Model-based clustering of high-dimensional data in Astrophysics

Charles Bouveyron

Model-based clustering of high-dimensional data in Astrophysics

Charles Bouveyron

Laboratoire MAP5, UMR CNRS 8145
Université Paris Descartes & Sorbonne Paris Cité

Abstract

The nature of data in Astrophysics has changed, as in other scientific fields, in the past decades due to the increase of the measurement capabilities. As a consequence, data are nowadays frequently of high dimensionality and available in mass or stream. Model-based techniques for clustering are popular tools which are renowned for their probabilistic foundations and their flexibility. However, classical model-based techniques show a disappointing behavior in high-dimensional spaces which is mainly due to their dramatical over-parametrization. The recent developments in model-based classification overcome these drawbacks and allow to efficiently classify high-dimensional data, even in the "small n / large p" situation. This work presents a comprehensive review of these recent approaches, including regularization-based techniques, parsimonious modeling, subspace classification methods and classification methods based on variable selection. The use of these model-based methods is also illustrated on real-world classification problems in Astrophysics using R packages.

1 Introduction

As noticed by Hubble back in 1936 [32]:

> *"The nebulae are so numerous that they cannot all be studied individually. Therefore, it is necessary to know whether a fair sample can be assembled from the most conspicuous objects [...]. The answer to this question [...] is sought in the classification of nebulae."*

clustering may be a powerful tool for Astrophysicists who have to face to mass of data. However, in Astrophysics and many other scientific fields, the recent technological developments have resulted in a dramatic increase of the measurement capabilities. In particular, it is nowadays frequent to observe high-dimensional data, *i.e.* the number p of measured variables is large, mass of data, *i.e.* the number of observations n is large, or even data streams, *i.e.* the observations arrive over the time and $n \to \infty$. Among clustering techniques, model-based approaches [27, 37]

are popular. They are renowned for their probabilistic foundations and their flexibility, as shown in [Girard and Saracco's Chapter]. One of the main advantages of these approaches is the fact that their models and results can be interpreted from both the statistical and practical points of view.

Unfortunately, model-based methods usually show a disappointing behavior in high-dimensional spaces. They suffer from the well-known *curse of dimensionality* [3] which is mainly due to the fact that model-based techniques are over-parametrized in high-dimensional spaces. Furthermore, in several applications, the number of available observations can be small compared to the number of variables and such a situation increases the problem difficulty. However and since the dimension of observed data is usually higher than their intrinsic dimension, it is theoretically possible to reduce the dimension of the original space without loosing any information. For this reason, dimension reduction methods are frequently used in practice to reduce the dimension of the data before the clustering step. Feature extraction methods, such as principal component analysis (PCA), or feature selection methods are very popular. However, dimension reduction usually does not consider the clustering task and provide a sub-optimal data representation for the classification step. Indeed, dimension reduction methods usually imply an information loss which could have been discriminative.

To avoid the drawbacks of dimension reduction, several approaches have been proposed in the last decade to allow model-based methods to efficiently classify high-dimensional data. Earliest approaches include constrained models or regularization. More recently, subspace clustering techniques and variable selection techniques have also been proposed. Subspace clustering techniques are mostly based on probabilistic versions of the factor analysis model and allow to classify the data in low-dimensional subspaces without reducing the dimension. Conversely, variable selection techniques do reduce the dimension of the data but select the variables to retain regarding the clustering objective. Both techniques turn out to be very efficient and their practical use will be discussed as well in this article.

This chapter is organized as follows. Section 2 introduces the curse of dimensionality in model-based clustering and also highlights some positive features. Earliest approaches for high-dimensional clustering are presented and discussed in Section 3. Then, Sections 4 and 5 respectively introduce some recent techniques for subspace clustering and variable selection. Some concluding remarks are given in Section 6. Before to move forward, let us notice that we present here only the models and, unless a specific note, the inference of those models is done via the EM algorithm (see [Girard and Saracco's Chapter]).

2 Curse and blessings of the dimensionality

Before to present classical and recent methods for classifying high-dimensional data, we focus in this section on the causes of the curse of dimensionality in model-based clustering. It will be also shown that high-dimensional spaces have interesting properties which may ease the clustering task.

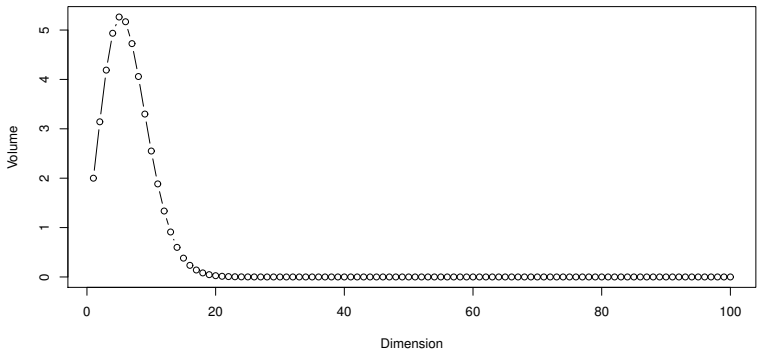


Figure 1: Volume of the unit hypersphere according to the dimension of the space.

2.1 Bellman's curse of the dimensionality

When reading research articles or books related to high-dimensional data, it is very likely to find the term "curse of dimensionality" to refer to problems caused by the analysis of high-dimensional data. This term was first used by R. Bellman in the preface of his book [3] promoting dynamic programming. Although the term "curse of dimensionality" used by Bellman is of course rather pessimistic, the paragraph of the preface in which the term first appeared is in fact more optimistic:

> *All this* [the problems linked to high dimension] *may be subsumed under the heading « the curse of dimensionality ». Since this is a curse,* [...]*, there is no need to feel discouraged about the possibility of obtaining significant results despite it.*

This paragraph will indeed show that the Bellman's thought was corrected since, at least for clustering, high dimensions have nice properties which do allow to obtain significant results.

Before moving to more optimistic things, it is important first to focus on some surprising features of high-dimensional spaces to correctly understand the difficulties of working with high-dimensional data. As revealed by several authors [23, 44, 50, 51], high-dimensional spaces are difficult to handle because simple ideas which are true and well-established in low-dimensional spaces (2D or 3D for instance) turn out to be wrong in high-dimensional spaces. A simple and classical example is the volume of the unit hypersphere which can be easily computed with respect to the dimension p of the space as follows:

$$V(p) = \frac{\pi^{p/2}}{\Gamma(p/2+1)},$$

where Γ is the usual Gamma function. Figure 1 shows the surprising behavior of the unit hypersphere volume according to the dimension of the space. It appears that, as expected, the volume of the sphere increases when moving from dimension 1 to

2, 2 to 3 and so forth. However, after the dimension 5, the volume stops increasing and very surprisingly, decreases very fast toward 0. Consequently, the volume of the unit hypersphere in a 30-dimensional space is 2×10^{-5}. This suggests that high-dimensional spaces have very different features than those of low-dimensional spaces.

2.2 The curse of dimensionality in model-based clustering

The curse of dimensionality takes a particular form in the context of model-based clustering. Indeed, model-based clustering methods require the estimation of a number of parameters which directly depends on the dimension of the observed space. If we consider the classical Gaussian mixture model with K groups, the total number of parameters to estimate is equal to:

$$\nu = (K-1) + Kp + Kp(p-1)/2,$$

where $(K-1)$, Kp and $Kp(p-1)/2$ are respectively the numbers of parameters to estimate for the proportions, the means and the covariance matrices. It turns out that the number of parameters to estimate is therefore a quadratic function of p in the case of the Gaussian mixture model and a large number of observations will be necessary to correctly estimate those model parameters. Furthermore, a more annoying problem occurs in the EM algorithm when computing the posterior probability $t_{ik} = E[Z = k|y_i, \theta]$ that observation $y_i \in \mathbb{R}^p$, $i = 1, ..., n$, belongs to cluster $k \in \{1, ..., K\}$. Indeed, this probability depends, in the GMM context, on the quantity $\Gamma_k(y) = -2\log(\pi_k \phi(y; \mu_k, \Sigma_k))$ which can be rewritten as:

$$\Gamma_k(y) = (y-\mu_k)^t \Sigma_k^{-1}(y-\mu_k) + \log(\det \Sigma_k) - 2\log(\pi_k) + p\log(2\pi),$$

and which requires the inversion of the covariance matrices Σ_k, for $k = 1, ..., K$. Consequently, if the number of observations n is small compared to p, the estimated covariance matrices $\hat{\Sigma}_k$ are ill-conditioned and their inversions conduce to unstable classification functions. In the worst case where $n < p$, the estimated covariance matrices $\hat{\Sigma}_k$ are singular and model-based clustering methods cannot be used at all. Unfortunately, this kind of situation tends to occur more and more frequently in many scientific fields such as Astrophysics or Biology for instance.

2.3 The blessing of dimensionality in clustering

Hopefully, as expected by Bellman, high-dimensional spaces have specific features which could also facilitate their exploration. In the context of clustering, high-dimensional spaces do have useful characteristics which ease the clustering of data in those spaces. In particular, Scott and Thompson [45] showed that high-dimensional spaces are mostly empty. A simple experience can once again illustrate this phenomenon. Let us consider the shell between the hypersphere of radius 0.9 and the unit hypersphere in a p-dimensional space. In order to study the behavior of the volume of this shell regarding the dimension of the space, let us consider the ratio between the volume of the two hyperspheres. Figure 2 presents the evolution

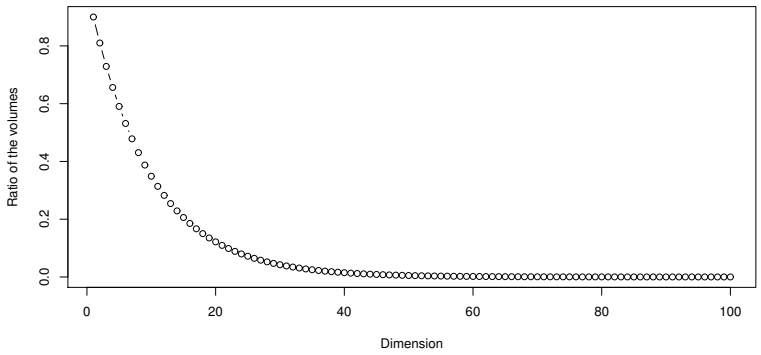


Figure 2: Ratio between the hypersphere of radius 0.9 and the unit hypersphere according to the dimension of the space.

of this ratio regarding the dimension p of the space. The ratio decreases quickly toward 0 and suggests that the p-dimensional shell between the two hyperspheres tends to have in fact an intrinsic dimension equals to $p-1$. A similar experiment, suggested by Huber [33], consists in drawing realizations of a p-dimensional random vector Y with uniform probability distribution on the hypersphere of radius 1. The probability that a realization y_i of this experiment belongs to the shell between the hypersphere of radius 0.9 and the unit hypersphere is therefore:

$$P\left(y_i \in S_{0.9}(p)\right) = 1 - 0.9^p.$$

In particular, the probability that y_i belongs to the shell between the hypersphere of radius 0.9 and the unit hypersphere in a 20-dimensional space is roughly equals to 0.88. Therefore, most of the realizations of the random vector Y live near a $p-1$ dimensional subspace and the remaining of the space is mostly empty. This suggests that clustering methods should model the groups in low-dimensional subspaces instead to model them in the whole observation space. Furthermore, it seems reasonable to expect that different groups live in different subspaces and if this may be a useful property for discriminating the groups. Subspace clustering methods, presented in Section 6, exploit this specific characteristic of high-dimensional spaces.

3 Earliest approaches for high-dimensional clustering

Earliest approaches to deal with the clustering of high-dimensional data can be split into three families: dimension reduction methods, regularization methods and parsimonious methods.

3.1 Dimension reduction

Approaches based on dimension reduction assume that the number p of measured variables is too large and, implicitly, that the data at hand live in a space of lower dimension, let us say $d < p$. A common practice is to project the data into a low-dimensional space and then to apply a clustering algorithm on the projected observations to obtain a partition of the original data.

The most popular linear method used for dimension reduction in this context is certainly principal component analysis (PCA). It was introduced by Pearson [39] who defines PCA as a linear projection that minimizes the average projection cost. In other words, PCA aims to find an orthogonal projection of the data set in a low-dimensional linear subspace, such that the variance of the projected data is maximum. This leads to the classical result where the principal axes $\{u_1, ..., u_d\}$ are the eigenvectors associated with the largest eigenvalues of the empirical covariance matrix S of the data. Interestingly, Tipping and Bishop [47] proposed, several decades after, a probabilistic view of PCA by assuming that the observations are independent realizations of a random variable $Y \in \mathbb{R}^p$ which is linked to a latent variable $X \in \mathbb{R}^d$ through the linear relation:

$$Y = \Lambda^t X + \varepsilon,$$

where X and ε are independent. It is further assume that $X \sim \mathcal{N}(\mu, I_d)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_p)$, such that the marginal distribution of Y is $\mathcal{N}(\Lambda\mu, \Lambda^t\Lambda + \sigma^2 I_p)$. The estimation of the parameters μ, Λ and σ^2 by maximum likelihood conduces in particular to estimate Λ by the eigenvectors associated with the largest eigenvalues of the empirical covariance matrix S of the data.

Notice that the probabilistic PCA (PPCA) model is in fact a particular case of the factor analysis (FA) model [46]. Indeed, the FA model makes the same assumption as the PPCA model except regarding the distribution of ε which is assumed to be $\mathcal{N}(0, \Psi)$, where Ψ is a diagonal covariance matrix. However, conversely to the PPCA model, the estimation of model parameters by maximum likelihood does not conduce to closed-form estimators in this case.

Despite the popularity of this approach, we would like to caution the reader that reducing the dimension without taking into consideration the clustering goal may be dangerous. Indeed, such a dimension reduction may yield a loss of information which could have been useful for discriminating the classes or groups. In particular, when PCA is used for reducing the data dimensionality, only the components associated with the largest eigenvalues are kept. Such a practice is disproved mathematically and practically by Chang [22] who showed that the first components do not necessary contain more discriminative information than the others. In addition, reducing the dimension of the data may not be a good idea since it is easier to discriminate groups of points in high-dimensional spaces than in lower dimensional spaces, assuming that one can build a good classifier in high-dimensional spaces.

Let us illustrate the possible disadvantage of PCA in the context of clustering high-dimensional data. For the purpose of illustration, we consider her biomedical data where a supervision is available. We consider the "prostate" data set which is

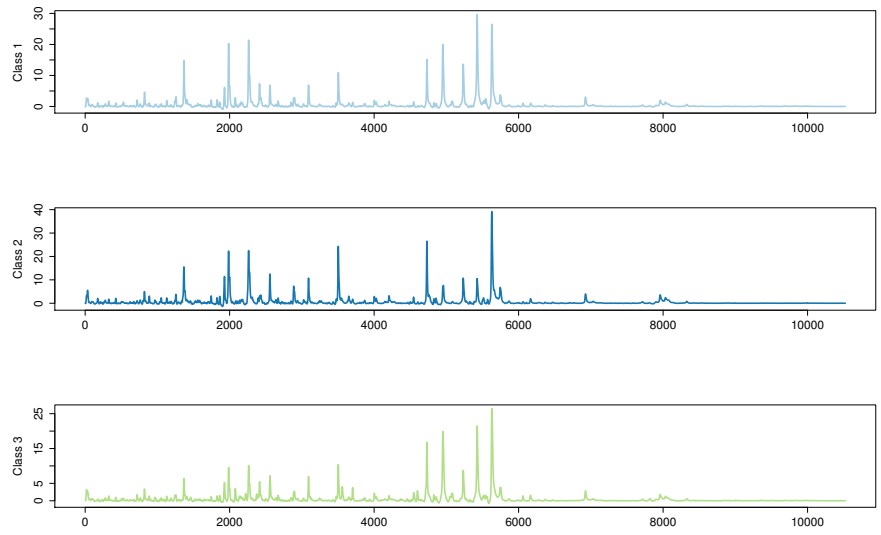


Figure 3: Mean spectra of the three classes for the prostate data set.

available in the *ChemometricsWithR* package [40] for R. The data were presented in [1]. The data set consists in 327 spectra of blood samples measured on 10523 variables from 2000 to 20000 Da. The samples come from patients with prostate cancer, benign prostatic hypertrophy and normal controls. Figure 3 presents the means of the three classes.

Listing 1 shows how to do a PCA on the prostate data and compares the classification ability of different principal subspaces.

Listing 1: Disavantages of PCA for clustering

```
# prostate data are stored in X
library(MASS)

# PCA of the data (using SVD since n < p)
U = svd(X)$v
par(mfrow=c(1,2))
plot(as.matrix(X) %*% U[,c(1,2)],col=cls,
        pch=19,xlab='PC axis 1',ylab='PC axis 2')
plot(as.matrix(X) %*% U[,c(2,3)],col=cls,
        pch=19,xlab='PC axis 2',ylab='PC axis 3')

# Supervised classification into principal subspaces
n = nrow(X); nb = 50
Tx = matrix(NA,2,nb)
X1 = as.matrix(X) %*% U[,1:2]
for (i in 1:nb){
  ind = sample(nrow(X),round(n/nb))
  Tx[1,i] = sum(predict(qda(X1[-ind,],cls[-ind]),
        X1[ind,])$cl == cls[ind]) / length(ind)
```

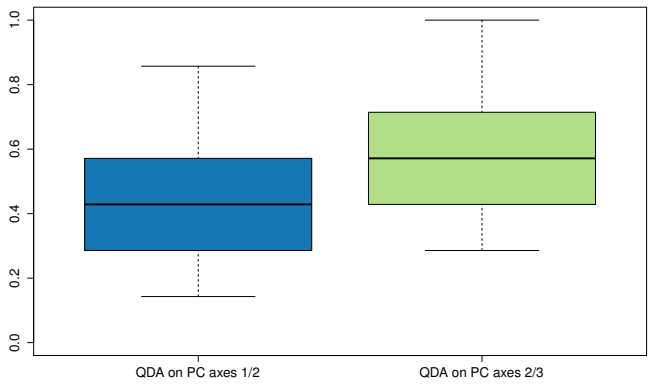


Figure 4: Classification performances of the supervised classification method QDA on two different principal subspaces.

```
}

X2 = as.matrix(X) %*% U[,c(2,3)]
for (i in 1:nb){
  ind = sample(nrow(X),round(n/nb))
  Tx[2,i] = sum(predict(qda(X2[-ind,],cls[-ind]),
        X2[ind,])$cl == cls[ind]) / length(ind)
}
boxplot(t(Tx),col=2:3,ylim=c(0,1),
        names=c('QDA on PC axes 1/2','QDA on PC axes 2/3'))
```

Results are presented on Figures 4 and 5. First, Figure 4 shows the classification performances (cross-validated on 50 folds) of the supervised classification method QDA (quadratic discriminant analysis) on two different principal subspaces. It turns out that the subspace spanned by the principal axes 2 and 3 better discriminates the 3 groups than the first principal plane. This is visually confirmed by looking on the projection of the prostate data into those two subspaces (Figure 5). As a summary, the reader should keep in mind that PCA may be a useful explanatory tool to visualize the data but should not be used as a preprocessing step for clustering or classification.

3.2 Regularization

It is also possible to see the curse of dimensionality in clustering as a numerical problem in the inversion of the covariance matrices Σ_k in Γ_k. From this point of view, a way to tackle the curse of dimensionality is to numerically regularize the estimates of the covariance matrices Σ_k before their inversion. As we will see, most of the regularization techniques have been proposed in the supervised classification framework, but they can be easily used for clustering as well. A simple way to regularize the estimation of Σ_k is to consider a ridge regularization which adds a

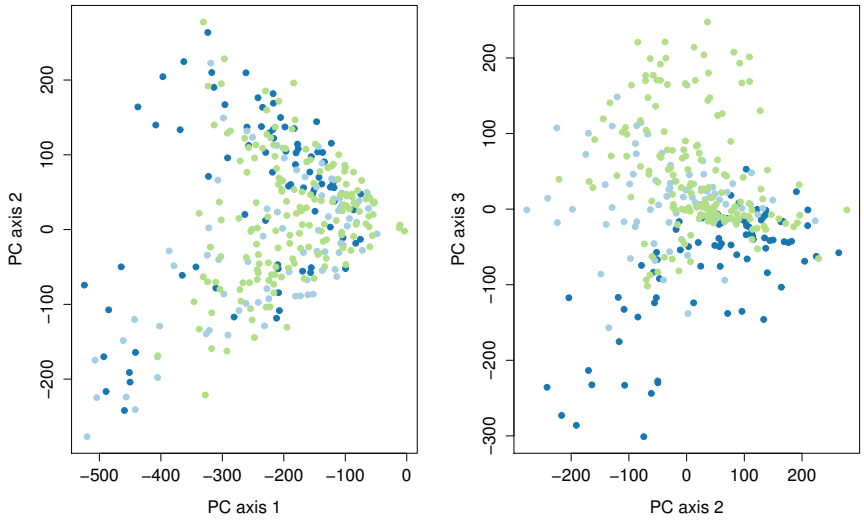


Figure 5: Projection of the prostate data into the subspaces spanned by PC1-2 (left) and PC2-3 (right).

positive quantity σ_k to the diagonal of the matrix:

$$\tilde{\Sigma}_k = \hat{\Sigma}_k + \sigma_k I_p.$$

Notice that this regularization is often implicitly used in statistical softwares, such as R [49] for performing a linear discriminant analysis (LDA) where, for instance, the `lda` function spheres the data before analysis. Friedman [29] also proposed, for his popular regularized discriminant analysis (RDA), the following regularization:

$$\hat{\Sigma}_k(\lambda, \gamma) = (1-\gamma)\hat{\Sigma}_k(\lambda) + \gamma \left(\frac{\mathrm{tr}(\hat{\Sigma}_k(\lambda))}{p} \right) I_p,$$

where :

$$\hat{\Sigma}_k(\lambda) = \frac{(1-\lambda)(n_k - 1)\hat{\Sigma}_k + \lambda(n-K)\hat{\Sigma}}{(1-\lambda)(n_k-1) + \lambda(n-K)}.$$

Thus, the parameter γ controls the ridge regularization whereas λ controls the contribution of the estimators $\hat{\Sigma}_k$ and $\hat{\Sigma}$, where $\hat{\Sigma}$ estimates the within covariance matrix. Finally, it is also possible to use the Moore–Penrose pseudo-inverse of $\hat{\Sigma}$ in Γ_k instead of the usual inverse $\hat{\Sigma}^{-1}$. The reader can also refer to [38] which provides a comprehensive overview of regularization techniques in classification.

The solution based on regularization does not have the same drawbacks than dimension reduction and can be used with less fear. However, all regularization techniques require the tuning of parameters which may be difficult in the unsupervised context, whereas this can be done easily in the supervised context using cross validation.

3.3 Parsimonious models

A third way to look at the the curse of dimensionality in clustering is to consider it as a problem of over-parameterized modeling. Indeed, the Gaussian model is known to be highly parameterized which naturally yields inference problems in high-dimensional spaces. Consequently, the use of constrained models is another solution to avoid the curse of dimensionality in model-based clustering.
A traditional way to reduce the number of free parameters of Gaussian models is to add constraints on the model through their parameters. Let us recall that the unconstrained Gaussian model (Full-GMM hereafter) requires the estimation of 20603 parameters when the number of components is $K = 4$ and the number of variables is $p = 100$. A first possible constraint for reducing the number of parameters to estimate is to constraint the K covariance matrices to be the same across all mixture components, *i.e.* $\Sigma_k = \Sigma$, $\forall k$. Notice that this model yields the famous linear discriminant analysis (LDA) [25] method in the supervised classification case.
In a similar spirit, Banfield & Raftery [2] and Celeux & Govaert [20] proposed, almost simultaneously, a parameterization of the Gaussian mixture model which yields a family of constrained models. To this end, they parametrize the covariance matrices from their eigenvalue decomposition:

$$\Sigma_k = \lambda_k D_k A_k D_k^t,$$

where D_k is the matrix of eigenvectors which determines the orientation of the cluster, A_k is a diagonal matrix proportional to the eigenvalues which explains its shape, and λ_k is a scalar which controls its volume. This model is referred to by the $[\lambda_k D_k A_k D_k^t]$ model in [20] and to by VVV in [2]. By constraining the parameters λ_k, D_k and A_k within and across the groups, 14 different parsimonious models can be enumerated. Among the 14 models, 4 models are highly parametrized as the Full-GMM model, 4 models have an intermediate level of parsimony as the Com-GMM model and, finally, 6 models are very parsimonious. Besides, this reformulation of the covariance matrices can be viewed as a generalization of the constrained models, presented previously. For example, the Com-GMM model is equivalent to the model $[\lambda DAD^t]$. The reader can refer to [20] for more details on these models.

The solution which introduces parsimony in the models is clearly a better solution in the context of model-based clustering since it proposes a trade-off between the perfect modeling and what one can correctly estimate in practice. We will see in the next sections that recent solutions for high-dimensional clustering are partially based on the idea of constrained modeling.

4 Subspace clustering methods

Conversely to previous solutions, subspace clustering methods exploit the "empty space" phenomenon to ease the discrimination between groups of points. To do so, they model the data in low-dimensional subspaces and introduce some restrictions

while keeping all dimensions. Subspace clustering methods are mostly related to the factor analysis [42] model which assumes that the observation space is linked to a latent space through a linear relationship. We focus here only on two models and we refer the reader to [13] for an extensive review on subspace clustering models.

4.1 Mixture of high-dimensional Gaussian mixture models

First, Bouveyron *et al.* [16, 17] proposed a family of 28 parsimonious and flexible Gaussian models to deal with high-dimensional data. Conversely to the other approaches, this family of GMM was directly proposed in both supervised and unsupervised classification contexts. In order to ease the designation of this family, we propose to refer to these Gaussian models for high-dimensional data by the acronym HD-GMM. Bouveyron *et al.* [16] proposed to constraint the GMM model through the eigen-decomposition of the covariance matrix Σ_k of the kth group:

$$\Sigma_k = Q_k \Lambda_k Q_k^t,$$

where Q_k is a $p \times p$ orthogonal matrix which contains the eigenvectors of Σ_k and Λ_k is a $p \times p$ diagonal matrix containing the associated eigenvalues (sorted in decreasing order). The key idea of the work of Bouveyron *et al.* is to reparametrize the matrix Λ_k, such as Σ_k has only $d_k + 1$ different eigenvalues:

$$\Delta_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{kd_k} \end{matrix}} & \mathbf{0} \\ \mathbf{0} & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & b_k \end{matrix}} \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} d_k \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} (p - d_k) \end{matrix}$$

where the d_k first values $a_{k1}, \ldots, a_{kd_k}$ parametrize the variance in the group-specific subspace and the $p - d_k$ last terms, the b_k's model the variance of the noise and $d_k < p$. With this parametrization, these parsimonious models assume that, conditionally to the groups, the noise variance of each cluster k is isotropic and is contained in a subspace which is orthogonal to the subspace of the kth group. Following the classical parsimony strategy, the authors proposed a family of parsimonious models from a very general model, the model $[a_{kj} b_k Q_k d_k]$, to very simple models.

Such an approach can be viewed in two different ways: on the one hand, these models enable to regularize the models in high-dimension. In particular, by constraining d_k such that $d_k = p - 1$ for $k = 1, ..., K$, the proposed approach can be viewed as a generalization of the works of [20, 28]. Indeed, the model $[a_{kj} b_k Q_k \, (p-1)]$ is equivalent to the Full-GMM model or the $[\lambda_k D_k A_k D_k]$ model in [20]. In the same manner, the model $[a_{kj} b_k Q \, (p-1)]$ is equivalent to the Diag-GMM and the $[a_j b Q \, (p-1)]$ is also the Com-Diag-GMM. On the other hand, this

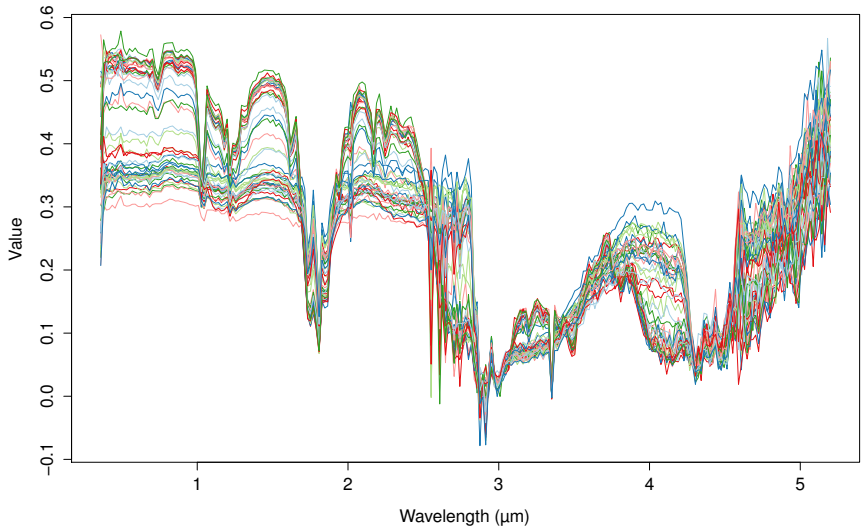


Figure 6: Some of the 38 400 measured spectra described on 256 wavelengths from 0.36 to 5.2 µm.

approach can also be viewed as an extension of the mixture of principal component analyzer (Mixt-PPCA) model [48] since it relaxes the equality assumption on d_k made in [48] and the model $[a_{kj}b_kQ_kd]$ is therefore equivalent to the Mixt-PPCA model.

For 16 of the 28 HD-GMM models, the inference can be done easily using the EM algorithm since update formula for mixture parameters are closed-form. The estimation of the intrinsic dimensions d_k, $k = 1, ..., K$, relies on the scree test of Cattell [19] which looks for a break in the eigenvalue scree of the empirical covariance matrix of each group. Let us finally notice that Bouveyron *et al.* [14] have demonstrated the surprising result that the maximum likelihood estimator of the intrinsic dimensions d_k is asymptotically consistent in the case of the model $[a_kb_kQ_kd_k]$.

Here, we propose to use HDDC algorithm to segment hyperspectral images of the Martian surface. This problem is indeed by its very nature an unsupervised classification problem. Visible and near infrared imaging spectroscopy is a key remote sensing technique to study the system of the planets. Imaging spectrometers, which are onboard of an increasing number of satellites, provide high-dimensional hyperspectral images. In March 2004, the OMEGA instrument (Mars Express, ESA) [5] has collected 310 Gbytes of raw images. The OMEGA imaging spectrometer has mapped the Martian surface with a spatial resolution varying between 300 to 3000 meters depending on the spacecraft altitude. It acquired for each resolved pixel the spectrum from 0.36 to 5.2 µm in 255 contiguous spectral channels. OMEGA is designed to characterize the composition of surface materials, discriminating between various classes of silicates, hydrated minerals, oxides and carbonates, organic

frosts and ices. For this experiment, a 300×128 image of the Martian surface is considered and a 255-dimensional spectral observation is therefore associated to each of the 38 400 pixels. Figure 6 shows some of the 38 400 measured spectra. According to the experts, there are $K = 5$ mineralogical classes to identify.

Listing 2 shows how to cluster Mars data with HDDC using the HDclassif package [4] for R.

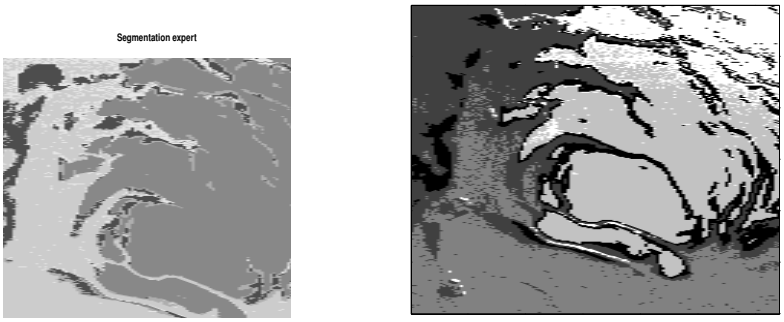
Listing 2: Clustering of Mars data with HDDC

```
# Mars data are stored in X
library(HDclassif)

# clustering of the data
out = hddc(X,5,model='AkBkQkDk',threshold=0.01)

# Displaying model parameters
out
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKBKQKDK
Posterior probabilities of groups
       1      2     3     4     5
   0.238 0.0892 0.243 0.334 0.095
Intrinsic dimensions of the classes:
       1 2 3  4 5
  dim: 8 6 6 11 6
         1       2       3       4       5
Ak: 0.0337 0.0407 0.0105 0.00933 0.0259
          1         2         3         4         5
Bk: 0.000138 0.000239 6.76e-05 0.000101 9.72e-05
BIC:  7743641
```

We have here applied HDDC with a specific model and for the expected number of groups. Notice that it is also possible to let the algorithm determine which model and number of groups are the most adapted for the data at hand. Regarding model parameters, one can see that HDDC estimates that the intrinsic dimensions of groups are all around 10 whereas, for recall, the original dimension is 255. Figure 7 shows the associated segmentation of the image and allows to compare it with an expert segmentation. Except the color shift, both segmentations look very similar and it confirms the interest of such model-based clustering techniques in this context.

4.2 The discriminative latent mixture models

Recently, Bouveyron & Brunet [10] proposed a family of mixture models which fit the data into a common and discriminative subspace. This mixture model, called the discriminative latent mixture (DLM) model, differs from the FA-based models in the fact that the latent subspace is common to all groups and is assumed to be the most discriminative subspace of dimension d. Indeed, roughly speaking, the FA-based models choose the latent subspace(s) maximizing the projected vari-



Expert segmentation HDDC segmentation

Figure 7: Segmentation of the hyperspectral image of the Martian surface using a physical model build by experts (left) and HDDC (right).

ance whereas the DLM model chooses the latent subspace which maximizes the separation between the groups.

Let $Y \in \mathbb{R}^p$ be the observed random vector and let $Z \in \{1, \ldots, K\}$ be once again the unobserved random variable to predict. The DLM model then assumes that Y is linked to a latent random vector $X \in \mathbb{E}$ through a linear relationship of the form:

$$Y = UX + \varepsilon,$$

where X and ε are independent, $\mathbb{E} \subset \mathbb{R}^p$ is assumed to be the most discriminative subspace of dimension $d \leq K-1$ such that $\mathbf{0} \in \mathbb{E}$, $K < p$, U is a $p \times d$ orthonormal matrix common to the K groups and satisfying $U^t U = \mathbf{I}_d$, and $\varepsilon \sim \mathcal{N}(0, \Psi)$ models the non discriminative information. Besides, within the latent space and conditionally to $Z = k$, X is assumed to be distributed as:

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

where $\mu_k \in \mathbb{R}^d$ and $\Sigma_k \in \mathbb{R}^{d \times d}$ are respectively the mean vector and the covariance matrix of the kth group. Given these distribution assumptions, the marginal distribution of Y is once again a mixture of Gaussians, *i.e.* $g(y) = \sum_{k=1}^{K} \pi_k \phi(y; m_k, S_k)$, where $m_k = U\mu_k$ and $S_k = U\Sigma_k U^t + \Psi$. Let $W = [U, V]$ be the $p \times p$ matrix such that $W^t W = W W^t = \mathbf{I}_p$ and V is an orthogonal complement of U. Finally, the noise covariance matrix Ψ is assumed to satisfy the conditions $V \Psi V^t = \beta \mathbf{I}_{p-d}$ and $U \Psi U^t = \mathbf{0}_d$, such that $\Delta_k = W^t S_k W$ is block-diagonal:

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

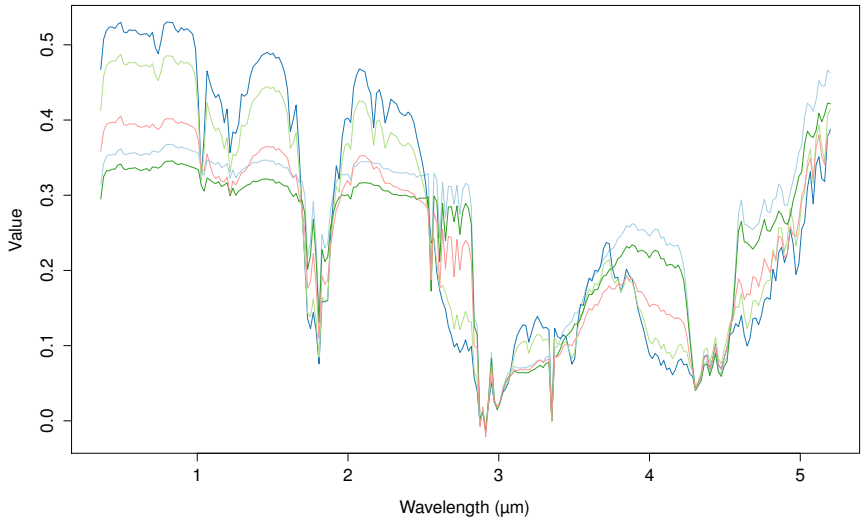


Figure 8: Mean spectra of the 5 groups formed by Fisher-EM on Mars data.

These last conditions imply that the discriminative and the non-discriminative subspaces are orthogonal, which suggests in practice that all the relevant classification information remains in the latent subspace. This model is referred to by DLM$_{[\Sigma_k\beta]}$ in [10]. Following the classical strategy, several other models can be obtained from the DLM$_{[\Sigma_k\beta]}$ model by relaxing or adding constraints on model parameters.

Conversely to most of the models based on mixture of FA models, the inference of the DLM models in the unsupervised context cannot be directly done using the EM algorithm because of the specific features of its latent subspace. To overcome this problem, an estimation procedure, called the Fisher-EM algorithm, is also proposed in [10] for estimating both the discriminative subspace and the parameters of the mixture model. This algorithm is based on the EM algorithm from which an additional step is introduced, between the E and the M-step. This additional step, named F-step, aims to compute the projection matrix U whose columns span the discriminative latent space. This step estimates at iteration q, the orientation matrix $U^{(q)}$ of the discriminative latent space by maximizing the Fisher's criterion [25, 30] under orthonormality constraints and conditionally to the posterior probabilities. This optimization problem is solved in [10] using the concept of orthonormal discriminant vector developed by [26] through a Gram-Schmidt procedure.Two additional procedures are proposed in [7] for the estimation of the latent subspace orientation. The convergence properties of the Fisher-EM algorithm were also studied in [11] from both the theoretical and the practical points of view. Let us finally notice that this modeling was also used in the context of supervised and semi-supervised classification and leads to the probabilistic Fisher discriminant analysis (pFDA) method [9].

We now present in Listing 3 an application of the Fisher-EM algorithm to the

Mars data (described in the previous section). The Fisher-EM algorithm is available in the R software through the FisherEM package [8].

Listing 3: Clustering of Mars data with Fisher-EM

```
# Mars data are stored in X
library(FisherEM)

# Clustering with Fisher-EM
out = fem(X,K=5,model='AkjB')

# Estimated model parameters
str(out)
List of 15
$ model: chr "AkjB"
$ cls  : int [1:38400] 2 2 2 2 2 2 2 2 2 5 ...
$ P    : num [1:38400, 1:5] 0.00 1.09e-302 0.00 0.00 ...
$ K    : int 5
$ p    : int 255
$ mean : num [1:5, 1:4] -0.181 -0.144 -0.131 -0.165 ...
$ my   : num [1:5, 1:255] 0.47 0.313 0.359 0.415  ...
$ prop : num [1:5] 0.263 0.185 0.145 0.166 0.241
$ D    : num [1:5, 1:255, 1:255] 2.61e-05 4.49e-05 ...
$ U    : num [1:255, 1:4] -0.00333 -0.00141 ...
$ aic  : num 28156951
$ bic  : num 28147159
$ loglik: num [1:50] 28160856 28160398 28160059 ...
$ ll   : num 28159240
```

Here also, Fisher-EM was used for a specific model and for $K = 5$, but the automatic selection of the model and K is possible. The object `out` contains several information which require some comments. First, `cls` and `P` contain respectively the partition into 5 groups of the data and the posterior probabilities that each observation belongs to the groups. The sub-object `prms` gathers all information about the learned mixture model. Astrophysicists will be mostly interested in visualizing the group means which are stored in `prms$my`. The estimated group means are plotted on Figure 8. Another parameter which is useful from the practical point of view is the loading matrix `U`. This matrix contains the coordinates of the discriminative axes and allows therefore to project the original data onto the discriminative subspace for further analyses. Figure 9 presents the projection of the clustered data on the estimated discriminative axes with Fisher-EM.

5 Variable selection for model-based clustering

Conversely to the approaches of the previous section, several recent works have been interested to simultaneously cluster data and reduce their dimensionality by selecting relevant variables for the clustering task.

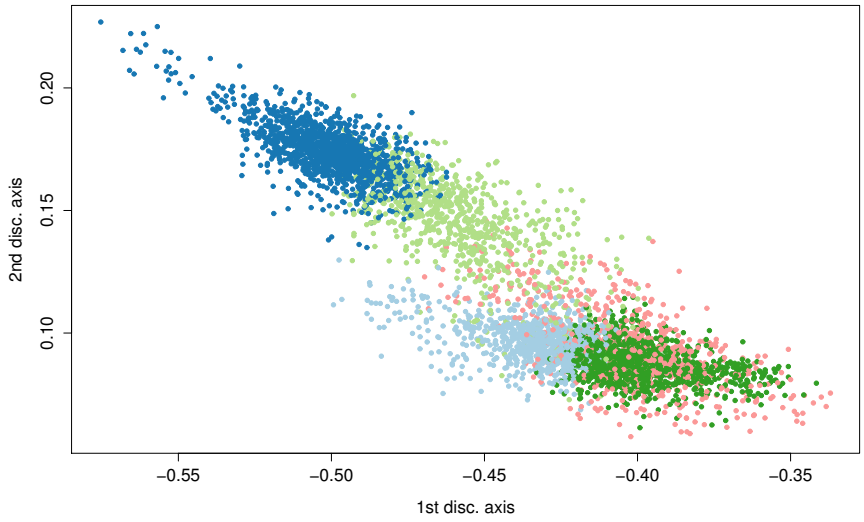


Figure 9: Mean spectra of the 5 groups formed by Fisher-EM on Mars data.

5.1 Variable selection as a model selection problem

The underlying idea of the works of Raftery and Dean [41] and Maugis *et al.* [35] is to find the variables which are relevant for the clustering task. The determination of the role of each variable is in particular apprehended in [35, 41] as a model selection problem in the GMM context. In the Raftery & Dean's approach, the authors define two different sets of variables: $\mathcal{S}$ which denotes the set of relevant variables and $\mathcal{S}_c$ which is the set containing the irrelevant variables. An interesting aspect of their approach is that they do not assume that the irrelevant variables are independent of the clustering variables. Maugis *et al.* [21, 35, 36] relax some restrictions of Raftery and Dean's model and propose a more general variable role modeling. They define two subsets of variables: on the one hand, the relevant ones, which are grouped in $\mathcal{S}$ and, on the other hand, its complementary $\mathcal{S}_c$, which is formed by the irrelevant variables. Maugis *et al.* consider two types of behaviors among these irrelevant variables: a subset $\mathcal{U}$ of irrelevant variables which can be explained by a linear regression from a subset $\mathcal{R}$ of the clustering variables and a subset $\mathcal{W}$ of irrelevant variables which are totally independent of all relevant variables. It is referred to by the model collection SRUW. From this characterization, the authors also recast the variable selection problem into a model selection problem through an approximation of the integrated log-likelihood. Then the selected model satisfies:

$$\arg \max_{(K,m,r,h,V)} \left\{ \mathrm{BIC}_{\mathrm{clust}}(\mathbf{y}^{\mathcal{S}}|\mathrm{K},\mathrm{m}) + \mathrm{BIC}_{\mathrm{reg}}(\mathbf{y}^{\mathcal{U}}|\mathrm{r},\mathbf{y}^{\mathcal{R}}) + \mathrm{BIC}_{\mathrm{ind}}(\mathbf{y}^{\mathcal{W}}|\mathrm{h}) \right\},$$

where $V = (\mathcal{S},\mathcal{R},\mathcal{U},\mathcal{W})$ stands for the variable partition. The first term of this expression, called $\mathrm{BIC}_{\mathrm{clust}}$, corresponds to the BIC criterion [43] for a Gaussian

mixture of K components on the relevant subset of variables $\mathcal{S}$. The model m belongs here to a collection of 28 parsimonious models which are available in the Mixmod software [6] and include the GMM family introduced by Celeux & Govaert [20]. The second term denoted by $\mathrm{BIC}_{\mathrm{reg}}$, is linked to the BIC criterion for a linear regression of the irrelevant variables $\mathcal{U}$ on a subset of clustering variables $\mathcal{R}$. Note that the index r stands for the structure of the covariance matrix which can be assumed to be spherical, diagonal or non-constrained. Finally, the last term depicts the BIC criterion for a Gaussian density on the variable subset $\mathcal{W}$ independent of the clustering variables. This Gaussian marginal distribution is characterized by a variance matrix Σ which is constrained to be either diagonal or spherical and is specified by the index h in the expression above.

Regarding the implementation, they propose an algorithm based on a backward-stepwise selection. It implies that all the variables are considered at the beginning of the procedure and only a block of variables is either included or excluded of the clustering relevant set of features at each iteration. Such an approach enables them to take into account variable block interactions, if they exist. Then a second algorithm is executed to select both the model and the number of components for the mixture model.

5.2 Variable selection by penalization of the loadings

An alternative approach for selecting the relevant variables through penalization is to directly apply the lasso penalty on the loading matrix of a MFA-based model. This has been achieved in particular in [12, 31, 52].

In the context of Fisher-EM, the direct penalization of the loading matrix U makes particularly sense since it is not estimated by likelihood maximization. The matrix U is indeed estimated in the F-step of the Fisher-EM algorithm by maximizing the Fisher criterion conditionally to the current partition of the data. To achieve the penalization of U, two solutions are proposed in [12]. The first solution is a two stage approach which first estimate U, at each iteration, with the F-step and then looks for its best sparse approximation as follows:

$$\min_{U} \left\| X^{(q)t} - Y^t U \right\|_F^2 + \lambda \sum_{j=1}^{d} \left\| u_j \right\|_1 ,$$

where u_j is the jth column vector of U, $X^{(q)} = \hat{U}^{(q)t} Y$ and $\| \|_F$ refers to the Frobenius norm. The solution of this penalized regression problem can be computed through the LARS algorithm [24]. The second solution directly recasts the maximization of the Fisher criterion as a regression problem and provides a sparse loading matrix by solving the lasso problem associated to this regression problem. However, solving this lasso problem is not direct in this case and requires the use of an iterative algorithm. Regarding the implementation details, it is proposed in [12] to initialize the sparseFEM algorithm with the result of the Fisher-EM algorithm and to determine the value of λ by model selection through a modified BIC criterion.

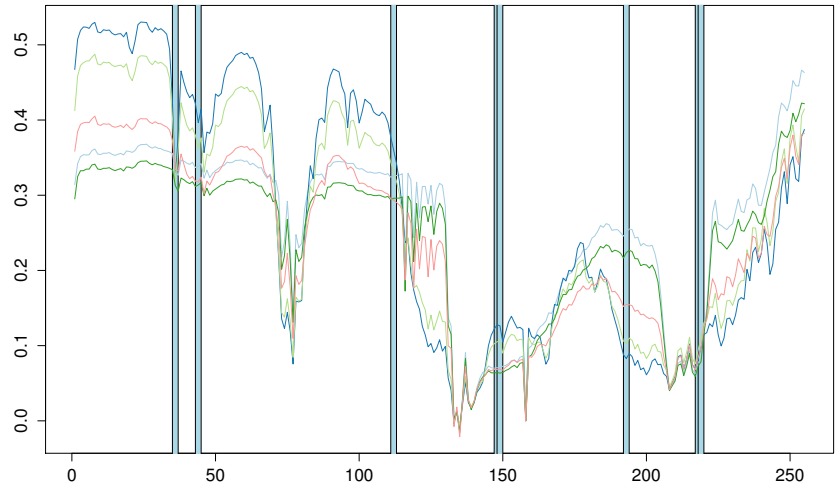


Figure 10: Mean spectra of the 5 groups formed by sparse FEM on Mars data and selection of the discriminative wavelengths (indicated by blue rectangles).

Listing 4 provides the R code to apply sparseFEM on the Mars data. The `sfem()` function is also provided in the FisherEM package.

Listing 4: Clustering of Mars data with sparse FEM

```
# Mars data are stored in X
library(FisherEM)

# Clustering with sparse FEM
out = sfem(X,K=5,model='AkjB',l1=0.1)
```

The level of sparsity is controlled here using the `l1` optional parameter: the smaller the l1 parameter is, the strongest the sparsity is. Here, a value of 0.1 therefore means that the output will be very sparse. Figure 10 allows to visualize the wavelengths which have been selected as discriminative ones. A possible interest of such a selection could be the measurement of only tens of wavelengths for future acquisitions instead of the 255 current ones for a result expected to be similar. This could reduce the acquisition time for each pixel from a few tens of seconds to less than one second.

6 Conclusion

This work has presented a comprehensive overview of some recent model-based methods for the unsupervised classification of data from Astrophysics. We have emphasized the interest of using subspace clustering methods and variable selection methods designed for clustering instead of preprocessing the data with dimension

reduction methods. The few practical examples offered here may help astrophysicists in applying recent model-based techniques to their own data. Let us finally notice that some recent works [15, 18, 34] have extended model-based clustering methods to functional data.

References

[1] B.L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Y. Yasui, Z. Feng, and G.L. Wright. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62(13):3609–3614, 2002.

[2] J. Banfield and A.E. Raftery. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.

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

[4] L. Bergé, C. Bouveyron, and S. Girard. Hdclassif: An R package for model-based clustering and discriminant analysis of high-dimensional data. *Journal of Statistical Software*, 46(6), 2012.

[5] J.-P. Bibring and 42 co-authors. Mars Surface Diversity as Revealed by the OMEGA/Mars Express Observations. *Science*, 307(5715):1576–1581, 2005.

[6] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics and Data Analysis*, 51:587–600, 2006.

[7] C. Bouveyron and C. Brunet. On the estimation of the latent discriminative subspace in the Fisher-EM algorithm. *Journal de la Société Française de Statistique*, 152(3):98–115, 2011.

[8] C. Bouveyron and C. Brunet. FisherEM: An R package for model-based clustering and visualization of high-dimensional data, 2012. https://cran.r-project.org/web/packages/FisherEM/.

[9] C. Bouveyron and C. Brunet. Probabilistic Fisher discriminant analysis: A robust and flexible alternative to Fisher discriminant analysis. *Neurocomputing*, 90(1):12–22, 2012.

[10] C. Bouveyron and C. Brunet. Simultaneous model-based clustering and visualization in the Fisher discriminative subspace. *Statistics and Computing*, 22(1):301–324, 2012.

[11] C. Bouveyron and C. Brunet. Theoretical and practical considerations on the convergence properties of the Fisher-EM algorithm. *Journal of Multivariate Analysis*, 109:29–41, 2012.

[12] C. Bouveyron and C. Brunet. Discriminative variable selection for clustering with the sparse Fisher-EM algorithm. *Computational Statistics*, 29(3-4):489–513, 2014.

[13] C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: A review. *Computational Statistics & Data Analysis*, 71:52–78, 2014.

[14] C. Bouveyron, G. Celeux, and S. Girard. Intrinsic Dimension Estimation by Maximum Likelihood in Isotropic Probabilistic PCA. *Pattern Recognition Letters*, 32(14):1706–1713, 2011.

[15] C. Bouveyron, E. Côme, and J. Jacques. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760, 2015.

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

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

[18] C. Bouveyron and J. Jacques. Model-based Clustering of Time Series in Group- specific Functional Subspaces. *Advances in Data Analysis and Classification*, 5(4):281–300, 2011.

[19] R. Cattell. The scree test for the number of factors. *Multivariate Behavioral Research*, 1(2):145–276, 1966.

[20] G. Celeux and G. Govaert. Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793, 1995.

[21] G. Celeux, M.-L. Martin-Magniette, C. Maugis, and A.E. Raftery. Letter to the editor. *Journal of the American Statistical Association*, 106(493), 2011.

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

[23] D. Donoho. High-dimensional data analysis: the curses and blessings of dimensionality. In *Math Challenges of the 21st Century*. American Mathematical Society, 2000.

[24] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32:407–499, 2004.

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

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

[27] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis and density estimation. *Journal of American Statistical Association*, 97:611–631, 2002.

[28] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458), 2002.

[29] J.H. Friedman. Regularized discriminant analysis. *The Journal of the American Statistical Association*, 84:165–175, 1989.

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

[31] G. Galimberti, A. Montanari, and C. Viroli. Penalized factor mixture analysis for variable selection in clustered data. *Computational Statistics and Data Analysis*, 53(12):4301–4310, 2009.

[32] E. Hubble. *The Realm of the Nebulae*. Yale University Press, London, 1936.

[33] P. Huber. Projection pursuit. *The Annals of Statistics*, 13(2):435–525, 1985.

[34] J. Jacques and C. Preda. Model-based clustering of multivariate functional data. *Computational Statistics and Data Analysis*, 8(3):231–255, 2014.

[35] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection for Clustering with Gaussian Mixture Models. *Biometrics*, 65(3):701–709, 2009.

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

[37] G.J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Interscience, New York, 2000.

[38] A. Mkhadri, G. Celeux, and A. Nasrollah. Regularization in discriminant analysis: a survey. *Computational Statistics and Data Analysis*, 23:403–423, 1997.

[39] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 6(2):559–572, 1901.

[40] R. Wehrens Pinheiro. *Chemometrics With R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*. Springer, Heidelberg, 2012.

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

[42] D. Rubin and D. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47(1):69–76, 1982.

[43] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[44] D. Scott. *Multivariate density estimation*. Wiley & Sons, New York, 1992.

[45] D. Scott and J. Thompson. Probability density estimation in higher dimensions. In *Fifteenth Symposium in the Interface*, pages 173–179, 1983.

[46] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[47] M.E. Tipping and C.M. Bishop. Probabilistic principal component analysis. Technical Report NCRG-97-010, Neural Computing Research Group, Aston University, 1997.

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

[49] W.N. Venables and B.D. Ripley. *Modern Applied Statistics with S*. Springer, 2002.

[50] M. Verleysen. *Learning high-dimensional data*, pages 141–162. Limitations and Future Trends in Neural Computations. IOS Press, 2003.

[51] M. Verleysen and D. François. The curse of dimensionality in data mining and time series prediction. *IWANN*, 2005.

[52] B. Xie, W. Pan, and X. Shen. Penalized mixtures of factor analyzers with application to clustering high-dimensional microarray data. *Bioinformatics*, 26(4):501–508, 2010.