $\leftarrow L[t-1] + \frac{1}{1-c} (L[t] - L[t-1])$
 if $| \text{aitken}[t] - \text{aitken}[t-1] | < \epsilon_M$ **then Break;**
 end
end
Return $q, \boldsymbol{\vartheta} = (\mathbf{U}, \boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$

3.6 Model selection

Choosing the latent space dimension As in the supervised case, the rank of $\tilde{\mathbf{S}}_B^{(t)}$ is at most $K-1$, hence $d \leq K-1$. We recommend setting it to $d = K-1$ for inference, as it is preferable to have redundant information than to lose discriminant directions. This presents the advantage to report the problem of selecting d to the one of selecting K .

Moreover, following Okada and Tomita (1985), the discriminant vectors found by the ODV procedure may be ordered according to the value of their 1-dimensional fisher criterion

$$F(\mathbf{u}_r) = \frac{\mathbf{u}_r^\top \tilde{\mathbf{S}}_B^{(t)} \mathbf{u}_r}{\mathbf{u}_r^\top \mathbf{S}_T \mathbf{u}_r} = \gamma_r,$$

where γ_r is the largest eigenvalue of the r -th problem solved, leaving:

$$F(\mathbf{u}_1) \geq \dots \geq F(\mathbf{u}_d). \quad (32)$$

Thus, for visualization purpose, we can choose $d = 2$ or $d = 3$, and project the data onto the corresponding subspace. Another solution, as is commonly done in PCA, is to show several combinations of discriminant axes, for instance with a matrix of 2-dimensional scatter plots.

Choosing the number of clusters In a clustering perspective, we propose to rely on the integrated classification likelihood (ICL, Biernacki et al. 2000) to choose K and the submodel. Recall that the criterion of Biernacki et al. (2000) is defined as:

$$\text{ICL}_{BIC}(\mathcal{M}, K) = \log p(\mathbf{Y}, \hat{\mathbf{Z}} \mid \hat{\boldsymbol{\vartheta}}, \mathcal{M}, K) - \frac{\gamma_{\mathcal{M}, K}}{2} \log(n), \quad (33)$$

where we take $\hat{\boldsymbol{\vartheta}}$ to be the parameter estimates at the end of BFEM. Although the marginal likelihood is intractable as explained in Section 3.1, the first term above is the classification likelihood which is tractable in the BDLM models. Actually, it can be computed with the variational lower bound \mathcal{J} , replacing τ_{ik} by \hat{z}_{ik} in the formulas of Propositions 1 and 2. A detailed proof of this fact is given in Appendix A.6.

3.7 An alternative Fisher criterion

The ODV method proposed to maximize the criterion in Equation (26) can be viewed as a greedy method, sequentially solving 1-dimensional problems $\mathbf{u}_r = \arg \max_{\mathbf{u}} F(\mathbf{u})$ under the growing set of constraints $\{\mathbf{u}^\top \mathbf{u}_h = 0, \forall h < r\}$. Such a method is not guaranteed to globally converge as emphasized in Hamamoto et al. (1991). Moreover, relying on the Gram-Schmidt procedure can sometimes lead to numerical instabilities in the BFEM algorithm. Thus, Bouveyron and Brunet (2011) proposed an alternative Fisher criterion, searching for the orthogonal projection matrix $\mathbf{U} \in \mathbb{R}^{p \times d}$ minimizing the following reconstruction error:

$$\mathbf{U}^{(t)} = \arg \min_{\mathbf{U}^\top \mathbf{U} = \mathbf{I}_d} \|\mathbf{S}_T^{-1} \tilde{\mathbf{S}}_B^{(t)} - \mathbf{U} \mathbf{U}^\top \mathbf{S}_T^{-1} \tilde{\mathbf{S}}_B^{(t)}\|_F^2 \quad (34)$$

This optimization problem has a somewhat PCA like flavor except that the matrix we wish to reconstruct is not the original data \mathbf{Y} but the measure of class separability: $\mathbf{S}_T^{-1} \tilde{\mathbf{S}}_B^{(t)}$. The classical results still holds, and the optimal $\mathbf{U}^{(t)}$ is given as the leading d left singular-vectors of $\mathbf{S}_T^{-1} \tilde{\mathbf{S}}_B^{(t)} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{V}$. Note that, since the product of these two symmetric matrices is not symmetric, the singular value decomposition is different from its spectral decomposition.

This modified F-step can be used to replace the ODV procedure at step (t) . Thus, we only have to perform a partial singular-value decomposition (SVD) on $\mathbf{S}_T^{-1} \tilde{\mathbf{S}}_B^{(t)}$ at each step. Since \mathbf{S}_T^{-1} is computed only once, this is particularly efficient.

4 Numerical experiments

This section compares the performance of different subspace clustering on simulated and classical real data benchmarks. We considered 6 different algorithms:

1. The proposed BFEM algorithm, with the ODV procedure described in Section 3.3 for the F-step. The results are also displayed for the alternative Fisher criterion of Section 3.7, which we refer to as the SVD procedure,
2. The Fisher-EM algorithm of Bouveyron and Brunet (2012a), implemented in the **FisherEM** R Package (Bouveyron et al. 2020). Again, we also show results for the F-step using the ODV procedure as well as the SVD,
3. The EM algorithm for the PGMM of McNicholas and Murphy (2008) with model CCU, corresponding to the low-rank constraint $\mathbf{S}_k = \mathbf{U} \mathbf{U} + \boldsymbol{\Psi}$ of the covariance matrix, without orthonormality constraints on \mathbf{U} . An implementation is available in the eponymous R package **pgmm** (McNicholas et al. 2019),
4. The EM algorithm for MCFA of Baek et al. (2009) and implemented in the **EMMIXmfa** R package (Rathnayake et al. 2019),
5. The EM algorithm for the HDDC model of Bouveyron et al. (2007) with model $[\cdot Q D]$, so that the learned subspaces are common, as in the other methods. We used the implementation available in the **HDclassif** R package (Bergé et al. 2019),
6. A k -means baseline.

A total of 8 distinct algorithms are tested, since both models BFEM and FEM may have two distinct F-step procedures. In Sections 4.2 and 4.3, we use colors to distinguish between models, and line marker types to differentiate between

the ODV and SVD method. For the sake of readability, we do not show the results of the HFMA model of Montanari and Viroli (2010) in the following figures since it did not perform well on our experimental settings. This might be due to the different constraints on the subspace means and covariances, making it more distant to the BDLM model than other subspace clustering methods.

Throughout the rest of this section, unless stated otherwise, each method has the same 10 restarts consisting of 10 different k -means results. The one achieving the greatest likelihood is kept, and the clustering is done with a MAP estimate over the posterior of \mathbf{Z} . The maximum of iterations is set to 100 everywhere and the same absolute tolerance of 10^{-6} is used. The fixed point algorithm in the VE-step has a tolerance of 10^{-6} but a maximum number of iterations set to 3.

Concerning the choice of d , it is set to $K - 1$ for both FEM and BFEM and to the true value d^* whenever it is known in the simulations. When d is unknown, the HDDC model has an internal heuristic to choose the best intrinsic dimension d_k of each cluster, and we use the Bayesian Information Criterion for the MCFA and PGMM as suggested in their original papers. The clustering results are reported using the Adjusted Rand Index (ARI, Hubert and Arabie 1985), a label independent measure of statistical similarity between two partitions. An ARI of 0 means that the two partitions are statistically independent, while identical partitions (up to label switching) give an ARI of 1. Hence, the higher the ARI, the better.

4.1 An introductory example

In order to illustrate the interest of discriminative subspaces, we begin with the numerical setting of Chang (1983) discussed in the introduction and Figure 1. There are $n = 300$ observations and $K = 2$ clusters in the data, defined as follows:

$$\mathbf{y}_i = -0.5\mathbf{r} + \mathbf{r}\mathbb{1}_{\{z_{ik}=1\}} + \mathcal{N}_p(\mathbf{0}, \mathbf{S}),$$

with $\forall j = 1, \dots, 15$:

$$r_j = 0.95 - 0.05j,$$

$$\mathbf{S}_{jj} = 1 \text{ and } \forall j' \neq j, \mathbf{S}_{jj'} = -0.13f_j f_{j'} \text{ with } f_j = \begin{cases} -0.9 & j \leq 8 \\ 0.5 & j > 8 \end{cases}.$$

Thus, it is a 2-component Gaussian mixture in dimension $p = 15$, with $\mathbf{m}_1 = -0.5\mathbf{r}$ and $\mathbf{m}_2 = 0.5\mathbf{r}$ and homoscedastic covariance $\mathbf{S}_1 = \mathbf{S}_2 = \mathbf{S}$. We emphasize that this simulation is not favoring any of the tested methods, except maybe for the standard GMM since the simulation is according to this model.

We ran each method with the true number of clusters $K = 2$, and used model selection for the choice of d for the concerned methods. The average results over 100 simulated datasets are represented for each method in Table 2. We do not distinguish the ODV and SVD methods here, since they lead to the same results on this simple example. One can see that the proposed discriminative subspace approach yields a better clustering in this setting, with a slight advantage over the frequentist version. In particular, extensions of pPCA like MCFA or PGMM do not allow to recover the correct partitions. This highlights the interest of discriminative subspaces even in different scenarios. The HDDC algorithm exhibits the same performance as BFEM. However, we point out that it selects intrinsic dimensions $d_k = 14$ to do so, which are the maximum values in this model. In contrast, BFEM works with $d = 1$ enabling to visualize the latent space in Figure 3, with a clear separation of the two classes. Finally, since the dimension p is still reasonable compared to n , a standard GMM with spectral constraints may be fitted and performs well. Here, the BIC criterion selects the EEE model which means ellipsoidal, equal volume, shape and orientation $\mathbf{S}_k = \lambda \mathbf{D} \mathbf{\Delta} \mathbf{D}$, and corresponds to the true model.

Table 2: Mean ARI and standard errors for BFEM and competing methods over 100 simulations of Chang (1983)’s setting.

Kmeans	BFEM	FEM	HDDC	MCFA	PGMM	Mclust
0.24 \pm 0.1	1 \pm 0	0.98 \pm 0.11	1 \pm 0	0.62 \pm 0.07	0.42 \pm 0.22	0.97 \pm 0.12

Figure 3 shows the 1-dimensional discriminative subspace found by BFEM on one simulation, with colors indicating cluster membership. Moreover, the solid lines represent the Gaussian density in each cluster $p(\mathbf{x} | \tilde{\boldsymbol{\mu}}_k, \tilde{\boldsymbol{\Sigma}}_k)$, along with the empirical within-cluster distribution as a histogram. Unsurprisingly, the subspace induces well-separated clusters and the empirical within-cluster distribution has a Gaussian shape fitting the theoretical one. Finally, the evolution of

the evidence lower bound during the BFEM algorithm is displayed in Figure 4. As expected, it is not monotonically increasing, especially in the first step, although the evolution is quite smooth. In addition, convergence happens before the limit of 100 iterations is reached.

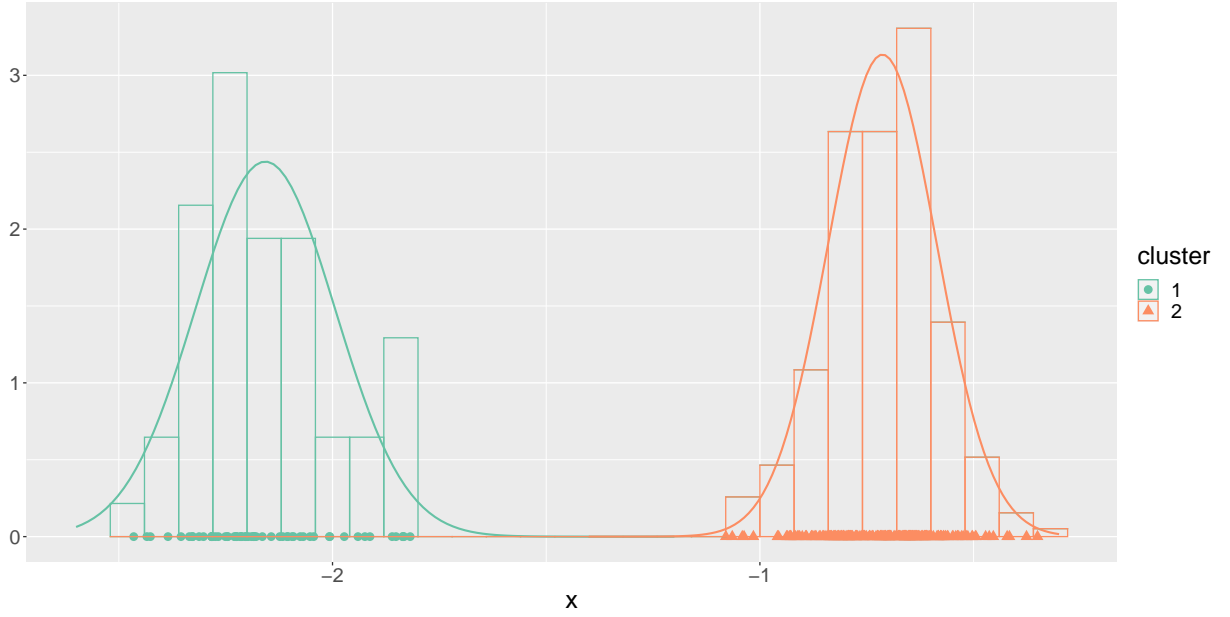


Figure 3: Projection of Chang's data in the 1-dimensional subspace found by BFEM, colors indicate the estimate cluster memberships. Solid lines represent the learned within-cluster Gaussian distributions, while the histograms represent the empirical ones.

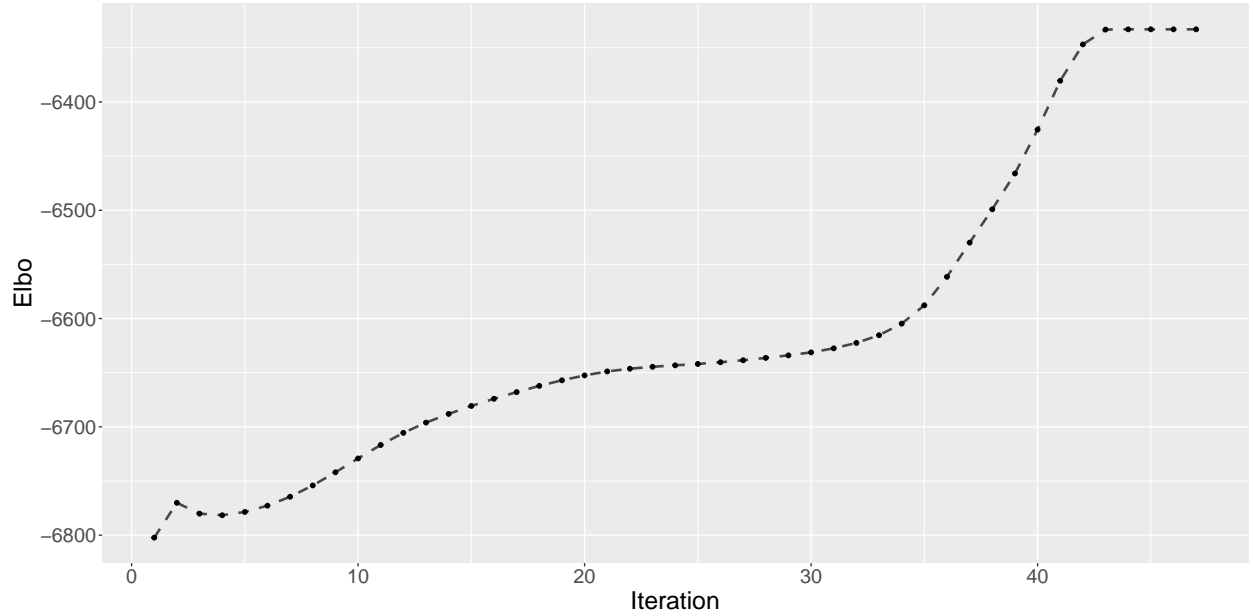


Figure 4: Evolution of the evidence lower bound during a run of the BFEM algorithm with ODV procedure on Chang's dataset.

4.2 Sensitivity to the dimension

We now propose to investigate the behavior of subspace clustering methods on increasingly high-dimensional problems. In this setting, we simulate \mathbf{X} from $K = 3$ Gaussian components in dimension $d = 2$. The respective means and covariance matrices are

$$\boldsymbol{\mu}_k = 3 \cdot (0, k)^\top \quad \boldsymbol{\Sigma}_k = \begin{pmatrix} 1.5 & 0.75 \\ 0.75 & 0.45 \end{pmatrix} \quad \boldsymbol{\pi} = (0.4, 0.3, 0.3)^\top.$$

Figure 5 illustrates a particular simulation of $n = 900$ data points. As can be seen, it corresponds to the particular case where the clusters are parallel Gaussian ellipses, differentiated with a mean-shift along the x-axis.

Next, we propose to simulate according to the DLM model. First, $\mathbf{D} = [\mathbf{U}, \mathbf{V}]$ is simulated from the Q-matrix of the QR decomposition of the matrix $\mathbf{A} \in \mathbb{R}^{p \times p}$, the latter being itself simulated according to $\mathbf{A} \sim \mathcal{N}_{p^2}(\mathbf{0}, 100\mathbf{I})$. Then, for each observation i , a $(p - d)$ dimensional standard Gaussian noise $\boldsymbol{\epsilon}_i$ is simulated. Finally, the data points are created as the linear transformation $\mathbf{y}_i = \mathbf{D}(\mathbf{x}_i^\top, \boldsymbol{\epsilon}_i^\top)^\top$. The first principal components are expected to behave poorly in terms of class separation in this scenario, which is illustrated in Figure 6 for $p = 50$. Indeed, the directions of greatest variations include noisy directions that contribute more to the variance than the second signal dimension.

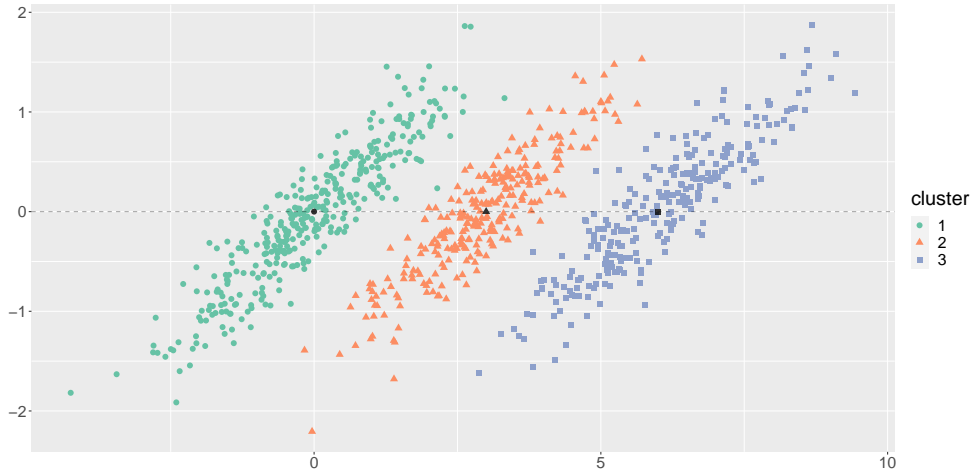


Figure 5: Simulation of the 3-components GMM of Section 4.2.

We investigate the behavior of each method as the dimension p increases from 5 to 155. Mean ARI and standard variations were computed on a 10-spaced linear grid, for 100 simulated datasets at each level p . As stated above, we use colors to differentiate between the BFEM and FEM, and line marker types to distinguish between OVD (solid) and SVD (dashed) for both algorithms. Concerning MCFA and PGMM, the subspace dimension was set to the true value $d = 2$. The results are displayed in Figure 7 and shows several things. First, the BFEM and FEM with the ODV method are very robust in this scenario, with a perfect recovery at each level p . Other subspace clustering methods quickly decrease beyond $p = 15$ with performances comparable, or below k -means. Thus, it underlines a limit of likelihood-based approach, as noisy dimensions are being fitted in the subspace when p increases. The discriminative approach, injecting clustering information in the search for the optimal subspace, is robust in this context as the optimal subspace is not necessarily aligned with greatest variance directions. Another interesting fact is the sensitivity to noise of the Fisher-EM with the SVD procedure, which displays a rather unstable behavior. We note that BFEM with the SVD procedure, while suboptimal, is still displaying a strong stability in high-dimensional settings, with an ARI decreasing only after $p = 85$.

4.3 Signal-to-noise ratio

We place ourselves in the same setting as section 4.2, only this time the dimension is fixed to a high-dimensional scenario $p = 150$. We propose to investigate the impact of the noise, with $\boldsymbol{\epsilon}_i$ now drawn from a centered Gaussian distribution with covariance $\beta \mathbf{I}_{p-d}$. The latter may be interpreted as controlling the signal-to-noise ratio (SNR) which



Figure 6: Subspaces found by PCA on a simulation with $p = 50$, color and shape indicate the true cluster membership. Once again, the most discriminative subspace is given by the first and last components, the other corresponding to noisy directions.

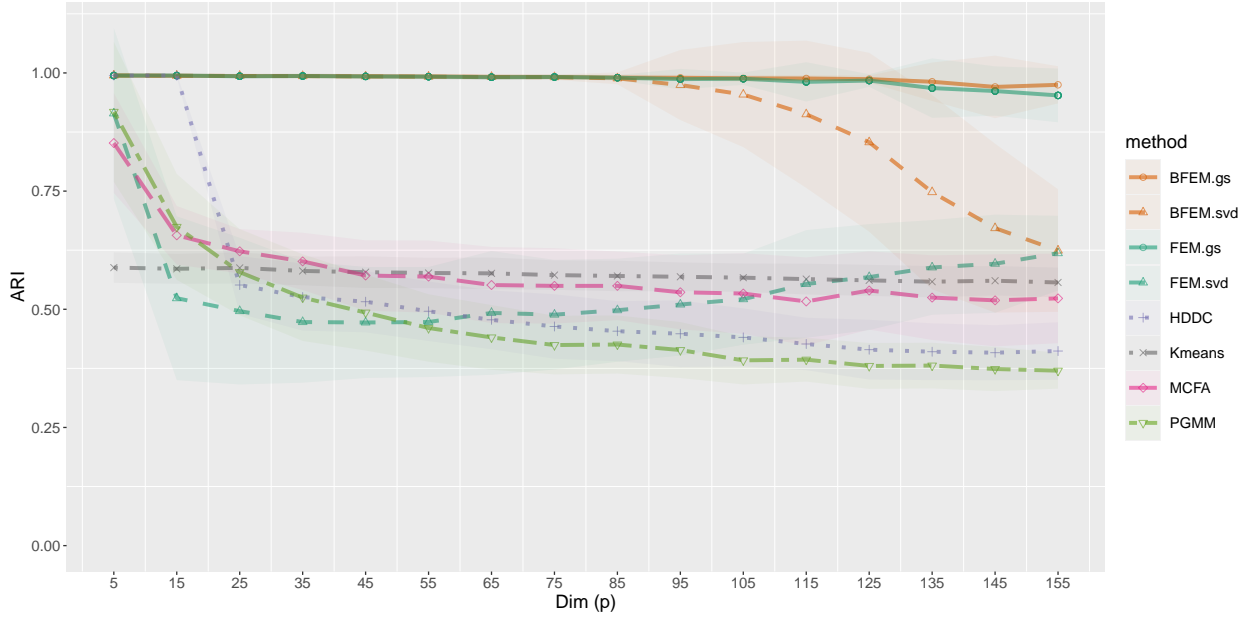


Figure 7: Evolution of the mean ARI over 100 different runs for each method, with an increasing dimensionality p and $n = 900$.

can be defined as the ratio between signal variance in the subspace of dimension $d = 2$ and the noise variance² β . Since all clusters have the same subspace covariance, we define the signal variance as $\text{Tr}[\Sigma]$, the inertia of a cluster cloud point in the latent subspace. The SNR is best expressed in decibels (dB), which corresponds to ten times the decimal logarithm of the variance ratio.

$$SNR = 10 \times \log_{10} \left(\frac{\text{Tr}[\Sigma]}{\beta} \right).$$

²One could use $(p - d)\beta$ as the actual variance of the signal, taking into account the fact that there are $(p - d)$ noisy directions. However, since $p - d$ is fixed here, it only acts as a scaling factor for the SNR.

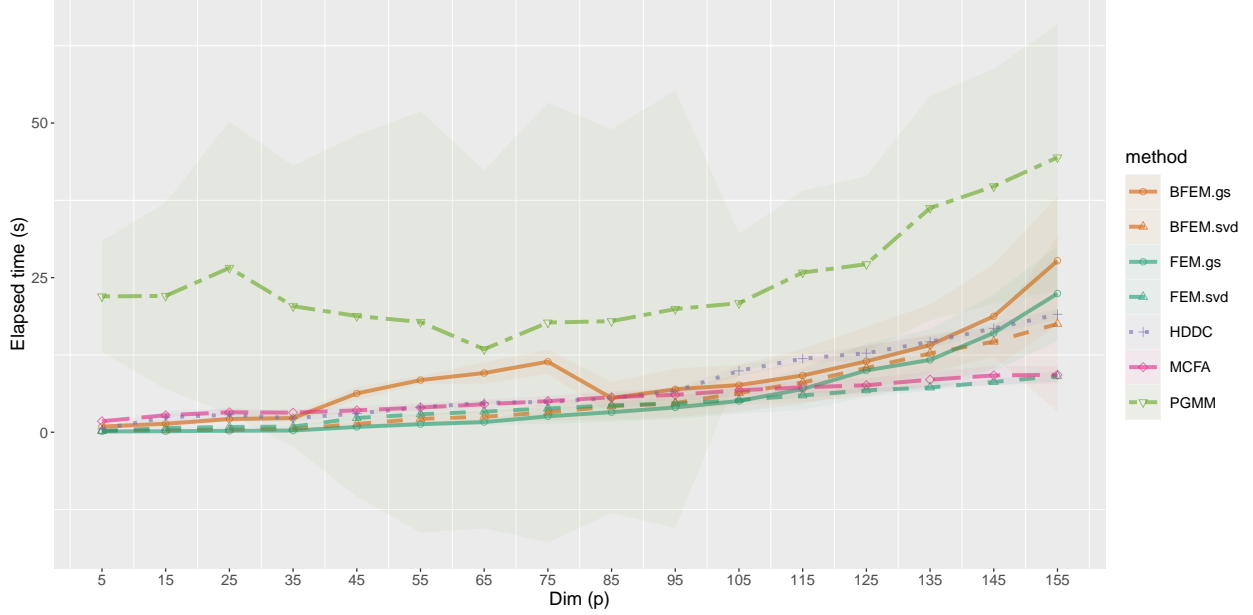


Figure 8: Mean elapsed time in seconds for one run of each method, computed on 100 datasets for each level p .

A value of 0 means that both variances are equal, and an increase (resp. decrease) of 3 dB means that the variance of the noise was divided (resp. multiplied) by 2. For example, a SNR of 3 dB means that the variance of the signal is 2 times that of the noise ($\beta \approx 1$), and a SNR of -6 means that β is 4 times greater than the signal variance ($\beta \approx 8$).

Figure 9 shows the mean ARI and standard deviations for an increasing SNR from -6 to 15 with a 0.5-spaced linear grid. Again 100 datasets are simulated for each level. Several comments are in order. First, for high values of the SNR, which we refer to as the noiseless regime, the clustering problem of Section 4.2 becomes trivial, except for k -means which is disadvantaged by the non-spherical shapes of clusters and Fisher-EM which seems to display a surprising instability in the noiseless limit. From the preceding section, we know that the Fisher-EM algorithm with the SVD procedure is not robust to high-dimension. However, this shows that the ODV procedure also suffers from instability in the frequentist setting. This may be due to poor conditioning of the soft between-class scatter matrix arising in this case. Apart from this somewhat surprising fact, the behaviors of other subspace clustering methods such as HDDC, MCFA and PGMM are expected. However, their performances quickly decrease, even for reasonable values of the SNR where the noise variance is orders of magnitude below the signal. A contrario, the BFEM displays a strong stability, and the SVD method seems to be applicable as long as the noise variance remains reasonably below the signal. In addition, BFEM with the ODV procedure is the most stable of all, with perfect recoveries even when the SNR is 0 dB, *i.e.* the equal variance case. Eventually, no clustering structure can be recovered below 0 dB, as the signal is completely overwhelmed by noisy directions, and the ARI of each method quickly decreases to 0.

4.4 Model selection

Here, we investigate the ability of the ICL criterion of Section 3.6 to choose both the number of clusters and the model. We use the setting of Section 4.2, which corresponds to $p = 150$, $K = 3$ and a model $\text{BDLM}_{[\Sigma\beta]}$. Two different levels of SNR are tried: 3dB ($\beta = 1$), which corresponds to the setting of Figure 7, and -2 dB ($\beta \approx 3$), which is a more complicated case for BFEM as shown in Figure 9. The results are shown in Table 3, we see that the performance of model selection are perfect for the first setting, and still very satisfying in the more difficult scenario with 90% of correct selection of the pair (K, \mathcal{M}) , and 98% of correct selection of K .

4.5 Real data benchmarks

Here, we consider classical real-data benchmarks studied in the Gaussian subspace clustering literature:

- Fisher’s *iris* is a traditional real dataset used to assess clustering algorithms, although it cannot be deemed as a high-dimensional problem. It consists in 150 observations of 3 iris species, 50 each, described by 4 variables.

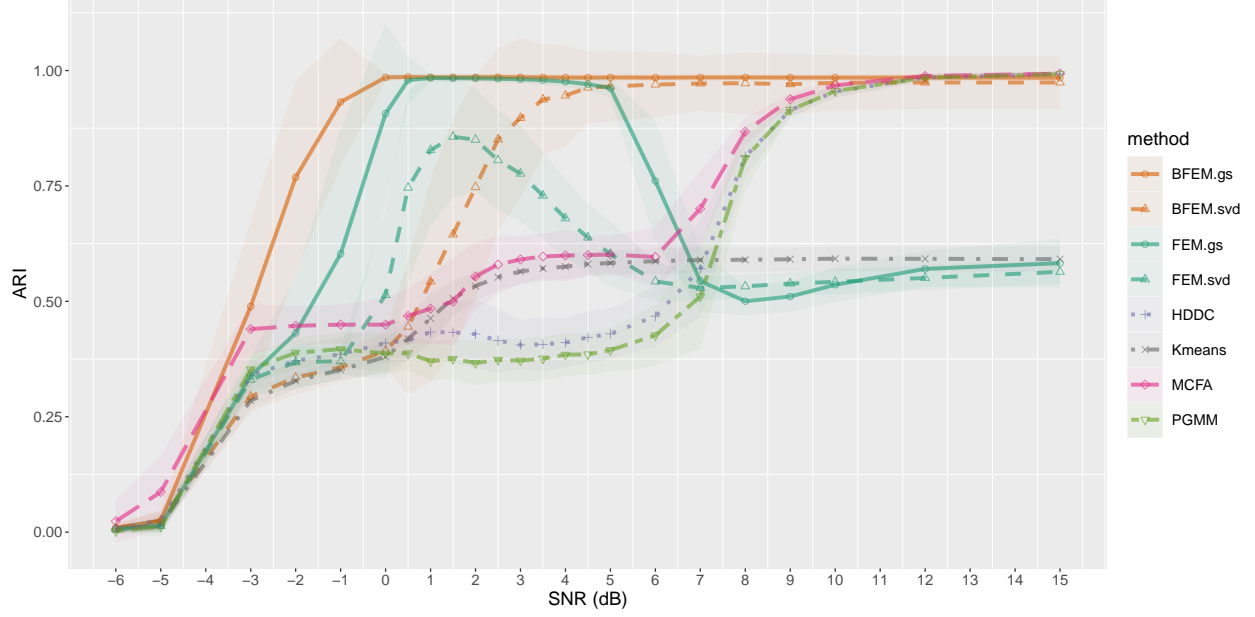


Figure 9: Evolution of the mean ARI over 100 different runs for an increasing signal-to-noise ratio, $p = 150$ and $n = 900$.

Table 3: Percentage of correct model selection for BFEM with $p = 150$ and varying SNR. The true model is $\text{BDLM}_{[\Sigma\beta]}$ with $K = 3$.

$SNR = 3$						
$K \setminus \mathcal{M}$	$[\Sigma_k\beta]$	$[\alpha_{kh}\beta]$	$[\alpha_k\beta]$	$[\Sigma\beta]$	$[\alpha\beta]$	$[\alpha_h\beta]$
2	0	0	0	0	0	0
3	0	0	0	100%	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0

$SNR = -2$						
$K \setminus \mathcal{M}$	$[\Sigma_k\beta]$	$[\alpha_{kh}\beta]$	$[\alpha_k\beta]$	$[\Sigma\beta]$	$[\alpha\beta]$	$[\alpha_h\beta]$
2	0	0	2%	0	0	0
3	8%	0	0	90%	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	0	0	0	0	0	0

- The *Italian Wine* dataset contains the description of 178 wines 27 variables related to *e.g.* color or alcohol (Weinen 1986). There are $K = 3$ types of wines, and we wish to know to which extent the $p=27$ variables can help relate to the type of wine. This dataset is also a famous introductory dataset for subspace clustering method (McNicholas and Murphy 2008; Bouveyron et al. 2019).

- The *Satellite* dataset consists in a satellite image decomposed into small sub-areas \mathbf{y}_i of 3×3 pixels. For each of these pixels, we have access to their value in 4 spectral bands, leaving a total of $p = 4 \times 9 = 36$ values describing an area. The goal is to recover the classification of the central pixel in \mathbf{y}_i which can be one of the $K = 7$ different type of soils.
- The *USPS358* dataset is a more realistic example of high-dimensional data clustering. It is a subset of the US postal dataset from UCI, which originally contained 16×16 images of scanned digits from 0 to 9, with only digits 3, 5 and 8 known to be the most difficult to discriminate. There are $n = 1,756$ images, described by pixels values indicating gray level in dimension $p = 16 \times 16 = 256$. In this scenario, we want to recover the 3 different digit classes.

These datasets are available in the **MBCbook** R package, except for the *Satellite* data which is from the UC Irvine repository³. Among them, only the *Satellite* and *USPS358* are high-dimensional problem in the sense that standard Gaussian mixtures fail to recover any relevant information. We ran each method with the true number of clusters of the corresponding dataset. The latent dimension of PGMM and MCFA was chosen by BIC with $d \in \{1, 2, 3\}$ for *Iris* and $d \in \{1, \dots, 20\}$ for *Wine27*, *USPS358* and *Satellite*. All submodels were allowed for HDDC, including the one with different subspaces, and we used the ICL as the model selection criterion. Table 4 shows the results and we can see that, while PGMM performs better on *Iris* and *Wine27*, BFEM displays a real interest on the high-dimensional settings of *Satellite* and *USPS358* in dimension $p = 256$, achieving the top performance. Moreover, for the *Wine27* dataset, MCFA and PGMM respectively chose $d = 6$ and $d = 4$ while HDDC selected a $[a_j b Q d]$ model with $d_k = d = 4$ in each class. A contrario, the FEM and BFEM works with $d = 2$ and the discriminative subspace can therefore be plotted entirely in a two-dimensional graphic, as shown in Figure 10. Once again, the data are well separated Gaussian ellipses according indicating the clustering quality. For the *Satellite* data, the problem is harder since there are 7 classes, the PGMM model and MCFA models select $d = 20$ which is the maximum value, while HDDC selects the $[a_{kj} b_k Q_k d]$ model with $d = 7$.

Table 4: Comparison of the BFEM algorithm with other methods on common real datasets benchmarks.

Dataset	p	Non-HD models		HD models				
		k -means	Mclust	HDDC	MCFA	PGMM	FEM	BFEM
<i>Iris</i>	4	0.73	0.90	0.90	0.92	0.94	0.88	0.90
<i>Wine 27</i>	27	0.90	0.93	0.95	0.96	0.98	0.93	0.93
<i>Satellite</i>	36	0.53	0.36	0.45	0.43	0.56	0.53	0.64
<i>USPS358</i>	256	0.64	0	0.35	0.28	0.38	0.66	0.76

5 An application to patch-based Gaussian single image denoising

In this section, we present an authentic unsupervised application of high-dimensional model-based clustering, contrasting with common applications with supervised datasets.

5.1 Single image denoising with the BFEM algorithm

Single image denoising consists in the restoration of an image degraded by an additive noise modeling the fluctuations induced by the physical process underlying recording devices such as camera (Buades et al. 2005). While classical methods include local filters, Fourier or Wavelet transforms in the frequency domain, or sparse representations, patch-based methods have been successfully applied to the denoising problem (Buades et al. 2011). Decomposing an image into a set of small sub-images or patches, of size $f \times f$, similar patches are supposed to be *i.i.d* from the same distribution and can be averaged to reduce their noise in a non-local fashion. The restored image pixels can then be reconstructed by averaging over all the denoised patches they belong to. Denoting by $\mathbf{t}_i \in \mathbb{R}^p$ a flattened original patch in dimension $p = f^2$, and \mathbf{y}_i its noisy observation, then patch-based image denoising assumes the following model:

$$\mathbf{y}_i = \mathbf{t}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_p(\mathbf{0}_p, \sigma^2 \mathbf{I}_p). \quad (35)$$

³Available at this address [https://archive.ics.uci.edu/ml/datasets/Statlog+\(Landsat+Satellite\)](https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite))

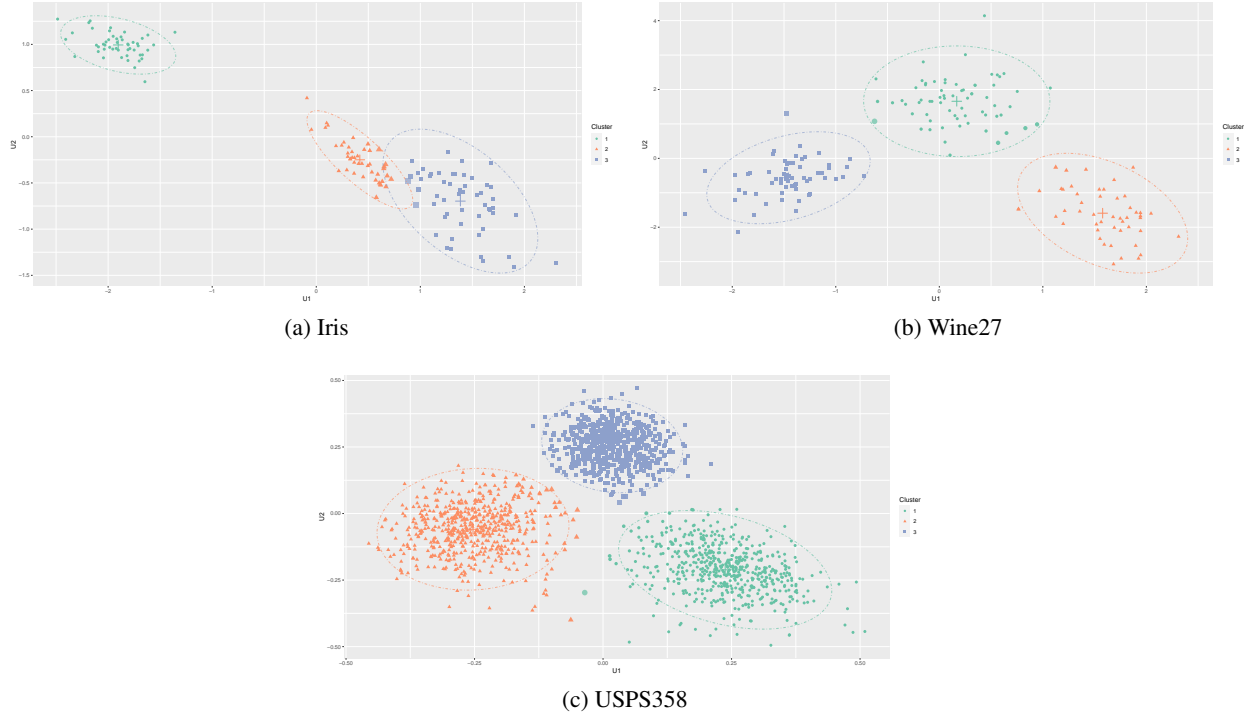


Figure 10: Iris, Wine27 and USPS358 datasets projected onto the two-dimensional subspace learned by BFEM.

Here, σ^2 denotes the variance of the additive Gaussian white noise, which is common over patches. In this context, Gaussian mixture models have been used extensively (Yu et al. 2011; Zoran and Weiss 2011) as they provide a sound probabilistic framework modeling each patch as:

$$p(\mathbf{t}_i) = \sum_{k=1}^K \pi_k \mathcal{N}_p(\mathbf{t}_i | \mathbf{m}_k, \Phi_k). \quad (36)$$

Thus, marginalizing on \mathbf{t}_i leaves the distribution:

$$p(\mathbf{y}_i) = \sum_{k=1}^K \pi_k \mathcal{N}_p(\mathbf{y}_i | \mathbf{m}_k, \mathbf{S}_k), \quad \mathbf{S}_k = \Phi_k + \sigma^2 \mathbf{I}_p,$$

and the conditional expectation of \mathbf{t}_i given \mathbf{y}_i may be used as the denoised patch estimate minimizing the expected mean squared error:

$$\hat{\mathbf{t}}_i = \mathbb{E}[\mathbf{t}_i | \mathbf{y}_i] = \sum_{k=1}^K \tau_{ik} \mathbb{E}[\mathbf{t}_i | \mathbf{y}_i, z_{ik} = 1]. \quad (37)$$

Here, $\tau_{ik} = p(z_{ik} = 1 | \mathbf{y}_i)$ denotes the posterior probabilities of the cluster membership z_i , and, $(\mathbf{t}_i, \mathbf{y}_i)$ being jointly Gaussian given z_i , the conditional expectation of a partitioned Gaussian vector leaves an exact formula for the right-hand side:

$$\mathbb{E}[\mathbf{t}_i | \mathbf{y}_i, z_{ik} = 1] = \mathbf{m}_k + (\mathbf{S}_k - \sigma^2 \mathbf{I}_p) \mathbf{S}_k^{-1} (\mathbf{y}_i - \mathbf{m}_k). \quad (38)$$

Thus, the final patch estimate $\hat{\mathbf{t}}_i$ is a patch-dependent combination of K linear *filters* (in the image processing terminology) of the noisy patch, and the denoising task amounts to compute the model parameters. However, the dimension being fixed by the patch size $p = f^2$, relatively high-dimensional problems arise even for common patch sizes found in the literature, *e.g.* $f = 8$ and $p = 64$. Thus, Gaussian subspace mixture models have been proposed to circumvent this issue. In their S-PLC algorithm, Wang and Morel (2013) used the mixture of principal component analyzers (Tipping and Bishop 1999, MPPCA) which amounts to the low-rank constraint $\mathbf{S}_k = \mathbf{U}_k \mathbf{U}_k^\top + \sigma^2 \mathbf{I}_p$. The authors also relaxed the original MPPCA model assumptions to allow each subspace \mathbf{U}_k to have its possibly different dimension d_k , and fixed the latter to be either 1, $p/2$ or $p - 1$. More recently, the high dimensional mixture model for image

denoising (HDMI, Houdard et al. 2018) used a slightly modified version of the HDDC model (Bouveyron et al. 2007), considering the specific decomposition $S_k = U_k \Lambda_k U_k^\top + \sigma^2 I_p$, with $\Lambda_k = \text{diag}(\lambda_{k1}, \dots, \lambda_{kd_k})$ and $U_k \in \mathbb{R}^{p \times d_k}$ column-orthonormal. This approach yields state-of-the-art performances in denoising applications, and includes a procedure for selecting each dimension d_k and the optimal level of noise through model selection. Here, we show that the BDLM model is also applicable in this context and that it yields satisfactory results compared to standard approaches.

Denoising with the BFEM algorithm Keeping the notations of Section 2, we introduce a low-dimensional patch representation $x_i \in \mathbb{R}^d$ and the observed patch is supposed to be observed from model (35) with the modification:

$$y_i = U x_i + \epsilon_i, \quad x_i \sim \mathcal{N}_d(\mu_k, \Lambda_k), \quad U^\top U = I_d. \quad (39)$$

Thus, a noisy patch y_i is supposed to be observed from the BDLM $[\Sigma_k, \sigma^2]$ model, with the decomposition:

$$S_k = D \Delta_k D^\top, \text{ with: } D = [U, V], \quad \Delta_k = \text{diag}(\Sigma_k, \sigma^2 I_{p-d}). \quad (40)$$

Note that this is a slightly modified version of the BDLM model since the matrix $\Sigma_k = \Lambda_k + \sigma^2 I_d$ now depends on the noise variance. This modification was also proposed for the HDDC model in the HDMI algorithm, assuming that Λ_k is diagonal but of variable dimension, and a specific EM was designed to keep the noise variance known and fixed to $\beta_k = \sigma^2$ throughout the estimation. Here, we propose to take an even simpler approach and to use the BFEM algorithm for the estimation of U and Σ_k . Having derived an estimate of these quantities, the true noise level σ^2 can be used for supervised denoising in Equation (38). Moreover, this formula may be further simplified in order to inverse the block matrix Δ_k instead of S_k as the following proposition describes.

Proposition 6 (Proof in Appendix B). *Under the BDLM model, the conditional patch expectation in cluster k may be written:*

$$\mathbb{E}[t_i | y_i, z_{ik} = 1] = U \tilde{\mu}_k + \tilde{D} \left(I_p - \sigma^2 \Delta_k^{-1} \right) \tilde{D}^\top (y_i - U \tilde{\mu}_k). \quad (41)$$

where $\tilde{D} = [U, \mathbf{0}_{p \times (p-d)}]$, $\Delta_k = \text{diag}(\hat{\Sigma}_k, \sigma^2 I_{p-d})$ and σ^2 is provided by the user.

Finally, we emphasize that even for basic image resolution, the number of observations n can be quite large. For example a 512×512 image divided in patch size $f = 8$, leaves us with $n = 255025$ patches in dimension $p = 64$. While this is good for estimation purposes, computations may be sped up by estimating the model parameters on a portion of the n initial patches only. Indeed, the variational posterior cluster membership distribution, $\tau_{new,k}$, of a new data point y_{new} , may be directly computed from Proposition 1 and used in the patch denoising formula in Equation (37).

5.2 Denoising results on natural images

We now proceed to illustrate the performances of BFEM for image denoising with the proposed methodology. The standard quantitative way to measure the quality of a restored image \hat{I} is the peak signal-to-noise ratio (PSNR), defined as:

$$PSNR(I, \hat{I}) = 10 \times \log_{10} \left(\frac{255}{\frac{1}{|\Omega|} \sum_{x \in \Omega} (I(x) - \hat{I}(x))^2} \right), \quad \Omega = \{1, \dots, n_r\} \times \{1, \dots, n_c\},$$

where n_r and n_c are the number of row and column pixels respectively. The denominator in the formula is the mean squared error and the higher the PSNR, the better the reconstruction.

As a point of comparison with other Gaussian subspace clustering methods in single image denoising, we report the results of Houdard et al. (2018, Table 4) on 3 natural grayscale images of resolution 512×512 : *Alley*, *Barbara* and *Simpson* which are displayed in Figure 11. The first two are benchmark images in the denoising literature as they capture natural scenes with highly structured regions (the brick in *Alley*, the tablecloth and scarf motives in *Barbara*). The last one was introduced in the HDMI paper. The methods considered are Non-Local Bayes (Lebrun et al. 2013, NL Bayes) which uses an unrestricted Gaussian model, the S-PLC (Wang and Morel 2013) and the HDMI (Houdard et al. 2018) described above. We emphasize that these methods are core-designed for single image denoising, with pre-processing and post-processing steps for NLBayes and S-PLC, which we do not include in our BFEM denoising methodology. As a matter of comparison, we also include a denoising with a standard HDDC clustering, which differs from HDMI since the noise β_k is not fixed in the EM. Table 5 gives the PSNR obtained for all the considered methods and two levels of standard error $\sigma \in \{20, 30\}$. In order to be consistent with the results taken from the HDMI paper, we use the same patch size of $f = 10$, leaving a problem in dimension $p = 100$. We tried similar values as the one

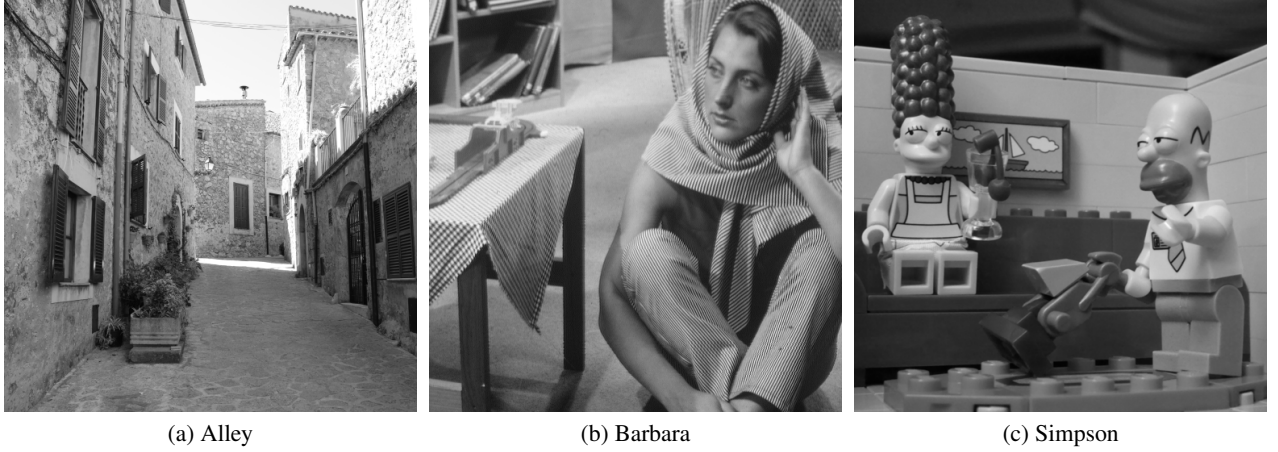


Figure 11: The three grayscale benchmark images used to illustrate the denoising algorithms.

used in the denoising literature, and we present the results for $K = 40$ and $K = 90$. The models are estimated on the whole images, and the BFEM algorithm uses a latent dimension fixed to $d = 39$. The BFEM results are of the same order of magnitude that state-of-the-art approaches, albeit not being core designed for denoising. Moreover, the denoising results also display qualitative properties.

Table 5: PSNR results comparison for HDDC and BFEM denoising with other methods for Gaussian single image denoising methods NLBayes (Lebrun et al. 2013) with and without flat patch tricks, S-PLE (Wang and Morel 2013).

Image	σ	Table 4 of Houdard et al. (2018)					New results			
		NL-Bayes		S-PLE	HDMI _{sup}		HDDC		BFEM	
		<i>original</i>	<i>no flat</i>		$K = 40$	$K = 90$	$K = 40$	$K = 90$	$K = 40$	$K = 90$
<i>Barbara</i>	20	31.52	31.29	30.37	31.32	31.61	30.93	31.35	29.88	30.82
	30	29.72	29.44	28.22	29.31	29.49	28.40	28.91	28.46	28.95
<i>Simpson</i>	20	34.74	33.72	34.08	35.05	34.91	34.45	34.17	34.32	34.54
	30	32.53	31.54	31.53	32.33	32.50	31.77	32.02	31.95	32.06
<i>Alley</i>	20	29.10	29.07	28.67	29.03	29.07	28.55	27.95	28.51	28.75
	30	27.43	27.37	26.92	27.31	27.39	26.71	27.06	27.05	27.16

While the PSNR is a good indicator of the reconstruction error, the denoising results may vary widely in term of restoration quality. Figure 12 shows the reconstructed images by each of the considered methods for the *Alley* image. We can see that the highly textured motives like the pavement of the alley are blurred, and high-frequency location such as the electric cables in the white sky are lost by S-PLE and NLBayes. A contrario, BFEM reconstructs these motives really well, with visual results comparable to HDMI denoising which achieves state-of-the-art performances in this context. This denoising is obtained without knowing the noise level and we emphasize that a refinement of the algorithm could be easily designed for denoising, keeping the noise variance fixed to $\beta_k = \sigma^2$ during estimation. The methodology also easily extends to color images, stacking the patch pixels in the red, green and blue channel and leaving a problem in dimension $p = 3f^2$.

6 Conclusion and perspectives

In this paper, we introduced a new clustering algorithm for high-dimensional continuous data. Relying on a Bayesian formulation of the discriminative latent mixture model, it relies on variational inference for the estimation of the model parameters. Building on an unsupervised version of Fisher discriminant analysis, the discriminative subspace is estimated iteratively throughout the algorithm with soft scatter matrices updated with the posterior membership probabilities at each step. An empirical Bayes strategy was proposed for setting the hyper-parameter λ , controlling between-class separation in the latent space, and an ICL criterion was derived for model selection. The simulations assessed the stability of the method with respect to both high-dimension and noise, and showed its significant improve-

ment over the frequentist version as well as other Gaussian subspace clustering methods. Finally, we trust that the image denoising application provides an interesting setting that could inspire the model-based clustering community in the design of specific algorithms.

In further works, we plan to consider two types of extensions for the BFEM algorithm. First, variable selection could be added in the model through sparsity constraints on the matrix U . A sparse version of the BFEM could be designed, in the line of Bouveyron and Brunet-Saumard (2014a), based on the regression formulation of the Fisher criterion with a Lasso-like penalty (Qiao et al. 2009). Second, the trace of ratio formulation in Equation (2) is itself a simplification of the ratio of trace problem:

$$\max_{U^T U = I_d} \frac{\text{Tr}[U^T S_B U]}{\text{Tr}[U^T S_W U]},$$

which does not accept a closed-form solution and necessitate iterative algorithms to solve. Recent works highlight the better discriminative power of this new formulation in the supervised context, and new iterative algorithms have been designed (Guo et al. 2003; Wang et al. 2007; Kokiopoulou et al. 2011). Based on these empirical results, it would be interesting to modify the F-step in order to maximize such a criterion. Given the diversity of possible algorithms and their need for calibration, we leave a detailed study to future work.

Aknowledgments

The authors wish to thank Antoine Houdard for the images and helpful discussions on image denoising.

This work has benefited from the support of the French government, through the 3IA Côte d’Azur Investment in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. This work was also supported by a DIM MathInnov grant from Région Ile-de-France. The authors are finally thankful for the support from fédération F2PM, CNRS FR 2036, Paris.

Conflict of interest

The authors declare that they have no conflict of interest.

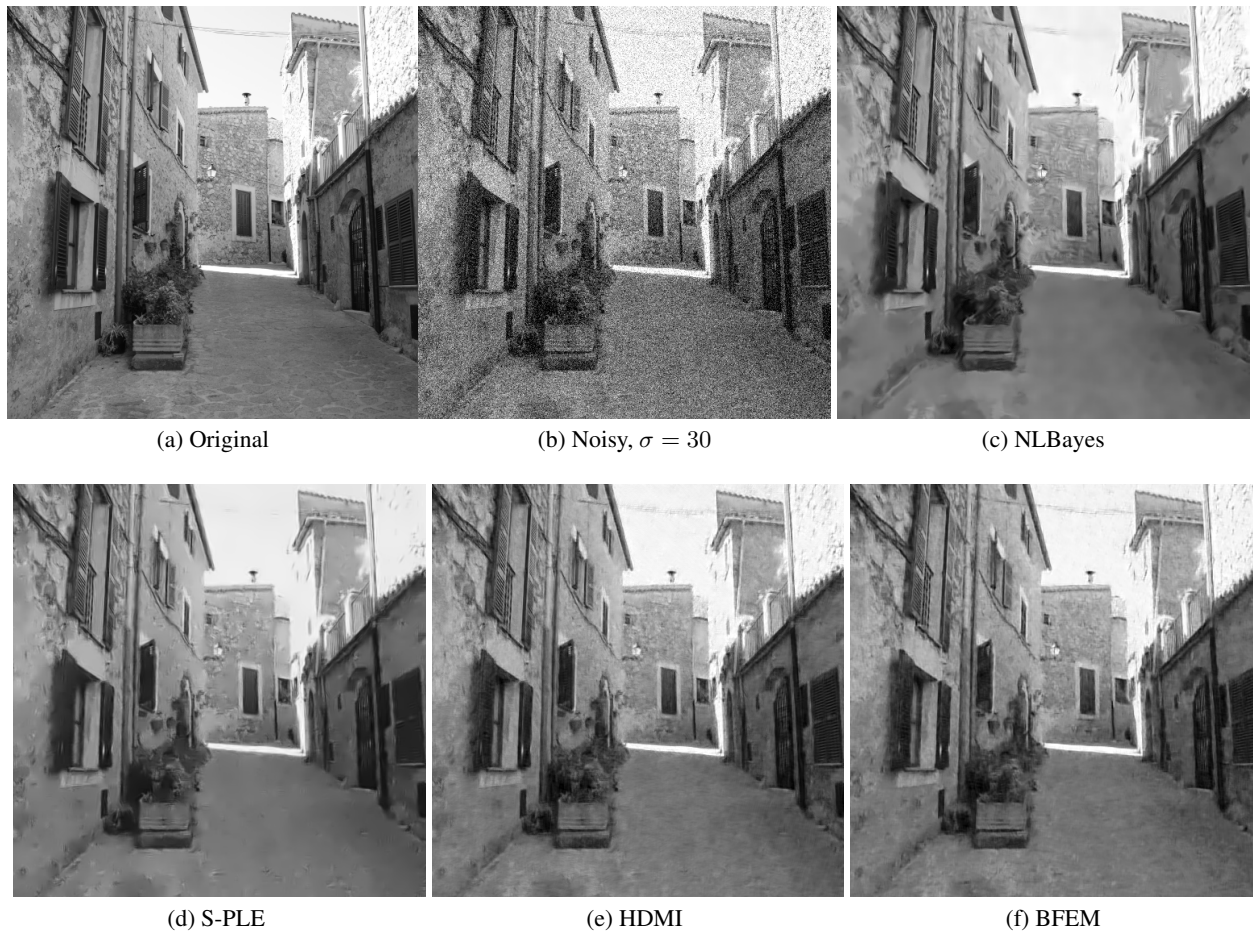


Figure 12: Qualitative results of each denoising methods at $\sigma = 30$ on the *Alley* gray-scale image. Images should be seen at full resolution on the electronic version of the paper.

A Proofs

A.1 Optimization of $q(\mathbf{Z})$

Proof of Proposition 1 on page 8. A classical result of the CAVI algorithm, see *e.g.* Blei et al. (2017, Equation (18)), states that the lower bound \mathcal{J} is maximized with respect to $q(\mathbf{z}_i)$ by:

$$q^*(\mathbf{z}_i) \propto \exp \left\{ \mathbb{E}_{\mathbf{z}_{-i}, \boldsymbol{\mu}} [\log p(\mathbf{Y}, \boldsymbol{\mu}, \mathbf{z}_i, \mathbf{z}_{-i} \mid \boldsymbol{\vartheta})] \right\}. \quad (42)$$

Here, the expectation is taken with respect to all the variables in $\boldsymbol{\mu}$ and \mathbf{Z} except for \mathbf{z}_i . Taking the log of this expression and leaving out everything that does not depend on \mathbf{z}_i leads to the following functional form:

$$\log q^*(\mathbf{z}_i) = \mathbb{E}_{\boldsymbol{\mu}_k} \left[\sum_k z_{ik} [\log(\pi_k) + \log \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{U} \boldsymbol{\mu}_k, \mathbf{S}_k)] \right] + \mathbf{C}, \quad (43)$$

$$= \sum_k z_{ik} [\log(\pi_k) + \mathbb{E}_{\boldsymbol{\mu}_k} [\log \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{U} \boldsymbol{\mu}_k, \mathbf{S}_k)]] + \mathbf{C}. \quad (44)$$

Here we recognize the functional form of a multinomial distribution:

$$q^*(\mathbf{z}_i) = \mathcal{M}_K(\mathbf{z}_i \mid 1, \boldsymbol{\tau}_i), \quad (45)$$

with:

$$\tau_{ik} \propto \pi_k \exp \{ \mathbb{E}_{\boldsymbol{\mu}_k} [\log \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{U} \boldsymbol{\mu}_k, \mathbf{S}_k)] \}. \quad (46)$$

□

A.2 Optimization of $q(\boldsymbol{\mu})$

Proof of Proposition 2 on page 8. Let $k \in \{1, \dots, K\}$, again the optimal $q(\boldsymbol{\mu}_k)$ verifies:

$$q^*(\boldsymbol{\mu}_k) \propto \exp \left\{ \mathbb{E}_{\mathbf{Z}, \boldsymbol{\mu}_{-k}} [\log p(\mathbf{Y}, \boldsymbol{\mu}_k, \boldsymbol{\mu}_{-k}, \mathbf{Z}) \mid \boldsymbol{\vartheta}] \right\}, \quad (47)$$

where the expectation is taken with respect to all variables in \mathbf{Z} and $\boldsymbol{\mu}$ except for $\boldsymbol{\mu}_k$. Taking the logarithm of this expression and leaving out everything that does not depend on $\boldsymbol{\mu}_k$ leads to the following functional form:

$$\log q^*(\boldsymbol{\mu}_k) = \log p(\boldsymbol{\mu}_k) + \sum_{i=1}^n \mathbb{E}_{\mathbf{z}_i, \boldsymbol{\mu}_{-k}} [\log p(\mathbf{y}_i \mid \mathbf{z}_i, \boldsymbol{\mu}_{-k})] + \mathbf{C}_1, \quad (48)$$

$$= \log \mathcal{N}_d(\boldsymbol{\mu}_k \mid \boldsymbol{\nu}, \lambda \mathbf{I}_d) + \sum_{i=1}^n \tau_{ik} \log \mathcal{N}_p(\mathbf{y}_i \mid \mathbf{U} \boldsymbol{\mu}_k, \mathbf{S}_k) + \mathbf{C}_2 \quad (49)$$

$$= -\frac{1}{2} \left[\lambda^{-1} (\boldsymbol{\mu}_k - \boldsymbol{\nu})^\top (\boldsymbol{\mu}_k - \boldsymbol{\nu}) + \sum_{i=1}^n \tau_{ik} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top \mathbf{S}_k^{-1} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k) \right] + \mathbf{C}_3, \quad (50)$$

$$= -\frac{1}{2} \left[\lambda^{-1} \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k + \sum_{i=1}^n \tau_{ik} \boldsymbol{\mu}_k^\top \mathbf{U}^\top \mathbf{S}_k^{-1} \mathbf{U} \boldsymbol{\mu}_k - 2 \left(\sum_{i=1}^n \tau_{ik} \mathbf{y}_i^\top \mathbf{S}_k^{-1} \mathbf{U} \boldsymbol{\mu}_k + \lambda^{-1} \boldsymbol{\nu}^\top \boldsymbol{\mu}_k \right) \right] + \mathbf{C}_4, \quad (51)$$

$$= -\frac{1}{2} \left[\boldsymbol{\mu}_k^\top (\lambda^{-1} \mathbf{I}_d + \tilde{n}_k \mathbf{U}^\top \mathbf{S}_k^{-1} \mathbf{U}) \boldsymbol{\mu}_k - 2 \left(\sum_{i=1}^n \tau_{ik} \mathbf{y}_i^\top \mathbf{S}_k^{-1} \mathbf{U} + \lambda^{-1} \boldsymbol{\nu}^\top \right) \boldsymbol{\mu}_k \right] + \mathbf{C}_4. \quad (52)$$

Putting:

$$\tilde{\mathbf{M}}_k = \left(\lambda^{-1} \mathbf{I}_d + \tilde{n}_k \mathbf{U}^\top \mathbf{S}_k^{-1} \mathbf{U} \right)^{-1}, \quad (53)$$

$$\tilde{\boldsymbol{\mu}}_k = \tilde{\mathbf{M}}_k \left(\sum_{i=1}^n \tau_{ik} \mathbf{U}^\top \mathbf{S}_k^{-1} \mathbf{y}_i + \lambda^{-1} \boldsymbol{\nu} \right) \quad (54)$$

we can then write:

$$\log q^*(\boldsymbol{\mu}_k) = -\frac{1}{2} \left[(\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k)^\top \tilde{\mathbf{M}}_k^{-1} (\boldsymbol{\mu}_k - \tilde{\boldsymbol{\mu}}_k) \right] + \text{const} . \quad (55)$$

We recognize the logarithm of a Gaussian density with mean $\tilde{\boldsymbol{\mu}}_k$ and covariance $\tilde{\mathbf{M}}_k$. Moreover, the expressions of $\tilde{\boldsymbol{\mu}}_k$ and $\tilde{\mathbf{M}}_k$ can be simplified, since $\mathbf{S}_k = \mathbf{D}\boldsymbol{\Delta}_k\mathbf{D}^\top$ with $\mathbf{D} = [\mathbf{U}, \mathbf{V}]$ and \mathbf{V} is the orthogonal complement of \mathbf{U} . Thus, $\mathbf{S}_k^{-1} = \mathbf{D}\boldsymbol{\Delta}_k^{-1}\mathbf{D}^\top$ and:

$$\mathbf{U}^\top \mathbf{S}_k^{-1} = \mathbf{U}^\top \mathbf{D}\boldsymbol{\Delta}_k^{-1}\mathbf{D}^\top = \boldsymbol{\Sigma}_k^{-1}\mathbf{U}^\top, \quad (56)$$

$$\mathbf{U}^\top \mathbf{S}_k^{-1}\mathbf{U} = \boldsymbol{\Sigma}_k^{-1}. \quad (57)$$

Thus, a use of Woodbury's identity gives

$$\begin{aligned} \tilde{\mathbf{M}}_k \lambda^{-1} \boldsymbol{\nu} &= \left(\mathbf{I}_d + \tilde{n}_k \mathbf{I}_d \lambda \mathbf{I}_d \boldsymbol{\Sigma}_k^{-1} \right)^{-1} \boldsymbol{\nu}, \\ &= \left[\mathbf{I}_d - \mathbf{I}_d \tilde{n}_k \left(\lambda^{-1} \mathbf{I}_d + \tilde{n}_k \boldsymbol{\Sigma}_k^{-1} \right)^{-1} \boldsymbol{\Sigma}_k^{-1} \mathbf{I}_d \right] \boldsymbol{\nu}, \\ &= \boldsymbol{\nu} - \tilde{\mathbf{M}}_k \boldsymbol{\Sigma}_k^{-1} \tilde{n}_k \boldsymbol{\nu}, \end{aligned}$$

and we finally get:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_k &= \tilde{\mathbf{M}}_k \left(\sum_{i=1}^n \tau_{ik} \mathbf{U}^\top \mathbf{S}_k^{-1} \mathbf{y}_i + \lambda^{-1} \boldsymbol{\nu} \right), \\ &= \tilde{\mathbf{M}}_k \boldsymbol{\Sigma}_k^{-1} \left(\sum_{i=1}^n \tau_{ik} \mathbf{U}^\top \mathbf{y}_i \right) + \tilde{\mathbf{M}}_k \lambda^{-1} \boldsymbol{\nu}, \\ &= \boldsymbol{\nu} + \tilde{\mathbf{M}}_k \boldsymbol{\Sigma}_k^{-1} \left(\sum_{i=1}^n \tau_{ik} \mathbf{U}^\top \mathbf{y}_i - \tilde{n}_k \boldsymbol{\nu} \right). \end{aligned} \quad (58)$$

□

A.3 Variational lower bound

We recall

$$\mathcal{J}(q, \boldsymbol{\vartheta}) = \mathbb{E}_q [\log p(\mathbf{Y}, \boldsymbol{\mu}, \mathbf{Z} \mid \boldsymbol{\vartheta})] - \mathbb{E}_q [\log q(\boldsymbol{\mu}, \mathbf{Z})].$$

Now that q^* have been derived in Propositions 1 and 2, we can compute the variational lower bound explicitly:

$$\begin{aligned} \mathcal{J}(q, \boldsymbol{\vartheta}) &= -\frac{1}{2} \sum_{i=1}^n \sum_{k=1}^K \tau_{ik} \left\{ p \log(2\pi) + \log |\mathbf{S}_k| + \mathbb{E}_{\boldsymbol{\mu}_k} \left[(\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top \mathbf{S}_k^{-1} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k) \right] \right\} \\ &\quad + \sum_{k=1}^K \tilde{n}_k \log(\pi_k) \\ &\quad - \frac{1}{2} \sum_{k=1}^K d \log(2\pi) + d \log(\lambda) + \frac{1}{\lambda} \mathbb{E}_{\boldsymbol{\mu}_k} \left[(\boldsymbol{\mu}_k - \boldsymbol{\nu})^\top (\boldsymbol{\mu}_k - \boldsymbol{\nu}) \right] \\ &\quad - \sum_i \sum_k \tau_{ik} \log(\tau_{ik}) \\ &\quad + \frac{Kd}{2} (\log(2\pi) + 1) + \frac{1}{2} \sum_k \log |\tilde{\mathbf{M}}_k|, \end{aligned} \quad (59)$$

with:

$$\begin{aligned} \mathbb{E}_{q^*} [\boldsymbol{\mu}_k] &= \tilde{\boldsymbol{\mu}}_k, \\ \mathbb{E}_{q^*} [\boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top] &= \tilde{\boldsymbol{\mu}}_k \tilde{\boldsymbol{\mu}}_k^\top + \tilde{\mathbf{M}}_k, \\ \log |\mathbf{S}_k| &= \log |\boldsymbol{\Sigma}_k| + (p - d) \log(\beta_k). \end{aligned} \quad (60)$$

Proof of Proposition 3 on page 8. Recall the form of the bound given in Proposition 3:

$$\begin{aligned} \mathcal{J}(\boldsymbol{\vartheta}) = \text{const} & - \frac{1}{2} \sum_{k=1}^K \tilde{n}_k \left\{ -2 \log(\pi_k) + \log |\boldsymbol{\Sigma}_k| + (p-d) \log(\beta_k) \right. \\ & \left. + \text{Tr} \left[\boldsymbol{\Sigma}_k^{-1} \mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U} \right] + \frac{1}{\beta_k} \left(\text{Tr} [\hat{\mathbf{C}}_k] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U}] \right) \right\}. \end{aligned} \quad (61)$$

Only the first two lines of Equation (59) depend on $\boldsymbol{\vartheta}$. Focusing on the first line:

$$\begin{aligned} & \mathbb{E}_q [\log p(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\vartheta})] \\ &= \sum_{i=1}^n \sum_{k=1}^K \tau_{ik} \left\{ p \log(2\pi) + \log |\mathbf{S}_k| + \mathbb{E}_{\boldsymbol{\mu}_k} \left[(\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top \mathbf{S}_k^{-1} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k) \right] \right\} \\ &= \gamma + \sum_{k=1}^K \tilde{n}_k \left\{ \log |\mathbf{S}_k| + \frac{1}{\tilde{n}_k} \mathbb{E}_{\boldsymbol{\mu}_k} \left[\sum_i \tau_{ik} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top \mathbf{S}_k^{-1} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k) \right] \right\}, \end{aligned}$$

where $\gamma = np \log(2\pi)$. The terms inside the expectation may be rearranged using the usual *trace trick* to make the empirical cluster covariance matrices appear:

$$\begin{aligned} & \frac{1}{\tilde{n}_k} \sum_i \tau_{ik} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top \mathbf{S}_k^{-1} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k), \\ &= \frac{1}{\tilde{n}_k} \sum_i \tau_{ik} \text{Tr} \left[(\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top \mathbf{S}_k^{-1} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k) \right], \\ &= \text{Tr} \left[\mathbf{S}_k^{-1} \frac{1}{\tilde{n}_k} \sum_i \tau_{ik} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k) (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top \right], \\ &= \text{Tr} \left[\mathbf{S}_k^{-1} \mathbf{C}_k \right], \end{aligned} \quad (62)$$

where

$$\mathbf{C}_k = \frac{1}{\tilde{n}_k} \sum_i \tau_{ik} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k) (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top,$$

is the empirical covariance matrices of cluster k . This trace may be further decomposed using the following lemma which relies on the particular form of \mathbf{S}_k^{-1} in the BDLM model.

Lemma 1. *For any square matrix $\mathbf{A} \in \mathcal{M}_{p \times p}(\mathbb{R})$, the following identity holds:*

$$\text{Tr} \left[\mathbf{S}_k^{-1} \mathbf{A} \right] = \text{Tr} \left[\boldsymbol{\Sigma}_k^{-1} \mathbf{U}^\top \mathbf{A} \mathbf{U} \right] + \frac{1}{\beta} \left(\text{Tr} [\mathbf{A}] - \text{Tr} [\mathbf{U}^\top \mathbf{A} \mathbf{U}] \right) \quad (63)$$

Proof. We can split \mathbf{S}_k^{-1} in two parts depending on the discriminative subspace and its orthogonal complement: take $\mathbf{D} = \tilde{\mathbf{D}} + \bar{\mathbf{D}}$, with $\tilde{\mathbf{D}} = [\mathbf{U}, \mathbf{0}_{p \times (p-d)}]$ and $\bar{\mathbf{D}} = [\mathbf{0}_{p \times d}, \mathbf{V}]$. Then, $\mathbf{S}_k^{-1} = \tilde{\mathbf{D}} \boldsymbol{\Delta}_k^{-1} \tilde{\mathbf{D}}^\top + \bar{\mathbf{D}} \boldsymbol{\Delta}_k^{-1} \bar{\mathbf{D}}^\top$, and:

$$\begin{aligned} \text{Tr} \left[\mathbf{S}_k^{-1} \mathbf{A} \right] &= \text{Tr} \left[\boldsymbol{\Delta}_k^{-1} \tilde{\mathbf{D}}^\top \mathbf{A} \tilde{\mathbf{D}} \right] + \text{Tr} \left[\boldsymbol{\Delta}_k^{-1} \bar{\mathbf{D}}^\top \mathbf{A} \bar{\mathbf{D}} \right], \\ &= \text{Tr} \left[\boldsymbol{\Sigma}_k^{-1} \mathbf{U}^\top \mathbf{A} \mathbf{U} \right] + \frac{1}{\beta} \text{Tr} \left[\mathbf{V}^\top \mathbf{A} \mathbf{V} \right] \end{aligned}$$

Moreover, $\mathbf{D} \mathbf{D}^\top = \mathbf{D}^\top \mathbf{D} = \mathbf{I}_p$ and $\mathbf{D} = \tilde{\mathbf{D}} + \bar{\mathbf{D}}$, hence:

$$\text{Tr} [\mathbf{A}] = \text{Tr} \left[\mathbf{D}^\top \mathbf{A} \mathbf{D} \right] = \text{Tr} \left[\mathbf{U}^\top \mathbf{A} \mathbf{U} \right] + \text{Tr} \left[\mathbf{V}^\top \mathbf{A} \mathbf{V} \right]$$

This concludes Lemma 1's proof. \square

Applying Lemma 1 to Equation (62) with $\mathbf{A} = \mathbb{E}_{\boldsymbol{\mu}_k}(\mathbf{C}_k)$ leaves:

$$\begin{aligned} \mathbb{E}_q[\log p(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\mu}, \boldsymbol{\vartheta})] = & \gamma + \sum_{k=1}^K \tilde{n}_k \left\{ \log |\boldsymbol{\Sigma}_k| + (p-d) \log(\beta_k) + \text{Tr} \left[\boldsymbol{\Sigma}_k^{-1} \mathbf{U}^\top \mathbb{E}(\mathbf{C}_k) \mathbf{U} \right] \right. \\ & \left. + \frac{1}{\beta} \left(\text{Tr} [\mathbb{E}(\mathbf{C}_k)] - \text{Tr} [\mathbf{U}^\top \mathbb{E}(\mathbf{C}_k) \mathbf{U}] \right) \right\}. \end{aligned} \quad (64)$$

The matrix $\mathbb{E}_{\boldsymbol{\mu}_k}(\mathbf{C}_k)$ is denoted as $\hat{\mathbf{C}}_k$ and, using Equation (60), one gets:

$$\begin{aligned} \hat{\mathbf{C}}_k &= \mathbb{E}_{\boldsymbol{\mu}_k}(\mathbf{C}_k), \\ &= \frac{1}{\tilde{n}_k} \sum_i \tau_{ik} \mathbb{E} \left[(\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)(\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k)^\top \right], \\ &= \frac{1}{\tilde{n}_k} \sum_i \tau_{ik} \left(\mathbf{y}_i \mathbf{y}_i^\top - \mathbf{U} \tilde{\boldsymbol{\mu}}_k \mathbf{y}_i^\top - \mathbf{y}_i (\mathbf{U} \tilde{\boldsymbol{\mu}}_k)^\top + \mathbf{U} \mathbb{E} \left[\boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top \right] \mathbf{U}^\top \right), \\ &= \frac{1}{\tilde{n}_k} \sum_{i=1}^n \tau_{ik} (\mathbf{y}_i - \mathbf{U} \tilde{\boldsymbol{\mu}}_k)(\mathbf{y}_i - \mathbf{U} \tilde{\boldsymbol{\mu}}_k)^\top + \mathbf{U} \tilde{\mathbf{M}}_k \mathbf{U}^\top. \end{aligned} \quad (65)$$

Finally, the second line of Equation (59) is simply

$$\mathbb{E}_q[\log p(\mathbf{Z} \mid \boldsymbol{\pi})] = -\frac{1}{2} \sum_{k=1}^K -2\tilde{n}_k \log(\pi_k). \quad (66)$$

This concludes the proof. \square

A.4 M-step

Proof of Proposition 4 on page 9. Although there are 12 different submodels, a lot of the proofs are the same.

Optimization of β . Let us start with the two possible cases for β , which are common regardless of the constraint on the latent covariance matrices:

- **Model BDL $\text{M}_{[(\cdot)\beta_k]}$:** In this case, the variational bound as a function of β_k is:

$$\mathcal{J}(\beta_k) = -\frac{1}{2} \tilde{n}_k \left[(p-d) \log(\beta_k) + \frac{1}{\beta_k} \left(\text{Tr} [\hat{\mathbf{C}}_k] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U}] \right) \right].$$

Thus, its only stationary point is:

$$\nabla_{\beta_k} \mathcal{J}(\hat{\beta}_k) = 0 \iff \hat{\beta}_k = \frac{\text{Tr} [\hat{\mathbf{C}}_k] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U}]}{p-d}.$$

- **Model BDL $\text{M}_{[(\cdot)\beta]}$:** In this case, the variational bound as a function of β is:

$$\begin{aligned} \mathcal{J}(\beta) &= -\frac{1}{2} \sum_{k=1}^K \tilde{n}_k \left\{ (p-d) \log(\beta) + \frac{1}{\beta} \left(\text{Tr} [\hat{\mathbf{C}}_k] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U}] \right) \right\}, \\ &= -\frac{1}{2} n \left\{ (p-d) \log(\beta) + \frac{1}{\beta} \left(\text{Tr} \left[\frac{1}{n} \sum_{k=1}^K \tilde{n}_k \hat{\mathbf{C}}_k \right] - \text{Tr} \left[\mathbf{U}^\top \left(\frac{1}{n} \sum_{k=1}^K \tilde{n}_k \hat{\mathbf{C}}_k \right) \mathbf{U} \right] \right) \right\}. \end{aligned}$$

And, again, its only stationary point is:

$$\nabla_{\beta} \mathcal{J}(\hat{\beta}) = 0 \iff \hat{\beta} = \frac{\text{Tr} [\hat{\mathbf{C}}] - \text{Tr} [\mathbf{U}^\top \hat{\mathbf{C}} \mathbf{U}]}{p-d},$$

with $\hat{\mathbf{C}} = \frac{1}{n} \sum_{k=1}^K \tilde{n}_k \hat{\mathbf{C}}_k$.

□

Optimization of Σ . There are now 6 cases to treat, which are the full, diagonal and isotropic covariance matrices where each case can be with or without homoscedasticity. We will need the two following formulas concerning matrix derivation. For any invertible matrix, $\mathbf{A} \in \mathbb{R}^{p \times p}$ we have:

$$\nabla_{\mathbf{A}} \log |\mathbf{A}| = \mathbf{A}^{-1}, \quad (67)$$

$$\nabla_{\mathbf{A}} \text{Tr} [\mathbf{A}^{-1} \mathbf{B}] = -(\mathbf{A}^{-1} \mathbf{B} \mathbf{A}^{-1})^{\top}. \quad (68)$$

- Model BDL $M_{[\Sigma_k(\cdot)]}$: We rewrite the bound of Equation (10) as a function of Σ_k :

$$\mathcal{J}(\Sigma_1, \dots, \Sigma_K) = -\frac{1}{2} \sum_{k=1}^K \tilde{n}_k \left\{ \log |\Sigma_k| + \text{Tr} [\Sigma_k^{-1} \mathbf{U}^{\top} \hat{\mathbf{C}}_k \mathbf{U}] \right\} + \text{const} . \quad (69)$$

Thus, using Equations (67) and (68) with $\mathbf{A} = \Sigma_k$ we get:

$$\nabla_{\Sigma_k} \mathcal{J}(\Sigma_k) = -\frac{\tilde{n}_k}{2} \left(\Sigma_k^{-1} - \Sigma_k^{-1} \mathbf{U}^{\top} \hat{\mathbf{C}}_k \mathbf{U} \Sigma_k^{-1} \right).$$

Then, a first order condition gives

$$\nabla_{\Sigma_k} \mathcal{J}(\Sigma_k) = 0 \iff \Sigma_k^{-1} = \Sigma_k^{-1} \mathbf{U}^{\top} \hat{\mathbf{C}}_k \mathbf{U} \Sigma_k^{-1}, \quad (70)$$

and we obtain the M-step estimate $\hat{\Sigma}_k$ by multiplying left and right by Σ_k :

$$\hat{\Sigma}_k = \mathbf{U}^{\top} \hat{\mathbf{C}}_k \mathbf{U}. \quad (71)$$

- Model BDL $M_{[\Sigma(\cdot)]}$: In this case, the variational bound can be rewritten as:

$$\mathcal{J}(\Sigma) = -\frac{n}{2} \left\{ \log |\Sigma| + \text{Tr} \left[\Sigma^{-1} \mathbf{U}^{\top} \left(\frac{1}{n} \sum_{k=1}^K \tilde{n}_k \hat{\mathbf{C}}_k \right) \mathbf{U} \right] \right\} + \text{const} . \quad (72)$$

And finding the root of the gradient leads to:

$$\hat{\Sigma} = \mathbf{U}^{\top} \hat{\mathbf{C}} \mathbf{U}. \quad (73)$$

- Model BDL $M_{[\alpha_{kh}(\cdot)]}$: In this model, the bound writes as a function of $\alpha = (\alpha_1, \dots, \alpha_K)$:

$$\mathcal{J}(\alpha) = -\frac{1}{2} \sum_{k=1}^K \tilde{n}_k \left\{ \sum_{h=1}^d \log(\alpha_{kh}) + \frac{\mathbf{u}_h^{\top} \hat{\mathbf{C}}_k \mathbf{u}_h}{\alpha_{kh}} \right\} + \text{const} . \quad (74)$$

Thus, the partial derivative with respect to α_{kh} is given as:

$$\nabla_{\alpha_{kh}} \mathcal{J}(\alpha_{kh}) = -\frac{\tilde{n}_k}{2} \left(\frac{1}{\alpha_{kh}} - \frac{\mathbf{u}_h^{\top} \hat{\mathbf{C}}_k \mathbf{u}_h}{\alpha_{kh}^2} \right),$$

and finding its root gives:

$$\hat{\alpha}_{kh} = \mathbf{u}_h^{\top} \hat{\mathbf{C}}_k \mathbf{u}_h. \quad (75)$$

- Model BDL $M_{[\alpha_h(\cdot)]}$: Put $\alpha = (\alpha_1, \dots, \alpha_d)$ and we have,

$$\mathcal{J}(\alpha) = -\frac{n}{2} \left\{ \sum_{h=1}^d \log(\alpha_h) + \frac{\mathbf{u}_h^{\top} \left(\frac{1}{n} \sum_{k=1}^K \tilde{n}_k \hat{\mathbf{C}}_k \right) \mathbf{u}_h}{\alpha_h} \right\} + \text{const} . \quad (76)$$

Its gradient with respect to α_h is computed in the same manner as above, and a first-order condition gives:

$$\hat{\alpha}_h = \mathbf{u}_h^{\top} \hat{\mathbf{C}} \mathbf{u}_h \quad (77)$$

- Model $\text{BDLM}_{[\alpha_k(\cdot)]}$: Introducing $\alpha = (\alpha_1, \dots, \alpha_K)$, the bound is written as:

$$\mathcal{J}(\alpha) = -\frac{1}{2} \sum_{k=1}^K \tilde{n}_k \left\{ d \log(\alpha_k) + \frac{1}{\alpha_k} \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U} \right] \right\} + \text{const} . \quad (78)$$

Again, its gradient with respect to α_k is easily computed as:

$$\nabla_{\alpha_k} \mathcal{J}(\alpha_k) = -\frac{\tilde{n}_k}{2} \left(\frac{d}{\alpha_k} - \frac{\text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U} \right]}{\alpha_k^2} \right) .$$

Finding its 0 point leaves the following M-step update for $\hat{\alpha}_k$

$$\hat{\alpha}_k = \frac{1}{d} \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}}_k \mathbf{U} \right] . \quad (79)$$

- Model $\text{BDLM}_{[\alpha(\cdot)]}$: For this final model, α is a positive scalar and the bound is:

$$\mathcal{J}(\alpha) = -\frac{n}{2} \left\{ d \log(\alpha) + \frac{1}{\alpha} \text{Tr} \left[\mathbf{U}^\top \left(\frac{1}{n} \sum_{k=1}^K \tilde{n}_k \hat{\mathbf{C}}_k \right) \mathbf{U} \right] \right\} + \text{const} . \quad (80)$$

$$\hat{\alpha} = \frac{1}{d} \text{Tr} \left[\mathbf{U}^\top \hat{\mathbf{C}} \mathbf{U} \right] . \quad (81)$$

□

□

A.5 Hyper-parameter estimation

To prove Proposition 5, we maximize the variational bound with respect to $(\boldsymbol{\nu}, \lambda)$. From Equation (59) we get:

$$\begin{aligned} \mathcal{J}(\boldsymbol{\nu}, \lambda) &= -\frac{1}{2} \sum_{k=1}^K d \log(2\pi) + d \log(\lambda) + \frac{1}{\lambda} \mathbb{E}_{\boldsymbol{\mu}_k} \left[\|\boldsymbol{\mu}_k - \boldsymbol{\nu}\|_2^2 \right] , \\ &= -\frac{1}{2} \sum_{k=1}^K d \log(2\pi) + d \log(\lambda) + \frac{1}{\lambda} \left[\|\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\nu}\|_2^2 + \text{Tr} \left[\tilde{\mathbf{M}}_k \right] \right] . \end{aligned}$$

Optimization with respect to $\boldsymbol{\nu}$.

$$\nabla_{\boldsymbol{\nu}} \mathcal{J}(\boldsymbol{\nu}) = -\frac{1}{\lambda} \sum_{k=1}^K (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\nu}) \quad (82)$$

Hence,

$$\begin{aligned} \nabla_{\boldsymbol{\nu}} \mathcal{J}(\hat{\boldsymbol{\nu}}) &= 0, \\ \iff \hat{\boldsymbol{\nu}} &= \frac{\sum_{k=1}^K \tilde{\boldsymbol{\mu}}_k}{K} \end{aligned}$$

Since \mathcal{J} is a concave function of $\boldsymbol{\nu}$ with a negative definite Hessian $\frac{-1}{2\lambda} \mathbf{I}_d$, this concludes the proof □

Optimization with respect to λ .

$$\nabla_{\lambda} \mathcal{J}(\lambda) = -\frac{1}{2} \sum_{k=1}^K \frac{d}{\lambda} - \frac{1}{\lambda^2} \left[\|\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\nu}\|_2^2 + \text{Tr} \left[\tilde{\mathbf{M}}_k \right] \right] .$$

Thus, the first-order condition gives:

$$\begin{aligned}\nabla_{\lambda} \mathcal{J}(\hat{\lambda}) &= 0, \\ \hat{\lambda} &= \frac{\sum_{k=1}^K \|\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\nu}\|_2^2 + \text{Tr}[\tilde{\mathbf{M}}_k]}{dK}.\end{aligned}$$

The second-order derivative gives us a condition for $\hat{\lambda}$ to be a maximum. Indeed, the following must hold for $\lambda = \hat{\lambda}$:

$$\nabla_{\lambda}^2 \mathcal{J}(\lambda) = \frac{\left(\sum_{k=1}^K \|\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\nu}\|_2^2 + \text{Tr}[\tilde{\mathbf{M}}_k]\right) \lambda - d}{2\lambda^3} < 0.$$

Hence, for positive λ , the latter is negative if and only if $\lambda < \frac{\sum_{k=1}^K \|\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\nu}\|_2^2 + \text{Tr}[\tilde{\mathbf{M}}_k]}{d} = K\hat{\lambda}$. Obviously, $\hat{\lambda}$ verifies this condition for $K \geq 2$ and is thus a maximum of \mathcal{J} . \square

A.6 Model selection

In this section, we briefly recall the key elements for the derivation of the ICL_{BIC} of Biernacki et al. (2000), given in Equation (33) on page 14. Then, we detail the computation of the latter via the variational lower bound. For the sake of notations, we drop the dependencies in \mathcal{M} and K here, since the discussion is independent of these quantities.

Derivation of the ICL Taking a Bayesian perspective on the model parameters $\boldsymbol{\vartheta} = (\boldsymbol{\pi}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \mathbf{U})$, we posit a prior distribution $p(\boldsymbol{\vartheta})$ and the ICL computes the following quantity:

$$\log p(\mathbf{Y}, \mathbf{Z}) = \log \int_{\boldsymbol{\vartheta}} \log p(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\vartheta}) d\boldsymbol{\vartheta}.$$

Considering a factorized prior $p(\boldsymbol{\vartheta}) = p(\boldsymbol{\pi})p(\boldsymbol{\Sigma}, \boldsymbol{\beta}, \mathbf{U})$, we can rewrite the ICL as the sum of two terms:

$$\log p(\mathbf{Y}, \mathbf{Z}) = \log p(\mathbf{Y} \mid \mathbf{Z}) + \log p(\mathbf{Z}).$$

Biernacki et al. (2000) proposed to approximate the first term with a Laplace approximation:

$$\log p(\mathbf{Y} \mid \mathbf{Z}) \approx \max_{\boldsymbol{\Sigma}, \boldsymbol{\beta}, \mathbf{U}} \log p(\mathbf{Y} \mid \mathbf{Z}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \mathbf{U}) - \frac{\gamma(\boldsymbol{\Sigma}, \boldsymbol{\beta}, \mathbf{U})}{2} \log(n),$$

where γ is the number of free parameters in $(\boldsymbol{\Sigma}, \boldsymbol{\beta}, \mathbf{U})$. Then, putting a symmetric Dirichlet prior on $\boldsymbol{\pi} \sim \mathcal{D}_K(\boldsymbol{\alpha} = (\alpha, \dots, \alpha))$, the second term can be computed exactly, involving Gamma functions. Fixing $\alpha = 1/2$, and assuming that each $n_k = \sum_i z_{ik}$ is large with n , a Stirling approximation on the Gamma functions is used to get:

$$\log p(\mathbf{Z}) \approx \max_{\boldsymbol{\pi}} \log p(\mathbf{Z} \mid \boldsymbol{\pi}) - \frac{K-1}{2} \log(n).$$

Finally, summing these two approximations, we get a BIC-like approximation of the integrated classification likelihood:

$$\text{ICL}_{BIC}(\mathcal{M}, K) = \max_{\boldsymbol{\vartheta}} \log p(\mathbf{Y}, \mathbf{Z} \mid \boldsymbol{\vartheta}) - \frac{\gamma_{\mathcal{M}, K}}{2} \log(n).$$

Computing the classification likelihood with the variational bound We now turn into the problem of computing $\log p(\mathbf{Y}, \hat{\mathbf{Z}} \mid \hat{\boldsymbol{\vartheta}})$ with the variational bound \mathcal{J} . We consider the parameters $\hat{\boldsymbol{\vartheta}}$ to be estimated via the BFEM algorithm and fixed. Consider $\mathbf{Z} = \hat{\mathbf{Z}}$ fixed and known. The result stems from the fact that the conditional posterior $p(\boldsymbol{\mu} \mid \hat{\mathbf{Z}}, \mathbf{Y})$ is tractable in the BDLM model and it is equal to $q(\boldsymbol{\mu}_k)$ if $\tau_{ik} = \hat{z}_{ik}$. Thus, the variational bound in is tight and equals the classification likelihood.

Formally, we want to show that the variational bound of Equation (59) is equal to the classification likelihood $\log p(\mathbf{Y}, \hat{\mathbf{Z}} \mid \hat{\boldsymbol{\vartheta}})$ when $\tau_{ik} = \hat{z}_{ik}$ in Propositions 1 and 2. When $\boldsymbol{\tau} \leftarrow \hat{\mathbf{Z}}$, we have:

$$q(\boldsymbol{\mu}, \mathbf{Z} \mid \hat{\mathbf{Z}}) = q\left(\boldsymbol{\mu} \mid \tilde{\boldsymbol{\mu}}(\hat{\mathbf{Z}}), \tilde{\mathbf{M}}(\hat{\mathbf{Z}})\right) \times \delta_{\hat{\mathbf{Z}}}(\mathbf{Z}), \quad (83)$$

with δ_x the Dirac mass at x , and $(\tilde{\mu}, \tilde{\mathbf{M}})$ computed with $\hat{\mathbf{Z}}$ instead of $\boldsymbol{\tau}$:

$$\begin{aligned}\hat{n}_k &= \sum_i \hat{z}_{ik}, \\ \tilde{\mathbf{M}}_k(\hat{\mathbf{Z}}) &= \left(\lambda^{-1} \mathbf{I}_d + \hat{n}_k \boldsymbol{\Sigma}_k^{-1} \right)^{-1}, \\ \tilde{\mu}_k(\hat{\mathbf{Z}}) &= \boldsymbol{\nu} + \tilde{\mathbf{M}}_k \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{U}^\top \sum_i \hat{z}_{ik} \mathbf{y}_i - \hat{n}_k \boldsymbol{\nu} \right).\end{aligned}$$

It happens that the conditional posterior distribution of $\boldsymbol{\mu}$: $p(\boldsymbol{\mu} \mid \hat{\mathbf{Z}}, \mathbf{Y})$ is tractable as a product of K distributions since:

$$\begin{aligned}p(\boldsymbol{\mu} \mid \hat{\mathbf{Z}}, \mathbf{Y}, \hat{\boldsymbol{\vartheta}}) &\propto p(\boldsymbol{\mu}) p(\mathbf{Y} \mid \hat{\mathbf{Z}}, \boldsymbol{\mu}, \hat{\boldsymbol{\vartheta}}), \\ &\propto \prod_{k=1}^K \mathcal{N}_d(\boldsymbol{\mu}_k \mid \boldsymbol{\nu}, \lambda) \times \prod_{i=1}^n \prod_{k=1}^K \mathcal{N}_p(\mathbf{y}_i \mid \hat{\mathbf{U}} \boldsymbol{\mu}_k, \hat{\mathbf{S}}_k)^{\hat{z}_{ik}}, \\ &\propto \prod_{k=1}^K \left\{ \mathcal{N}_d(\boldsymbol{\mu}_k \mid \boldsymbol{\nu}, \lambda) \times \prod_{i \in \hat{\mathcal{C}}_k} \mathcal{N}_p(\mathbf{y}_i \mid \hat{\mathbf{U}} \boldsymbol{\mu}_k, \hat{\mathbf{S}}_k) \right\}.\end{aligned}$$

Using the same reasoning as in Appendix A.2, the distributions into bracket happens to be (un-normalized) Gaussians with parameter $(\tilde{\mu}_k(\hat{\mathbf{Z}}), \tilde{\mathbf{M}}_k(\hat{\mathbf{Z}}))$. Thus, we have that:

$$p(\boldsymbol{\mu} \mid \hat{\mathbf{Z}}, \mathbf{Y}, \hat{\boldsymbol{\vartheta}}) = q(\boldsymbol{\mu} \mid \tilde{\mu}(\hat{\mathbf{Z}}), \tilde{\mathbf{M}}(\hat{\mathbf{Z}})). \quad (84)$$

Finally, denoting $\mathbf{H}(q) = \mathbb{E}_{\boldsymbol{\eta} \sim q}[q(\boldsymbol{\eta})]$, we may write the expression of the variational bound $\mathcal{J}(\hat{q}, \hat{\boldsymbol{\vartheta}})$ with \hat{q} defined in Equation (83), and get:

$$\begin{aligned}\mathcal{J}(\hat{q}, \hat{\boldsymbol{\vartheta}}) &= \mathbb{E}_{(\boldsymbol{\mu}, \mathbf{Z}) \sim \hat{q}} \left[\log(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\mu} \mid \hat{\boldsymbol{\vartheta}}) \right] + \mathbf{H}(\hat{q}), \\ &= \mathbb{E}_{\boldsymbol{\mu} \sim q(\boldsymbol{\mu} \mid \tilde{\mu}(\hat{\mathbf{Z}}), \tilde{\mathbf{M}}(\hat{\mathbf{Z}}))} \left[\mathbb{E}_{\mathbf{Z} \sim \delta_{\hat{\mathbf{Z}}}} \left[\log(\mathbf{Y}, \mathbf{Z}, \boldsymbol{\mu} \mid \hat{\boldsymbol{\vartheta}}) \right] \right] + \mathbf{H} \left(q(\boldsymbol{\mu} \mid \tilde{\mu}(\hat{\mathbf{Z}}), \tilde{\mathbf{M}}(\hat{\mathbf{Z}})) \right), \\ &= \mathbb{E}_{\boldsymbol{\mu} \sim p(\boldsymbol{\mu} \mid \hat{\mathbf{Z}}, \mathbf{Y})} \left[\log(\mathbf{Y}, \hat{\mathbf{Z}}, \boldsymbol{\mu} \mid \hat{\boldsymbol{\vartheta}}) \right] + \mathbf{H} \left(p(\boldsymbol{\mu} \mid \hat{\mathbf{Z}}, \mathbf{Y}, \hat{\boldsymbol{\vartheta}}) \right), \\ &= \log p(\mathbf{Y}, \hat{\mathbf{Z}} \mid \hat{\boldsymbol{\vartheta}}) - \text{KL}(p(\boldsymbol{\mu} \mid \hat{\mathbf{Z}}, \mathbf{Y}, \hat{\boldsymbol{\vartheta}}) \parallel p(\boldsymbol{\mu} \mid \hat{\mathbf{Z}}, \mathbf{Y}, \hat{\boldsymbol{\vartheta}})), \\ &= \log p(\mathbf{Y}, \hat{\mathbf{Z}} \mid \hat{\boldsymbol{\vartheta}}).\end{aligned}$$

Here, the second line used the fact that $\mathbf{H}(\delta_{\hat{\mathbf{Z}}}) = 0$, and the last equality used the tightness of the variational lower bound when q is equal to the posterior distribution.

B Image denoising

Proof of Proposition 6 on page 23. Denote $\psi_k(\mathbf{Y}, \boldsymbol{\mu}) = \mathbb{E}[\mathbf{t}_i \mid \mathbf{y}_i, \boldsymbol{\mu}_k, z_{ik} = 1]$ and $\mathbf{m}_k = \mathbf{U} \boldsymbol{\mu}_k$. We begin by recalling that Equation (38) is obtained from the fact that $(\mathbf{t}_i, \mathbf{y}_i) \mid \{z_{ik} = 1\}$ is jointly Gaussian, with covariance $\mathbf{S}_k - \sigma^2 \mathbf{I}_p$. Thus,

$$\begin{aligned}\psi_k(\mathbf{Y}, \boldsymbol{\mu}) &= \mathbf{U} \boldsymbol{\mu}_k + \left(\mathbf{S}_k - \sigma^2 \mathbf{I}_p \right) \mathbf{S}_k^{-1} (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k), \\ &= \mathbf{U} \boldsymbol{\mu}_k + \left(\mathbf{I}_p - \sigma^2 \mathbf{S}_k^{-1} \right) (\mathbf{y}_i - \mathbf{U} \boldsymbol{\mu}_k),\end{aligned}$$

This expression involves direct inverse of \mathbf{S}_k^{-1} which can be avoided thanks to the particular block structure made on the matrix \mathbf{S}_k in the BDL model. Keeping the same notations as in Lemma 1, we split \mathbf{S}_k^{-1} in two parts

depending on the discriminative subspace and its orthogonal complement: $S_k^{-1} = \tilde{D}\Delta_k^{-1}\tilde{D}^\top + \bar{D}\Delta_k^{-1}\bar{D}^\top$, with $\tilde{D} = [U, \mathbf{0}_{p \times (p-d)}]$ and $\bar{D} = [\mathbf{0}_{p \times d}, V]$. Then, we have:

$$\begin{cases} \bar{D}\Delta_k^{-1}\bar{D}^\top &= \frac{1}{\sigma^2}\bar{D}\bar{D}^\top \\ \bar{D}\bar{D}^\top + \tilde{D}\tilde{D}^\top &= I_p \end{cases} \implies \bar{D}\Delta_k^{-1}\bar{D}^\top = \frac{1}{\sigma^2}(I_p - \tilde{D}\tilde{D}^\top).$$

Thus,

$$\begin{aligned} \sigma^2 S_k^{-1} &= \sigma^2 [\tilde{D}\Delta_k^{-1}\tilde{D}^\top + \bar{D}\Delta_k^{-1}\bar{D}^\top], \\ &= I_p + \tilde{D}(\sigma^2\Delta_k^{-1} - I_p)\tilde{D}^\top, \end{aligned}$$

and we can thus simplify the expression in Equation (38) into:

$$\psi_k(Y, \mu) = U\mu_k + \tilde{D}(I_p - \sigma^2\Delta_k^{-1})\tilde{D}^\top(y_i - U\mu_k),$$

Finally, in BDLM, μ_k is a random variable and we can further take the expectation of $\psi_k(Y, \mu)$ according to the variational distribution $q(\mu_k)$, leaving:

$$\psi_k(Y) = \mathbb{E}_{q(\mu_k)} \left[\mathbb{E}[t_i | y_i, \mu_k, z_{ik} = 1] \right] = U\tilde{\mu}_k + \tilde{D}(I_p - \sigma^2\Delta_k^{-1})\tilde{D}^\top(y_i - U\tilde{\mu}_k) \quad (85)$$

Note that this Proposition closely resembles the Proposition 1 of Houdard et al. (2018), this is explained by the similar assumptions made by the HDDC and BDLM models. \square

References

- Aggarwal, Charu C and ChengXiang Zhai (2012). “A survey of text clustering algorithms”. In: *Mining text data*. Springer, pp. 77–128 (cit. on p. 1).
- Agrawal, Rakesh et al. (1998). “Automatic subspace clustering of high dimensional data for data mining applications”. In: *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pp. 94–105 (cit. on p. 2).
- Airoldi, Edoardo M et al. (2008). “Mixed membership stochastic blockmodels”. In: *Journal of machine learning research* 9.Sep, pp. 1981–2014 (cit. on p. 11).
- Baek, Jangsun, Geoffrey J McLachlan, and Lloyd K Flack (2009). “Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data”. In: *IEEE transactions on pattern analysis and machine intelligence* 32.7, pp. 1298–1309 (cit. on pp. 3, 4, 6, 14).
- Banfield, Jeffrey D and Adrian E Raftery (1993). “Model-based Gaussian and non-Gaussian clustering”. In: *Biometrics*, pp. 803–821 (cit. on pp. 2, 5).
- Baudry, Jean-Patrick et al. (2010). “Combining Mixture Components for Clustering.” In: *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 9 2, pp. 332–353 (cit. on p. 4).
- Bellman, Richard (1957). *Dynamic Programming*. Princeton University Press (cit. on p. 1).
- Bergé, Laurent, Charles Bouveyron, and Stéphane Girard (2019). *HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data*. R package version 2.2.0 (cit. on p. 14).
- Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (2000). “Assessing a mixture model for clustering with the integrated completed likelihood”. In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725 (cit. on pp. 3, 14, 33).
- Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer (cit. on pp. 7, 11).
- Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). “Variational inference: A review for statisticians”. In: *Journal of the American Statistical Association* 112.518, pp. 859–877 (cit. on pp. 7, 8, 27).
- Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan, pp. 993–1022 (cit. on p. 11).
- Bouveyron, Charles and Camille Brunet (2011). “On the estimation of the latent discriminative subspace in the Fisher-EM algorithm”. In: (cit. on p. 14).
- Bouveyron, Charles and Camille Brunet (2012a). “Simultaneous model-based clustering and visualization in the Fisher discriminative subspace”. In: *Statistics and Computing* 22.1, pp. 301–324 (cit. on pp. 1, 3–5, 7, 10–12, 14).
- Bouveyron, Charles and Camille Brunet (2012b). “Theoretical and practical considerations on the convergence properties of the Fisher-EM algorithm”. In: *Journal of Multivariate Analysis* 109, pp. 29–41 (cit. on p. 12).

- Bouveyron, Charles, Camille Brunet, and Nicolas Jouvin (2020). *FisherEM: The FisherEM Algorithm to Simultaneously Cluster and Visualize High-Dimensional Data*. R package version 1.5.2 (cit. on p. 14).
- Bouveyron, Charles and Camille Brunet-Saumard (2014a). “Discriminative variable selection for clustering with the sparse Fisher-EM algorithm”. In: *Computational Statistics* 29.3-4, pp. 489–513 (cit. on p. 25).
- Bouveyron, Charles and Camille Brunet-Saumard (2014b). “Model-based clustering of high-dimensional data: A review”. In: *Computational Statistics & Data Analysis* 71, pp. 52–78 (cit. on pp. 3, 6).
- Bouveyron, Charles, Stéphane Girard, and Cordelia Schmid (2007). “High-dimensional data clustering”. In: *Computational Statistics & Data Analysis* 52.1, pp. 502–519 (cit. on pp. 3, 6, 14, 23).
- Bouveyron, Charles et al. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Vol. 50. Cambridge University Press (cit. on pp. 2, 20).
- Buades, Antoni, Bartomeu Coll, and Jean-Michel Morel (2005). “A review of image denoising algorithms, with a new one”. In: *Multiscale Modeling & Simulation* 4.2, pp. 490–530 (cit. on p. 21).
- Buades, Antoni, Bartomeu Coll, and Jean-Michel Morel (2011). “Non-local means denoising”. In: *Image Processing On Line* 1, pp. 208–212 (cit. on p. 21).
- Celeux, Gilles and Gérard Govaert (1992). “A classification EM algorithm for clustering and two stochastic versions”. In: *Computational statistics & Data analysis* 14.3, pp. 315–332 (cit. on p. 3).
- Celeux, Gilles and Gérard Govaert (1995). “Gaussian parsimonious clustering models”. In: *Pattern recognition* 28.5, pp. 781–793 (cit. on p. 2).
- Chang, Wei-Chien (1983). “On using principal components before separating a mixture of two multivariate normal distributions”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 32.3, pp. 267–275 (cit. on pp. 2, 15).
- Coleman, Guy Barrett and Harry C Andrews (1979). “Image segmentation by clustering”. In: *Proceedings of the IEEE* 67.5, pp. 773–785 (cit. on p. 1).
- Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). “Maximum likelihood from incomplete data via the EM algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22 (cit. on p. 2).
- Duda, Richard O, Peter E Hart, and David G Stork (2000). *Pattern classification*. John Wiley & Sons (cit. on p. 3).
- Dy, Jennifer G and Carla E Brodley (2004). “Feature selection for unsupervised learning”. In: *Journal of machine learning research* 5.Aug, pp. 845–889 (cit. on p. 3).
- Fisher, Ronald A (1936). “The use of multiple measurements in taxonomic problems”. In: *Annals of eugenics* 7.2, pp. 179–188 (cit. on p. 3).
- Foley, Donald H. and John W Sammon (1975). “An optimal set of discriminant vectors”. In: *IEEE Transactions on computers* 100.3, pp. 281–289 (cit. on p. 11).
- Fruhwirth-Schnatter, Sylvia, Gilles Celeux, and Christian P Robert (2019). *Handbook of mixture analysis*. Chapman and Hall/CRC (cit. on p. 7).
- Fukunaga, Keinosuke (1990). *Introduction to Statistical Pattern Recognition (2nd Ed.)* USA: Academic Press Professional, Inc. (cit. on pp. 3, 10).
- Ge, Rong et al. (2016). “Efficient algorithms for large-scale generalized eigenvector computation and canonical correlation analysis”. In: *International Conference on Machine Learning*, pp. 2741–2750 (cit. on p. 12).
- Ghahramani, Zoubin and Geoffrey E Hinton (1996). *The EM algorithm for mixtures of factor analyzers*. Tech. rep. Technical Report CRG-TR-96-1, University of Toronto (cit. on p. 3).
- Ghojogh, Benyamin, Fakhri Karray, and Mark Crowley (2019). “Eigenvalue and generalized eigenvalue problems: Tutorial”. In: *arXiv preprint arXiv:1903.11240* (cit. on p. 10).
- Ghosh, Debashis and Arul M Chinnaiyan (2002). “Mixture modelling of gene expression data from microarray experiments”. In: *Bioinformatics* 18.2, pp. 275–286 (cit. on pp. 1, 2).
- Giraud, Christophe (2014). *Introduction to high-dimensional statistics*. Vol. 138. CRC Press (cit. on p. 1).
- Guo, Yue-Fei et al. (2003). “A generalized Foley–Sammon transform based on generalized fisher discriminant criterion and its application to face recognition”. In: *Pattern Recognition Letters* 24.1-3, pp. 147–158 (cit. on p. 25).
- Hamamoto, Yoshihiko et al. (1991). “A note on the orthonormal discriminant vector method for feature extraction”. In: *Pattern recognition* 24.7, pp. 681–684 (cit. on pp. 11, 14).
- Houdard, Antoine, Charles Bouveyron, and Julie Delon (2018). “High-dimensional mixture models for unsupervised image denoising (HDMI)”. In: *SIAM Journal on Imaging Sciences* 11.4, pp. 2815–2846 (cit. on pp. 23, 24, 35).
- Hubert, Lawrence and Phipps Arabie (1985). “Comparing partitions”. In: *Journal of classification* 2.1, pp. 193–218 (cit. on p. 15).
- Jaakkola, Tommi S and Michael I Jordan (2000). “Bayesian parameter estimation via variational methods”. In: *Statistics and Computing* 10.1, pp. 25–37 (cit. on p. 7).
- Jégou, Hervé et al. (2010). “Aggregating local descriptors into a compact image representation”. In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, pp. 3304–3311 (cit. on p. 1).
- Kokiopoulou, Effrosini, Jie Chen, and Yousef Saad (2011). “Trace optimization and eigenproblems in dimension reduction methods”. In: *Numerical Linear Algebra with Applications* 18.3, pp. 565–602 (cit. on p. 25).

- Lebrun, Marc, Antoni Buades, and Jean-Michel Morel (2013). “A nonlocal bayesian image denoising algorithm”. In: *SIAM Journal on Imaging Sciences* 6.3, pp. 1665–1688 (cit. on pp. 23, 24).
- Lu, Gui-Fu, Jian Zou, and Yong Wang (2016). “A new and fast implementation of orthogonal LDA algorithm and its incremental extension”. In: *Neural Processing Letters* 43.3, pp. 687–707 (cit. on p. 11).
- McLachlan, Geoffrey J and Thriyambakam Krishnan (2007). *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons (cit. on p. 12).
- McLachlan, Geoffrey J and David Peel (2004). *Finite mixture models*. John Wiley & Sons (cit. on p. 2).
- McNicholas, Paul D. et al. (2019). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.2.4 (cit. on p. 14).
- McNicholas, Paul David and Thomas Brendan Murphy (2008). “Parsimonious Gaussian mixture models”. In: *Statistics and Computing* 18.3, pp. 285–296 (cit. on pp. 3, 6, 14, 20).
- Montanari, Angela and Cinzia Viroli (2010). “Heteroscedastic factor mixture analysis”. In: *Statistical Modelling* 10.4, pp. 441–460 (cit. on pp. 3, 6, 15).
- Morris, Carl N (1983). “Parametric empirical Bayes inference: theory and applications”. In: *Journal of the American statistical Association* 78.381, pp. 47–55 (cit. on p. 11).
- Nie, Feiping et al. (2008). “Trace ratio criterion for feature selection.” In: *AAAI*. Vol. 2, pp. 671–676 (cit. on p. 3).
- Okada, Toshiniko and Shingo Tomita (1985). “An optimal orthonormal system for discriminant analysis”. In: *Pattern Recognition* 18.2, pp. 139–144 (cit. on pp. 11, 13).
- Parsons, Lance, Ehtesham Haque, and Huan Liu (2004). “Subspace clustering for high dimensional data: a review”. In: *Acm Sigkdd Explorations Newsletter* 6.1, pp. 90–105 (cit. on p. 2).
- Qiao, Zhihua, Lan Zhou, and Jianhua Z Huang (2009). “Sparse Linear Discriminant Analysis with Applications to High Dimensional Low Sample Size Data.” In: *International Journal of Applied Mathematics* 39.1 (cit. on p. 25).
- Rathnayake, Suren et al. (2019). *EMMIXmfa: Mixture Models with Component-Wise Factor Analyzers*. R package version 2.0.11 (cit. on p. 14).
- Scrucca, Luca (2010). “Dimension reduction for model-based clustering”. In: *Statistics and Computing* 20.4, pp. 471–484 (cit. on p. 3).
- Scrucca, Luca et al. (2016). “mclust 5: clustering, classification and density estimation using Gaussian finite mixture models”. In: *The R journal* 8.1, p. 289 (cit. on p. 3).
- Tipping, Michael E and Christopher M Bishop (1999). “Mixtures of probabilistic principal component analyzers”. In: *Neural computation* 11.2, pp. 443–482 (cit. on pp. 3, 22).
- Torre, Fernando De la and Takeo Kanade (2006). “Discriminative cluster analysis”. In: *Proceedings of the 23rd international conference on Machine learning*, pp. 241–248 (cit. on p. 3).
- Wang, Huan et al. (2007). “Trace ratio vs. ratio trace for dimensionality reduction”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1–8 (cit. on p. 25).
- Wang, Yi-Qing and Jean-Michel Morel (2013). “SURE guided Gaussian mixture image denoising”. In: *SIAM Journal on Imaging Sciences* 6.2, pp. 999–1034 (cit. on pp. 22–24).
- Wei, Xin and Chunguang Li (2013). “Bayesian mixtures of common factor analyzers: Model, variational inference, and applications”. In: *Signal processing* 93.11, pp. 2894–2905 (cit. on p. 6).
- Weinen, Multivariable Datenanalyse zur Sortenklassifizierung von (1986). “Multivariate data analysis as a discriminating method of the origin of wines”. In: *Vitis* 25, pp. 189–201 (cit. on p. 20).
- Ye, Jieping (2005). “Characterization of a family of algorithms for generalized discriminant analysis on undersampled problems”. In: *Journal of Machine Learning Research* 6.Apr, pp. 483–502 (cit. on pp. 3, 10, 11).
- Yoshida, Ryo, Tomoyuki Higuchi, and Seiya Imoto (2004). “A mixed factors model for dimension reduction and extraction of a group structure in gene expression data”. In: *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004*. IEEE, pp. 161–172 (cit. on p. 3).
- Yu, Guoshen, Guillermo Sapiro, and Stéphane Mallat (2011). “Solving inverse problems with piecewise linear estimators: From Gaussian mixture models to structured sparsity”. In: *IEEE Transactions on Image Processing* 21.5, pp. 2481–2499 (cit. on p. 22).
- Zoran, Daniel and Yair Weiss (2011). “From learning models of natural image patches to whole image restoration”. In: *2011 International Conference on Computer Vision*. IEEE, pp. 479–486 (cit. on p. 22).