High-dimensional clustering

Christophe Biernacki, Cathy Maugis

Contents

Chapter 2

High-dimensional clustering

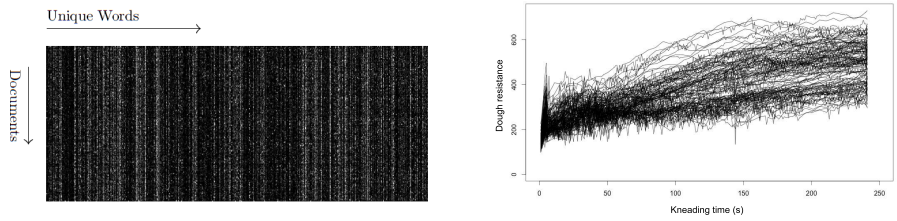
Christophe Biernacki and Cathy Maugis-Rabusseau

2.1 Introduction

High-dimensional (HD) data sets are now frequent, mostly motivated by technological reasons which concern automation in variable acquisition, cheaper availability of data storage and more powerful standard computers for quick data management possibility. All fields are impacted by this general phenomenon of variable number inflation, only the definition of "high" being domain dependent. In marketing, this number can be of order 10^2, in microarray gene expression between 10^2 and 10^4, in text mining 10^3 or more, of order 10^6 for single nucleotide polymorphism (SNP) data, *etc.* Note also that sometimes much more variables can be involved, what can be typically the case with discretized curves, for instance curves coming from temporal sequences.

Here are two related illustrations. Figure 2.1(a) displays a text mining example[1]. It mixes Medline (1033 medical abstracts) and Cranfield (1398 aeronautical abstracts) making a total of 2431 documents. Furthermore, all the words (excluding stop words) are considered as features making a total of 9275 unique words. The data matrix consists of documents on the rows and words on the columns with each entry giving the term frequency, that is the number of occurrences of corresponding word in corresponding document. Figure 2.1(b) displays a curve example. This Kneading data set comes from Danone Vitapole Paris Research Center and concerns the quality of cookies and the relationship with the flour kneading process (Lévéder *et al.* [2004]). It is composed by 115 different flours for which the dough resistance is measured during the kneading process for 480 seconds. We notice that the equispaced instants of time in the interval [0; 480] (here 241 measures) could be much more large than 241 if measures were more frequently recorded.

[1] This data set is publicly available at `ftp://ftp.cs.cornell.edu/pub/smart`.



(a) (b)

Figure 2.1: Examples of high-dimensional data sets: (a) Text mining: $n = 2431$ documents and the frequency that $d = 9275$ unique words occurs in each document (a whiter cell indicates a higher frequency); (b) Curves: $n = 115$ kneading curves observed at $d = 241$ equispaced instants of time in the interval $[0; 480]$.

Such a technological revolution has a huge impact in other scientific fields, as societal or also mathematical ones. In particular, high-dimensional data management brings some new challenges to statisticians since standard (low-dimensional) data analysis methods struggle to directly apply to the new (high-dimensional) data sets. The reason can be twofold, sometimes linked, involving either combinatorial difficulties or disastrously large estimate variance increase. Data analysis methods are essential for providing a synthetic view of data sets, allowing data summary and data exploratory for future decision making for instance. This need is even more acute in the high-dimensional setting since on the one hand the large number of variables suggests that a lot of information is conveyed by data but, in the other hand, such information may be hidden behind their volume.

Cluster analysis is one of the main data analysis method. It aims at partitioning a data set $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, composed by n individuals and lying in a space $\mathcal{X}$ of dimension d into K groups $G_1, \ldots, G_K$. This partition is denoted by $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, lying in a space $\mathcal{Z}$, where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$ is a vector of $\{0, 1\}^K$ such that $z_{ik} = 1$ if individual $\mathbf{x}_i$ belongs to the kth group G_k, and $z_{ik} = 0$ otherwise ($i = 1, \ldots, n$, $k = 1, \ldots, K$). Figure 2.2 gives an illustration of this principle when $d = 2$. Model-based clustering allows to reformulate cluster analysis as a well-posed estimation problem both for the partition $\mathbf{z}$ and for the number K of groups. It considers data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ as n i.i.d. realizations of a mixture pdf $f(\cdot; \boldsymbol{\theta}_K) = \sum_{k=1}^{K} \pi_k f(\cdot; \boldsymbol{\alpha}_k)$, where $f(\cdot; \boldsymbol{\alpha}_k)$ indicates the pdf, parameterized by $\boldsymbol{\alpha}_k$, associated to the group k, where π_k indicates the mixture proportion of this component ($\sum_{k=1}^{K} \pi_k = 1$, $\pi_k \geq 0$) and where $\boldsymbol{\theta}_K = (\pi_k, \boldsymbol{\alpha}_k, k = 1, \ldots, K)$ indicates the whole mixture parameters. From the whole data set $\mathbf{x}$ it is then possible to obtain a mixture parameter estimate $\hat{\boldsymbol{\theta}}_K$ to deduce a partition estimate $\hat{\mathbf{z}}$ from the conditional probability $f(\mathbf{z}|\mathbf{x}; \hat{\boldsymbol{\theta}}_K)$.

It is also possible to derive an estimate $\hat{K}$ from an estimate of the marginal probability $\hat{f}(\mathbf{x}|K)$. More details on mixture models, related estimation of $\boldsymbol{\theta}_K$, $\mathbf{z}$ and K are given throughout Chapter **??**.

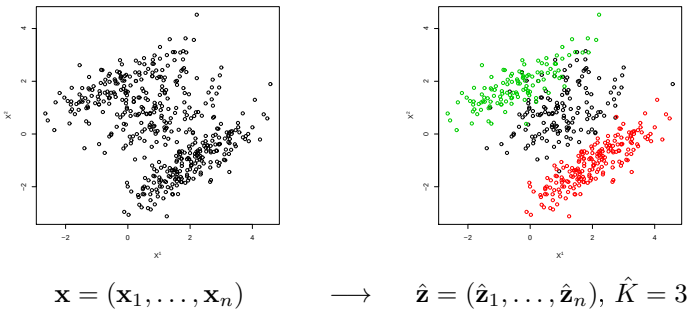


Figure 2.2: The clustering purpose illustrated in the two-dimensional setting.

Beyond the nice mathematical background it provides, model-based clustering has led also to numerous and significant practical successes in the "low-dimensional" setting as Chapter **??** relates, with references therein. Extending the general framework of model-based clustering to the "high-dimensional" setting is thus a natural and desirable purpose. In principle, the more information we have about each individual, the better a clustering method is expected to perform. However the structure of interest may often be contained in a subset of the available variables and a lot of variables may be useless or even harmful to detect a reasonable clustering structure. It is thus important to select the relevant variables from the cluster analysis view point. It is a recent research topic in contrast to variable selection in regression and classification models (Kohavi and John [1997]; Guyon and Elisseeff [2003]; Miller [1990]). This new interest for variable selection in clustering comes from the increasingly frequent use of these methods on high-dimensional data sets, such as transcriptome data sets.

Three types of approaches dealing with variable selection in clustering have been proposed. The first one includes clustering methods with weighted variables (see for instance Friedman and Meulman [2004]) and dimension reduction methods. For this later, McLachlan *et al.* [2002] use a mixture of factor analyzers to reduce the extremely high dimensionality of a gene expression problem. A suitable Gaussian mixture family is considered in Bouveyron *et al.* [2007] to take into account the dimension reduction and the data clustering simultaneously. In contrast to this first method type, the last two approaches select explicitly relevant variables. The so-called "filter" approaches select the variables before a clustering analysis (see for instance Dash *et al.* [2002]; Jouve and Nicoloyannis [2005]). Their main weakness is the influence of independent selection step of the clustering results. In contrast, the so-called "wrapper" approaches combine

variable selection and clustering. For distance-based methods, one can cite Fowlkes *et al.* [1988] for a forward selection approach with complete linkage hierarchical clustering, Devaney and Ram [1997] who propose a stepwise algorithm where the quality of the feature subsets is measured with the COBWEB algorithm or the method of Brusco and Cradit [2001] based on the adjusted Rand index for K-means clustering. There exists also wrapper methods in the model-based clustering setting. When the number of variables is greater than the number of individuals, Tadesse *et al.* [2005] propose a fully Bayesian method using a reversible jump algorithm to simultaneously choose the number of mixture components and select variables. Kim *et al.* [2006] use a similar approach by formulating clustering in terms of Dirichlet process mixtures. In Gaussian mixture model clustering, Law *et al.* [2004] propose to evaluate the importance of the variables in the clustering process via "feature saliencies" and use the *Minimum Message Length* criterion. Raftery and Dean [2006] recast the problem of comparing two nested variable subsets as a model comparison problem and address it using Bayes factor. An interesting aspect of their model formulation is that irrelevant variables are not required to be independent of the clustering variables. They avoid thus the unrealistic independence assumption between the relevant and irrelevant variables for the clustering, considered in Tadesse *et al.* [2005], Kim *et al.* [2006] and Law *et al.* [2004]. In their model, the whole irrelevant variable subset depends on the whole relevant variables through a linear regression equation. However, some relevant variables are not necessarily required to explain all irrelevant variables in the linear regression and their introduction involves additional parameters without a significant increase of the loglikelihood. The related extensions proposed by Maugis *et al.* [2009a,b] follow this remark.

Many model proposals already exist, including associated parameter estimation and, sometimes, specific model selection strategies. We will divide these models into canonical and non-canonical ones, indicating if parameter constraints are respectively defined relatively to the initial data space or relatively to a transformation (a factorial mapping typically). Before presenting such models, and their related model selection process, we draw what are the pros (blessing) and the cons (curse) of having many variables for performing a cluster analysis process.

2.2 HD clustering: Curse or blessing?

2.2.1 HD density estimation: Curse

In the previous section, we provided some examples of high-dimensional data sets. In the present section, the aim is to give a somewhat more theoretical definition of what a high-dimensional data set should be in a density estimation setting. Such a definition will dramatically depends on the non-parametric and

on the parametric cases. It also relies on some asymptotic arguments. Remind that we consider a data set $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, $\mathbf{x}_i$ being described by d variables.

Non-parametric case

In the non-parametric situation, usually $\mathbf{x}_i$ is considered to rely in a high-dimensional space as soon as $n = o\left(e^d\right)$, thus as soon as the logarithm of the sample size, $\ln n$, is negligible beside the space dimension d. A first justification of this claim is given by Bellman [1961]: To approximate within error $\epsilon > 0$ a (Lipschitz) function of d variables, about $(1/\epsilon)^d$ evaluations (provided by the sample size $n \ldots$) on a grid are required. A second justification is also given by Silverman [1986]: Approximating a Gaussian distribution with fixed Gaussian kernels and with approximate error of about 10% requires a sample size $\log_{10} n(d) \approx 0.6(d - 0.25)$. For instance, with $d = 10$, $n(10) \approx 7.10^5$, implying already a huge sample size for a quite moderate dimensional setting.

Parametric case

In the parametric situation, let $\mathcal{S}_{\mathbf{m}}$ be a model described by $D_{\mathbf{m}}$ continuous parameters, likely depending on the dimension d. In such a case, the data set $\mathbf{x}$ is said to rely in a high-dimensional space as soon as n is small in comparison to a particular function g of $D_{\mathbf{m}}$, namely $n = o(g(D_{\mathbf{m}}))$. As an illustration for g, we consider the heteroscedastic Gaussian mixture with true parameter $\boldsymbol{\theta}^*$ and K components. We note $\hat{\boldsymbol{\theta}}_K$ the Gaussian MLE with K components. In that situation, g is a linear function from the following result (Maugis and Michel [2012]): It exists positive constants κ and A such that

$$\mathbb{E}_{\mathbf{x}}[d_H^2(f(\cdot;\boldsymbol{\theta}^*), f(\cdot|\hat{\boldsymbol{\theta}}_{\hat{K}}))] \leq \kappa \left[\inf_K \{\mathrm{KL}(f(\cdot;\boldsymbol{\theta}^*), f(\cdot;\hat{\boldsymbol{\theta}}_K)) + \mathrm{pen}(K)\} + \frac{1}{n}\right]$$

where d_H denotes the Hellinger distance, KL the Kullback-Leibler divergence and

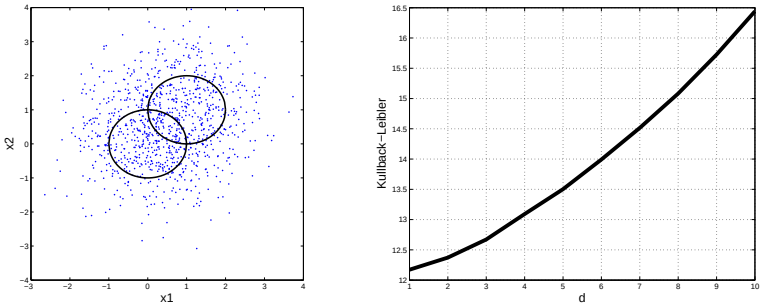
$$\mathrm{pen}(K) \geq \kappa \frac{D_K}{n} \left\{2A \ln d + 1 - \ln\left(1 \wedge \left[\frac{D_K}{n} A \ln d\right]\right)\right\}.$$

Thus the HD non-parametric and parametric situations are drastically different in magnitude. However, in practice, D_K can be high since $D_K \sim d^2/2$ in this Gaussian situation, combined with potentially large constants. For highlighting this fact, consider the following two-component multivariate Gaussian mixture:

$$\pi_1 = \pi_2 = \frac{1}{2}, \quad \mathbf{X}_1|Z_{11} = 1 \sim \mathsf{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{X}_1|Z_{12} = 1 \sim \mathsf{N}(\mathbf{1}, \mathbf{I}), \tag{2.1}$$

with $\mathbf{a} = (a \ldots a)'$ a real vector of size d. An illustration of this setting is displayed in Figure 2.3(a). Note that the two components are more and more separated when d grows since $\|\mathbf{1} - \mathbf{0}\|_{\mathbf{I}} = \sqrt{d}$. However, the quality of the

mixture density estimate degrades (the Kullback-Leibler divergence increases) when dimension increases as it is illustrated in Figure 2.3(b) with a homoscedastic model and with equal mixing proportions.



(a) (b)

Figure 2.3: HD curse in the parametric density estimation context: (a) A bivariate data set example with isodensity of each component and (b) the Kullback-Leibler divergence of the density estimate when d increases.

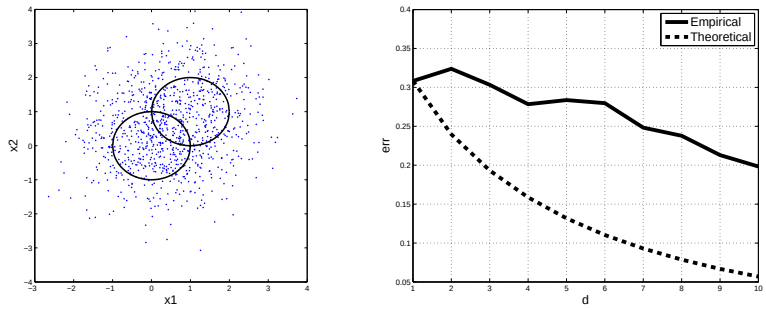
2.2.2 HD clustering: A mix of curse and blessing

Contrary to density estimation where increasing dimension has a clear negative effect, dimension may have both positive and negative effects on the clustering task. We distinguish now which factors favor such "blessing" or "curse" outcomes.

Blessing factors

We retrieve the model design (2.1). We display again a corresponding sample in Figure 2.4(a). We have already mentioned that the two components are more and more separated when d increases. The reason is that each variable uniformly provides its own separation information such that the associated theoretical error decreases when d grows. Indeed, this error is equal to $\mathsf{err}_{theo} = \Phi(-\sqrt{d}/2)$, where Φ is the cdf of $\mathsf{N}(0,1)$. We can see this decrease with d by a dash line in Figure 2.4(b). An interesting consequence is then that the empirical error rate decreases also with d as it could be noticed in continuous line in Figure 2.4(b). It means that increasing dimension may have a positive effect on the clustering task as soon as all variables convey meaningful information on the hidden partition.

We propose now to illustrate more drastically this positive effect through a simple factorial mapping visualization. We consider the three following Gaus-



(a) (b)

Figure 2.4: HD blessing in the clustering context when most variables convey independent partitioning information: (a) A bivariate data set example with isodensity of each component and (b) the theoretical (dash line) and the empirical (continuous line) error rate when d increases.

sians, all more and more separated when d increases:

$$\pi_1 = \pi_2 = \pi_3 = \tfrac{1}{3},$$
$$\mathbf{X}_1|Z_{11} = 1 \sim \mathsf{N}(\mathbf{0}, \mathbf{I}), \quad \mathbf{X}_1|Z_{12} = 1 \sim \mathsf{N}(\mathbf{2}, \mathbf{I}), \quad \mathbf{X}_1|Z_{13} = 1 \sim \mathsf{N}(-\mathbf{2}, \mathbf{I}), \quad .$$

Then Figure 2.5(a)-(d) displays a related sample of size $n = 1000$ for different dimensions on the main two axes of the Factorial Discriminant Analysis (FDA) mapping. It clearly appears that components are more and more easily recognized when dimension increases, although it is a simple visualization process. At the limit, no complex clustering algorithm would be enough to identify clusters...

Curse factors

In fact, increasing dimension may have a positive effect on clustering retrieval only if variables inject some partioning information. In addition, such information has to be not redundant. We illustrate now these two particular features.

Firstly, we consider many variables which provide no separation information. We retrieve the same parameter setting as (2.1) except that the components are not more separated when d grows since $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\mathbf{I}} = 1$, where $\boldsymbol{\mu}_1 = \mathbf{0}$ is the center of the first Gaussian and where $\boldsymbol{\mu}_2 = (1\ 0\ \ldots\ 0)'$ is the one of the second, thus $(k = 1, 2)$

$$\mathbf{X}_1|Z_{1k} = 1 \sim \mathsf{N}(\boldsymbol{\mu}_k, \mathbf{I}). \tag{2.2}$$

A sample is displayed on Figure 2.6(a). Figure 2.6(b) shows in dash line that the theoretical error rate is constant (it corresponds to $\mathsf{err}_{theo} = \Phi(-\frac{1}{2})$) when the dimension increases, as expected. Consequently, the empirical error rate degrades in this situation (continuous line of the same figure).

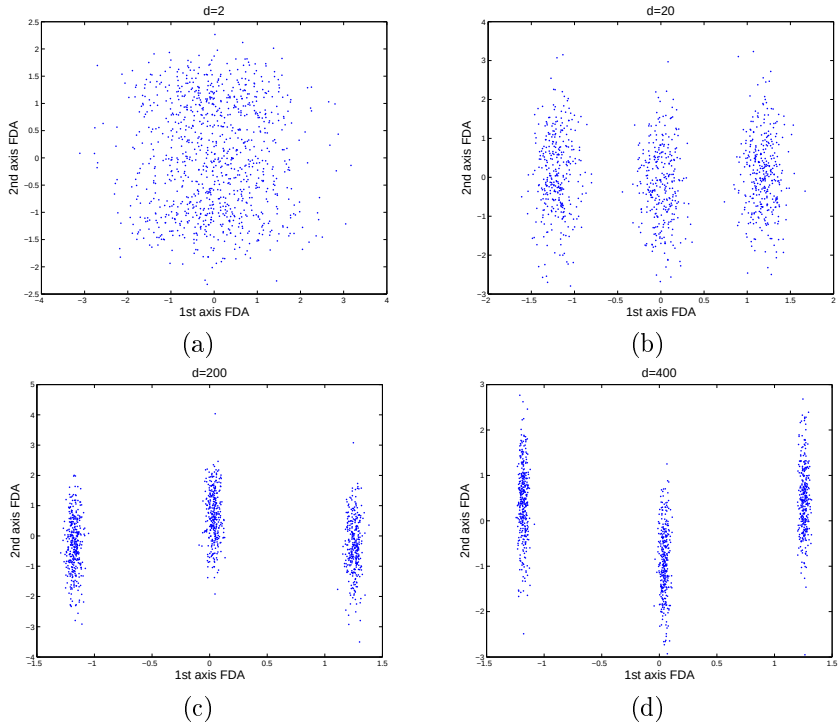


Figure 2.5: Factorial Discriminant Analysis (FDA) on the main two factorial axes of three Gaussian components more and more separated when the space dimension increases: (a) $d = 2$, (b) $d = 20$, (c) $d = 200$ and $d = 400$.

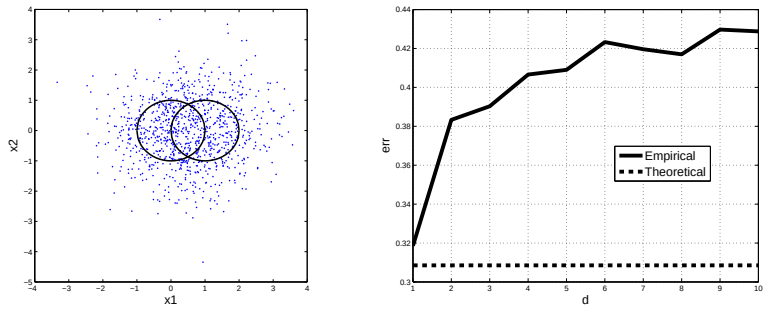
Secondly, we consider a case where many variables provide separation, but redundant information, in the following sense: It is the same parameter setting as before for the first dimension except for all other ones

$$\mathbf{X}_{1j} = \mathbf{X}_{11} + \varepsilon_j, \quad \text{where} \quad \varepsilon_j \overset{iid}{\sim} \mathsf{N}(0,1) \quad (j = 2, \ldots, d). \tag{2.3}$$

See a data example in Figure 2.7(a). Thus, components are not more separated when d grows since $\|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|_{\boldsymbol{\Sigma}} = 1$, $\boldsymbol{\Sigma}$ denoting the common covariance matrix of each Gaussian component, and $\boldsymbol{\mu}_k$ denoting the center of the component $k = 1, 2$ (note that both $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}$ could be easily computed from Equation (2.2) and (2.3)). Consequently, $\mathsf{err}_{theo} = \Phi(-\frac{1}{2})$ is constant and the empirical error increases with d, as illustrated in Figure 2.7(b) with previous conventions.

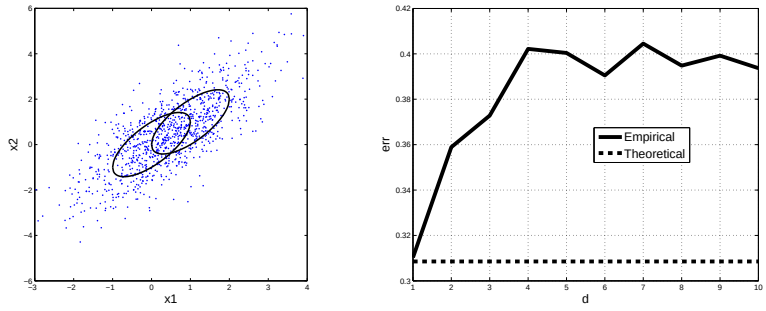
2.2.3 Intermediate conclusion

In case where variables have important blessing consequences for the clustering performance, it is important to perform the clustering task in the whole data



(a) (b)

Figure 2.6: HD curse in the clustering context when variables convey no partitioning information: (a) A bivariate data set example with isodensity of each component and (b) the theoretical (dash line) and the empirical (continuous line) error rate when d increases.



(a) (b)

Figure 2.7: HD curse in the clustering context when variables convey redundant partitioning information: (a) A bivariate data set example with isodensity of each component and (b) the theoretical (dash line) and the empirical (continuous line) error rate when d increases.

space. In particular, "filter" methods performing variable selection before the clustering task have to be excluded, the risk of removing discriminant features being too large. The remaining question is then which "wrapper" methods to be used? Such methods should "manage" with priority the fact that some variables have negative effects for clustering. The general answer is to design specific parsimonious models for clustering, the most emblematic ones relying on some variable selection principle. We will see also several alternative strategies, in particular variable clustering (to not be mingled with individual clustering, our primary task), aiming at assigning different roles ("clusters") to the variables. Such a principle is quite widespread in fact (in the canonical data space or in

a transformed space) even if it is not often initially described with this point of view.

Behing this model design which is the first step of high-dimensional model-based clustering, the question of model selection is then asked. In some situations, traditional model selection criteria could be directly applied. However, in many cases, two kinds of difficulties may happen. Firstly, the number of competing models avoids to enumerate all possible models which compete. Typically, in a variable selection context the number of possibilities is combinatorial. In such a case, strategies for designing an intelligent path in a relevant subset of models is a possible answer. Secondly, validity of traditional model selection criteria themselves can be challenged, requiring some original proposals.

In the rest of this chapter, we will give an overview of the main high-dimensional clustering methods. We will systematically highlight novelty of the proposed models, possible connections between them (variable selection or variable clustering, initial space or non-canonical space) and issues for model selection (criteria and strategies of use).

2.3 Non-canonical models

As discussed previously, models designed for high-dimensional clustering rely on parsimonious definition of related parameters. In this section, we focus on situations where parsimony is injected through parameters defined in a transformed feature space, called here non-canonical feature space. We consider this case before the canonical feature space situation (next section) since it is somewhat related to the pioneering idea of filtering. Indeed, factorial analysis (for instance principal component analysis in the continuous case) was first conducted for selecting (new) variables before applying any clustering method on them. Here, ideas are related but with a wrapper point of view. Most situations address continuous features.

2.3.1 Gaussian mixture of factor analysers

In Gaussian model-based clustering, increasing the number of variables has its main effect on the number of parameters included in the covariance matrices $\mathbf{\Sigma}_k$, since it is of quadratic order. Consequently, most methods aim at introducing parsimony first on $\mathbf{\Sigma}_k$. History and details could be found in Bouveyron and Brunet [2014]. In particular, Ghahramani and Hinton [1997] and McLachlan [2003] design the following reparameterization of $\mathbf{\Sigma}_k$:

$$\mathbf{\Sigma}_k = \mathbf{B}_k\mathbf{B}_k' + \omega_k\mathbf{\Lambda}_k$$

where $\mathbf{B}_k$ is a *loadings* $d \times q$ non-square real matrix ($1 \leq q \leq q_{\max}$, $q_{\max} < d$), ω_k is a positive real number and $\mathbf{\Lambda}_k$ is a $d \times d$ diagonal positive definite matrix such that $|\mathbf{\Lambda}_k| = 1$. For a well understanding of the underlined motivation,

it is equivalent to assuming $\mathbf{X}_1 \in \mathbb{R}^d$ to be generated by the following latent variable $\mathbf{Y}_1 \in \mathbb{R}^q$ lying in a smaller (latent) space than $\mathbb{R}^d$

$$\mathbf{X}_1 | \mathbf{Y}_1, Z_{1k} = 1 = \mathbf{B}_k \mathbf{Y}_1 + \boldsymbol{\mu}_k + \boldsymbol{\varepsilon}_k$$

where $\mathbf{Y}_1 \perp \boldsymbol{\varepsilon}_k$ ($\perp$ denoting independence), $\mathbf{Y}_1 \sim \mathsf{N}(\mathbf{0}, \mathbf{I})$ and $\boldsymbol{\varepsilon}_k \sim \mathsf{N}(\mathbf{0}, \omega_k \boldsymbol{\Lambda}_k)$. In this layout, $\mathbf{Y}_1$ is called the *factor*, by straightforward analogy to factor analysis methods. Estimation is performed through an alternating expectation-condition maximization (AECM) algorithm (Meng and van Dyke [1997]).

Complexity of such a model is equal to $D_{\mathbf{m}} = (K-1) + Kd + Kq[d - (q-1)/2] + Kd$, where it can be seen that the quadratic part has vanished. In fact, it corresponds to the most complex model of a whole family, McNicholas and Murphy [2008] having defined 12 associated parsimonious versions, including for instance inter-class equality between $\mathbf{B}_k$, identity of $\boldsymbol{\Lambda}_k = \mathbf{I}$, *etc.* Finally, models in competition $(\mathcal{S}_{\mathbf{m}})_{\mathbf{m} \in \mathcal{M}}$ gather the combinations non only of these 12 parsimonious versions but also of the couples (q, K) of the latent dimension and of the number of components. In practice, $q_{\max}$ is expected to be quite low for parsimonious reasons and thus the cardinal of $\mathcal{M}$ is not excessively high. Traditional model selection criteria (as BIC) can then be directly applied on this collection. The R package PGMM[2] provides an implementation of this method.

2.3.2 HD Gaussian mixture models

Bouveyron *et al.* [2007] propose another way for obtaining parsimony on the covariance matrices $\boldsymbol{\Sigma}_k$. It relays on the following spectral decomposition

$$\boldsymbol{\Sigma}_k = \mathbf{D}_k \boldsymbol{\Delta}_k \mathbf{D}_k'$$

where $\mathbf{D}_k$ is the orthogonal matrix of the eigenvectors of $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Delta}_k$ is a diagonal matrix containing the related eigenvalues. They impose $\boldsymbol{\Delta}_k$ to follow the parsimonious structure

$$\boldsymbol{\Delta}_k = \begin{pmatrix} \boxed{\begin{matrix} a_{k1} & & 0 \\ & \ddots & \\ 0 & & a_{kq_k} \end{matrix}} & & \mathbf{0} \\ & \boxed{\begin{matrix} b_k & & 0 \\ & \ddots & \\ 0 & & b_k \end{matrix}} & \\ \mathbf{0} & & \end{pmatrix} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} q_k \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} (d - q_k) \end{matrix}$$

with $a_{kj} \geq b_k > 0$, for $j = 1, ..., q_k$ and $q_k < d$. Such an assumption can be somewhat related to a kind of principal component analysis per Gaussian group. It could also be viewed as a kind of variable clustering selection, the

[2] `http://cran.r-project.org/web/packages/pgmm/index.html`

$d - q_k$ remaining variables of $\mathbf{\Delta}_k$ corresponding to a group of "noisy" features. Figure 2.8 illustrates a three dimensional ($d = 3$) and two components situation ($K = 2$) where both subspace dimensions q_1 and q_2 are equal ($q_1 = q_2 = 2$) but differ in orientation. Estimation can easily performed through an EM algorithm for instance.

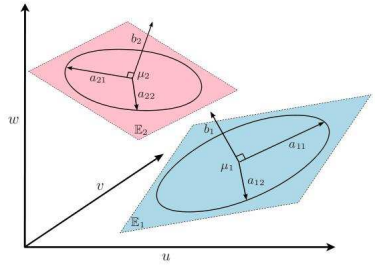


Figure 2.8: Illustration of the HD clustering mixture Gaussian model in a two components situation (provided by Bouveyron *et al.* [2007]).

Complexity of such a model is given $D_{\mathbf{m}} = (K-1) + Kd + \sum_{k=1}^{K} q_k[d - (q_k+1)/2] + \sum_{k=1}^{K} q_k + 2K$. In addition, Bouveyron *et al.* [2007] propose eight parsimonious versions by imposing for instance equality between subspace dimensions ($q_k = q$, for all k), *etc.* Finally, the whole model family $(S_m)_{\mathbf{m}\in\mathcal{M}}$ includes couples $((q_1, \ldots, q_K), K)$ of subspace dimension and number of components, combined with the eight models. Since q_k may depend on the component, contrary to the Gaussian mixture of factor analysers described in the previous section, the number of models becomes combinatorial. Then, it may be difficult in the HD setting to browse all models for applying a BIC-like criterion for instance. Consequently, Bouveyron *et al.* [2007] propose a kind of rule of thumb criterion for selecting each q_k, looking for a break in the eigenvalue scree of the empirical covariance matrix for each group component, the so-called scree test of Cattell [Cattell, 1966]. The RMIXMOD package[3] (Lebret *et al.* [2015]) implements these models.

2.3.3 Functional data

Functional and discretized data

Strictly speaking, real functional data (Ramsay and Silverman [2005], Ferraty and Vieu [2006]) correspond to $i = 1, \ldots, n$ curves which are realizations of n random variables linked to n L^2-continuous real-valued stochastic processes $\mathbf{Y}_i = \{\mathbf{Y}_i(\mathrm{t}) \in \mathbb{R}, \mathrm{t} \in [0, T]\}$ taking values in a Hilbert space H of functions defined on the (time) interval $[0, T]$. Thus, it corresponds to an infinite dimensional space. Since most functional data are longitudinal, we adopt here

[3] `http://cran.r-project.org/web/packages/Rmixmod/index.html`

the convention of parameterizing models in terms of time. However, it applies equally well with any other features as angle, length, *etc.* In addition, extensions are possible for multivariate curves, it means that individual i is described by several curves (see for instance James and Sugar [2003] or Jacques and Preda [2014b]).

In practice, each $\mathbf{Y}_i$ is unobserved for two, essentially technological, reasons. Firstly, the n curves $\mathbf{Y}_i$ are discretized each in m_i time-points $\{Y_i(\mathrm{t}_{is}), 0 \leq s \leq m_i, \mathrm{t}_{is} \in [0,T]\}$. Secondly, an error on observation is usually present such that only m_i ordered time-points $\{X_i(\mathrm{t}_{is}), 0 \leq s \leq m_i, \mathrm{t}_{is} \in [0,T]\}$ $(i = 1,\ldots,n)$ are available for each curve. For instance, the following relationship between discretized (unobserved) values $Y_i(\mathrm{t}_{is})$ and noisy (observed) values $X_i(\mathrm{t}_{is})$ could be assumed:

$$X_i(\mathrm{t}_{is}) = Y_i(\mathrm{t}_{is}) + \boldsymbol{\varepsilon}_{is}, \tag{2.4}$$

where $\boldsymbol{\varepsilon}_{is}$ has zero mean and is uncorrelated with each other and $Y_i(\mathrm{t}_{is})$. Other assumptions are possible as we will see below.

We refer to Jacques and Preda [2014a] for a general review on clustering for functional data, including the model-based one. Difficulty of performing unanimous clustering on generative distributions comes from the fact that, contrary to the finite-dimensional setting, the notion of density probability is generally not defined for functional random variable (Delaigle and Hall [2010]). Consequently, related techniques require defining density probabilities in a finite-dimensional space, leading to multiple and different implementations.

In this chapter, we divide model-based clustering techniques into two different categories: these ones where the generative model is explicitly defined on the observed values $\mathbf{X}_i = \{X(\mathrm{t}_{is}), 0 \leq s \leq m_i, \mathrm{t}_{is} \in [0,T]\}$, $i = 1,\ldots,n$, and these ones for which it is not the case. Indeed, this split will have important consequences for some aspects concerning model selection.

Clustering with *no explicit* distribution on $\mathbf{X}_i$

Usually, the first step before a clustering method is to reconstruct the initial functional form of data. It can then be viewed as a preprocessing step ("filtering" method). It often relies on the assumption that the unobserved curve $\mathbf{Y}_i$ can be expressed in a basis of d functions $\{\phi_j\}_{j=1,\ldots,d}$, for instance B-splines or wavelets, in the following form:

$$Y_i(\mathrm{t}) = \sum_{j=1}^{d} \gamma_{ij}\phi_j(\mathrm{t}).$$

Using then the regression (2.4) hypothesis, traditional least squared coefficients estimates are obtained by

$$\hat{\boldsymbol{\gamma}}_i = (\boldsymbol{\Phi}_i'\boldsymbol{\Phi}_i)^{-1}\boldsymbol{\Phi}_i'\mathbf{X}_i$$

where $\mathbf{\Phi}_i = (\phi_j(\mathrm{t}_{is}))$ is a $m_i \times d$ matrix gathering the value of each basis function for each time discretization knot. Finally, standard model-based clustering techniques (typically multivariate Gaussian mixtures, eventually HD variants previously described in Sections 2.3.1 and 2.3.2) can be directly applied on the estimated coefficients $\hat{\boldsymbol{\gamma}}_i$. The partition on individuals $\mathbf{X}_i$ is obtained as a simple by-product, being the same as this one of individuals $\hat{\boldsymbol{\gamma}}_i$.

Instead of partitioning the basis coefficients $\hat{\boldsymbol{\gamma}}_i$, a model-based clustering technique can be alternatively applied to some principal component scores resulting from functional principal component analysis (FPCA) of the previous reconstructed curves. In practice, the computational process for implementing FPCA consists of performing a standard (centered) PCA to the matrix $\tilde{\mathbf{\Gamma}}\mathbf{W}\tilde{\mathbf{\Gamma}}'\mathbf{T}$, where $\mathbf{\Gamma} = (\hat{\gamma}_{ij})$ is the $n \times d$ matrix of estimated coefficients, $\mathbf{T} = \frac{1}{n}\mathbf{I}$ is the $n \times n$ matrix of weights for curves, $\tilde{\mathbf{\Gamma}}$ is the $n \times d$ matrix of centered coefficients of $\mathbf{\Gamma}$ and $\mathbf{W}$ is the $d \times d$ matrix of the inner products $w_{jj'} = \int_0^T \phi_j(\mathrm{t})\phi_{j'}(\mathrm{t})d\mathrm{t}$ $(1 \leq j, j' \leq d)$ (it acts like a metric). Thus, the jth principal component score $\mathbf{C}_j$ is the jth eigenvector associated to the largest jth eigenvalue:

$$\tilde{\mathbf{\Gamma}}\mathbf{W}\tilde{\mathbf{\Gamma}}'\mathbf{T}\mathbf{C}_j = \alpha_j\mathbf{C}_j.$$

As usual with PCA, FPCA performs a kind of variable ordering. Finally, clustering is performed on a truncating principal component scores $\mathbf{C}_1, \ldots, \mathbf{C}_q$, with $q \leq d$.

From a model selection point of view, both previous methods allow to use some information criteria like BIC for selecting the number K of components. However, it is not really possible to use them for selecting other parts of the model which are the functional basis $\{\phi_j\}_{j=1,\ldots,d}$ and, specifically to FPCA, the truncation of order q.

Clustering with *explicit* distribution on $\mathbf{X}_i$

Ideally, for beneficing from the whole mathematical statistics corpus, model-based clustering techniques would require a distribution on all $\mathbf{X}_i = (X_i(\mathrm{t}_{is}), 0 \leq s \leq m_i, \mathrm{t}_{is} \in [0, T])$, $i = 1, \ldots, n$. First of all, it is important to notice that performing the clustering task directly with observed values $\mathbf{X}_i$'s as if they would correspond to classical multivariate values is not desirable, even if it could meet this goal. The first reason is that each $\mathbf{X}_i$ does not necessarily rely in the same space dimension (here m_i for each), even if in practice it could be often the case. The second and the most important reason is that working with such raw data wastes order information on them.

Contrary to the raw data case, several techniques propose distributions on $\mathbf{X}_i$ which take all the functional data specificity into account. Jacques and Preda [2013] perform FPCA by group, leading to principal components per group noted C_{ijk}. In addition, they assume a Gaussian distribution of the C_{ijk}, leading to conditional independence of them since being already uncorrelated.

It leads to the following Gaussian mixture model, relying on a truncation of order $1 \leq q_k \leq d$ for each component:

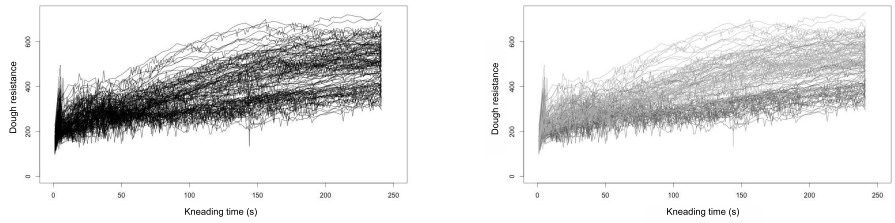
$$f(\mathbf{x}_i;\boldsymbol{\theta}) \approx \sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} \phi(C_{ijk};0,\alpha_{jk})$$

where $\phi(\cdot;0,\alpha_{jk})$ is the univariate Gaussian density of mean zero (scores C_{ijk} are centered) and variance α_{kj} (corresponding also to eigenvalues). Then, parameter estimation is provided through an EM-like algorithm for maximizing the (pseudo) log-likelihood, where both steps are the following:

E-step Compute conditional probabilities $t_{ik} \propto \pi_k \prod_{j=1}^{q_k} \phi(C_{ijk};0,\alpha_{kj})$ as usual.

M-step First, principal scores are updated. Notice that weights $\mathbf{T}_k$ depend now on t_{ik}'s, $\boldsymbol{\Gamma}_k$ too. Second, perform the q_k truncation order selection by detecting a kind of elbow in the eigenvalues by the scree test of Cattell (Cattell [1966]). Finally, parameters π_k are computed as usual and parameters α_k are already given from previous conditional FPCA.

This process is implemented in the R FUNCLUST[4] package. As an illustration, this package is applied to kneading curves, which are described in Section 2.1, in Figure 2.9. From a model selection point of view, there are some important remarks. Strickly speaking, it is just a pseudo likelihood method since data C_{ijk} are changing at each iteration step of EM. Consequently, using selection criteria like BIC could be hazardous for choosing K, q_k or the functional basis. However, in practice, BIC works well for choosing K. However, it is not used for selecting q_k, as previous said, for limiting computing time. No attempt for choosing the basis is performed.



(a) (b)

Figure 2.9: $n = 115$ kneading curves observed at $d = 241$ equispaced instants of time in the interval [0; 480]: (a) raw curves, (b) three groups partioning curves with the FUNCLUST package.

Alternatively, James and Sugar [2003] consider randomness directly on the basis coefficients $\boldsymbol{\gamma}_i$. They assume that $\boldsymbol{\gamma}_i$ arises from a homoscedastic Gaussian

[4] `http://cran.r-project.org/web/packages/Funclustering/index.html`

multivariate model which, coupling with (2.4), provides the following regression model, conditionally on the ith curve belonging to the kth cluster (so conditional to $Z_{ik} = 1$) :

$$\mathbf{X}_i = \mathbf{\Phi}_i(\boldsymbol{\mu}_k + \boldsymbol{\epsilon}_i) + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\epsilon}_i \sim \mathsf{N}(\mathbf{0}, \mathbf{\Sigma})$ and $\boldsymbol{\varepsilon}_i \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Also, some parsimonious assumptions are made on centers $\boldsymbol{\mu}_k$. Then, an EM algorithm allows to estimate all parameters. Contrary to the model of Jacques and Preda [2013] described just above, we are now faced to an unambiguous generative approach allowing straightforward model selection with any classical criterion for choosing every quantity of interest (the number K of clusters, the basis $\{\phi_j\}$ and the parsimony of all means $\boldsymbol{\mu}_k$), even if the authors prefer to use a so-called "distortion function" criterion for selecting K faster since avoiding EM computations for all K values.

In the same spirit as James and Sugar [2003], Samé *et al.* [2011] give another regression model providing a full generative, flexible and parsimonious distribution on the $\mathbf{X}_i$'s. They assume that the curves arise from a mixture of regressions on a basis of polynomial functions (the order to be given by model selection), with possible changes in regime at each instant of time. The mixing proportions are defined by logistic functions for allowing segmentation in time. An EM procedure is performed for estimation and several parsimonious versions are described. This full generative distribution allows again full model selection (number of clusters, polynomial order of the basis function and number of regime changes) in any standard way. However, as in many previous settings, the number of competing models can increase drastically. For instance, the basic functions can change by regime, multiplying combinations.

2.3.4 Intermediate conclusion

Many parsimonious modelling solutions exist for dealing with HD data, concerning as well independent and functional data, even if some gaps remain to be filled like categorical functional data or also mixed (continuous and categorical typically) multivariate functional data. Most of existing models rely on a generative distribution on the data space, allowing direct use of standard selection criteria. However, the crucial question is focused on the multiplicity of models to be compared. It is the reason why some authors favor some more empirical, but fast, rules for model selection.

We guess that future researches should address new advances for fast selection of multiple models in a short allocated time. In the next section, devoted to canonical model setting, we will see early several attempts for this purpose, for instance by designing a particular strategy in the model space, avoiding all model evaluation.

2.4 Canonical models

We address now models for HD data which position parsimony assumptions directly on the initial (or canonical) variable space. Advantage of such approaches, beside non-canonical ones, is a great model readibility for the practitioner. Indeed, this one is usually more accustomed to his variable set than to a somewhat more artificial set, as the factorial features could be sometimes.

In this context, this chapter tackles important notions: variable selection, variable clustering, model selection validity and also strategies for dealing with model multiplicity.

2.4.1 Parsimonious mixture models

Classical mixture models have already been presented in Chapter **??**, Section **??**. It gathers in particular the Gaussian mixture model for the continuous case and the latent multinomial mixture model for the categorical case, including also many parsimonious variants. Dealing with HD data impose to consider essentially some of the most parsimonious ones thus there is a need to provide more details in this section. Then, extension to the mixed case (merging continuous and categorical features) is presented as a straightforward extension. All these models are implemented in the R package RMIXMOD[5]. Finally, we will present a new attempt for variable selection in the continuous, categorical and mixed situations.

Spherical and diagonal Gaussian mixtures for continuous variables

We consider data sets $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, with $\mathbf{x}_i \in \mathbb{R}^d$. The most parsimonious Gaussian mixture models defined by Celeux and Govaert [1995] belong to the so-called spherical and diagonal families. An example of diagonal model is given in Figure 2.10. Using notations already provided in Section **??** of Chapter **??**, their most complex versions respectively correspond to constraints $\mathbf{\Sigma}_k = \lambda_k \mathbf{I}$ and $\mathbf{\Sigma}_k = \lambda_k \mathbf{B}_k$ on the covariance matrix $\mathbf{\Sigma}_k$ of the kth component, where $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$ and $\mathbf{B}_k$ diagonal with $|\mathbf{B}_k| = 1$. Including some parsimonious versions, which allow some parts to vary or not between components, a total of two spherical and four diagonal models are available. All models, and their respective number of parameters, are displayed in Table 2.1. Model selection can be easily performed by traditional criteria, like BIC.

Latent class model for categorical variables

We consider now data sets $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, each $\mathbf{x}_i$ containing d categorical variables, the jth having m_j response levels. The coding $\mathbf{x}_i = (x_i^{jh}; j =$

[5] `http://cran.r-project.org/web/packages/Rmixmod/index.html`

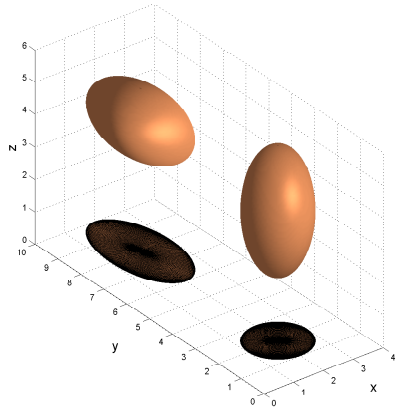


Figure 2.10: Isodensity of a two-components diagonal Gaussian mixture in the three-dimensional space.

Family	Model	Number of parameters
diagonal	$[\lambda \mathbf{B}]$	$\dim(\boldsymbol{\pi}) + Kd + d$
	$[\lambda_k \mathbf{B}]$	$\dim(\boldsymbol{\pi}) + Kd + d + K - 1$
	$[\lambda \mathbf{B}_k]$	$\dim(\boldsymbol{\pi}) + 2Kd - K + 1$
	$[\lambda_k \mathbf{B}_k]$	$\dim(\boldsymbol{\pi}) + 2Kd$
spherical	$[\lambda \mathbf{I}]$	$\dim(\boldsymbol{\pi}) + Kd + 1$
	$[\lambda_k \mathbf{I}]$	$\dim(\boldsymbol{\pi}) + Kd + K$

Table 2.1: Some characteristics of the two spherical and the four diagonal models. We have $\dim(\boldsymbol{\pi}) = K - 1$ in the case of free proportions and $\dim(\boldsymbol{\pi}) = 0$ in the case of equal proportions.

$1, \ldots, d; h = 1, \ldots, m_j)$ indicates that $x_i^{jh} = 1$ if i has response level h for variable j and $x_i^{jh} = 0$ otherwise. The standard model for clustering observations described through categorical variables is the so-called latent class model (see for instance Goodman [1974]). Data are assumed to arise independently from a mixture of K multivariate multinomial distributions with pdf

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}, \tag{2.5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ denotes the vector parameter of the latent class model to be estimated, with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$ and $\boldsymbol{\alpha}_k = (\alpha_k^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$, α_k^{jh} denoting the probability that variable j has level h if object i is in cluster k. Thus, the latent class model assumes that the variables are *conditionally independent* knowing the latent groups.

Lebret *et al.* [2015] propose four parsimonious versions, with thus a total of five models. They correspond to an extension of the parameterization of Bernoulli distributions used by Celeux and Govaert [1991] for clustering and also by Aitchinson and Aitken [1976] for kernel discriminant analysis. The basic idea is to impose the vector $\boldsymbol{\alpha}_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$ to take the form $(\beta_k^j, \ldots, \beta_k^j, \gamma_k^j, \beta_k^j, \ldots, \beta_k^j)$ with $\gamma_k^j > \beta_k^j$. Since $\sum_{h=1}^{m_j} \alpha_k^{jh} = 1$, we have $(m_j - 1)\beta_k^j + \gamma_k^j = 1$ and, consequently, $\beta_k^j = (1 - \gamma_k^j)/(m_j - 1)$. The constraint $\gamma_k^j > \beta_k^j$ becomes finally $\gamma_k^j > 1/m_j$. Then, the vector $\boldsymbol{\alpha}_k^j$ can be broken up into the two following parameters:

- $\mathbf{a}_k^j = (a_k^{j1}, \ldots, a_k^{jm_j})$ where $a_k^{jh} = 1$ if h corresponds to the rank of γ_k^j (in the following, this rank will be noted $h(k, j)$), 0 otherwise;
- $\varepsilon_k^j = 1 - \gamma_k^j$ which corresponds to the probability that the data $\mathbf{x}_i$ arising from the kth component are such that $x_i^{jh(k,j)} \neq 1$.

In other words, the multinomial distribution associated to the jth variable of the kth component is reparameterized by a center $\mathbf{a}_k^j$ and the dispersion ε_k^j around this center. Thus, it allows us to give an interpretation similar to the center and the variance matrix used for continuous data in the Gaussian mixture context. Finally, the relationship between the initial parameterization and the new one is given by:

$$\alpha_k^{jh} = \begin{cases} 1 - \varepsilon_k^j & \text{if } h = h(k, j) \\ \varepsilon_k^j / (m_j - 1) & \text{otherwise.} \end{cases} \tag{2.6}$$

In the following, this model will be denoted by $[\varepsilon_k^j]$. In this context, three other models can be easily deduced. We note $[\varepsilon_k]$ the model where ε_k^j is independent of the variable j, $[\varepsilon^j]$ the model where ε_k^j is independent of the component k and, finally, $[\varepsilon]$ the model where ε_k^j is independent of both the variable j and the component k. In order to maintain some unity in the notation, we will denote also $[\varepsilon_k^{jh}]$ the most general model initially introduced. The number of free parameters associated to each model is given in Table 2.2. Again, model selection can be easily performed by traditional criteria, like BIC.

Mixed data models

It is frequent in practice to mix continuous and categorical data. Thus the ith individual is composed by two parts, $\mathbf{x}_i = (\mathbf{x}_i^{cont}, \mathbf{x}_i^{cat})$, $\mathbf{x}_i^{cont}$ and $\mathbf{x}_i^{cat}$ designing the continuous and the categorical ones respectively. In that case, it is easy to combine (diagonal) parsimonious Gaussian mixture and latent class model by conditional independence [Moustaki and Papageorgiou, 2005]:

$$f(\mathbf{x}; \boldsymbol{\alpha}_k) = f(\mathbf{x}^{cont}; \boldsymbol{\alpha}_k^{cont}) \times f(\mathbf{x}^{cat}; \boldsymbol{\alpha}_k^{cat})$$

Model	Number of parameters
$[\varepsilon]$	$\dim(\boldsymbol{\pi}) + 1$
$[\varepsilon^j]$	$\dim(\boldsymbol{\pi}) + d$
$[\varepsilon_k]$	$\dim(\boldsymbol{\pi}) + K$
$[\varepsilon_k^j]$	$\dim(\boldsymbol{\pi}) + Kd$
$[\varepsilon_k^{jh}]$	$\dim(\boldsymbol{\pi}) + K\sum_{j=1}^{d}(m_j - 1)$

Table 2.2: Number of free parameters of the five multinomial models. We have $\dim(\boldsymbol{\pi}) = K - 1$ in the case of free proportions and $\dim(\boldsymbol{\pi}) = 0$ in the case of equal proportions.

with $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_k^{cont}, \boldsymbol{\alpha}_k^{cat})$ (see also Section ?? in Chapter ??). Then, the previous six Gaussian mixture models and the five multinomial mixture models can be combined, defining straightforwardly 30 new mixed models. Classical criteria can be used for selecting them, with also the number of clusters K.

Although previously described models, in the continuous, categorical or mixed data situations, are the most parsimonious ones in their respective families, they are not really designed for realistic HD situations involving several thousands of variables for instance. Indeed, their parameter number remains too high in such cases.

Variable selection has always been a natural answer for HD clustering as already discussed in the beginning of this chapter. Typically, filtering methods relying on a preliminary factorial analysis step then cut the number of factorial variables to be retained. However, in model-based clustering involving a full wrapping approach, the difficulty is to integrate properly this selection step in the model itself. Thus, we discuss now more suitable methods for the HD situation.

2.4.2 Variable selection through regularization

In this section, we focus on the variable selection problem in the Gaussian mixture clustering context.

ℓ_1-penalization procedures

Inspired by the success of the Lasso regression, Pan and Shen [2007] propose to take advantage of the sparsity property of ℓ_1-penalization of the likelihood to perform automatic variable selection for high-dimensional model-based clustering. Their procedure, called PS-Lasso in the sequel, consists of using a Lasso method to select relevant clustering variables and estimate mixture parameters in the same exercise. The covariance matrices are assumed to be identical and diagonal ($\boldsymbol{\Sigma}_k = \mathbf{V} = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2)$) and an ℓ_1 penalty is considered on mean

parameters. For any $K \in \mathbb{N}^*$, the following function has to be maximized:

$$\boldsymbol{\theta}_K \mapsto \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} \pi_k \phi \left(\bar{\mathbf{x}}_i ; \boldsymbol{\mu}_k, \mathbf{V} \right) \right] - \lambda \sum_{k=1}^{K} \left\| \boldsymbol{\mu}_k \right\|_1 , \tag{2.7}$$

where $\boldsymbol{\theta}_K = (\boldsymbol{\pi}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \mathbf{V})$, $\left\| \boldsymbol{\mu}_k \right\|_1 = \sum_{j=1}^{d} \left| \mu_{kj} \right|$, $\bar{\mathbf{x}}_i = (x_{ij} - \bar{\mathbf{x}}_j)_{1 \leq j \leq p}$ with $\bar{\mathbf{x}}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$, λ is a non-negative regularization parameter and $\phi(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian density of center $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. An EM-algorithm is proposed to solve this parameter estimation problem. Next, a modified BIC criterion is used to select K and λ:

$$\mathrm{BIC}_{(K,\lambda)} = -2 \ln \left[\prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i ; \boldsymbol{\mu}_k, \mathbf{V}) \right] + \ln(n) D_{(K,\lambda)}$$

where $D_{(K,\lambda)} = (K-1) + Kd + d - q$, q denoting the number of the maximum penalized likelihood estimate mean components that are equal to 0.

This approach was successively extended in Zhou *et al.* [2009] (Gaussian mixtures with diagonal covariance matrices) and finally in Zhou *et al.* [2009]. In this last paper, a regularized Gaussian mixture model with unconstrained covariance matrices is proposed. They employ a ℓ_1 penalty on mean parameters and on covariance matrices as follows:

$$\boldsymbol{\theta}_K \mapsto \sum_{i=1}^{n} \ln \left[\sum_{k=1}^{K} \pi_k \phi \left(\bar{\mathbf{x}}_i ; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \right) \right] - \lambda \sum_{k=1}^{K} \left\| \boldsymbol{\mu}_k \right\|_1 - \rho \sum_{k=1}^{K} \left\| \boldsymbol{\Sigma}_k^{-1} \right\|_1 , \tag{2.8}$$

where

$$\left\| \boldsymbol{\mu}_k \right\|_1 = \sum_{j=1}^{d} \left| \mu_{kj} \right|, \left\| \boldsymbol{\Sigma}_k^{-1} \right\|_1 = \sum_{\substack{j,j'=1 \\ j \neq j'}}^{d} \left| (\boldsymbol{\Sigma}_k^{-1})_{jj'} \right| ,$$

and where λ and ρ are two non-negative regularization parameters. This parameter estimation problem is solved using an EM algorithm where the so-called glasso algorithm (Friedman *et al.* [2007]) is used to estimate sparse precision matrices $\boldsymbol{\Sigma}_k^{-1}$.

Lasso-MLE procedure

In Meynet [2012] and Meynet and Maugis-Rabusseau [2012], they highlight that the ℓ_1-penalization induces shrinkage of the coefficients and thus biased estimators with high estimation risk. Moreover, the use of a BIC-type criterion for the model selection can be unsuitable for high-dimensional data. Consequently, they propose to only use an ℓ_1-penalized likelihood approach to determine potential sets of relevant variables. This allows to efficiently construct a data-driven model subcollection with reasonable complexity, even for

high-dimensional situations. The evaluation of the MLE rather than the ℓ_1-penalized estimator for each model is considered to avoid estimation problems due to ℓ_1-penalization shrinkage. More precisely, the data $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ are assumed to have a null expectation (in practice, empirical centering of the data is performed to ensure this assumption) and their unknown density f is estimated by a finite spherical Gaussian mixture. The clusters are characterized by the mean parameters $(\boldsymbol{\mu}_k)_{1\leq k\leq K}$ and a variable j is called *irrelevant* for the clustering if $\mu_{kj} = 0$ for all $k = 1, \ldots, K$; otherwise it is called *relevant*. The relevant variable subset (resp. irrelevant variable subset) is denoted by $\boldsymbol{J}_r$ (resp. $\boldsymbol{J}_r^c = \{1, \ldots, d\} \setminus \boldsymbol{J}_r$). Consequently, the variable selection problem is recast into a model selection problem, where the model collection is $(S_{(K,\boldsymbol{J}_r)})$ with

$$S_{(K,\boldsymbol{J}_r)} = \left\{ \begin{array}{l} \mathbf{x}_i \in \mathbb{R}^d \mapsto f(\mathbf{x}_i;\boldsymbol{\theta}) = \left[\sum_{k=1}^{K} \pi_k \, \phi(\mathbf{x}_i^{\boldsymbol{J}_r};\boldsymbol{\mu}_k,\sigma^2\mathbf{I})\right] \phi(\mathbf{x}_i^{\boldsymbol{J}_r^c};\mathbf{0},\sigma^2\mathbf{I}) \\ \boldsymbol{\theta} = \left(\pi_1,\ldots,\pi_K,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K,\sigma^2\right) \in \Pi_K \times \left(\mathbb{R}^{|\boldsymbol{J}_r|}\right)^K \times \mathbb{R}_+^* \end{array} \right\},$$

$\mathbf{x}_i^{\boldsymbol{J}_r}$ denoting the restriction of $\mathbf{x}_i$ on $\boldsymbol{J}_r$, $|\boldsymbol{J}_r|$ corresponding to the cardinal of $\boldsymbol{J}_r$ and Π_K denoting the simplex related to parameters $(\pi_1, \ldots, \pi_K)$. The dimension of a model $S_{(K,\boldsymbol{J}_r)}$ corresponds to the total number of free parameters estimated in the model: $D_{(K,\boldsymbol{J}_r)} = K(1 + |\boldsymbol{J}_r|)$.

The so-called Lasso-MLE procedure proposed in Meynet and Maugis-Rabusseau [2012] is decomposed into three main steps. In the first step, as Pan and Shen [2007], an ℓ_1-approach is considered: For each $(K, \lambda) \in \mathbb{N}^* \times \mathcal{G}_\lambda$ ($\mathcal{G}_\lambda$ is a given grid on λ), the Lasso estimator $\hat{\boldsymbol{\theta}}^L_{(K,\lambda)}$ is computed by maximizing (2.7) and the associated relevant variable subset is

$$\boldsymbol{J}_{(K,\lambda)} = \{j \in \{1, \ldots, d\} : \ \exists\, k \in \{1, \ldots, K\} \text{ such that } \hat{\mu}_{kj} \neq 0\}.$$

Thus a random model subcollection $\{\mathcal{S}_{(K,\boldsymbol{J}_r)} : \ (K, \boldsymbol{J}_r) \in \mathcal{M}^L\}$ is obtained, where

$$\mathcal{M}^L = \{(K, \boldsymbol{J}_r) : \ K \in \mathbb{N}^*, \boldsymbol{J}_r \in \bigcup_{\lambda\in\mathcal{G}_\lambda} \boldsymbol{J}_{(K,\lambda)}\}.$$

The second step consists of computing the MLE $\hat{\boldsymbol{\theta}}_{(K,\boldsymbol{J}_r)}$ using the standard EM algorithm for each model $(K, \boldsymbol{J}_r) \in \mathcal{M}^L$. The third step is devoted to model selection. As in Maugis and Michel [2012], a non asymptotic penalized criterion is proposed to solve the model selection problem. By extending the general model selection theorem of Massart [2007] (Theorem 7.11) (see also Section ?? in Chapter??) (demander ref à Pascal dans le book), Meynet [2012] proves that the penalty is

$$\text{pen}_{(K,\boldsymbol{J}_r)} = \kappa_1 \frac{D_{(K,\boldsymbol{J}_r)}}{n} \left[1 + \kappa_2 \ln\left(\frac{d}{D_{(K,\boldsymbol{J}_r)}}\right)\right], \tag{2.9}$$

where κ_1 and κ_2 are two unknown constants. As expected, the penalty is proportional to the model dimension. The logarithmic term quantifies the

model collection complexity by taking into account the possible large number of models with identical dimension. Nevertheless this logarithm term becomes unnecessary if the number of models with the same dimension is small enough. For instance, for finite Gaussian mixture models in a low-dimensional setting, a penalty proportional to the dimension is sufficient to select a model close to the oracle (Maugis and Michel [2011]). But in the high-dimensional context, the number of models having the same dimension is expected to grow. Nonetheless, thanks to the random preselection of relevant variables subsets, a complete variable selection is not performed here. Thus, if the random model subcollection is much poorer than the whole model collection and contains few models with the same dimension, a penalty proportional to the dimension

$$\text{pen}_{(K,\boldsymbol{J}_r)} = \frac{D_{(K,\boldsymbol{J}_r)}}{n} \tag{2.10}$$

might be sufficient to select a model with proper dimension. Next, the penalty depending on unknown multiplicative constants is calibrated using the so-called slope heuristics [Birgé and Massart, 2007; Baudry *et al.*, 2012].

Comparing PS-Lasso and Lasso-MLE

To compare the Lasso-MLE and PS-Lasso procedures, the following simulated example is proposed in Meynet and Maugis-Rabusseau [2012]. The data set consists of $n = 200$ observations described by $d = 1\,000$ variables. The data are simulated according to a mixture of two Gaussian distributions $\pi_1\,\phi(\cdot; \mathbf{0}_d, \mathbf{I}) + (1-\pi_1)\phi(\cdot; \boldsymbol{\mu}_2, \mathbf{I})$ where $\boldsymbol{\mu}_2 = (1.5, \ldots, 1.5, \mathbf{0}_{950})$ and $\pi_1 = 0.85$. The relevant variables are the first fifty variables ($\boldsymbol{J}_r^\star = \{1, \ldots, 50\}$). 20 simulations of the data set are performed. For each simulation, models with $K \in \{1, 2, 3\}$ clusters are considered. The results are summarized in Table 2.3. Table 2.3 shows that

Procedure	Estimator	TR	FR	$\hat{K}$ 1	2	3	ARI
PS-Lasso	oracle	50.3 (0.2)	214.6 (79.0)	0	16	4	0.90 (0.03)
	BIC	49.7 (0.8)	14.3 (3.4)	0	18	2	0.86 (0.02)
Lasso-MLE	oracle	50.0 (0.0)	0.2 (0.2)	0	20	0	0.95 (0.02)
	AIC	50.0 (0.0)	17.1 (4.2)	0	14	6	0.90 (0.04)
	BIC	49.8 (0.4)	4.4 (2.2)	0	20	0	0.92 (0.02)
	DDSE	50.0 (0.0)	2.4 (1.7)	0	20	0	0.94 (0.02)

Table 2.3: Averaged number of true relevant (TR) and false relevant (FR) variables ($\pm$ standard deviation); number of times a clustering with $\hat{K} = 1$, 2 and 3 components is selected; Averaged ARI ($\pm$ standard deviation) over the 20 simulations. DDSE stands for data-driven slope estimation.

the PS-Lasso oracle model, and to a lesser extend the model selected by BIC, contain many false relevant variables and may overestimate the number of

mixture components. This confirms that the PS-Lasso procedure is not suited to recover the true model and the true relevant variables. Moreover, BIC data clustering is disappointing. In contrast, the Lasso-MLE oracle model always coincides with the true model and leads to a very good data clustering. The data-driven slope estimation (2.10) achieves better performance than BIC and AIC.

2.4.3 Variable role modelling

SRUW modelling

In this section, we focus on variable selection procedures in model-based clustering which are based on variable role modelling without variable transformation. After a series of papers (Law *et al.* [2004]; Tadesse *et al.* [2005]; Raftery and Dean [2006]; Maugis *et al.* [2009a]), Maugis *et al.* [2009c] propose a general model for selecting variables for clustering with Gaussian mixtures. This model, called SRUW, distinguishes between relevant variables ($\boldsymbol{S}$) and irrelevant variables ($\boldsymbol{S}^c$) for clustering. In addition, the irrelevant variables are divided into two categories. A part of the irrelevant variables ($\boldsymbol{U}$) may be dependent on a subset $\boldsymbol{R}$ of the relevant variables and another part ($\boldsymbol{W}$) are independent of other variables. Thus the data density is assumed to be decomposed into three parts as follows:

$$f(\mathbf{x}_i|\mathbf{m};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i^{\boldsymbol{S}};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) \times \phi(\mathbf{x}_i^{\boldsymbol{U}};\mathbf{a}+\mathbf{x}_i^{\boldsymbol{R}}\mathbf{b},\boldsymbol{\Omega}) \times \phi(\mathbf{x}_i^{\boldsymbol{W}};\boldsymbol{\gamma},\boldsymbol{\Gamma})$$

where $\mathbf{x}_i^{\boldsymbol{S}}$ designates the restriction of $\mathbf{x}_i$ in the set of variables $\boldsymbol{S}$ (similarly for $\boldsymbol{U}$, $\boldsymbol{R}$ and $\boldsymbol{W}$), $\boldsymbol{\theta} = \left((\pi_k,\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)_{k=1}^K,\mathbf{a},\mathbf{b},\boldsymbol{\Omega},\boldsymbol{\gamma},\boldsymbol{\Gamma}\right)$ is the full parameter vector (with straightforward dimensions for each of its components) and $\mathbf{m} = (K,\mathbf{m}_{\boldsymbol{\Sigma}},\mathbf{m}_{\boldsymbol{\Omega}},\mathbf{m}_{\boldsymbol{\Gamma}},\boldsymbol{S},\boldsymbol{R},\boldsymbol{U},\boldsymbol{W})$ is the full model index with $\mathbf{m}_{\boldsymbol{\Sigma}}$, $\mathbf{m}_{\boldsymbol{\Omega}}$ and $\mathbf{m}_{\Gamma}$ denoting the form of the relevant covariance matrices $(\boldsymbol{\Sigma}_k)_{k=1}^K$, the form of the regression variance matrix $\boldsymbol{\Omega}$ and the form of the covariance matrix $\boldsymbol{\Gamma}$ of the independent variables $\boldsymbol{W}$ respectively. It can be any structure defined by Celeux and Govaert [1995] for $\mathbf{m}_{\boldsymbol{\Sigma}}$, a spherical, diagonal or general structure for $\mathbf{m}_{\boldsymbol{\Omega}}$ and a spherical or diagonal structure for $\mathbf{m}_{\boldsymbol{\Gamma}}$.

The SRUW model generalizes several previous model selection methods. The procedure of Law *et al.* [2004], where irrelevant variables are assumed to be independent of all the relevant variables, corresponds to $\boldsymbol{W} = \boldsymbol{S}^c$, $\boldsymbol{R} = \emptyset$, $\boldsymbol{U} = \emptyset$. The variable selection procedure of Raftery and Dean [2006], available in the R package CLUSTVARSEL[6], assumes that the irrelevant variables are regressed on the whole relevant variable set ($\boldsymbol{W} = \emptyset$, $\boldsymbol{U} = \boldsymbol{S}^c$ and $\boldsymbol{R} = \boldsymbol{S}$). The generalization of Maugis *et al.* [2009a] enriches this model by allowing the irrelevant variables to be explained by only a subset of the relevant variables

[6] `https://cran.r-project.org/web/packages/clustvarsel/index.html`

$\boldsymbol{R} \subset \boldsymbol{S}$ ($\boldsymbol{W} = \emptyset, \boldsymbol{U} = \boldsymbol{S}^c$); this method is implemented in the SELVARCLUST software[7].

The SRUW method recasts the variable selection problem for model-based clustering as a model selection problem. It is solved maximizing the following BIC-type criterion:

$$\text{crit}_{\mathbf{m}} = \text{BIC}^{\text{clust}}_{(K,\mathbf{m}_{\boldsymbol{\Sigma}})}(\mathbf{x}^{\boldsymbol{S}}) + \text{BIC}^{\text{reg}}_{\mathbf{m}_{\boldsymbol{\Omega}}}(\mathbf{x}^{\boldsymbol{U}} \mid \mathbf{x}^{\boldsymbol{R}}) + \text{BIC}^{\text{indep}}_{\mathbf{m}_{\boldsymbol{\Gamma}}}(\mathbf{x}^{\boldsymbol{W}}), \tag{2.11}$$

where $\text{BIC}^{\text{clust}}_{(\cdot,\cdot,\boldsymbol{S})}$ represents the BIC criterion of the Gaussian mixture model with the variables $\boldsymbol{S}$, $\text{BIC}^{\text{reg}}_{(\cdot,\boldsymbol{U},\boldsymbol{R})}$ represents the BIC criterion of the regression model of the variables $\boldsymbol{U}$ on the variables $\boldsymbol{R}$ and $\text{BIC}^{\text{indep}}_{(\cdot,\boldsymbol{W})}$ represents the BIC criterion of the Gaussian model with the variables $\boldsymbol{W}$.

Since the SRUW model collection is large, two embedded backward or forward stepwise algorithms for variable selection, one for the clustering and one for the linear regression, are considered to solve this model selection problem. A backward algorithm allows one to start with all variables in order to take variable interactions into account. A forward procedure, starting with an empty clustering variable set or a small variable subset, could be preferred for numerical reasons if there are numerous variables. The method is implemented in the SELVARCLUSTINDEP software.[8] The two embedded stepwise variable selection algorithms are used to identify the SRUW sets. It leads to compare two models at each step in order to determine which variable should be excluded or included in the set $\boldsymbol{S}$, $\boldsymbol{R}$, $\boldsymbol{U}$ or $\boldsymbol{W}$. But in a high-dimensional setting, even the variable selection method with the two forward stepwise algorithms becomes painfully slow and alternative methods are desirable.

SelvarMix procedure

In order to avoid the highly CPU-time consuming of stepwise algorithms of SELVARCLUSTINDEP, an alternative variable selection procedure in two steps is proposed by Sedki *et al.* [2014]. This variable selection procedure is implemented in the R package SELVARMIX[9].

In the first step, the variables are ranked through the Lasso-like procedure of Zhou *et al.* [2009] (see Section 2.4.2). For any $K \in \mathbb{N}^\star$ and two non-negative regularization parameters λ and ρ on two grids of values $\mathcal{G}_\lambda$ and $\mathcal{G}_\rho$, the criterion defined in Equation (2.8) is maximized. The estimated mixture parameters $\widehat{\boldsymbol{\theta}}_K(\lambda,\rho) = ((\hat{\boldsymbol{\pi}}_k(\lambda,\rho)), (\hat{\boldsymbol{\mu}}_k(\lambda,\rho)), (\hat{\boldsymbol{\Sigma}}_k(\lambda,\rho)))_{k=1}^K$ are computed with the EM algorithm of Zhou *et al.* [2009]. It is worth noting that this Lasso-like criterion does not take into account the typology of the variables induced by the SRUW model. Strictly speaking, it only distinguishes two possible roles for the variables: a variable is declared related or independent of the clustering.

[7] SELVARCLUST is available at `http://www.math.univ-toulouse.fr/~maugis/`

[8] SELVARCLUSTINDEP is available at `http://www.math.univ-toulouse.fr/~maugis/`

[9] `https://cran.r-project.org/web/packages/SelvarMix/index.html`

Varying the regularization parameters (λ, ρ) in $\mathcal{G}_\lambda \times \mathcal{G}_\rho$, a score is defined for each variable $j \in \{1, \ldots, d\}$ and for fixed K:

$$\mathcal{O}_K(j) = \sum_{(\lambda,\rho)\in\mathcal{G}_\lambda\times\mathcal{G}_\rho} \left(1 - \mathbb{1}_{\widehat{\mu}_{1j}(\lambda,\rho)=\ldots=\widehat{\mu}_{Kj}(\lambda,\rho)=0}\right).$$

The larger $\mathcal{O}_K(j)$, the more related for the clustering the variable j is expected to be. The variables are thus ranked by their decreasing values on $\mathcal{O}_K(j)$, this variable ranking being noted $\mathcal{I}_K = (j_1, \ldots, j_d)$.

Conditional to a model $(K, \mathbf{m_\Sigma}, \mathbf{m_\Omega}, \mathbf{m_\Gamma})$ composed by the number of groups and all the structures of covariance matrices, the relevant clustering variable set $\boldsymbol{S}$ is first determined. The variable set is scanned according to the $\mathcal{I}_K$ order. One variable is added to $\boldsymbol{S}$ if

$$\begin{aligned}\mathrm{BIC}^{\mathrm{diff}}(j_v) &= \mathrm{BIC}^{\mathrm{clust}}_{(K,\mathbf{m_\Sigma})}\left(\mathbf{x}^{\boldsymbol{S}}, \mathbf{x}^{j_v}\right) \\ &\quad -\mathrm{BIC}^{\mathrm{clust}}_{(K,\mathbf{m_\Sigma})}\left(\mathbf{x}^{\boldsymbol{S}}\right) - \mathrm{BIC}^{\mathrm{reg}}_{\mathbf{m_\Omega}}\left(\mathbf{x}^{j_v} \mid \mathbf{x}^{\boldsymbol{R}[j_v]}\right)\end{aligned}$$

is positive, $\boldsymbol{R}[j_v]$ being the variables of $\boldsymbol{S}$ required to linearly explain $\mathbf{x}^{j_v}$. The scanning of $\mathcal{I}_K$ is stopped as soon as c successive variables have a non positive $\mathrm{BIC}^{\mathrm{diff}}$ value, c being a fixed positive integer. Next the independent variable set $\boldsymbol{W}$ is determined as follows: Scanning the variable set according to the reverse order of $\mathcal{I}_K$, a variable j_v is added to $\boldsymbol{W}$ if the subset $\boldsymbol{R}[j_v]$ of $\boldsymbol{S}$ (derived from the backward stepwise algorithm) is empty. The algorithm stops as soon as c successive variables are not declared independent. The redundant variables are thus declared to be $\boldsymbol{U} = \{1, \ldots, d\} \setminus \{\boldsymbol{S} \cup \boldsymbol{W}\}$ and the subset $\boldsymbol{R}$ of $\boldsymbol{S}$ required to linearly explain $\mathbf{x}^{\boldsymbol{U}}$ is derived from the backward stepwise algorithm. Finally, the model $(K, \mathbf{m_\Sigma}, \mathbf{m_\Omega}, \mathbf{m_\Gamma})$ maximizing the criterion (2.11) is selected.

Variable selection without multiple parameter estimation

Altough some strategies design such reduced deterministic paths for limiting the number of model evaluations, this number remains too high for fast model selection. Indeed, each model comparison requires to estimate model parameters which are needed for any model selection criterion like BIC. Marbac and Sedki [2015] propose an original strategy avoiding parameter estimation for all models which compete, thus limiting the computing time. Then a parameter estimation is just performed for the retained model at the end of their process. Their strategy is applied in the diagonal Gaussian mixture but could be easily extended to the multinomial or the mixed situations also.

In their context, a variable is said to be *irrelevant* for the clustering task if its one-dimensional marginal distributions are equal between components. In the Gaussian diagonal situation for instance, and noting $\boldsymbol{\Sigma}_k = \mathrm{diag}(\sigma^2_{k1}, \ldots, \sigma^2_{kd})$, a variable j is thus *irrelevant* if

$$\mu_{1j} = \ldots = \mu_{Kj} \quad \text{and} \quad \sigma^2_{1j} = \ldots = \sigma^2_{Kj}.$$

By introducing a variable w_j such that $w_j = 0$ if variable j is irrelevant for the clustering and $w_j = 1$ otherwise, selecting $\mathbf{w} = (w_1, \ldots, w_d)$ is thus strictly equivalent to selecting a given probabilistic model. Then any model selection criterion, like BIC, could be used for selecting the pair $\mathbf{m} = (K, \mathbf{w})$.

Their strategy relies on a variant of the ICL criterion of Biernacki *et al.* [2000]. The ICL criterion (see Section **??** in Chapter **??**) is defined by $\text{ICL}_\mathbf{m} = \ln f(\mathbf{x}, \hat{\mathbf{z}}_\mathbf{m}|\mathbf{m})$, where $\hat{\mathbf{z}}_\mathbf{m}$ is the MAP of the MLE of $\boldsymbol{\theta}$ with the model $\mathcal{S}_\mathbf{m}$. The proposed variant is the so-called MICL criterion (*Maximum Integrated Complete-data Likelihood*) defined by

$$\text{MICL}_\mathbf{m} = \ln f(\mathbf{x}, \mathbf{z}^*_\mathbf{m}|\mathbf{m}) \quad \text{with} \quad \mathbf{z}^*_\mathbf{m} = \arg\max_{\mathbf{z}\in\mathcal{Z}} \ln f(\mathbf{x}, \mathbf{z}|\mathbf{m}).$$

Then, the model $\mathcal{S}_{\mathbf{m}^*}$ maximizing $\text{MICL}_\mathbf{m}$ is retained:

$$\mathbf{m}^* = \arg\max_{\mathbf{m}\in\mathcal{M}} \text{MICL}_\mathbf{m}.$$

Marbac and Sedki [2015] prove that MICL, like ICL, is consistent for choosing $\mathbf{w}$ when the number K of components is known. Nevertheless, like ICL (see again Section **??** in Chapter **??**), MICL is consistent for choosing K only when clusters do not too much overlap. In addition, closed-form expression of MICL is available when there exists conjuguate priors, what is the case for Gaussian and multinomial mixtures. For instance, see Equation (**??**) of Chapter **??** for the exact expression of ICL in the multinomial case.

The question of maximizing MICL on $\mathbf{w}$ is obviously the crucial difficulty. Marbac and Sedki [2015] implement the following simple alternate procedure, for a fixed K value (thus this algorithm has to be run for different candidate values of K). Starting from a value $\mathbf{w}^{(0)}$ (thus $\mathcal{S}_{\mathbf{m}^{(0)}}$) uniformly sampled in the corresponding space and then a value $\mathbf{z}^{(0)}$ being deduced from the MAP rule of the associated MLE, an iteration of the algorithm is composed by the following two steps ($q \leq 0$):

Partition step Fix $\mathbf{z}^{(q+1)}$ such that

$$\ln f(\mathbf{x}, \mathbf{z}^{(q+1)}|\mathbf{m}^{(q)}) \geq \ln f(\mathbf{x}, \mathbf{z}^{(q)}|\mathbf{m}^{(q)}).$$

Model step Fix $\mathbf{m}^{(q+1)} = \arg\max_{\mathbf{m}\in\mathcal{M}} \ln f(\mathbf{x}, \mathbf{z}^{(q+1)}|\mathbf{m})$ such that $\mathbf{m}^{(q+1)} = (K, \mathbf{w}^{(q+1)})$ with $(j = 1, \ldots, d)$

$$w_j^{(q+1)} = \arg\max_{w_j\in\{0,1\}} \ln f(x_1^j, \ldots, x_n^j|K, w_j, \mathbf{z}^{(q+1)}).$$

This procedure can be trapped in local maxima and thus several run are required. In addition, it can be time consuming when the sample size increases, due to the so-called "model step". However, it is a very promising first attempt for dealing with model multiplicity in variable selection, without systematic

parameter estimation which corresponds in fact, for most current approaches, to a huge time consuming waste. The algorithm of these authors is available as an R package called VARSELLCM[10].

2.4.4 Co-clustering

Definition and historical utility

Simultaneous clustering of rows and columns, usually designated by bi-clustering, co-clustering or block clustering, is an important technique in two way data analysis. They consider the two sets simultaneously and organize the data into homogeneous blocks. Two partition representations are thus now needed. First, as usual, a partition of n individuals (lines of the data matrix $\mathbf{x}$) into K clusters still noticed $\mathbf{z} = (z_{11}, \ldots, z_{nK})$ with $z_{ik} = 1$ if i belongs to cluster k and $z_{ik} = 0$ otherwise (we note as well $z_i = k$ if $z_{ik} = 1$). Second, and symmetrically, a partition of d variables (columns of the data matrix $\mathbf{x}$) into L clusters is denoted by $\mathbf{w} = (w_{11}, \ldots, w_{dL})$ with $w_{jl} = 1$ if j belongs to cluster l and $w_{jl} = 0$ otherwise (we note as well $w_j = l$ if $w_{jl} = 1$). Both space partitions are respectively denoted by $\mathcal{Z}$ and $\mathcal{W}$. Figure 2.11 gives an illustration of this purpose.

In recent years, co-clustering have found numerous applications in the fields ranging from data mining, information retrieval, biology, computer vision and so forth. Dhillon [2001] publishes an article on text data mining by simultaneously clustering the documents and content (words) using bipartite spectral graph partitioning. This is a quite useful technique for instance to manage huge corpus of unlabeled documents. Xu *et al.* [2010] present another co-clustering application (again using bipartite spectral graph) to understand subset aggregates of web users by simultaneously clustering the users (sessions) and the page view information. Giannakidou *et al.* [2008] employ a similarity metric based co-clustering technique for social tagging system. In field of bio-informatics, co-clustering is mainly used to find structures in gene expression data. This is useful for instance to find sets of genes which correspond to a particular kind of disease. Some of the pioneer material in this context can be found in Kluger *et al.* [2003]. Recently many model-based co-clustering algorithms have also been developed to target computer vision applications. For instance, Qiu [2004] demonstrates the utility of co-clustering in image grouping by simultaneously clustering images with their low-level visual features. Guan *et al.* [2005] extend this work and present opportunity to develop a novel content based image retrieval system. Similarly, Rasiwasia and Vasconcelos [2009] use co-clustering to model scenes.

[10] `https://cran.r-project.org/web/packages/VarSelLCM/index.html`

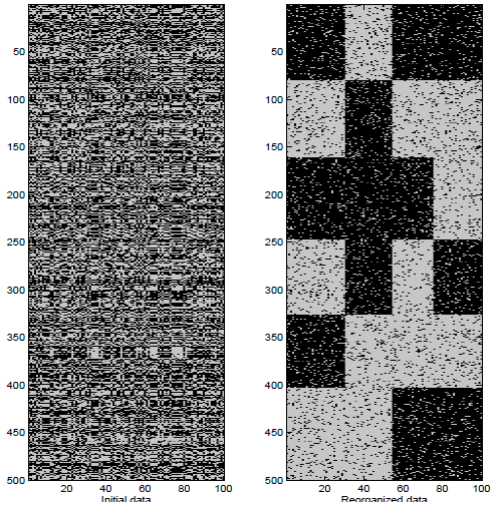


Figure 2.11: Co-clustering principle illustrated on a binary data set: On the left, the initial data set ($n = 500$ and $d = 100$); On the right, the reorganized data set with a simultaneous partitioning of rows and columns ($K = 6$ and $L = 4$).

Probabilistic formulation and use in HD clustering

We refer to the book of Govaert and Nadif [2013] for providing more details on co-clustering techniques, probabilistic or not. Here, we focus on model-based co-clustering as being often a generalization of non-probabilistic methods and allowing coherent formulation from estimation to model selection. In the following set, sum or product on i, j, k and l stands for ranges $\{1,\ldots,n\}$, $\{1,\ldots,d\}$, $\{1,\ldots,K\}$ and $\{1,\ldots,L\}$ respectively.

Block model-based clustering can be seen as an extension of the traditional mixture model-based clustering (see Chapter **??**). The basic idea is to extend the latent class principle of local (or conditional) independence. Each data point x_i^j is assumed to be independent once z_i and w_j are fixed:

$$f(\mathbf{x}|\mathbf{z},\mathbf{w};\boldsymbol{\theta}) = \prod_{i,j} f(x_i^j;\boldsymbol{\alpha}_{z_i w_j}).$$

We have noted $\boldsymbol{\theta} = (\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\alpha})$, where $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_{kl})$, $\boldsymbol{\pi} = (\pi_k)$ and $\boldsymbol{\rho} = (\rho_k)$ are the vectors of probabilities π_k and ρ_l that a row and a column belong to the kth row component and to the lth column component respectively. Assuming also independence between all z_i and w_j, the latent block mixture model has final pdf

$$f(\mathbf{x};\boldsymbol{\theta}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,j} \pi_{z_i}\rho_{w_j} f(x_i^j;\boldsymbol{\alpha}_{z_i w_j}). \tag{2.12}$$

The pdf $f(\cdot;\boldsymbol{\alpha}_{z_i w_j})$ depends on the kind of data x_i^j:

- In the binary case ($x_i^{jh} \in \{0,1\}^2$, with $\sum_{h=1}^2 x_i^{jh} = 1$), $f(\cdot;\boldsymbol{\alpha}_{kl})$ corresponds to the Bernoulli distribution $\mathsf{B}(\boldsymbol{\alpha}_{kl})$ of parameter $\boldsymbol{\alpha}_{kl} = p(X_i^j = 1)$ (see Govaert and Nadif [2008]).
- In the categorical case with m levels ($x_i^{jh} \in \{0,1\}^m$, with $\sum_{h=1}^m x_i^{jh} = 1$), $f(\cdot;\boldsymbol{\alpha}_{kl})$ corresponds to the multinomial distribution $\mathsf{M}(\boldsymbol{\alpha}_{kl})$ of parameter $\boldsymbol{\alpha}_{kl} = (\alpha_{kl}^1,\ldots,\alpha_{kl}^m)$ with $\alpha_{kl}^h = p(X_i^j = h)$ for $h = 1,\ldots,m$ (see Keribin *et al.* [2015]).
- In the contingency table case ($x_i^j \in \mathbb{N}$), $f(\cdot;\boldsymbol{\alpha}_{kl})$ corresponds to the Poisson distribution $\mathsf{P}(\mu_k\nu_l\gamma_{kl})$ of parameter $\boldsymbol{\alpha}_{kl} = (\mu_k,\nu_l,\gamma_{kl})$. The Poisson parameter is here split into μ_k and ν_l the effects of the row k and the column l respectively and γ_{kl} the effect of the block kl (see Govaert and Nadif [2010]). Unfortunately, this parameterization is not identifiable. It is therefore not possible to estimate simultaneously μ_k, ν_l and γ_{kl} without imposing further constraints. Constraints $\sum_k \pi_k\gamma_{kl} = \sum_l \rho_l\gamma_{kl} = 1$ and $\sum_k \mu_k = 1, \sum_l \nu_l = 1$ are a possibility.
- In the continous case ($x_i^j \in \mathbb{R}$), $f(\cdot;\boldsymbol{\alpha}_{kl})$ corresponds to the Gaussian distribution $\mathsf{N}(\mu_{kl},\sigma_{kl}^2)$ of parameter $\boldsymbol{\alpha}_{kl} = (\mu_{kl},\sigma_{kl}^2)$, denoting respectively the mean and the variance (see Govaert and Nadif [2013]).

Such models can be very parsimonious[11] even in the HD setting provided that L is quite low, as it is shown in Table 2.4. The number of parameters of this table has to be compared to this one of Tables 2.1 and 2.2. Consequently, these block clustering models could be good candidates for performing HD clustering even if they are not exactly designed for this aim initially. In such a case, clustering of columns can just be seen as an instrumental strategy for obtaining HD parsimonious models. Indeed, the HD clustering purpose only concerns n and not d in our case. However, column clustering has advantage to provide an easy readability of the model to the practitioner.

Model	Number of parameters
Binary	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL$
Categorical	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL(m-1)$
Contingency	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + KL$
Continuous	$\dim(\boldsymbol{\pi}) + \dim(\boldsymbol{\rho}) + 2KL$

Table 2.4: Number of parameters of the block clustering models. We have $\dim(\boldsymbol{\pi}) = K-1$ in the case of free proportions in lines and $\dim(\boldsymbol{\pi}) = 0$ in the case of equal proportions. Symmetrically, we have $\dim(\boldsymbol{\rho}) = L-1$ in the case of free proportions in columns and $\dim(\boldsymbol{\rho}) = 0$ in the case of equal proportions.

Parameter estimation

EM-based algorithms are the standard approach to estimate model parameters by maximizing the observed log-likelihood. Here, the complete data is represented as a vector $(\mathbf{x}, \mathbf{z}, \mathbf{w})$ where unobservable vectors $\mathbf{z}$ and $\mathbf{w}$ are the labels. The *complete* log-likelihood can then be written

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}, \mathbf{w}) = \sum_k (\sum_i z_{ik}) \log \pi_k + \sum_l (\sum_j w_{jl}) \log \rho_l + \sum_{i,j,k,l} z_{ik} w_{jl} \log f(x_i^j; \boldsymbol{\alpha}_{kl}).$$

Then, from Section ?? of Chapter ??, the expected complete log-likelihood $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)})$ involved at the qth iteration of the EM algorithm is expressed by

$$\begin{aligned} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(q)}) &= \sum_{i,k} p(Z_i = k | \mathbf{x}; \boldsymbol{\theta}^{(q)}) \ln \pi_k + \sum_{j,l} p(W_i = l | \mathbf{x}; \boldsymbol{\theta}^{(q)}) \ln \rho_l \\ &\quad + \sum_{i,j,k,l} p(Z_i = k, W_j = l | \mathbf{x}; \boldsymbol{\theta}^{(q)}) \ln f(x_i^j; \boldsymbol{\alpha}_{kl}). \end{aligned} \tag{2.13}$$

Unfortunately, difficulties arise owing to the dependence structure in the model, and more precisely in the combinatorial difficulty for evaluating the terms $p(Z_i = k, W_j = l | \mathbf{x}; \boldsymbol{\theta}^{(q)})$. Several solutions exist for skirting this difficulty (see Govaert and Nadif [2013] for more details), including:

[11] Some more parsimonious versions are also defined (see references).

- The so-called *variational approach* which constraints the problematic joint probability to satisfy the relation

$$f(\mathbf{z}, \mathbf{w}|\mathbf{x}; \boldsymbol{\theta}) \approx f(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}) f(\mathbf{w}|\mathbf{x}; \boldsymbol{\theta}).$$

- To replace the E-step by a S-step, so using a SEM algorithm instead of EM (see details on SEM in Section ?? of Chapter ??). In the S-step, random couples $(\mathbf{z}, \mathbf{w})$ (conditionnally to $\mathbf{x}$) are drawn sequentially by the following two-step Gibbs algorithm (see more details in Keribin *et al.* [2015])

$$\mathbf{Z}|\mathbf{x}, \mathbf{w}; \boldsymbol{\theta} \quad \text{and} \quad \boldsymbol{W}|\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}.$$

Several estimation algorithms are implemented in the R package BLOCKCLUSTER[12].

For finishing this estimation description, it is important to note two important features. Firstly, many local maxima of the likelihood may exist in the block clustering model, more than in the standard mixture context, probably owing to the latent data multiplicity. In practice, many runs should then be launched to avoid traps in local maxima. Secondly, computing the (observed) log-likelihood value $\ell(\boldsymbol{\theta}; \mathbf{x})$ itself is difficult for the same combinatorial reasons that previously. Such an unavailability can have important consequences on model selection also.

Model selection

Models in competition are indexed by the number of clusters in line and column, thus $\mathcal{S} = (K, L)$. It is crucial to notice that model selection in block clustering has to be performed with caution since some traditional criteria cannot be used straightforwardly. In particular, it is hazardous to use asymptotic criteria like BIC since asymptotic is now double with both quantities n and d. In addition, using non asymptotic evaluation of the integrated likelihood $f(\mathbf{x})$ has to be given up because of the combinatorial difficulty involved by the latent variables $\mathbf{z}$ and $\mathbf{w}$.

Avoiding both asymptotic problems and combinatorial difficulties is possible by using exact expression of the ICL criterion (Biernacki *et al.* [2000], Biernacki *et al.* [2011]). In the block clustering context, ICL is written

$$\text{ICL}_{\mathbf{m}} = \ln f(\mathbf{x}, \hat{\mathbf{z}}_{\mathbf{m}}, \hat{\mathbf{w}}_{\mathbf{m}}) = \ln f(\mathbf{x}|\hat{\mathbf{z}}_{\mathbf{m}}, \hat{\mathbf{w}}_{\mathbf{m}}) + \ln f(\hat{\mathbf{z}}_{\mathbf{m}}) + \ln f(\hat{\mathbf{w}}_{\mathbf{m}}),$$

$\hat{\mathbf{z}}_{\mathbf{m}}$ and $\hat{\mathbf{w}}_{\mathbf{m}}$ being the MAP estimate of $\mathbf{z}$ and $\mathbf{w}$ respectively obtained from the MLE $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$. Lomet *et al.* [2012] provide the corresponding closed-form expression of ICL for the Gaussian situation and Keribin *et al.* [2015] similarly for the Bernoulli/multinomial case. We refer the reader to these references for detailed discussion about the Bayesian hyperparameter choice.

[12] `http://cran.r-project.org/web/packages/blockcluster/index.html`

In addition, in this multinomial setting with m levels, Keribin *et al.* [2015] use their non-asymptotic expression to derive the new following asymptotic one, called ICLbic:

$$\text{ICLbic}_{\mathbf{m}} = \ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \mathbf{x}, \hat{\mathbf{z}}_{\mathbf{m}}, \hat{\mathbf{w}}_{\mathbf{m}}) - \frac{K-1}{2}\ln(n) - \frac{L-1}{2}\ln(d) - \frac{KL(m-1)}{2}\ln(nd).$$

It is interesting to notice that, in comparison to the ICLbic formula in the simple mixture context (see Equation (**??**) in Chapter **??**), now both the row number n and the column number d are involved in the penalty. Using then the straightforward link $\text{ICL}_{\mathbf{m}} = \ln f(\hat{\mathbf{z}}_{\mathbf{m}}, \hat{\mathbf{w}}_{\mathbf{m}}|\mathbf{x}; \hat{\boldsymbol{\theta}}_{\mathbf{m}}) + \text{BIC}_{\mathbf{m}}$ between ICLbic and ICL, they propose the following block clustering specific asymptotic version of BIC

$$\text{BIC}_{\mathbf{m}} = \ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \mathbf{x}) - \frac{K-1}{2}\ln(n) - \frac{L-1}{2}\ln(d) - \frac{KL(m-1)}{2}\ln(nd).$$

Again, it is interesting to observe the way that both n and d are present in the penalty. Neverthess, the BIC calculus remains unattainable since it relies on the unvailable value of the log-likelihood $\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}}; \mathbf{x})$.

Finally, Keribin *et al.* [2015] make the conjecture, corroborated with experiments, that BIC and ICL are asymptotically equivalent and thus have the same asymptotic behaviour. As a consequence, the ICL criterion is expected to be consistent for selecting both K and L in block clustering, for any true parameter setting. It is totally different from row clustering where consistency is only true for sufficiently separated clusters (see Baudry [2012] and also Section **??** in Chapter **??**). Such a remark is crucial because it is linked to the blessing of HD clustering we have discussed in length earlier in Section 2.2.2.

Return on the blessing in HD clustering

We illustrate now, in the binary block clustering setting, that HD situations are a whole blessing for row clustering. Denoting by $p(X_i^{j1} = 1|Z_i = k) = \tau_k = \sum_{l=1}^{L} \alpha_{kl}\rho_l$, then the marginal distribution of X_i^j on j is the following mixture of binomial distributions $\mathsf{B}(\cdot,\cdot)$

$$\left\{\sum_j X_i^{j1}\right\} | Z_i = k \sim \mathsf{B}(d, \tau_k).$$

In that case, Brault [2014] provides the following control of partition error $\mathbf{z}$ of this mixture, $\mathbf{z}^*$ denoting the true row partition:

$$p(\hat{\mathbf{z}} \neq \mathbf{z}^*) \leq 2n \exp\left\{-\frac{1}{8}d\left[\min_{k\neq k'}|\tau_k - \tau_{k'}|\right]\right\} + K(1 - \min_k \pi_k)^n.$$

It implies the important fact that row clustering is consistent in high-dimension provided some asymptotic constraints between n and d, for instance that

$$\ln(n) = o(d).$$

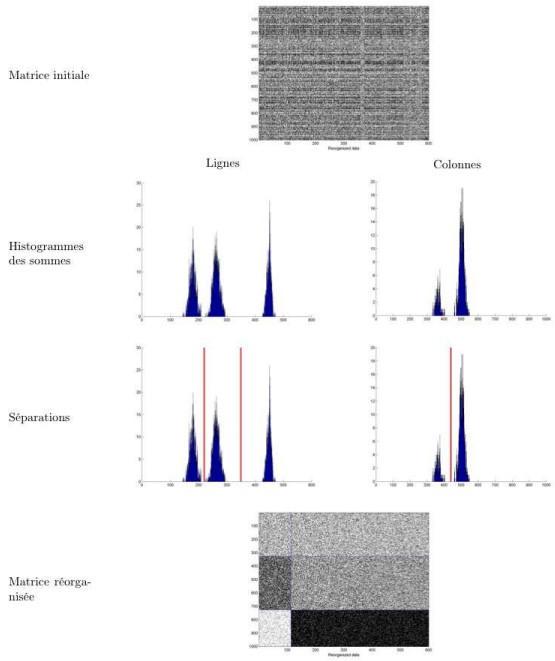


Figure 2.12: Illustration of the low row cluster overlap in the binary HD setting: The initial data matrix is at the top; Histogram of the sum of columns is displayed at the second line (first column); The third line underlines that three row clusters are clearly present (first column); The reorganized matrix (in row and columns) is available at the last line of the figure. Symmetrical comments could be made on column cluster overlap (second column on the figure). This figure has been provided by Brault [2014].

Figure 2.12 illustrates this low row cluster overlap in a HD setting. Note that the same comment could be made on column cluster overlap when n increases, even if it is not the first topic of this chaper focused on row HD clustering.

In the same spirit, Mariadassou and Matias [2013] show the following more general result in the binary case, on the consistency of the couple $(\hat{\mathbf{z}}, \hat{\mathbf{w}})$:

$$\hat{\boldsymbol{\theta}} \overset{n,d\to\infty}{\longrightarrow} \boldsymbol{\theta}^* \quad \Rightarrow \quad p(\hat{\mathbf{z}} = \mathbf{z}^*, \hat{\mathbf{w}} = \mathbf{w}^* | \mathbf{x}; \hat{\boldsymbol{\theta}}) \overset{n,d\to\infty}{\longrightarrow} 1,$$

where $\boldsymbol{\theta}^*$ and $\mathbf{w}^*$ respectively design the true $\boldsymbol{\theta}$ and $\mathbf{w}$.

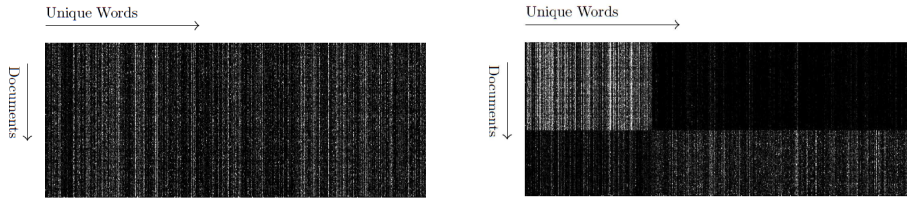
Contingency table illustration: document clustering

We retrieve the text mining example introduced in Section 2.3.2. Since it concerns a contingency table (cross counting documents and words) we apply

	Medline	Cranfield
Medline	1033	·
Cranfield	·	1398

Figure 2.13: Confusion table by applying block clustering for text partitioning.

a Poisson block clustering model. The "true" block partitioning involves $K = 2$ document clusters (row) and $L = 2$ word clusters (column). Table 2.13 displays the confusion table for documents by using 2×2 blocks. We show that we exactly retrieve the underlying document structure, what is expected by the blessing effect of HD clustering, the data set being here with $d = 9275$. Figure 2.14 gives a view of the data set before and after reorganization by block-clustering. We also distinguish clear partitioning in rows and columns.



(a) (b)

Figure 2.14: Text mining example: (a) the initial data set; (b) the reorganized data set with $(K, L) = (2, 2)$.

2.4.5 Intermediate conclusion

Designing parsimonious models in the canonical space for HD data has the expected advantage of being more meaningful for the practitioner than non-canonical ones. In this context, several specific contributions exist, that could be split into variable selection-like and variable clustering-like approaches. Beyond their apparent difference, they share the common property to recast a particular, but simple, role for the variables in a generative and very parsimonious way when the dimension of the feature space increases. However, although only generative approaches are involved, it is not always straightforward to use classical model selection criteria. Indeed, some questions about either their asymptotic validity, their explicit calculus (the likelihood is not always calculable) or their use in case of a huge number of competing models is posed. Nevertheless, recent advances in this active field of research suggest possibility to progressively overcome these scientific locks. Beyond these model

selection questionings, the important task consisting of designing specific generative models for HD *mixed* features should be also undertaken since it is currently poorly developed albeit more and more present in nowadays data sets.

2.5 Future methodological challenges

Two strong trends are highly expected to hold in a near future, that should be addressed by specific researches. Firstly, data sets will be described by a constantly increasing number of features, these features being possibly themselves of very different kinds. For instance, (high-dimensional) multivariate categorical functional data could be mixed with (high-dimensional) multivariate counting data, *etc.* Secondly, the number of model candidates for dealing with these kinds of data sets will constantly increase, leading to a unmanageable number of models estimation in practice. Such a situation will address the question to design some specific strategies in model selection, for avoiding ineffective and unnecessary estimation of a too large number of models.

Bibliography

Aitchinson, J. and Aitken, C. G. G. [1976]. Multivariate Binary Discrimination by the Kernel Method. *Biometrika*, **63**, 413–420.

Baudry, J.-P. [2012]. Estimation and Model Selection for Model-Based Clustering with the Conditional Classification Likelihood. URL `http://hal.upmc.fr/hal-00699578`.

Baudry, J.-P., Maugis, C. and Michel, B. [2012]. Slope Heuristics: overview and implementation. *Statistics and Computing*, **22**(2), 455–470.

Bellman, R. [1961]. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ.

Biernacki, C., Celeux, G. and Govaert, G. [2000]. Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725.

Biernacki, C., Celeux, G. and Govaert, G. [2011]. Exact and Monte Carlo Calculations of Integrated Likelihoods for the Latent Class Model. *Journal of Statistical Planning and Inference*, **140**(11), 2991.

Birgé, L. and Massart, P. [2007]. Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, **138**(1-2), 33–73.

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

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

Brault, V. [2014]. *Estimation et sélection de modèle pour le modèle des blocs latents*. Thèse de doctorat, Université Paris Sud. URL `http://www.math.u-psud.fr/~brault/Article/These.pdf`.

Brusco, M. J. and Cradit, J. D. [2001]. A variable selection heuristic for k-means clustering. *Psychometrika*, **66**(2), 249–270.

Cattell, R. [1966]. The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, **1**(2), 245–276.

Celeux, G. and Govaert, G. [1991]. Clustering Criteria for Discrete Data and Latent Class Models. *Journal of Classification*, **8**, 157–176.

Celeux, G. and Govaert, G. [1995]. Gaussian Parsimonious Models. *Pattern Recognition*, **28**(5), 781–793.

Dash, M., Choi, K., Scheuermann, P. and Liu, H. [2002]. Feature Selection for Clustering - A Filter Solution. *Proceedings of the Second IEEE International Conference on Data Mining*, 115–122.

Delaigle, A. and Hall, P. [2010]. Defining probability density for a distribution of random functions. *The Annals of Statistics*, **38**, 1171–1193.

Devaney, M. and Ram, A. [1997]. Efficient Feature Selection in Conceptual Clustering. *Machine Learning: Proceedings of the Fourteenth International Conference*, 92–97.

Dhillon, I. S. [2001]. Co-Clustering Documents and Words Using Bipartite Spectral Graph Partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 269–274. ACM, New York, NY, USA. ISBN 1-58113-391-X. doi:10.1145/502512.502550. URL `http://doi.acm.org/10.1145/502512.502550`.

Ferraty, F. and Vieu, P. [2006]. *NonParametric Functional Data Analysis*. Series in Statistics. Springer, New York.

Fowlkes, E. B., Gnanadesikan, R. and Kettenring, J. R. [1988]. Variable selection in clustering. *Journal of Classification*, **5**(2), 205–228.

Friedman, J., Hastie, T. and Tibshirani, R. [2007]. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.

Friedman, J. H. and Meulman, J. J. [2004]. Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society, Series B*, **66**(4), 815–849.

Ghahramani, Z. and Hinton, G. [1997]. The EM algorithm for factor analyzers. *Technical report*, University of Toronto.

Giannakidou, E., Koutsonikola, V., Vakali, A. and Kompatsiaris, Y. [2008]. Co-Clustering Tags and Social Data Sources. In *Web-Age Information Management, 2008. WAIM'08. The Ninth International Conference on*, 317–324. IEEE.

Goodman, L. A. [1974]. Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

Govaert, G. and Nadif, M. [2008]. Block Clustering with Bernoulli Mixture Models: Comparison of Different Approaches. *Computational Statistics and Data Analysis*, **52**(6), 3233–3245.

Govaert, G. and Nadif, M. [2010]. Latent Block Model for Contingency Table. *Communications in Statistics - Theory and Methods*, **39**(3), 416–425.

Govaert, G. and Nadif, M. [2013]. *Co-Clustering*. Wiley.

Guan, J., Qiu, G. and Xue, X. [2005]. Spectral Images and Features Co-Clustering with Application to Content-Based Image Retrieval. In *Multimedia Signal Processing, 2005 IEEE 7th Workshop on*, 1–4. IEEE.

Guyon, I. and Elisseeff, A. [2003]. An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

Jacques, J. and Preda, C. [2013]. Funclust: a curves clustering method using functional random variable density approximation. *Neurocomputing*, **112**, 164–171.

Jacques, J. and Preda, C. [2014a]. Functional data clustering: a survey. *Advances in Data Analysis and Classification*, **8**(3), 231–255.

Jacques, J. and Preda, C. [2014b]. Model-based clustering of multivariate functional data. *Computational Statistics and Data Analysis*, **71**, 92–106.

James, G. and Sugar, C. [2003]. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, **98**(462), 397–408.

Jouve, P.-E. and Nicoloyannis, N. [2005]. A Filter Feature Selection Method for Clustering. *Proceedings of International Symposium on Methodologies for Intelligent Systems*, 583–593.

Keribin, C., Brault, V., Celeux, G. and Govaert, G. [2015]. Estimation and selection for the latent block model on categorical data. *Statistics and Computing*, 1201–1216. ISSN 0960-3174. doi:10.1007/s11222-014-9472-2.

Kim, S., Tadesse, M. G. and Vannucci, M. [2006]. Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, **93**(4), 877–893.

Kluger, Y., Basri, R., Chang, J. and Gerstein, M. [2003]. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome research*, **13**(4), 703–716.

Kohavi, R. and John, G. H. [1997]. Wrappers for Feature Subset Selection. *Artificial Intelligence*, **97**(1-2), 273–324.

Law, M. H., Figueiredo, M. A. T. and Jain, A. K. [2004]. Simultaneous Feature Selection and Clustering Using Mixture Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **26**(9), 1154–1166.

Lebret, R., Iovleff, S., Langrognet, F., Biernacki, C., Celeux, G. and Govaert, G. [2015]. Rmixmod: The R Package of the Model-Based Unsupervised, Supervised and Semi-Supervised Classification Mixmod Library. *Journal of Statistical Software*, **in press**.

Lévéder, C., Abraham, C., Cornillon, P. A., Matzner-Lober, E. and Molinari, N. [2004]. Discrimination de courbes de pétrissage. *Chimiometrie*, 37–43.

Lomet, A., Govaert, G. and Grandvalet, Y. [2012]. Model selection in block clustering by the integrated classification likelihood. In *20th International Conference on Computational Statistics (COMPSTAT 2012)*, 519–530. Lymassol, France. URL `https://hal.archives-ouvertes.fr/hal-00730829`.

Marbac, M. and Sedki, M. [2015]. Variable selection for model-based clustering using the integrated complete-data likelihood. *arXiv:1501.06314*.

Mariadassou, M. and Matias, C. [2013]. Convergence of the groups posterior distribution in latent or stochastic block models. *Bernoulli*, **21**(1), 537–573.

Massart, P. [2007]. *Concentration inequalities and model selection*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

Maugis, C., Celeux, G. and Martin-Magniette, M. [2009a]. Variable selection for clustering with Gaussian mixture models. *Biometrics*, **65**(3), 701–709.

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

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

Maugis, C. and Michel, B. [2011]. Data-driven penalty calibration: a case study for Gaussian mixture model selection. *ESAIM Probability and Statistics*, **15**, 320–339.

Maugis, C. and Michel, B. [2012]. A non asymptotic penalized criterion for Gaussian mixture model selection. *ESAIM: Probability and Statistics*, **15**, 41–68.

McLachlan, G. [2003]. The EMMIX-MFA software. URL `http://www.maths.uq.edu.au/~gjm/mix_soft/mfa/`.

McLachlan, G., Bean, R. and Peel, D. [2002]. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**(3), 413–422.

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

Meng, X. and van Dyke, D. [1997]. The EM algorithm – An old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B*, **59**(3), 511.

Meynet, C. [2012]. *Sélection de variables pour la classification non supervisée en grande dimension*. Thèse de doctorat, Université Paris-Sud 11. URL `https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxjYXJvbGluZW1leW5ldHxneDo2YjA2Y2NlMmJiNDQwMjE3`.

Meynet, C. and Maugis-Rabusseau, C. [2012]. A sparse variable selection procedure in model-based clustering. *Research report.* URL `http://hal.inria.fr/hal-00734316`.

Miller, A. J. [1990]. *Subset Selection in Regression.* Chapman and Hall, London.

Moustaki, I. and Papageorgiou, I. [2005]. Latent class models for mixed variables with applications in Archaeometry. *Computational Statistics and Data Analysis*, **48**(3), 65—675.

Pan, W. and Shen, X. [2007]. Penalized Model-Based Clustering with Application to Variable Selection. *Journal of Machine Learning Research*, **8**, 1145–1164.

Qiu, G. [2004]. Image and Feature Co-Clustering. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, **4**, 991–994. IEEE.

Raftery, A. E. and Dean, N. [2006]. Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, **101**(473), 168–178.

Ramsay, J. O. and Silverman, B. W. [2005]. *Functional data analysis.* Series in Statistics. Springer, New York, 2nd edition edition.

Rasiwasia, N. and Vasconcelos, N. [2009]. Holistic Context Modeling Using Semantic Co-Occurrences. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 1889–1895. IEEE.

Samé, A., Chamroukhi, F., Govaert, G. and P., A. [2011]. Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification*, **5**(4), 301–321.

Sedki, M., Celeux, G. and Maugis-Rabusseau, C. [2014]. SelvarMix: A R package for variable selection in model-based clustering and discriminant analysis with a regularization approach. *Research report.* URL `https://hal.inria.fr/hal-01053784`.

Silverman, B. W. [1986]. *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Tadesse, M. G., Sha, N. and Vannucci, M. [2005]. Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, **100**(470), 602–617.

Xu, G., Zong, Y., Dolog, P. and Zhang, Y. [2010]. Co-Clustering Analysis of Weblogs Using Bipartite Spectral Projection Approach. *Knowledge-Based and Intelligent Information and Engineering Systems*, 398–407.

Zhou, H., Pan, W. and Shen, X. [2009]. Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, **3**, 1473–1496.