Mixture models

Christophe Biernacki

Contents

1 MIXTURE MODELS
Christophe Biernacki **1**

1.1 Mixture models as a many-purpose tool 1

1.1.1 Starting from applications 1

1.1.2 The mixture model answer 4

1.1.3 Classical mixture models 6

1.1.4 Other models 11

1.2 Estimation 11

1.2.1 Overview 11

1.2.2 Maximum likelihood and variants 12

1.2.3 Theoretical difficulties related to the likelihood 14

1.2.4 Estimation algorithms 16

1.3 Model selection in density estimation 22

1.3.1 Need to select a model 22

1.3.2 Frequentist approach and deviance 24

1.3.3 Bayesian approach and integrated likelihood 30

1.4 Model selection in (semi-)supervised classification 36

1.4.1 Need to select a model 36

1.4.2 Error rates-based criteria 37

1.4.3 A predictive deviance criterion 39

1.5 Model selection in clustering 43

1.5.1 Need to select a model 43

1.5.2 Partition-based criteria 44

1.5.3 The Integrated Completed Likelihood criterion 46

1.6 Experiments on real data sets 51

1.6.1 BIC: extra-solar planets 52

1.6.2 AIC$_{cond}$/BIC/AIC/BEC/$\hat{e}^{cv}$: benchmark data sets . . . 54

1.6.3 AIC$_{cond}$/$\hat{e}^{cv}_{V}$: textile data set 55

1.6.4 BIC: social comparison theory 57

1.6.5 NEC: marketing data 59

1.6.6 ICL: prostate cancer data 60

1.6.7 BIC: density estimation in the steel industry 62

1.6.8 BIC: partitioning communes of Wallonia 64

1.6.9 ICLbic/BIC: acoustic emission control 66

1.6.10 ICLbic/ICL/BIC/ILbayes: a seabird data set 67

1.7 Future methodological challenges 68

Bibliography **71**

Chapter 1

MIXTURE MODELS

Christophe Biernacki

1.1 Mixture models as a many-purpose tool

Finite mixture models are one of the probabilistic frameworks which reach an especially diverse community of people, including statisticians and practitioners (scientific or not). Initial reasons for being confronted with mixtures may be different for impacted communities but lead finally to close interconnections between them. Indeed, applied statisticians and practitioners usually discover finite mixture models from the numerous application fields where they meet numerous successes. It typically gathers $\{\emptyset$,un,semi-$\}$ supervised *classification* and *density estimation*. The keys of these successes are both their high meaningfulness and flexibility. However, flexibility is in return a matter of algorithmic and mathematical questionings for methodological and theoretical statisticians. In particular, it addresses *estimation* and *model selection* issues, on both computational and mathematical aspects. But, solutions to be provided to these issues highly beneficiate to depend on initial related application fields.

1.1.1 Starting from applications

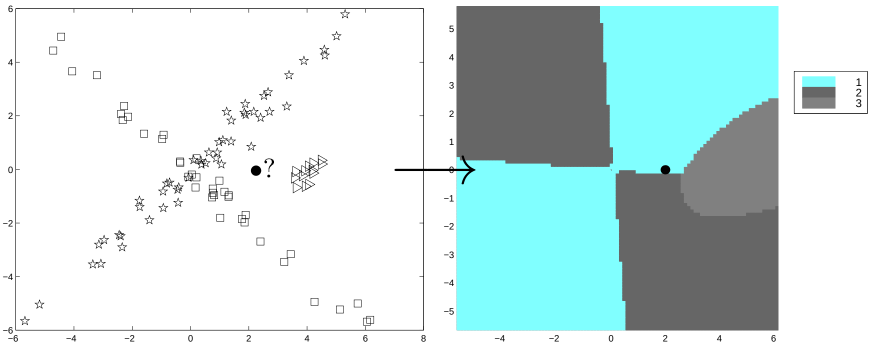
Supervised classification

In *supervised* classification, data are composed of n individuals $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ belonging to a space $\mathcal{X}$ of dimension d, and also of an associated partition in K groups $G_1, \ldots, G_K$. This partition is denoted by $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$, 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 ($i = 1, \ldots, n$, $k = 1, \ldots, K$). The data set is thus composed of all pairs $\mathcal{D} = (\mathbf{x}, \mathbf{z}) = ((\mathbf{x}_1, \mathbf{z}_1), \ldots, (\mathbf{x}_n, \mathbf{z}_n))$.

It is generally denoted as the learning data set. The aim is to estimate the group $\mathbf{z}_{n+1}$ of any new individual $\mathbf{x}_{n+1}$ in $\mathcal{X}$ for which the group would be unknown. This aim can be reformulated as the estimation of an allocation rule r from $\mathcal{D}$ and defined as follows:

$$\begin{array}{rccc} r: & \mathcal{X} & \longrightarrow & \{1,\dots,K\} \\ & \mathbf{x}_{n+1} & \longmapsto & r(\mathbf{x}_{n+1}). \end{array} \tag{1.1}$$

An illustration is given in Figure 1.1. Note that the space of individuals $\mathcal{X}$ usually corresponds to $\mathbb{R}^d$ in the continuous case or also to $\{0,1\}^d$ in the binary situations. Other examples of $\mathcal{X}$ will be exhibited in Section 1.1.4.



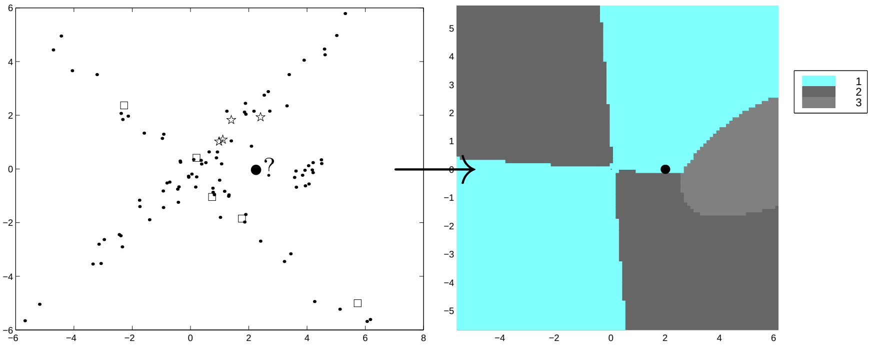
$(\mathbf{x}, \mathbf{z})$ and $\mathbf{x}_{n+1}$

$\hat{r}$ and $\hat{\mathbf{z}}_{n+1}$

Figure 1.1: Supervised classification purpose: illustration with a learning data set $(\mathbf{x}, \mathbf{z})$ in $\mathbb{R}^2$ with three groups. The new individual to be classified is denoted by $\mathbf{x}_{n+1}$ and is displayed by a "•".

Semi-supervised classification

In *semi-supervised* classification, the aim is the same as in supervised classification but the data set is composed of n^l individuals $(0 \leq n^l \leq n)$ $\mathbf{x}^l = (\mathbf{x}_1, \dots, \mathbf{x}_{n^l})$ for which group memberships $\mathbf{z}^l = (\mathbf{z}_1, \dots, \mathbf{z}^l_n)$ are known, whereas the $n^u = n - n^l$ remaining individuals $\mathbf{x}^u = (\mathbf{x}_{n^l+1}, \dots, \mathbf{x}_n)$ have unknown labels $\mathbf{z}^u = (\mathbf{z}_{n^l+1}, \dots, \mathbf{z}_n)$. We will note $\mathcal{D} = (\mathcal{D}_l, \mathcal{D}_u)$ with $\mathcal{D}_l = (\mathbf{x}^l, \mathbf{z}^l)$ and $\mathcal{D}_u = \mathbf{x}^u$. The main idea is thus that the unlabelled individuals may be useful to learn an allocation rule (see McLachlan [1992] p. 37–43). Usually, unlabelled individuals are expected to be more numerous than the labelled ones since the latter are clearly cheaper to obtain. An illustration of the semi-supervised setting is given in Figure 1.2.

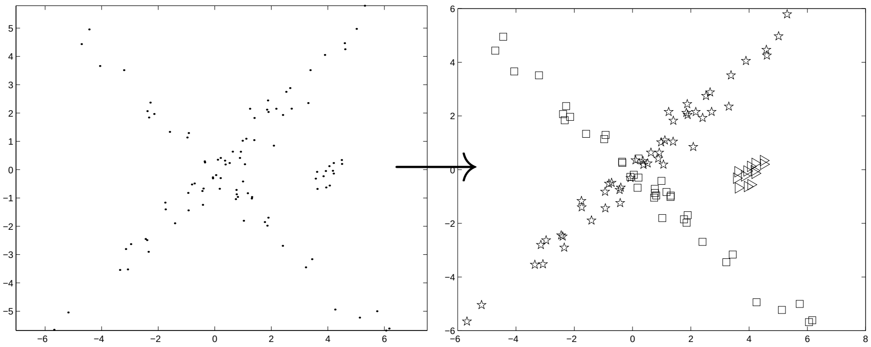


$(\mathbf{x}, \mathbf{z}^l)$ and $\mathbf{x}_{n+1}$ $\hat{r}$ and $\hat{\mathbf{z}}_{n+1}$

Figure 1.2: Semi-supervised classification purpose: illustration with a learning data set $(\mathbf{x}, \mathbf{z}^l)$ in $\mathbb{R}^2$ with three groups. The new individual to be classified is denoted by $\mathbf{x}_{n+1}$ and is displayed by a "$\bullet$".

Unsupervised classification

In *unsupervised* classification, or clustering, only individuals $\mathbf{x}$ are known and thus observed data are restricted to $\mathcal{D} = \mathbf{x}$. The aim is focused to estimating the partition $\mathbf{z}$ related to $\mathbf{x}$ and not to estimate a partition of all the space $\mathcal{X}$. However, in some cases like mixtures (as we will seen later), a partition of all the space $\mathcal{X}$ can be given as a simple by-product. In its more general, but also more difficult, version, the number of groups K is unknown and thus has also to estimated. An illustration of the clustering setting is displayed in Figure 1.3.



$\mathbf{x}$ $(\mathbf{x}, \hat{\mathbf{z}})$

Figure 1.3: Clustering purpose: illustration for data $\mathbf{x}$ in $\mathbb{R}^2$ and an estimated partition $\hat{\mathbf{z}}$ with three groups.

Density estimation

In density estimation, data are composed by individuals $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ belonging to a space $\mathcal{X}$ of dimension d and the aim is to estimate the distribution $\mathbf{x} \in \mathcal{X} \mapsto f(\mathbf{x})$ from which the sample arises. Then f can be used for multi-purposes like hypothesis testing. An illustration of the density estimation setting is given in Figure 1.4

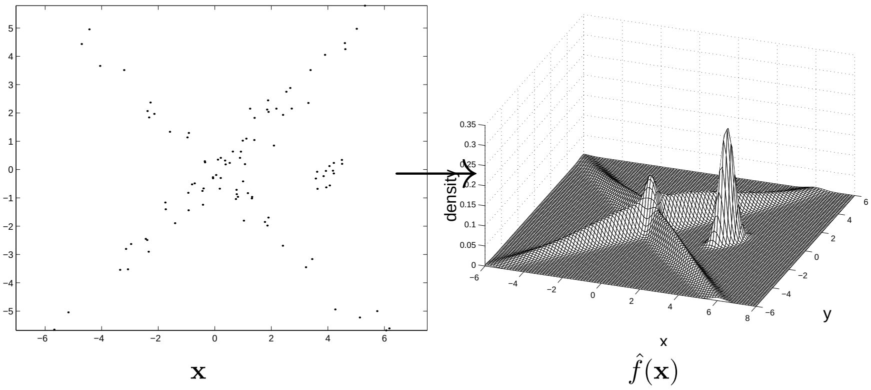


Figure 1.4: Density estimation purpose: illustration for data $\mathbf{x}$ in $\mathbb{R}^2$.

1.1.2 The mixture model answer

{∅,un,semi-} supervised classifications

The keystone to solve classification questions relies on the rigorous definition of a group. Intuitively, a group gathers elements which resemble each other. In a probabilistic framework, the resemblance between elements belonging to the same group may result by the fact that they arise from the same probability distribution function (pdf). Then, juxtaposing distributions associated to each group leads to a so-called mixture of distributions.

Thus, the individual $\mathbf{x}_1 \in \mathcal{X}$ belongs to the group G_k if and only if this individual is a realization of a random variable (rv) $\mathbf{X}_1 \in \mathcal{X}$ conditionally to the fact that $\{Z_{1k} = 1\}$, where $\mathbf{Z}_1 = (Z_{11}, \ldots, Z_{1K})'$ is a vector of $\{0,1\}^K$ indicating the group membership of $\mathbf{X}_1$. We still use the notation $Z_{ik} = 1$ if the individual $\mathbf{X}_1$ belongs to the kth group G_k, and $Z_{1k} = 0$ otherwise ($k = 1, \ldots, K$). The distribution of $\mathbf{X}_1$ conditionally to the group G_k, or equivalently the pdf of the rv $\mathbf{X}_1 | Z_{1k} = 1$, is written

$$\mathbf{X}_1 | Z_{1k} = 1 \sim f_k. \tag{1.2}$$

In addition, the pdf of $\mathbf{Z}_1$ corresponds to the multinomial distribution of order 1

$$\mathbf{Z}_1 \sim \mathsf{M}(\boldsymbol{\pi}), \tag{1.3}$$

with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ where π_k $(k = 1, \ldots, K)$ designates the mixing proportion of the component k in the mixture or equivalently the unconditional probability that an individual arises from this component, it means ($\sum_{k=1}^K \pi_k = 1$ and $\pi_k \geq 0$):

$$\pi_k = p(Z_{1k} = 1). \tag{1.4}$$

It means also that each group $G_1, \ldots, G_K$ is present with proportions $\pi_1, \ldots, \pi_K$, respectively. The joint pdf of the couple $(\mathbf{X}_1, \mathbf{Z}_1)$ is thus written

$$f(\mathbf{x}_1, \mathbf{x}_1) = \prod_{k=1}^K [\pi_k f_k(\mathbf{x}_1)]^{z_{1k}}, \tag{1.5}$$

and the marginal pdf of $\mathbf{X}_1$ is straightforwardly deduced. It corresponds to the so-called mixture pdf f:

$$\mathbf{X}_1 \sim f = \sum_{k=1}^K \pi_k f_k. \tag{1.6}$$

From this model, the pdf of $\mathbf{Z}_1$ conditional to $\{\mathbf{X}_1 = \mathbf{x}_1\}$, it means of the rv $\mathbf{Z}_1 | \mathbf{X}_1 = \mathbf{x}_1$, is given by

$$\mathbf{Z}_1 | \mathbf{X}_1 = \mathbf{x}_1 \sim \mathsf{M}(\mathbf{t}_1), \tag{1.7}$$

where $\mathbf{t}_1 = (t_{11}, \ldots, t_{1K})$ et t_{1k} $(k = 1, \ldots, K)$ is a conditional probability easily obtained by the Bayes theorem

$$\begin{aligned} t_{1k} &= p(Z_{1k} = 1 | \mathbf{X}_1 = \mathbf{x}_1) \\ &= \frac{\pi_k f_k(\mathbf{x}_1)}{f(\mathbf{x}_1)}. \end{aligned} \tag{1.8}$$

Thanks to these conditional probabilities, an allocation rule r can be proposed for each individual $\mathbf{x}_1$ of $\mathcal{X}$ by the so-called *maximum a posteriori* method (denoted now by MAP). It simultaneously gives a united answer to all issues addressed by supervised classification, semi-supervised classification and clustering. This simply consists of assigning an individual to the group with the largest conditional probability:

$$\forall \mathbf{x}_1 \in \mathcal{X} \quad r(\mathbf{x}_1) = k \text{ if } t_{1k} \geq t_{1h} \text{ for } h = 1, \ldots, K. \tag{1.9}$$

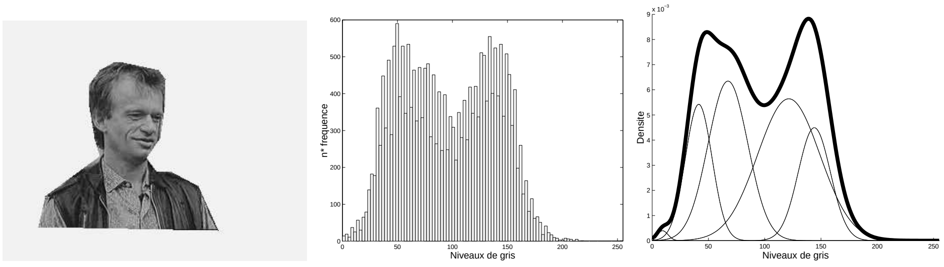
Beyond the intuitive appearance of such an allocation rule, a more subtle notion is hidden. Indeed, considering equal wrong assignment costs for each group (it is often a realistic case), using the MAP rule is strictly equivalent to minimize the classification error probability $e(r)$ associated to every rule r and defined by

$$
\begin{aligned}
e(r) &= \sum_{k=1}^{K} \pi_k \sum_{h=1,h\neq k}^{K} p(r(\mathbf{X}_1)=h|Z_{ik}=1) \qquad (1.10)\\
&= 1-\mathbb{E}_{(\mathbf{X}_1,\mathbf{Z}_1)}[Z_{1r(\mathbf{X}_1)}]. \qquad (1.11)
\end{aligned}
$$

This optimal rule is often designated as the *Bayes rule* in decision theory. It can also be extended to the case of unbalanced costs. All details can be found in numerous references as McLachlan [1992] (Chap. 1) or Flury [1997] (Chap. 7).

Density estimation

Mixture models design also an extremely flexible family of distributions. It is illustrated in Figure 1.5 where a Gaussian mixture is used to approximate the distribution of the grey scale distribution of an image.



(a) (b) (c)

Figure 1.5: Illustration of the flexibility of mixtures for the density estimation purpose: (a) a grayscale image, (b) the grayscale histogram associated to the character and (c) its estimation by a univariate Gaussian mixture.

1.1.3 Classical mixture models

Independence and parametric assumptions

From the mixture point of view, all classification purposes rely first on calculating conditional probabilities and then on using the optimal MAP rule. Since the conditional probabilities are expressed in function of the mixing proportions $\pi_1,\ldots,\pi_K$ and of the conditional pdfs $f_1,\ldots,f_K$, such quantities have to be estimated not only from available data $\mathcal{D}$ but also by means of more or less realistic assumptions, in any case often simplistic, which are available on the mixture model.

A first assumption concerns the sampling type. Pairs individuals-labels $(\mathbf{x}_1,\mathbf{z}_1)$,..., $(\mathbf{x}_n,\mathbf{z}_n)$ are assumed to i.i.d. (independently and identically distributed) arise from n pairs of rv $(\mathbf{X}_1,\mathbf{Z}_1),\ldots,(\mathbf{X}_n,\mathbf{Z}_n)$ following the same distribution as $(\mathbf{X}_1, \mathbf{Z}_1)$, distribution defined by (1.5). Such an hypothesis is performed both in clustering and in (semi-)supervised classification even if labels are not observed in the former situation. Note that this independence assumption may be relaxed like in hidden Markov models where independence between conditional rv $\mathbf{X}_{1|Z_{1k}=1}, \ldots, \mathbf{X}_{n|Z_{nk}=1}$ is preserved whereas it is relaxed between rv $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ (see for instance Besag [1986], McLachlan and Peel [2000] Chap. 13).

A second assumption concerns conditional pdf $f_1, \ldots, f_K$. It is also possible to perform non-parametric pdf (Silverman [1986], McLachlan [1992] Chap. 9, Benaglia *et al.* [2011]), or even semi-parametric pdf (Bordes *et al.* [2007]). However, it is more often assumed that f_k is wholly defined with a finite vectorial parameter $\boldsymbol{\alpha}_k$ and thus $(k = 1, \ldots, K)$

$$f_k = f(\cdot; \boldsymbol{\alpha}_k). \tag{1.12}$$

This assumption is quite weak since parametric mixture models are highly flexible. Denoting by $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ the mixture parameter with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ and $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_K)$, the mixture pdf is then given by

$$\begin{aligned} f &= f(\cdot; \boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \pi_k f(\cdot; \boldsymbol{\alpha}_k), \end{aligned} \tag{1.13}$$

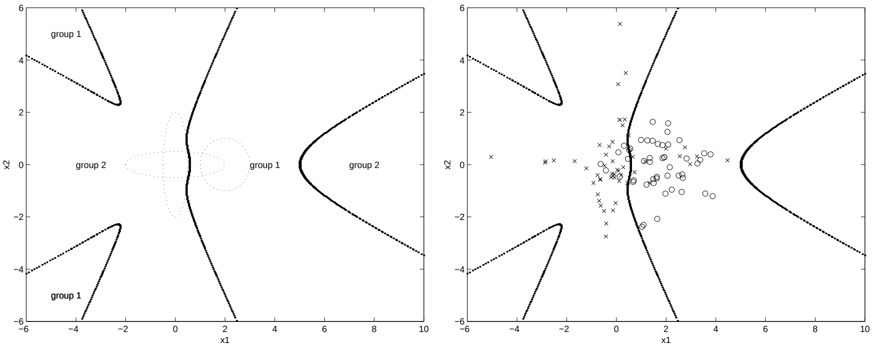
and the conditional probability is also parameterized by $\boldsymbol{\theta}$: $t_{1k} = t_{1k}(\boldsymbol{\theta})$. Thus, the couple composed by the parametric pdf $f(\cdot; \boldsymbol{\theta})$ and a space $\Theta_{\mathbf{m}}$ where evolves this parameter defines a so-called *model*, denoted now by $\mathcal{S}_{\mathbf{m}}$:

$$\mathcal{S}_{\mathbf{m}} = \{\mathbf{x}_1 \in \mathcal{X} \mapsto f(\mathbf{x}_1; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\}. \tag{1.14}$$

Moreover, $D_{\mathbf{m}} = \dim(\Theta_{\mathbf{m}})$ will designate the number of *continuous* parameters in $S_{\mathbf{m}}$. Note that, in the following, we will sometimes use the convenient language shortcut which confounds the index $\mathbf{m}$ and the corresponding model $\mathcal{S}_{\mathbf{m}}$.

In the following, we will assume that the mixture families of interest are identifiable, up to a label numbering permutation. It means that two different mixture parameters, even with label numbering permutation, lead to two different mixture pdfs (McLachlan and Peel [2000] Section 1.14).

Note that a component distribution f_k may be itself defined by a mixture of distributions, in particular in the supervised or in the semi-supervised setting. It corresponds thus to a so-called *mixture of mixture* (see for instance Hastie and Tibshirani [1996] and Miller and Browning [2003]). An illustration is displayed in Figure 1.6 with $\mathcal{X} = \mathbb{R}^2$ and $K = 2$ main components of same mixing proportions ($\pi_1 = \pi_2 = 0.5$), the first one f_1 being a Gaussian $\mathsf{N}((2,0)', \mathbf{I})$ and the second one f_2 being a mixture of two Gaussian subcomponents $\mathsf{N}((0,0)', \mathrm{diag}(0.25, 4))$ and $\mathsf{N}((0,0)', \mathrm{diag}(4, 0.25))$ with same proportions. The borderline between the two main components is also given on this figure to illustrate its great flexibility with such mixtures.



(a) (b)

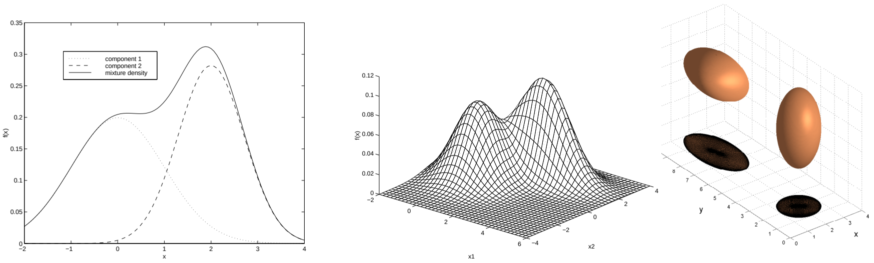
Figure 1.6: Mixture of a Gaussian component (group 1) and of a mixture of two Gaussian components (group 2): (a) classification borderline with associated isodensities and (b) classification borderline with a sample.

Gaussian mixtures

The multivariate mixture model is certainly the most known and used model for continuous data. It has a long history of use in clustering (see for instance Wolfe [1971], Bock [1981]) and in supervised classification (see numerous references in McLachlan [1992]). In that case, $\mathbf{x}_i$ $(i = 1, \ldots, n)$ are continuous variables $\mathbb{R}^d$ and the conditional density of components is written $(k = 1, \ldots, K)$

$$f(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \phi(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)\right), \tag{1.15}$$

with $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, $\boldsymbol{\mu}_k \in \mathbb{R}^d$ the component mean (or centre) and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ its variance-covariance matrix. Figures 1.7 (a), (b) and (c) respectively display univariate, bivariate and trivariate Gaussian mixtures.



(a) (b) (c)

Figure 1.7: Gaussian mixtures in (a) univariate, (b) bivariate and (c) trivariate situations.

At this stage, it is quite common to impose constraints on the parameter $\boldsymbol{\theta}$ through the space Θ. It is motivated by two essential reasons: either a prior information is available and is taken into account in this way, or the sample size is too small for providing a good estimation of the most general model. Indeed, the better is estimation of $\boldsymbol{\theta}$, the better is estimation of conditional probabilities and the associated MAP partition. See Section 1.3, 1.4 and 1.5 for detailed discussions about model selection. More precisely, it is possible to fix not only simple constraints on mixing proportions (equal or free) but also some more specific constraints on covariance matrices. Following the seminal approach of Banfield and Raftery [1993], Celeux and Govaert [1995] propose a spectral decomposition of the covariance matrices which allows a simple and useful meaning. Each covariance matrix is decomposed by $\mathbf{\Sigma}_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}'_k$, with $\lambda_k = |\mathbf{\Sigma}_k|^{1/d}$ the so-called volume of the component k, $\mathbf{D}_k$ the orthogonal matrix gathering the eigenvectors of $\mathbf{\Sigma}_k$ and corresponding to so-called orientation of this component, and $\mathbf{A}_k$ the diagonal matrix of normalized eigenvalues sorted by decreasing order on the diagonal and of determinant one, corresponding to the so-called shape of this component. By allowing some parameters, but not necessarily all, to vary or not between components, Celeux and Govaert [1995] obtain fourteen different models which they group into three families: the spherical family where the shape is equal to the identity matrix and thus only the volume has a role, the diagonal family where the covariance matrix is diagonal, and finally the general family which gathers all other situations (for instance the homoscedastic case where covariance matrices are equal or the heteroscedastic case corresponding to the most general situation with no constraints on covariance matrices). Combining these constraints with too standard constraints on mixing proportions (equal or free) leads then to 28 particular Gaussian mixture models.

Competitor parsimonious models have also been proposed since these previous seminal ones. In particular, we can note the variance-correlation decomposition $\mathbf{\Sigma}_k = \mathbf{T}_k \mathbf{R}_k \mathbf{T}_k$ of the covariance matrices (Biernacki and Lourme [2013]) where $\mathbf{T}_k$ is the corresponding diagonal matrix of conditional standard deviations and $\mathbf{R}_k$ the associated matrix of conditional correlations. Parsimonious models are obtained by combining simple constraints on matrices $\mathbf{T}_k$ and/or $\mathbf{R}_k$. These new models are stable when projected into the canonical planes and, so, faithfully representable in low dimension. They are also stable by modification of the measurement units of the data and such a modification does not change the model selection based on likelihood criteria. We can mention also Biecek *et al.* [2012] who permit not only inter-component constraints between covariance matrices, but also particular intra-component constraints like equality between variances or equality between covariances. Both last family models permit also some constraints on the centres of the Gaussians.

Latent class mixtures

Using categorical data is very frequent in statistics also. The standard model for clustering observations described through categorical variables is the so-called latent class model (see for instance Goodman [1974]). This model is assuming that the observations arose from a mixture of multivariate distributions and that the variables are conditionally independent knowing the groups. It has been proved to be successful in many practical situations (see for instance Aitkin *et al.* [1981]).

Observations to be classified are described with d discrete variables. Each variable j has m_j response levels. Data are $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ where $\mathbf{x}_i = (x_i^{jh}; j = 1, \ldots, d; h = 1, \ldots, m_j)$ with $x_i^{jh} = 1$ if i has response level h for variable j and $x_i^{jh} = 0$ otherwise. Data are supposed to arise independently from a mixture of K multivariate multinomial distributions with pdf

$$f(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x}_i; \boldsymbol{\alpha}_k) \tag{1.16}$$

with

$$f(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} \prod_{h=1}^{m_j} (\alpha_k^{jh})^{x_i^{jh}}, \tag{1.17}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ is denoting 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. As previously said, the latent class model is assuming that the variables are *conditionally independent* knowing the latent groups.

Analysing multivariate categorical data is difficult because of the curse of dimensionality. The standard latent class model which requires $(K-1) + K\sum_j (m_j - 1)$ parameters to be estimated is an answer to the dimensionality problem. It is much more parsimonious than the saturated log linear model which requires $\prod_j m_j$ parameters. For instance, with $K = 5$, $d = 10$, $m_j = 4$ for all variables, the latent class model is characterised with 154 parameters whereas the saturated log linear model requires about 10^6 parameters. Moreover, the latent class model can appear to produce a better fit than unsaturated log linear models while demanding less parameters.

In the binary case, some parsimonious alternatives have been also proposed by Celeux and Govaert [1991] by using the following reparameterization:

$$f(\mathbf{x}_i; \boldsymbol{\alpha}_k) = \prod_{j=1}^{d} (\varepsilon_{kj})^{|x_{ij} - \delta_{kj}|} (1 - \varepsilon_{kj})^{1 - |x_{ij} - \delta_{kj}|} \tag{1.18}$$

where $(\delta_{kj}, \varepsilon_{kj}) = (0, \alpha_{kj})$ if $\alpha_{kj} < 1/2$ and $(\delta_{kj}, \varepsilon_{kj}) = (1, 1 - \alpha_{kj})$ otherwise. Thus parameters $\boldsymbol{\alpha}_k$ are defined by $\boldsymbol{\alpha}_k = (\boldsymbol{\delta}_k, \boldsymbol{\varepsilon}_k)$ with $\boldsymbol{\delta}_k = (\delta_{k1}, \ldots, \delta_{kd})'$ a binary vector of dimension d acting as the center of the group since it corresponds to the modal value, and with $\boldsymbol{\varepsilon}_k = (\varepsilon_{k1}, \ldots, \varepsilon_{kd})'$ a vector belonging to the set $]0, 1/2[^d$ and acting as the dispersion of the component since it corresponds to the probability of each variable to have a different value from the center. It allows to retrieve the parameterization used by Aitchinson and Aitken [1976] in non-parametric supervised classification on nominal variables by the kernel method.

From such a decomposition, it is possible to draw parsimonious situations by imposing varying constraints on dispersions $\boldsymbol{\varepsilon}_k$. Three parsimonious models are thus proposed: the simplest one is independent of both the group and the variable; another model depends only on the group; the last one depends only on the variable. Combining with two constraints on mixing proportions (equal or free), it leads to finally eight particular mixture models for categorical data.

1.1.4 Other models

We presented previously the Gaussian and the latent class model since they correspond to the more widespread ones for continuous and categorical data, respectively. However, many other component distributions are possible, depending on the data and the hypotheses at hand. Kinds of data, and associated models, may be numerous (see also McLachlan and Peel [2000]): ranking data (Marden [1995], Jacques and Biernacki [2014]), directional data (Mardia and Jupp [2000]), ordinal data (Biernacki and Jacques [2015]), high dimensional continuous data (Bouveyron *et al.* [2007], McNicholas and Browne [2013]), graphical data (Nowicki and Snijders [2001]), functional data (Jacques and Preda [2014]),. . . Some recent works propose also models relaxing the conditional independence assumption for categorical and for mixed data while preserving identifiability, parsimony and parameter interpretation. The reader can refer for instance to Marbac *et al.* [2013] and Marbac *et al.* [2014], respectively, and many references therein.

1.2 Estimation

1.2.1 Overview

In density estimation, the central question is to estimate the parameter $\boldsymbol{\theta}_{\mathbf{m}}$, the model $\mathcal{S}_{\mathbf{m}}$ being fixed. The estimation of the model $\mathcal{S}_{\mathbf{m}}$, or equivalently of its index $\mathbf{m}$, will be discussed later and designed as the *model selection* problem which is the central thema of this book. Consequently, we will usually omit the index $\mathbf{m}$ thorough Section 1.2.

In the semi-supervised and unsupervised settings, the most simple and widespread estimation strategy is the *plug-in* one. It consists in estimating first $\boldsymbol{\theta}$, subject to constraints of $\mathcal{S}$, and then to directly use its estimate $\hat{\boldsymbol{\theta}}$ for estimating finally related conditional probabilities, useful for obtaining the

MAP rule. Then, we do not tackle alternative strategies which would directly estimate conditional probabilities: in (semi-)supervised classification, it concerns either the Bayesian predictive method of Ripley [1996] (p. 45–55), or the logistic regression (Ripley [1996], p. 43–45); in clustering, it concerns the Bayesian unsupervised clustering of Binder [1978].

Following the *plug-in* principle, Pearson [1894] initially used the method of moments for estimating the mixture parameter for a two component univariate Gaussian mixture model. Despite some renewed popularity of such an approach (see for instance Monfrini [2003] or also some references in McLachlan and Peel [2000] Chap. 1), it is globally abandoned nowadays. We do not consider either in this chapter Bayesian techniques for estimating $\boldsymbol{\theta}$ (see Robert [1994]) because we focus on the maximum likelihood method for its popularity, its simplicity and its relevant estimators properties under some quite general conditions, typically unicity and existence (see Lehmann [1983] Chap. 6).

1.2.2 Maximum likelihood and variants

Definition

Denoting by $\ell(\boldsymbol{\theta};\mathcal{D}) = \ln f(\mathcal{D};\boldsymbol{\theta})$ the *observed-data* log-likelihood of $\boldsymbol{\theta}$ (simply denoted sometimes as the observed log-likelihood or also as the log-likelihood), the maximum likelihood estimate (MLE) is given by

$$\hat{\boldsymbol{\theta}}_{\mathcal{D}} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta};\mathcal{D}). \tag{1.19}$$

In the following, we will note also $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathcal{D}}$ for simplicity when no confusion is possible. The log-likelihood is easily expressed thanks to the data independence hypothesis. It is written

$$\begin{aligned} \ell(\boldsymbol{\theta};\mathcal{D}) &= \ell(\boldsymbol{\theta};\mathcal{D}_l) + \ell(\boldsymbol{\theta};\mathcal{D}_u) && (1.20)\\ &= \sum_{i=1}^{n^l}\sum_{k=1}^{K} z_{ik}\ln\left(\pi_k f(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right) + \sum_{i=n^l+1}^{n} \ln\left(f(\mathbf{x}_i;\boldsymbol{\theta})\right). && (1.21)\end{aligned}$$

The log-likelihood $\ell(\boldsymbol{\theta};\mathbf{x},\mathbf{z}) = \ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u)$ is called *complete-data log-likelihood* (simply denoted sometimes as complete the log-likelihood) since it involves complete data $\mathbf{x}$ and $\mathbf{z}$. It is usually more simple to maximize that the log-likelihood $\ell(\boldsymbol{\theta};\mathcal{D})$ since it vanishes the initial mixture problem.

Theoretical properties

We give here two results which generalize respectively Proposition 2.2 and 2.3 of Ripley [1996] (p. 32–34) to our particular data set $\mathcal{D}$ which depends on the ratio of non-missing data n^l/n. We assume now that $n^l/n \to \beta$ with $\beta \in [0,1]$ when $n \to \infty$. Taking $\mathcal{D}'$ an independent data copy of $\mathcal{D}$, we will note also in the following $\boldsymbol{\theta}^*$ the value of $\boldsymbol{\theta}$ which minimizes the Kullback-Leibler divergence between the true (unknown) distribution $f(\mathcal{D}')$ and the candidate mixture distribution $f(\mathcal{D}';\boldsymbol{\theta})$:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\mathcal{D}'}[\ln f(\mathcal{D}') - \ln f(\mathcal{D}';\boldsymbol{\theta})]. \tag{1.22}$$

Under some standard regularity conditions, the first result concerns the pointwise consistency with $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}^*$ (see for instance White [1982]). If the true distribution is included in the candidate parametric family, we retrieve thus $f(\mathcal{D}') = f(\mathcal{D}';\boldsymbol{\theta}^*)$. The second result concerns the distributional consistency. We express it now and give also a proof since it involves new Fisher information matrices depending on β.

Proposition 1.1 *Under standard regularities conditions,*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} N\left(\mathbf{0}, \mathbf{J}_\beta^{-1}\mathbf{K}_\beta\mathbf{J}_\beta^{-1}\right) \tag{1.23}$$

with $N(\mathbf{0},\mathbf{V})$ the multivariate Gaussian distribution of zero mean and of covariance matrix $\mathbf{V}$, with $n^l/n \to \beta \in [0,1]$ when $n \to \infty$, and with $\mathbf{J}_\beta = (\beta\mathbf{J}_c + (1-\beta)\mathbf{J})$ and $\mathbf{K}_\beta = (\beta\mathbf{K}_c + (1-\beta)\mathbf{K})$ where

$$\mathbf{J}_c = -\mathbb{E}_{(\mathbf{X}_1,\mathbf{Z}_1)}\nabla^2 \ln f(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}^*), \qquad \mathbf{J} = -\mathbb{E}_{\mathbf{X}_1}\nabla^2 \ln f(\mathbf{X}_1;\boldsymbol{\theta}^*), \tag{1.24}$$
$$\mathbf{K}_c = \mathbb{V}_{(\mathbf{X}_1,\mathbf{Z}_1)}\nabla \ln f(\mathbf{X}_1,\mathbf{Z}_1;\boldsymbol{\theta}^*), \qquad \mathbf{K} = \mathbb{V}_{\mathbf{X}_1}\nabla \ln f(\mathbf{X}_1;\boldsymbol{\theta}^*). \tag{1.25}$$

Expectation is taken relatively to the true joint distribution $f(\mathbf{x}_1,\mathbf{z}_1)$ for the Fisher information matrices $\mathbf{J}_c$ and $\mathbf{K}_c$, and relatively to the true marginal distribution $f(\mathbf{x}_1)$ for the other information matrices $\mathbf{J}$ and $\mathbf{K}$. First and second derivatives concern $\boldsymbol{\theta}$. Note that if the true distribution is included in the candidate parametric family, then we retrieve the other classical results since $\mathbf{J}_c = \mathbf{K}_c$ and $\mathbf{J} = \mathbf{K}$.

Proof The maximum likelihood estimate verifies $\nabla\ell(\hat{\boldsymbol{\theta}};\mathcal{D}) = \mathbf{0}$. A Taylor expansion at the first order gives

$$\mathbf{0} = \nabla\ell(\hat{\boldsymbol{\theta}};\mathcal{D}) = \nabla\ell(\boldsymbol{\theta}^*;\mathcal{D}) + \nabla^2\ell(\tilde{\boldsymbol{\theta}};\mathcal{D})\ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \tag{1.26}$$

with $\tilde{\boldsymbol{\theta}}$ a vector "between" $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$, through the multidimensional meaning. Using now the central limit theorem and the strong law of large numbers, we obtain

$$\begin{aligned} &\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \\ &= \quad [-\frac{1}{n}\nabla^2\ell(\tilde{\boldsymbol{\theta}};\mathcal{D}_l) - \frac{1}{n}\nabla^2\ell(\tilde{\boldsymbol{\theta}};\mathcal{D}_u)]^{-1}[\frac{1}{\sqrt{n}}\nabla\ell(\boldsymbol{\theta}^*;\mathcal{D}_l) + \frac{1}{\sqrt{n}}\nabla\ell(\boldsymbol{\theta}^*;\mathcal{D}_u)] \quad (1.27) \\ &\xrightarrow{d} \quad [\beta\mathbf{J}_c + (1-\beta)\mathbf{J}]^{-1}\mathrm{N}\left(\mathbf{0}, [\beta\mathbf{K}_c + (1-\beta)\mathbf{K}]\right) = \mathrm{N}\left(\mathbf{0}, \mathbf{J}_\beta^{-1}\mathbf{K}_\beta\mathbf{J}_\beta^{-1}\right). \quad (1.28) \end{aligned}$$

The fact that $\mathbf{J}_c = \mathbf{K}_c$ and $\mathbf{J} = \mathbf{K}$ when the model is true is already a well-known property (see for instance Lehmann [1983] p. 118). □

Variants

We can also note that, in clustering, there exists a specific estimation method, sometimes called *classification approach* in contrast to this one of maximum likelihood sometimes called *mixture approach* (Celeux and Govaert [1993]). It consists in maximizing the complete-data log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{z}^u)$ on the couple $(\boldsymbol{\theta}, \mathbf{z}^u)$:

$$(\hat{\boldsymbol{\theta}}_c, \hat{\mathbf{z}}_c^u) = \arg \max_{\boldsymbol{\theta} \in \Theta, \mathbf{z}^u \in \mathcal{Z}^u} \ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{z}^u), \tag{1.29}$$

where $\mathcal{Z}^u$ denotes the space where $\mathbf{z}^u$ stands. The interest of this approach is to take explicitly into account the clustering purpose without sacrificing the simplicity of the plug-in principle. Indeed, in the small sample case, it can be observed that the estimated partition is better with the classification approach than with the mixture approach (see for instance Biernacki [1997] p. 52). But, complete-data maximum likelihood $\hat{\boldsymbol{\theta}}_c$ can be biased, even asymptotically, in particular if components have quite strong overlap (Bryant and Williamson [1978]).

Nevertheless, another positive point of the classification approach is the ability to retrieve some standard, and initially non-probabilistic, clustering criteria (Celeux and Govaert [1993]). For instance in the Gaussian case, Celeux and Govaert [1992] exhibited that maximizing the complete-data likelihood allows to retrieve, depending on the model at hand, some distance-based classical criteria. Thus, in the equal mixing proportion case, the K-means criterion (Ward [1963]) is equivalent to assume a spherical model with identical volume; this one of Friedman and Rubin [1967] is equivalent to an homoscedastic model; this one of Scott and Symons [1971] is equivalent to the most general model. In the latent class model for binary data, the most simple model corresponds to a χ^2-type criterion initially established without any reference to a probabilistic framework (see for instance Gower [1974]).

1.2.3 Theoretical difficulties related to the likelihood

Multiple roots

Maximum likelihood, in the mixture setting or not, is often faced to the existence of multiple roots of the log-likelihood. Roots correspond to the $\boldsymbol{\theta}$ values verifying

$$\nabla \ell(\boldsymbol{\theta}; \mathcal{D}) = \mathbf{0}. \tag{1.30}$$

Obviously, under some standard regularity conditions, the theory asserts existence of a unique consistent root of this equation (see for instance Cramér [1946] or also its multivariate extension by Tarone and Gruenhage [1975]). However, poor guidance is generally given for choosing this consistent root in case of multiple roots even if the bibliographical paper of Small *et al.* [2000] discusses of several approaches (see also an anterior discussion in Lehmann [1983] Chap. 6).

It includes for instance an iterated method based on consistent estimates, the use of a bootstrap method or also a technique relying on the asymptotic properties of the roots, when these properties can be explicitly expressed. Another possibility simply consists in selecting the root associated to the maximum likelihood value since Wald [1949] established consistency of the global maximum likelihood on some conditions. The Wald's properties of this MLE have been then extended by White [1982] in the very realistic situation of a misspecified model (see Section 1.3.2 and in particular Proposition 1.1). Consequently, the strategy consisting in retaining the maximum value of the maximum likelihood function is often adopted.

Pathological cases

It exists some situations where this global maximum is not consistent as illustrated in Neyman and Scott [1948], Ferguson [1982] or Stefanski and Carroll [1987]. In the heteroscedastic Gaussian case (but also in some non-Gaussian cases), it exists also a difficulty since the global maximum is not bounding above as noted first by Kiefer and Wolfowitz [1956] (note that this maximum is not a root of (1.30)). It corresponds to so-called *degenerated solutions*. It happens for instance by positioning a Dirac distribution at a particular data point (it corresponds to a specific degenerated Gaussian), while imposing the generalized variance (*i.e.* the determinant of the covariance matrix) to be non-null for at least one of the other Gaussians. In addition, among other local maxima of the likelihood, some of them may correspond to *spurious maximizers* as called by McLachlan and Peel [2000] Section 3.10. It corresponds to non-degenerated solutions where one or many covariance matrices are close to degeneracy, providing potentially large *finite* values of the likelihood although they do not correspond to some reality about the "true" parameter.

Practical difficulty for finding a suitable root

More details could be found in Redner and Walker [1984] Section 2.2 or McLachlan and Peel [2000] Section 1.18 for a detailed historical review on methods aiming at maximizing the likelihood in mixtures of distributions.

In the mixture context, solving the highly non-linear Equation (1.30) is generally impossible in closed-form. However, the increase of computing facilities helped to gradually overcome this difficulty. Thus, some simple mixture situations have been successfully solved by iterative methods. For instance, Rao [1948] used the *scoring* method of Fisher for studying a mixture of two univariate homoscedastic Gaussians, Mendenhall and Hader [1958] used a Newton method for a simpler situation with a unique scalar parameter. Then, Day [1969] for a multivariate mixture of two Gaussians, and Wolfe [1971] (and other references of the same author) with any number of heteroscedastic Gaussians, all used at similar periods some optimizing methods already close to the EM algorithm of Dempster *et al.* [1977]. This algorithm, and its numerous variants, is certainly today the most widespread estimation method for mixtures.

Although such algorithms allow to provide simple and relevant solutions for maximizing the likelihood, they are usually faced to the previous theoretical problems related to the likelihood: multiple roots and other stationary solutions, degeneracy, spurious solutions. Sometimes, it is added some difficulties related to the retained optimization method such some relative slow convergence or initial parameter dependency for EM that we describe now.

1.2.4 Estimation algorithms

The EM algorithm

For optimizing $\ell(\boldsymbol{\theta};\mathcal{D})$ in the general setting, the EM algorithm of Dempster *et al.* [1977] is often performed. It is a general algorithm for optimizing incomplete data (thus no restricted to mixtures) for maximizing the likelihood. Since the seminal paper of Dempster *et al.* [1977], numerous authors described its properties and its variants (see for instance McLachlan and Krishnam [1997] or Redner and Walker [1984] for the mixture context). In the mixture framework, missing data correspond to unknown labels $\mathbf{z}^u$. Starting from an initial parameter $\boldsymbol{\theta}^{(0)}$, EM proceeds in two sequential steps, the so-called E-step (*Expectation*) and the so-called M-step (*Maximization*). Noting

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{Z}^u)|\mathcal{D}] \tag{1.31}$$

the expectation of the complete-data log-likelihood $\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{Z}^u)$ with respect to the conditional distribution $f(\mathbf{z}^u|\mathcal{D};\boldsymbol{\theta}^{(q)})$, these two steps are expressed by:

E-step Calculate $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$;

M-step Choose $\boldsymbol{\theta}^{(q+1)} \in \Theta$ such that $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)})$.

If it exists several possible values $\boldsymbol{\theta}^{(q+1)}$ at the M-step, we retain simply one of them. Finally, the algorithm stops as soon as the log-likelihood reaches stationarity:

$$|\ell(\boldsymbol{\theta}^{(q+1)};\mathcal{D}) - \ell(\boldsymbol{\theta}^{(q)};\mathcal{D})| \leq \varepsilon, \tag{1.32}$$

with ε a fixed small non-negative value. It is also possible to stop EM after a predefined iteration number.

EM properties

A first important property of EM is that the log-likelihood monotonically increases along the run: $\ell(\boldsymbol{\theta}^{(q+1)};\mathcal{D}) \geq \ell(\boldsymbol{\theta}^{(q)};\mathcal{D})$ pour $q \geq 0$. Proving this point (see for instance McLachlan and Krishnam [1997] Chap. 3) relies on the following decomposition of the log-likelihood into a term of complete-data-like log-likelihood and an entropy-like term:

$$\ell(\boldsymbol{\theta};\mathcal{D}) = \ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u) + \xi(\boldsymbol{\theta};\mathbf{z}^u), \tag{1.33}$$

where the complete-data-like log-likelihood is

$$\begin{aligned}
&\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{z}^u)\\
&= \ell(\boldsymbol{\theta};\mathcal{D}_l) + \ell(\boldsymbol{\theta};\mathcal{D}_u,\mathbf{z}^u) &(1.34)\\
&= \sum_{i=1}^{n^l}\sum_{k=1}^{K} z_{ik}\ln\left(\pi_k f(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right) + \sum_{i=n^l+1}^{n}\sum_{k=1}^{K} z_{ik}\ln\left(\pi_k f(\mathbf{x}_i;\boldsymbol{\alpha}_k)\right) &(1.35)
\end{aligned}$$

and the entropy-like term is

$$\xi(\boldsymbol{\theta};\mathbf{z}^u) = -\sum_{i=n^l+1}^{n}\sum_{k=1}^{K} z_{ik}\ln(t_{ik}(\boldsymbol{\theta})) \geq 0. \tag{1.36}$$

This last term varies between 0 and $n^u\ln(K)$.

Taking expectation of both members of this equation subject to $f(\mathbf{z}^u|\mathcal{D};\boldsymbol{\theta}^{(q)})$, we obtain

$$\begin{aligned}
\ell(\boldsymbol{\theta};\mathcal{D}) &= \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[\ell(\boldsymbol{\theta};\mathcal{D},\mathbf{Z}^u)|\mathcal{D}] + \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[\xi(\boldsymbol{\theta};\mathbf{Z}^u)|\mathcal{D}] &(1.37)\\
&= Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(q)}) + \xi(\boldsymbol{\theta};\mathbf{t}^u(\boldsymbol{\theta}^{(q)})), &(1.38)
\end{aligned}$$

where $\mathbf{t}^u(\boldsymbol{\theta}) = (\mathbf{t}_{n^l+1}(\boldsymbol{\theta}),\ldots,\mathbf{t}_n(\boldsymbol{\theta}))$. The transformation of the entropy term is a consequence of $t_{ik}(\boldsymbol{\theta}^{(q)}) = p(Z_{ik} = 1|\mathbf{X}_i = \mathbf{x}_i;\boldsymbol{\theta}^{(q)}) = \mathbb{E}_{\boldsymbol{\theta}^{(q)}}[Z_{ik}|\mathbf{X}_i = \mathbf{x}_i]$ $(i = n^l+1,\ldots,n,\ k = 1,\ldots,K)$. We thus deduce that

$$\begin{aligned}
\ell(\boldsymbol{\theta}^{(q+1)};\mathcal{D}) - \ell(\boldsymbol{\theta}^{(q)};\mathcal{D}) &= \{\xi(\boldsymbol{\theta}^{(q+1)};\mathbf{t}^u(\boldsymbol{\theta}^{(q)})) - \xi(\boldsymbol{\theta}^{(q)};\mathbf{t}^u(\boldsymbol{\theta}^{(q)}))\}\\
&\quad + \{Q(\boldsymbol{\theta}^{(q+1)};\boldsymbol{\theta}^{(q)}) - Q(\boldsymbol{\theta}^{(q)};\boldsymbol{\theta}^{(q)})\}. &(1.39)
\end{aligned}$$

The first term of the second member of this equation is non-negative as defined in the M-step. We then conclude by noting that the second term is also non-negative since

$$\begin{aligned}
&\xi(\boldsymbol{\theta}^{(q+1)};\mathbf{t}^u(\boldsymbol{\theta}^{(q)})) - \xi(\boldsymbol{\theta}^{(q)};\mathbf{t}^u(\boldsymbol{\theta}^{(q)})) =\\
&\sum_{i=n^l+1}^{n}\left\{\sum_{k=1}^{K} t_{ik}(\boldsymbol{\theta}^{(q)})\ln\left(\frac{t_{ik}(\boldsymbol{\theta}^{(q)})}{t_{ik}(\boldsymbol{\theta}^{(q+1)})}\right)\right\} \geq 0. &(1.40)
\end{aligned}$$

Indeed, we recognize, for each $i = n^l+1,\ldots,n$, the Kullback-Leibler divergence between distributions $\mathbf{t}_i(\boldsymbol{\theta}^{(q)})$ and $\mathbf{t}_i(\boldsymbol{\theta}^{(q+1)})$.

A second EM property is its speed of convergence towards a stationary value of the likelihood. This convergence rate is usually considered as low since it is linear around a stationary parameter $\boldsymbol{\theta}^*$ of the likelihood (see McLachlan and Krishnam [1997] Chap. 3.9), contrary to Newton-like methods which benefit from a local quadratic convergence. Each EM iteration is a mapping g of Θ into Θ such that $\boldsymbol{\theta}^{(q+1)} = g(\boldsymbol{\theta}^{(q)})$. If $\boldsymbol{\theta}^{(q)}$ converges towards a parameter $\boldsymbol{\theta}^*$ and that g is a continuous mapping also, then $\boldsymbol{\theta}^* = g(\boldsymbol{\theta}^*)$. A Taylor expansion of $g(\boldsymbol{\theta}^{(q)})$ around $\boldsymbol{\theta}^*$ allows to write

$$\boldsymbol{\theta}^{(q+1)} - \boldsymbol{\theta}^* \approx \mathbf{H}(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^{(q)} - \boldsymbol{\theta}^*), \tag{1.41}$$

with $\mathbf{H}(\boldsymbol{\theta}^*)$ the Jacobian matrix $D \times D$ of $g(\boldsymbol{\theta})$, D being the number of continuous parameters in Θ. Thus, an EM iteration is nearly linear around convergence with convergence matrix equal to $\mathbf{H}(\boldsymbol{\theta}^*)$. In addition, since the global convergence rate is given by

$$\gamma = \lim_{q \to \infty} \frac{\|\boldsymbol{\theta}^{(q+1)} - \boldsymbol{\theta}^*\|}{\|\boldsymbol{\theta}^{(q)} - \boldsymbol{\theta}^*\|} \tag{1.42}$$

for any norm $\|\cdot\|$ of $\mathbb{R}^D$, it also corresponds to the largest eigenvalue of $\mathbf{H}(\boldsymbol{\theta}^*)$. The speed of convergence of EM then depends on the value of γ, a large value leading to a slow convergence rate.

Beyond its theoretical properties, EM is widely appreciated for its ease of implementation, its generally computationally light iterations (no Hessian matrix to compute), the low memory requirement to make it work (it requires little storage) and finally it quite appealing principle. All these previous points can be easily guessed when having a precise look at its two steps. The E-step finally consists, for the mixture case, to compute conditional probabilities $t_{ik}(\boldsymbol{\theta}^{(q)})$ ($i = n^l+1, \ldots, n$, $k = 1, \ldots, K$) since the complete-data log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{z}^u)$ is linear with respect to missing data $\mathbf{z}^u$. In other words, we have the identity $Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(q)}) = \ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{t}^u(\boldsymbol{\theta}^{(q)}))$. The M-step allows to find the parameter $\boldsymbol{\theta}^{(q+1)}$ in closed form for many standard mixture models. Indeed, it is often easy to obtain the maximum likelihood estimate with complete data and the M-step finally consists of maximizing the complete-data likelihood where missing data have been replaced by their expectation, thus the previous conditional probabilities. Mixing proportions are given by

$$\pi_k^{(q+1)} = \frac{n_k^{(q)}}{n}, \tag{1.43}$$

where $n_k^{(q)} = \sum_{i=1}^{n^l} z_{ik} + \sum_{i=n^l+1}^{n} t_{ik}(\boldsymbol{\theta}^{(q)})$ corresponds to the "fuzzy" population of the component k. Other parameter estimates depend on the parametric model at hand. For instance, in the general heteroscedastic Gaussian case, we retrieve familiar expressions for centres and covariance matrices estimates ($k = 1, \ldots, K$):

$$\boldsymbol{\mu}_k^{(q+1)} = \frac{1}{n_k^{(q)}} \left(\sum_{i=1}^{n^l} z_{ik}\mathbf{x}_i + \sum_{i=n^l+1}^{n} t_{ik}(\boldsymbol{\theta}^{(q)})\mathbf{x}_i \right) \tag{1.44}$$

$$\begin{aligned}\boldsymbol{\Sigma}_k^{(q+1)} &= \frac{1}{n_k^{(q)}}\left(\sum_{i=1}^{n^l} z_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})' \right. \\ &\quad \left. + \sum_{i=n^l+1}^{n} t_{ik}(\boldsymbol{\theta}^{(q)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})(\mathbf{x}_i - \boldsymbol{\mu}_k^{(q+1)})'\right). \qquad (1.45)\end{aligned}$$

In the more restricted homoscedastic case we have

$$\boldsymbol{\Sigma}^{(q+1)} = \frac{1}{n}\sum_{k=1}^{K} n_k^{(q)} \boldsymbol{\Sigma}_k^{(q+1)}. \qquad (1.46)$$

Celeux and Govaert [1995] described the M-step for each of the fourteen Gaussian models already described in Section 1.1.3. For the classical latent class model considered in Section 1.1.3, Celeux and Govaert [1991] produced also corresponding E-steps.

Variants of EM

Since EM may be quite slow in some cases, numerous authors proposed modified versions of EM aiming at accelerating its convergence while preserving its simplicity. In this context, Liu and Sun [1997] consider, in the mixture context, the ECME algorithm (*Expectation Conditional Maximization of Either*) of Liu and Rubin [1994]. In ECME, the E-step of EM is unchanged but its M-step is replaced by the CM-step (*Conditional Maximization*) which maximizes, a choice based on parameters, either the expectation of complete-data log-likelihood as in the initial EM, either directly the log-likelihood. Alternatively, modifying the E-step, Ueda and Nakano [1998] propose a deterministic version of EM involving *simulated annealing*. It corresponds to the so-called DAEM algorithm (*Deterministic Annealing EM*) and it aims to overcome the problem of local maxima. More precisely, at the E-step, the conditional probabilities of the groups are raised to a given power, similar to a temperature, which tends towards unity when the number of iterations increases. Pilla and Lindsay [2001] suggested a new definition of the missing data in order to reduce their number. In that case, the convergence rate is improved in some parametric directions which can depend on the iteration number of EM. Another way for exploring in depth the parameter space in various directions, Celeux *et al.* [2001] proposed also an EM algorithm with sequential update of parameters for each component.

Other alternatives propose stochastic versions of EM. Their fundamental motivation is to avoid local maxima of the likelihood. In this way, the SEM algorithm (*Stochastic EM*) of Celeux and Diebolt [1985] incorporates an additional random S-step (*Stochastic*) between the E-step and the M-step. This new step consists of drawing the group memberships from a multinomial distribution of order one with the group conditional probabilities as parameters, instead of taken their expectation as initially in EM. starting from $\boldsymbol{\theta}^{(0)}$, SEM is expressed by

E-step As the E-step of EM;

S-step For each $i = n^l + 1, \ldots, n$, draw $\mathbf{z}_i(\boldsymbol{\theta}^{(q)})) \sim \mathsf{M}(\mathbf{t}_i(\boldsymbol{\theta}^{(q)}))$;

M-step Choose $\boldsymbol{\theta}^{(q+1)} \in \Theta$ such that $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{z}^u(\boldsymbol{\theta}^{(q)}))$.

Since the parameter sequence $(\boldsymbol{\theta}^{(q)})$ generated by SEM does not punctually converges, due to the S-step definition, the algorithm generally stops after a predefined number of iterations. This sequence converges in distribution towards the unique stationary distribution. Asymptotically, the average of this distribution provides a sensible local estimate of the likelihood. Thus, SEM allows to be less dependent on the initial value $\boldsymbol{\theta}^{(0)}$ if a "sufficient" iteration number is performed. Noting also that there exists a simulated annealing version of SEM, SAEM (*Simulated Algorithm EM*) of Celeux and Diebolt [1990], which allows to start with SEM and which allows to finish with EM while controlling a given temperature. SAEM has the advantage to punctually converge and simultaneously to be less dependent on the starting position.

Optimizing the complete-data log-likelihood can be performed with the CEM (*Classification* EM) algorithm which is a clustering version of EM proposed by Celeux and Govaert [1992]. CEM consists of adding a C-step (*Classification*) between the E-step and the M-step of EM. It simply corresponds to a MAP of the group conditional probabilities previously calculated at the E-step. The detail of CEM is the following:

E-step As the E-step of EM;

C-step Defined $\mathbf{z}(\boldsymbol{\theta}^{(q)})$ as the MAP of $\mathbf{t}(\boldsymbol{\theta}^{(q)})$;

Step- M Choose $\boldsymbol{\theta}^{(q+1)} \in \Theta$ such that $\boldsymbol{\theta}^{(q+1)} = \arg\max_{\boldsymbol{\theta}\in\Theta} \ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{z}^u(\boldsymbol{\theta}^{(q)}))$.

Remind that CEM does not optimize the observed-data log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{D})$ but the complete-data log-likelihood $\ell(\boldsymbol{\theta}; \mathcal{D}, \mathbf{z}^u)$ on the couple $(\boldsymbol{\theta}, \mathbf{z}^u)$.

Initializing EM

Instead of introducing randomness in the iterations of the EM algorithm itself (like SEM), it is possible to introduce randomness through the starting value $\boldsymbol{\theta}^{(0)}$. The underlying idea is that a sensible starting value $\boldsymbol{\theta}^{(0)}$ could be able to solve at the same time the problem of slow convergence rate and also the problem of local maxima. In practice, it is recommended to run EM from several initial parameters and then to retain the best run. However, the question to choose such initial parameters has to be addressed. In this aim, Coleman and Woodruff [2000] used a clustering method starting from a random partition of a subsample. McLachlan and Peel [2000] proposed, in the Gaussian case, to start with equal mixing proportions, with K centres drawn from a multivariate Gaussian with empirical mean and empirical covariance matrix of the whole data set, with homoscedastic covariances matrices equal to the empirical covariance matrix of the whole data set. Markatou *et al.* [1998] used a bootstrap method to preselect a sensible parameter subspace. Alternatively, Biernacki *et al.* [2003] formalized the following three step strategy:

Search-step It provides several starting values for EM;

Iteration-step EM is run from each previous starting values;

Selection-step Retain the previous run providing the highest likelihood.

Originality relies on the Search-step which can involved CEM, SEM or small preliminary runs of EM itself. Nevertheless, as rightly underlined by Meila and Heckerman [2001], choosing a starting parameter is essentially a trade-off between its relevance and its computational cost.

Impact of estimation on model selection

The EM solution can highly depend on its starting position especially in a multivariate context. This jeopardizes statistical analysis of mixture for two reasons. Firstly, as we have just discussed above, ML estimation is expected to provide sensible estimates of the mixture parameters. Secondly, the highest maximized likelihood enters the definition of numerous criteria (see Section 1.3 and the next sections) aiming to select a good mixture model and especially to choose a relevant number of mixture components. Thus, it is important to get the highest criterion value when estimating the parameters of a mixture through maximum likelihood.

Let us illustrate this fact with a simple example. We consider a sample of size $n = 50$ from a two-component univariate Gaussian mixture with proportions $\pi_1 = \pi_2 = 0.5$, means $\mu_1 = -0.8$, $\mu_2 = 0.8$ and variances $\sigma_1^2 = 1$, $\sigma_2^2 = 1.5$. All the parameters are supposed to be known, except the means μ_1 and μ_2. The likelihood has two local maxima as shown in Figure 1.8. If the lowest likelihood maximum is selected, it can have consequence for choosing the number of components K. For instance, Table 1.1 gives the AIC criterion values (Akaike [1974] and Section 1.3 below) for $K = 1$ and for the two different ML solutions for $K = 2$. Thus, despite its marked tendency to favour too complex models (see below again), AIC concludes wrongly for a single Gaussian distribution when the lowest local maximum likelihood is selected.

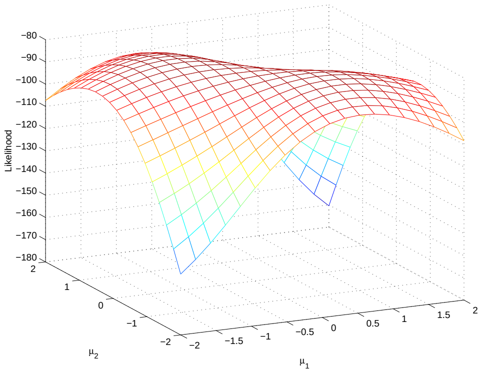


Figure 1.8: A two-mode likelihood surface.

	$K = 1$	$K = 2$ (highest ML)	$K = 2$ (lowest ML)
AIC	-85.29	-84.88	-85.95

Table 1.1: AIC criterion values for different MLE values.

1.3 Model selection in density estimation

1.3.1 Need to select a model

The bias/variance trade-off

Prefixing a parametric model $\mathcal{S}_{\mathbf{m}} = \{\mathbf{x}_1 \in \mathbb{R}^d \mapsto f(\mathbf{x}_1; \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta_{\mathbf{m}}\}$ as a candidate for the true, but unknown, distribution f allowed to stand in a simplified framework, where powerful parametric inference tools are available (see the previous section). However, this parametric hypothesis is binding since this true distribution can highly differ from the candidate one. For instance, the true component densities are not Gaussians or the true number of components is larger than this one involved in the model at hand. As a consequence, the estimated distribution is a *biased* estimate of f. There exists also a more subtle notion of "wrong" model through the idea of over-parameterized model. For instance, using a general heteroscedastic Gaussian model whereas the true components are spherical Gaussians would lead, for small sample sizes at least, to poor estimates in comparison to the use of a candidate model with spherical Gaussians. The same harmful behaviour would appear by involving for instance a number of components in the model which is larger than in the true distribution. Such situations are a consequence of *too large variance* estimates.

In order to formalize this bias/variance trade-off we consider now a family of model index collection $\mathcal{M} = \{\mathbf{m}\}$ corresponding to a family of model collection

$\{\mathcal{S}_{\mathbf{m}} : \mathbf{m} \in \mathcal{M}\}$. We denote by

$$\mathsf{KL}(f, f_{\boldsymbol{\theta}_{\mathbf{m}}}) = \mathbb{E}_{\mathcal{D}'}[\ln f(\mathcal{D}') - \ln f(\mathcal{D}'; \boldsymbol{\theta}_{\mathbf{m}})] \tag{1.47}$$

the Kullback-Leibler divergence between the true distribution f and any proposed distribution $f_{\boldsymbol{\theta}_{\mathbf{m}}} = f(\cdot; \boldsymbol{\theta}_{\mathbf{m}})$ corresponding to a model (index) $\mathbf{m}$ in $\mathcal{M}$, where $\mathcal{D}'$ is a sample independent of $\mathcal{D}$ but with the same distribution. Sometimes, we refer also to $2\mathsf{KL}(f, f_{\boldsymbol{\theta}_{\mathbf{m}}})$ as the *deviance* of $\mathcal{S}_{\mathbf{m}}$. The following reasoning could be applied to any other contrast than the Kullback-Leibler divergence; this remark will be useful in Section 1.4 and 1.5. We note also $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ the MLE of $\boldsymbol{\theta}_{\mathbf{m}}$ and $\boldsymbol{\theta}^*_{\mathbf{m}}$ the best parameter $\boldsymbol{\theta}_{\mathbf{m}}$ with the Kullback-Leibler divergence

$$\boldsymbol{\theta}^*_{\mathbf{m}} = \arg \inf_{\boldsymbol{\theta} \in \Theta_{\mathbf{m}}} \mathsf{KL}(f, f_{\boldsymbol{\theta}_{\mathbf{m}}}). \tag{1.48}$$

Then, we have the following straightforward but fundamental decomposition of $\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}})$, where we have noted $\hat{\boldsymbol{\theta}}_{\mathbf{m}} = \hat{\boldsymbol{\theta}}_{\mathcal{D},\mathbf{m}}$:

$$\begin{aligned} &\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) \\ &= \Big\{\mathsf{KL}(f, f_{\boldsymbol{\theta}^*_{\mathbf{m}}}) - \mathsf{KL}(f, f)\Big\} + \Big\{\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) - \mathsf{KL}(f, f_{\boldsymbol{\theta}^*_{\mathbf{m}}})\Big\} &(1.49) \\ &= \Big\{\text{bias}_{\mathbf{m}}\Big\} + \Big\{\text{variance}_{\mathbf{m}}\Big\}. &(1.50) \end{aligned}$$

The bias corresponds to the so-called *error of approximation* and the variance to the so-called *error of estimation.*

In order to illustrate the variance effect on the accuracy estimate of the mixture parameter, we generate 30 samples of size 40 and 200 from the following bivariate mixture with two components: $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = (0,0)'$, $\boldsymbol{\mu}_2 = (2,2)'$, $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$. The parameter $\boldsymbol{\theta}$ is then estimated by an EM algorithm with both a simple spherical and a more complex general Gaussian mixture of two components. Table 1.2 illustrates that the Kullback-Leibler divergence increases with the more complex model, revealing the effect of the variance. We note also that the variance decreases with the sample size.

n	$\mathbf{m}$	$\hat{\mathbb{E}}_{\mathcal{D}}\mathsf{KL}(f_{\boldsymbol{\theta}}, f_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}})$
40	spherical	0.0760
	general	0.1929
200	spherical	0.0116
	general	0.0245

Table 1.2: Effect of the variance of $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ on the density estimation quality.

What about hypothesis testing?

Tests of hypothesis, like the famous Likelihood Ratio Test (LRT), are often not really suitable in a model selection purpose for several important reasons.

Firstly, they induce a dissymmetry in the models comparisons through the null hypothesis and the alternative hypothesis. Secondly, selecting between more than two models leads to sequential testing which generates a lack of control on the global type I error rate. Thirdly, general tests like the LRT are able only to test nested models, what is quite restricted. A last reason, specific to the mixture case, is that the asymptotic distribution of the LRT is not necessarily a χ^2 distribution with the usual number of freedom (see for instance Aitkin and Rubin [1985] or Everitt [1981]) since the so-called standard regularity conditions do not hold. Indeed, in the case of the number of groups selection, two of these regularity conditions collapse: the model is not identifiable and also the borderline of the parameter space is reached for mixing proportions (one component situation corresponds to a two component situation with one empty component). However, some proposals exist for overcoming this problem like heuristic asymptotic distributions in Wolfe [1971], like marginalization over mixing parameters in Aitkin and Rubin [1985] or like a bootstrap non asymptotic estimation of the LRT distribution in Mclachlan [1987].

Model selection criteria that we present now will overcome most previous difficulties encountered by hypothesis testing, even if a particular attention should be paid to the number of components selection. They are also generally expressed as a penalization of the maximum log-likelihood by a measure of the model complexity. The list of described criteria is not exhaustive since the aim is to provide only the probably most important families of them.

1.3.2 Frequentist approach and deviance

The frequentist point of view consists of selecting the model $\mathbf{m} \in \mathcal{M}$ by using the *deviance* $2\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}})$ or alternatively the *expected deviance* $2\mathbb{E}_{\mathcal{D}}\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}})$. Approaches can be asymptotic or not.

In the following, we will remove the indices $\mathbf{m}$ and/or $\mathcal{D}$ when no ambiguity is possible. For instance, $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathcal{D},\mathbf{m}}$ denotes the MLE of $\boldsymbol{\theta}$ with the data set $\mathcal{D}$ and model $\mathcal{S}_{\mathbf{m}}$. Similarly, we use $\boldsymbol{\theta}^* = \boldsymbol{\theta}^*_{\mathbf{m}}$ for the best theoretical parameter, $D = D_{\mathbf{m}}$ for the number of parameters, $\mathcal{S} = \mathcal{S}_{\mathbf{m}}$ for the model, *etc.*

Expected deviance and related AIC-like criteria

The ideal model $\mathcal{S}_{\mathbf{m}^*}$ to be retained is this one minimizing the expected deviance

$$\mathrm{D}_{\mathbf{m}} = 2\mathbb{E}_{\mathcal{D}}\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}), \tag{1.51}$$

thus

$$\mathbf{m}^* \in \arg\min_{\mathbf{m}\in\mathcal{M}} \mathrm{D}_{\mathbf{m}}. \tag{1.52}$$

The main task is to estimating $\mathrm{D}_{\mathbf{m}}$ first, to then estimating $\mathbf{m}^*$. Its asymptotic approximation essentially relies on the following proposition.

Proposition 1.2 *Noting* $D^* = \mathrm{tr}[\mathbf{K}_\beta \mathbf{J}_\beta^{-1}]$, D *can be expressed by*

$$\mathrm{D} = 2\{\ln f(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} + 2D^* + O_p(\sqrt{n}). \tag{1.53}$$

Moreover, if the true distribution is included in the parametric distribution family described by $\mathcal{S}$*, then* $D^* = D$*, where* D *is the number of parameters in* Θ.

Proof We start with a Taylor expansion of order two around $\boldsymbol{\theta}^*$ of twice the log-likelihood $2\ell(\hat{\boldsymbol{\theta}};\mathcal{D}')$:

$$\begin{align*}
&2\ell(\hat{\boldsymbol{\theta}};\mathcal{D}') \\
&\approx\ 2\ell(\boldsymbol{\theta}^*;\mathcal{D}') + 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\nabla\ell(\boldsymbol{\theta}^*;\mathcal{D}') + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\nabla^2\ell(\boldsymbol{\theta}^*;\mathcal{D}')(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \tag{1.54}\\
&=\ 2\ell(\boldsymbol{\theta}^*;\mathcal{D}') + 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\nabla\ell(\boldsymbol{\theta}^*;\mathcal{D}') + \mathrm{tr}[\nabla^2\ell(\boldsymbol{\theta}^*;\mathcal{D}')(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)']. \tag{1.55}
\end{align*}$$

This result, associated to the fact that $\mathbb{E}_{\mathcal{D}'}\nabla\ell(\boldsymbol{\theta}^*;\mathcal{D}') = \mathbf{0}$ and also to independence between $\mathcal{D}$ and $\mathcal{D}'$, allows to write

$$\begin{align*}
\mathrm{D} &\approx 2\mathbb{E}_{\mathcal{D}'}[\ln f(\mathcal{D}') - \ell(\boldsymbol{\theta}^*;\mathcal{D}')] - 2\mathbb{E}_{\mathcal{D}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\mathbb{E}_{\mathcal{D}'}\nabla\ell(\boldsymbol{\theta}^*;\mathcal{D}') \\
&\quad -\mathrm{tr}[\mathbb{E}_{\mathcal{D}'}\nabla^2\ell(\boldsymbol{\theta}^*;\mathcal{D}')\mathbb{E}_{\mathcal{D}}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'] \tag{1.56}\\
&\approx 2\mathbb{E}_{\mathcal{D}'}[\ln f(\mathcal{D}') - \ell(\boldsymbol{\theta}^*;\mathcal{D}')] \\
&\quad -\mathrm{tr}[\{\mathbb{E}_{\mathcal{D}'_l}\nabla^2\ell(\boldsymbol{\theta}^*;\mathcal{D}'_l) + \mathbb{E}_{\mathcal{D}'_u}\nabla^2\ell(\boldsymbol{\theta}^*;\mathcal{D}'_u)\}\mathbb{V}_{\mathcal{D}}\hat{\boldsymbol{\theta}}] \tag{1.57}\\
&= 2\mathbb{E}_{\mathcal{D}'}[\ln f(\mathcal{D}') - \ell(\boldsymbol{\theta}^*;\mathcal{D}')] - \mathrm{tr}[(-n\mathbf{J}_\beta)(\mathbf{J}_\beta^{-1}\mathbf{K}_\beta\mathbf{J}_\beta^{-1}/n)] \tag{1.58}\\
&= 2\mathbb{E}_{\mathcal{D}'}[\ln f(\mathcal{D}') - \ell(\boldsymbol{\theta}^*;\mathcal{D}')] + \mathrm{tr}[\mathbf{K}_\beta\mathbf{J}_\beta^{-1}]. \tag{1.59}
\end{align*}$$

The error in this expression is of order $O(1/\sqrt{n})$. It remains to estimate the first term from the observed sample $\mathcal{D}$ to conclude:

$$\begin{align*}
&2\mathbb{E}_{\mathcal{D}'}[\ln f(\mathcal{D}') - \ell(\boldsymbol{\theta}^*;\mathcal{D}')] \\
&\approx\ 2\{\ln f(\mathcal{D}) - \ell(\boldsymbol{\theta}^*;\mathcal{D})\} \tag{1.60}\\
&\approx\ 2\{\ln f(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} - 2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'\nabla\ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \mathrm{tr}[\nabla^2\ell(\hat{\boldsymbol{\theta}};\mathcal{D})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'] \tag{1.61}\\
&=\ 2\{\ln f(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} - \mathrm{tr}[\{\nabla^2\ell(\hat{\boldsymbol{\theta}};\mathcal{D}_l) + \nabla^2\ell(\hat{\boldsymbol{\theta}};\mathcal{D}_u)\}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)'] \tag{1.62}\\
&\approx\ 2\{\ln f(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} - \mathrm{tr}[\{-n\mathbf{J}_\beta\}\mathbb{V}_{\mathcal{D}}\hat{\boldsymbol{\theta}}] = 2\{\ln f(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\} + D^*. \tag{1.63}
\end{align*}$$

Error in this last approximation is of order $O_p(\sqrt{n})$, and thus becomes the new global order of approximation for $\mathrm{D}_{\mathcal{S}}$. Noticing that **I**is the identity matrix of dimension $D \times D$, we deduce then that if the true distribution belongs to the parametric family described by the model candidate $\mathcal{S}$, thus $D^* = \mathrm{tr}[\mathbf{K}_\beta\mathbf{K}_\beta^{-1}] = \mathrm{tr}[\mathbf{I}_D] = D$, where $\mathbf{I}_D$ designates the identity matrix of dimension D. □

Thus, the theoretical expected deviance D can be expressed in function of the *observed deviance* $2\{\ln f(\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\}$ penalized by a measure of the model complexity, D^*. We obtain the so-called NIC criterion (*Network Information Criterion*) of Murata *et al.* [1991, 1993, 1994]:

$$\mathrm{NIC} = \ell(\hat{\boldsymbol{\theta}};\mathcal{D})\ - D^*, \tag{1.64}$$

and we retain the model $\mathcal{S}$ leading to the largest NIC value. When the true distribution is included in the parametric family described by $\mathcal{S}$, we retrieve also the so-called AIC criterion (*An Information Criterion*) of Akaike [1973, 1974]:

$$\mathrm{AIC} = \ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) \ - D. \tag{1.65}$$

Practical implementation of NIC is quite restricted since it is difficult to estimate (pseudo) Fisher matrices $\mathbf{J}_\beta$ and $\mathbf{K}_\beta$. Alternatively, we can prefer using its simpler variant AIC but with the crude assumption that the true distribution is included in the model $\mathcal{S}$ at hand. Strictly speaking, it would also impose to compare only nested models $\mathcal{S}$, for the same reason.

Alternatively, it is also possible to obtain non-asymptotic approximation of D by dividing the whole sample into two disjoint parts, so-called *learning* set and *test* set. Independence of both data sets assures that the related estimate is unbiased. The unbiased property is preserved while the variance is reduced if we make the average of estimates providing from different cuttings learning/test of the whole sample. In this light, Stone [1977] has obtained (a non-asymptotic version of) the NIC criterion before Murata *et al.* [1994] because establishing a asymptotic link between NIC and the following *Cross Validation* criterion (denoted by CV) defined by

$$\mathrm{CV} = \sum_{i=1}^{n^l} \ln f(\mathbf{x}_i; \hat{\boldsymbol{\theta}}_{\{i\}}) + \sum_{i=n^l+1}^{n} \ln f(\mathbf{x}_i, \mathbf{z}_i; \hat{\boldsymbol{\theta}}_{\{i\}}), \tag{1.66}$$

where $\hat{\boldsymbol{\theta}}_{\{i\}}$ is the MLE of $\boldsymbol{\theta}$ obtained from the whole data set $\mathcal{D}$ excepted the ith individual. Indeed, such a criterion asymptotically converges towards NIC. Note that Smyth [2000] suggests rather a coarse cross validation process by involving test samples with more than a unique individual. See also recent results about cross-validation in Arlot and Celisse [2010].

Properties of AIC-like criteria

It is proved that NIC is an inconsistent model selection criterion since it retains too complex models with non-null probability, even asymptotically. Let for instance two nested models $\mathcal{S}_1$ and $\mathcal{S}_2$ with $\Delta D = D_2 - D_1 > 0$ and let the additional hypothesis that the more parsimonious model $\mathcal{S}_1$ is the true one. We note also $\Delta\ell = \ell(\hat{\boldsymbol{\theta}}_2; \mathcal{D}) - \ell(\hat{\boldsymbol{\theta}}_1; \mathcal{D})$. Then, the following development establishes that it is possible to wrongly retain the more complex model for large sample sizes:

$$2(\mathrm{AIC}_2 - \mathrm{AIC}_1) + 2\Delta D = 2\Delta\ell \xrightarrow{d} \chi^2_{\Delta D}, \tag{1.67}$$

since $p(\chi^2_{\Delta D} > 2\Delta D) > 0$. In fact, when models in competition consist of choosing the number of components in a mixture, the asymptotic distribution of the ratio of the maximum likelihoods is not well-established, as explained in the beginning of the current section. Consequently, non-consistency of AIC is not really well-established in that case, even if it attested by numerical experiments (see illustration below). In addition, the expression of AIC/NIC itself is not totally valid since it relies on Taylor expansions not really justified again for the number of component situation. This is the reason why Bozdogan [1981, 1983], using the conjecture of Wolfe [1971], proposes a slight over-penalization of AIC:

$$\text{AIC3} = \ell(\hat{\boldsymbol{\theta}}; \mathcal{D}) \ - 1.5D. \tag{1.68}$$

However, this new criterion does not solve the non-consistency problem of AIC, even if it will select more parsimonious models than AIC because of its over-penalization.

We numerically illustrate that AIC and AIC3 criteria tend to select too complex models, even in the very simple situation of well-separated components with the very parsimonious spherical Gaussian model. We consider 30 samples of size $n = 200$ generated from a bivariate Gaussian mixture of two well-separated components with mixing proportions $\pi_1 = \pi_2 = 0.5$, with centres $\boldsymbol{\mu}_1 = (0,0)'$ and $\boldsymbol{\mu}_2 = (3.3.0)'$, and with covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$. A sample is displayed on Figure 1.9. A spherical model with equal proportions is estimated by an EM algorithm for different numbers of components $K \in \{1, \ldots, 5\}$ and the frequency of choosing K by the AIC and AIC3 criteria is displayed in Table 1.3.

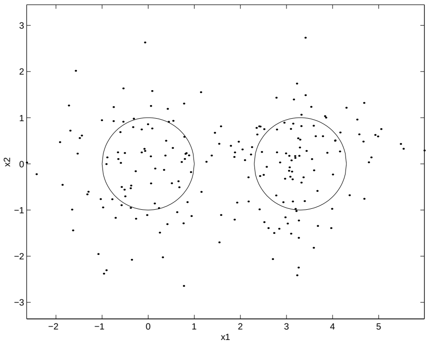


Figure 1.9: A sample of two well-separated bivariate Gaussian components with associated isodensities.

K	1	2	3	4	5
AIC	.	87	7	3	3
AIC3	.	97	3	.	.

Table 1.3: Frequency of the selected number of components with AIC and AIC3 for two well-separated bivariate components.

Deviance and related slope heuristics criteria

The ideal model $\mathcal{S}_{\hat{\mathbf{m}}^*}$ to be retained now is the one minimizing the deviance (no longer the expected one)

$$\hat{\mathbf{m}}^* \in \arg\min_{\mathbf{m}\in\mathcal{M}} 2\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_\mathbf{m}}). \tag{1.69}$$

The main task is thus to estimating the deviance $2\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_\mathbf{m}})$. A non-asymptotic approach is now presented. This presentation is inspired by the work of Baudry *et al.* [2012b].

We first derive the following straightforward but meaningful deviance decomposition:

$$\begin{aligned}
\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_\mathbf{m}}) &= -\ell(\hat{\boldsymbol{\theta}}_\mathbf{m};\mathcal{D}) + \ln f(\mathcal{D}) \\
&\quad + \Big\{\mathsf{KL}(f, f_{\hat{\boldsymbol{\theta}}_\mathbf{m}}) - \mathsf{KL}(f, f_{\boldsymbol{\theta}^*_\mathbf{m}})\Big\} + \Big\{\ell(\hat{\boldsymbol{\theta}}_\mathbf{m};\mathcal{D}) - \ell(\boldsymbol{\theta}_\mathbf{m};\mathcal{D})\Big\} \\
&\quad + \Big\{\mathsf{KL}(f, f_{\boldsymbol{\theta}^*_\mathbf{m}}) - \mathsf{KL}(f, f)\Big\} - \Big\{\ln f(\mathcal{D}) - \ell(\boldsymbol{\theta}_\mathbf{m};\mathcal{D})\Big\} \quad (1.70) \\
&= -\ell(\hat{\boldsymbol{\theta}}_\mathbf{m};\mathcal{D}) + \text{constant} \\
&\quad + \Big\{\text{variance}_\mathbf{m}\Big\} + \Big\{\widehat{\text{variance}}_\mathbf{m}\Big\} \\
&\quad + \Big\{\text{bias}_\mathbf{m}\Big\} - \Big\{\widehat{\text{bias}}_\mathbf{m}\Big\}, \quad (1.71)
\end{aligned}$$

where “constant” is independent of $\mathbf{m}$, where “$\text{variance}_\mathbf{m}$” and “$\widehat{\text{variance}}_\mathbf{m}$” respectively denote a variance-like term and its empirical version, and where “$\text{bias}_\mathbf{m}$” and “$\widehat{\text{bias}}_\mathbf{m}$” respectively denote a bias-like term and its empirical version. The second and the third lines of Equation (1.71) can be seen as an ideal penalty of the maximum log-likelihood. In order to estimate this penalty, the *slope heuristics* principle (Birgé and Massart [2007]) establishes some links between such quantities though the following two assumptions. The first assumption is to expect that both the theoretical and the empirical version of the variance are similar, thus “$\text{variance}_\mathbf{m} \approx \widehat{\text{variance}}_\mathbf{m}$”. The second assumption is to expect that the theoretical and the empirical bias are similar, thus “$\text{bias}_\mathbf{m} - \widehat{\text{bias}}_\mathbf{m} \approx 0$”. It then produces the following SH criterion (*Slope Heuristics*) penalizing the maximum log-likelihood

$$\text{SH}_\mathbf{m} = \ell(\hat{\boldsymbol{\theta}}_\mathbf{m};\mathcal{D}) - 2\widehat{\text{variance}}_\mathbf{m}. \tag{1.72}$$

The model with the highest SH value has to be retained. The question is now to estimate this new penalty.

The key relies on the fact that most optimal penalties shapes can be seen as linear functions of the complexity number, so the number of parameters $D_\mathbf{m}$ in our parametric case (see for instance Maugis and Michel [2012] for the Gaussian mixture case). Thus, the optimal penalty is now known up to a multiplicative constant κ:

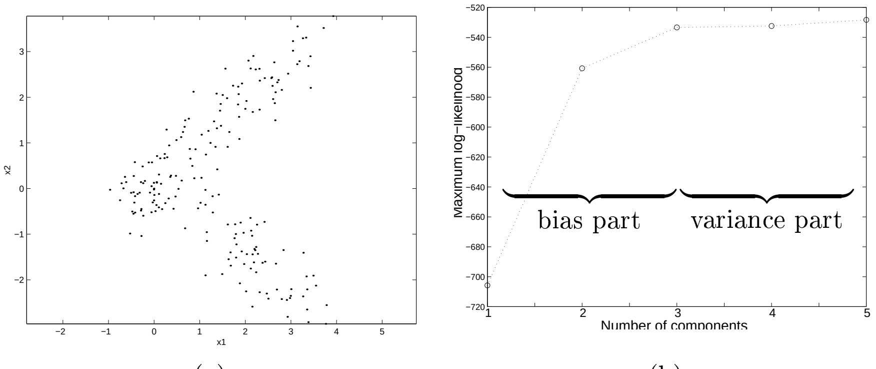
$$2\widehat{\text{variance}}_\mathbf{m} = \kappa D_\mathbf{m}. \tag{1.73}$$

The value of κ can be then estimated either by the so-called *dimension jump* principle, or by the so-called *slope* estimation principle. The slope estimation principle relies firstly on the following decomposition of $2\widehat{\text{variance}}_{\mathbf{m}}$:

$$2\widehat{\text{variance}}_{\mathbf{m}} = 2\Big\{\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}};\mathcal{D}) - f(\mathcal{D})\Big\} + 2\Big\{f(\mathcal{D}) - \ell(\boldsymbol{\theta}_{\mathbf{m}};\mathcal{D})\Big\}. \tag{1.74}$$

For the most complex models, we expect secondly the bias-like term $f(\mathcal{D}) - \ell(\boldsymbol{\theta}_{\mathbf{m}};\mathcal{D})$ to become nearly constant. Thus, the proportionality $\kappa D_{\mathbf{m}}$ can only be expressed through the log-likelihood term $\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}};\mathcal{D})$. In other words, it means that for complex enough models, $\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}};\mathcal{D})$ behaves linearly with $D_{\mathbf{m}}$ and the corresponding slope is $\kappa/2$. Then $\kappa/2$ can be estimated by a linear regression of $\ell(\hat{\boldsymbol{\theta}}_{\mathbf{m}};\mathcal{D})$ on $\frac{\kappa}{2}D_{\mathbf{m}}$. Thus, the involved penalty is here data-driven, contrary to this one used in AIC for instance. Note also that this method requires the estimation of a quite large number of "too" complex models to be involved. We can notice that this method formalizes some classical rules of thumb strategies aiming to detect an elbow directly in the maximum log-likelihood curve, like the so-called EL criterion (*Elbow Likelihood*) of Cutler and Windham [1993].

An illustration of the bias-variance trade-off on the log-likelihood function is given in Figure 1.10. It is apparent through an elbow in the curve of the maximum log-likelihood. We guess also the linearly part of the maximum log-likelihood beyond three components.



(a) (b)

Figure 1.10: Illustration of the bias and the variance parts with the maximum log-likelihood contrast: (a) sample from a mixture with three bivariate Gaussian components and (b) maximum log-likelihood for different numbers of components candidates.

In practice, the graphical user interface CAPUSHE[1] (CAlibrated Penalty Using Slope HEuristics), implements in R both the dimension jump and the slope estimation methods.

[1] http://cran.r-project.org/web/packages/capushe/

1.3.3 Bayesian approach and integrated likelihood

Integrated likelihood

In a Bayesian context, the key point is to retain the model $\mathcal{S}_{\hat{\mathbf{m}}^*}$ associated to the largest posterior probability[2]. This probability is expressed by

$$f(\mathbf{m}|\mathcal{D}) \propto f(\mathcal{D}|\mathbf{m})f(\mathbf{m}). \tag{1.75}$$

Thus, $\hat{\mathbf{m}}^* \in \arg\max_{\mathbf{m}\in\mathcal{M}} f(\mathbf{m}|\mathcal{D})$. In the case where all models have the same prior probabilities, this is equivalent to select the model maximizing $f(\mathcal{D}|\mathbf{m})$. This quantity, usually called *integrated likelihood* or also *marginal likelihood*, is expressed by

$$f(\mathcal{D}|\mathbf{m}) = \int_{\Theta_{\mathbf{m}}} f(\mathcal{D}|\boldsymbol{\theta},\mathbf{m})f(\boldsymbol{\theta}|\mathbf{m})d\boldsymbol{\theta}, \tag{1.76}$$

where $f(\mathcal{D}|\boldsymbol{\theta},\mathbf{m}) = f(\mathcal{D};\boldsymbol{\theta})$. Evaluating this probability relies on the definition of a prior distribution $f(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ (we note also for simplicity $f(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{m})$) and also on the computation of the integral. The integral computation is possible only in some restricted situations (typically with conjugate priors). Otherwise, several very different methods to approximate it are available (see for instance Kass and Raftery [1995]): numerical methods (but their are unstable in high dimension), Monte Carlo methods like the Gibbs or the Metropolis-Hastings samplers, the asymptotic Laplace-Metropolis approximation obtained from a Taylor expansion at the second order of the integral. We describe first the BIC criterion which is derived from the Laplace-Metropolis approximation. We then present a Monte Carlo evaluation in the latent model case, where conjugate non-informative priors are available.

Asymptotic approximation

The Laplace-Metropolis approximation allows in particular to express the integrated log-likelihood as the maximum log-likelihood penalized by the number of parameters D and also the sample size n. It thus provides a simple expression which allows also to avoid defining the prior distribution on $\boldsymbol{\theta}$. The following proposition details this important property (see for instance Kass and Wasserman [1995], Raftery [1995] p. 130–133 or also Ripley [1996] p. 62-65). We prove it in the general setting where the model at hand does not necessary include the true distribution.

Proposition 1.3 *Under standard regularity conditions, we have*[3]

$$\ln f(\mathcal{D}) = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \frac{D}{2}\ln(n) + O_p(1). \tag{1.78}$$

[2] There exists also another approach combining the frequentist deviance and the Bayesian posterior distributions. It leads to the so-called DIC criterion (*Deviance Information Criterion*), proposed by Spiegelhalter *et al.* [2002].

[3] Such an approximation is quite crude since of high order. Raftery [1995] (p. 130–133) proposed to retain the particular prior distribution $f(\boldsymbol{\theta}) = \mathsf{N}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1})$, which provides,

Proof The posterior distribution $f(\boldsymbol{\theta}|\mathcal{D})$ of $\boldsymbol{\theta}$ is assumed to be approximatively a Gaussian $\mathsf{N}(\tilde{\boldsymbol{\theta}}, \mathbf{V})$. In that case, its mean corresponds also to its mode, thus $\tilde{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta}|\mathcal{D})$, and the covariance matrix corresponds to the inverse of the Hessian of $-\ln f(\tilde{\boldsymbol{\theta}}|\mathcal{D})$, thus $\mathbf{V} = [-\nabla^2 \ln f(\tilde{\boldsymbol{\theta}}|\mathcal{D})]^{-1}$ (some simple algebra are used). When the sample size is large enough, the distribution $f(\boldsymbol{\theta}|\mathcal{D})$ is concentrated around its mode, so $g(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta};\mathcal{D}) - \ln f(\boldsymbol{\theta})$ is also concentrated around $\tilde{\boldsymbol{\theta}}$ since $f(\boldsymbol{\theta}|\mathcal{D}) \propto \exp\{-g(\boldsymbol{\theta})\}$. Consequently, the Taylor expansion at the second order of $g(\boldsymbol{\theta})$ around $\tilde{\boldsymbol{\theta}}$ is valid for large sample sizes, what allows to write, after noticing that $\nabla g(\tilde{\boldsymbol{\theta}}) = \mathbf{0}$,

$$\begin{aligned} f(\mathcal{D}) &= \int_\Theta f(\mathcal{D};\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta} = \int_\Theta \exp\{-g(\boldsymbol{\theta})\} d\boldsymbol{\theta} && (1.79)\\ &\approx \exp\{-g(\tilde{\boldsymbol{\theta}})\} \int_\Theta \exp\left\{-\frac{1}{2}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})'\mathbf{V}^{-1}(\boldsymbol{\theta}-\tilde{\boldsymbol{\theta}})\right\} d\boldsymbol{\theta} && (1.80)\\ &= \exp\{-g(\tilde{\boldsymbol{\theta}})\}(2\pi)^{D/2}|\mathbf{V}|^{1/2}. && (1.81)\end{aligned}$$

The last equation is due to the fact that the integral is equal to the normalize constant of a Gaussian distribution $\mathsf{N}(\tilde{\boldsymbol{\theta}}, \mathbf{V})$. The associated error being of order $O_p(1/n)$ (see Tierney and Kadane [1986]), we obtain

$$\ln f(\mathcal{D}) = \ell(\tilde{\boldsymbol{\theta}};\mathcal{D}) + \ln f(\tilde{\boldsymbol{\theta}}) + \frac{D}{2}\ln(2\pi) + \frac{1}{2}\ln|\mathbf{V}| + O_p(1/n). \qquad (1.82)$$

For large sample sets, we make the following two approximations $\tilde{\boldsymbol{\theta}} \approx \hat{\boldsymbol{\theta}}$ et $\mathbf{V} \approx \frac{1}{n}\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1}$ where $\hat{\mathbf{J}}_\beta$ et $\hat{\mathbf{K}}_\beta$ are respectively $\mathbf{J}_\beta$ et $\mathbf{K}_\beta$ where is replaced $\boldsymbol{\theta}^*$ by $\hat{\boldsymbol{\theta}}$ inside expectations (see Proposition 1.1), and thus $|\mathbf{V}| \approx n^{-D}|\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1}|$. An error of order $O_p(1/\sqrt{n})$ being induced by these last approximations, we obtain

$$\ln f(\mathcal{D}) = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) + \ln f(\hat{\boldsymbol{\theta}}) + \frac{D}{2}\ln(2\pi) - \frac{D}{2}\ln(n) + \frac{1}{2}\ln|\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1}| + O_p(1/\sqrt{n}). \qquad (1.83)$$

In this equation, the first term is of order $O_p(n)$, the fourth one of order $O_p(\ln(n))$ and all other ones of order equal or less than à $O_p(1)$. Removing all terms of order less or equal to $O_p(1)$, it gives than:

$$\ln f(\mathcal{D}) = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) - \frac{D}{2}\ln(n) + O_p(1). \qquad (1.84)$$

□

Such an approximation leads to maximize the so-called BIC criterion (*Bayesian Information Criterion*) of Schwarz [1978]:

$$\text{BIC} = \ell(\hat{\boldsymbol{\theta}};\mathcal{D}) \; - \frac{D}{2}\ln(n). \qquad (1.85)$$

Unlike the NIC criterion, the BIC penalty is simply expressed by a function of the number of parameters, the candidate $\mathcal{S}$ corresponding or not to the true model. Thus, the difficulty to estimating D^* in NIC is no more present. We can notice that the BIC criterion has been also proposed in the coding theory setting by Rissanen [1989] with the name MDL for *Minimum Description Length.* The

in average, the same information quantity as a unique observation. Thus,

$$\ln f(\hat{\boldsymbol{\theta}}) = -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\hat{\mathbf{J}}_\beta^{-1}\hat{\mathbf{K}}_\beta\hat{\mathbf{J}}_\beta^{-1}| \qquad (1.77)$$

and then, combining with Equation (1.83), some terms vanish and the greatest order becomes now $O_p(1/\sqrt{n})$.

BIC penalty being heavier than this one of AIC as soon as $\ln(n) > 2$ (so $n > 8$), BIC is expected to select more parsimonious models than AIC. In fact, it can be proven even that BIC is consistent. For instance, for two nested models $\mathcal{S}_1$ and $\mathcal{S}_2$, $\mathcal{S}_1$ being the true one, we have, in a similar way as Equation (1.67),

$$2(\mathrm{BIC}_2 - \mathrm{BIC}_1) + \Delta D \ln(n) = 2\Delta\ell \xrightarrow{d} \chi^2_{\Delta D}, \tag{1.86}$$

where $\Delta D = D_2 - D_1$, $\Delta\ell = \ell(\hat{\boldsymbol{\theta}}_2; \mathcal{D}) - \ell(\hat{\boldsymbol{\theta}}_1; \mathcal{D})$. Noting $\mu = \Delta D$ and $\sigma^2 = 2\Delta D$ respectively the mean and the variance of the rv $\chi^2_{\Delta D}$ and using also the Chebyschev inequality, we can write

$$p(\chi^2_{\Delta D} > \Delta D \ln(n)) \leq p(|\chi^2_{\Delta D} - \mu| > \Delta D \ln(n) - \mu) \leq \frac{\sigma^2}{(\Delta D \ln(n) - \mu)^2} \overset{n\to\infty}{\longrightarrow} 0. \tag{1.87}$$

It means that, asymptotically, BIC will select the simplest model $\mathcal{S}_1$, which corresponds to the true one. We could show also that BIC do not underestimate the order of the model. Thus, if the true model was $\mathcal{S}_2$, BIC will retain it with probability one. The proof still relies on the distribution of the ratio of the maximum likelihood, which is a non-central χ^2 (see Biernacki [1997] p. 74–75).

However, since all these consistency proofs rely on the fact that the model parameter is not on the borderline of the parameters space Θ, validity of such results can be hazardous in the mixture context for selecting a number of components. Some specific works on this problem exist. Leroux [1992] proved that BIC does not asymptotically underestimate the true number of components. Roeder and Wasserman [1997] proved, in the Gaussian mixture context to estimate a density in a non-parametric manner, that using BIC to select the number of components leads to consistent density estimate. Keribin [2000] generalizes these results by proving, under some conditions and by using a locally canonical reparameterization in order to obtain valid Taylor expansions, that BIC does not either overestimate the number of components, asymptotically.

Moreover, when the true model is not present in the family at hand, BIC will asymptotically select the model in the model family being the closest to the true one (see Lebarbier and Mary-Huard [2004]). It corresponds then to the case $f \neq f_{\boldsymbol{\theta}^*_{\mathbf{m}^*}}$ where $\mathbf{m}^*$ is the best model $\mathbf{m}$ in the set $\mathcal{M}$

$$\mathbf{m}^* = \arg \inf_{\mathbf{m}\in\mathcal{M}} \mathsf{KL}(f, f_{\boldsymbol{\theta}^*_{\mathbf{m}}}). \tag{1.88}$$

Non-asymptotic approximation for the latent class model

In the Gaussian mixture context, the BIC criterion appears to give a reasonable answer to the important problem of choosing the number of mixture components (see for instance Fraley and Raftery [2002]). However, some previous works dealing with the latent class model (see for instance Nadif and Govaert [1998]) for the binary case suggest that BIC needs particular large sample size to reach its expected asymptotic behaviour in practical situations. In this section, we take profit from the possibility to avoid asymptotic approximation of the observed integrated likelihood to propose an alternative non-asymptotic criterion[4].

Actually, a conjugate Jeffreys non informative prior distribution is available for the latent class model parameters (contrary to what happens for Gaussian mixture models; see for instance Marin *et al.* [2005]) and integrating the complete-data likelihood leads to a closed form formula. Defined in a Bayesian perspective, the integrated complete-data likelihood of a mixture is defined by

$$f(\mathbf{x},\mathbf{z}) = \int_{\Theta} f(\mathbf{x},\mathbf{z};\boldsymbol{\theta}) f(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{1.90}$$

Classical Jeffreys non informative Dirichlet prior distributions for the mixing proportions and the latent class parameters (respectively of order K and m_j) are given by

$$f(\boldsymbol{\pi}) = \mathsf{D}(\tfrac{1}{2},\ldots,\tfrac{1}{2}) \quad \text{and} \quad f(\boldsymbol{\alpha}_k^j) = \mathsf{D}(\tfrac{1}{2},\ldots,\tfrac{1}{2}). \tag{1.91}$$

Assuming independence between prior distributions of the mixing proportions $\boldsymbol{\pi}$ and the latent class parameters $\boldsymbol{\alpha}_k^j$ ($k = 1,\ldots,g$; $j = 1,\ldots,d$), we get, since the Dirichlet prior distribution is conjugate for the multinomial model (see for instance Robert [2001] Section 3.3.3), that

$$f(\mathbf{x},\mathbf{z}) = \frac{\Gamma(\frac{K}{2})}{\Gamma(\frac{1}{2})^K} \frac{\prod_{k=1}^K \Gamma(n_k + \frac{1}{2})}{\Gamma(n + \frac{K}{2})} \prod_{k=1}^K \prod_{j=1}^d \frac{\Gamma(\frac{m_j}{2})}{\Gamma(\frac{1}{2})^{m_j}} \frac{\prod_{h=1}^{m_j} \Gamma\left(n_k^{jh} + \frac{1}{2}\right)}{\Gamma(n_k + \frac{m_j}{2})}, \tag{1.92}$$

where $n_k = \#\{i : z_{ik} = 1\}$ and $n_k^{jh} = \#\{i : z_{ik} = 1, x_i^{jh} = 1\}$.

Denoting now by $\mathcal{Z}^u$ all possible combinations of labels $\mathbf{z}^u$, Equation (1.76) can be written (see Frühwirth-Schnatter [2006] p. 140)

$$f(\mathcal{D}) = \sum_{\mathbf{z}^u \in \mathcal{Z}^u} f(\mathbf{x},\mathbf{z}), \tag{1.93}$$

and thus the integrated likelihood $f(\mathcal{D})$ is explicit since the integrated complete-data likelihood $f(\mathbf{x},\mathbf{z})$ can be exactly calculated for the latent class model as just seen before.

Unfortunately, the sum over $\mathcal{Z}^u$ includes generally two many terms to be exactly computed. Following Casella *et al.* [2000], an importance sampling procedure can solve this problem. The importance sampling function, denoted by

[4] Notice that general non asymptotic approximation of $f(\mathcal{D})$ is possible (see Chib [1995]) by using the identity, for any $\boldsymbol{\theta}$ value,

$$f(\mathcal{D}) = \frac{f(\mathcal{D};\boldsymbol{\theta}) f(\boldsymbol{\theta})}{f(\boldsymbol{\theta}|\mathcal{D})}. \tag{1.89}$$

The denominator has then to be estimated from a MCMC (*Monte Carlo Markov Chain*) sampler for instance. However, this general method suffers from instabilities.

$I_{\mathcal{D}}(\mathbf{z}^u)$, is a pdf on $\mathbf{z}^u$ ($\sum_{\mathbf{z}^u \in \mathcal{Z}^u} I_{\mathcal{D}}(\mathbf{z}^u) = 1$ and $I_{\mathcal{D}}(\mathbf{z}^u) \geq 0$) which can depend on $\mathcal{D}$, its support necessarily including the support of $f(\mathbf{x}, \mathbf{z})$. Denoting by $\mathbf{z}^{u(1)}, \ldots, \mathbf{z}^{u(S)}$ an i.i.d. sample of size S from $I_{\mathcal{D}}(\mathbf{z}^u)$, $f(\mathcal{D})$ can be consistently estimated by the following Monte Carlo approximation

$$\hat{f}(\mathcal{D}) = \frac{1}{S} \sum_{s=1}^{S} \frac{f(\mathcal{D}, \mathbf{z}^{u(s)})}{I_{\mathcal{D}}(\mathbf{z}^{u(s)})}. \tag{1.94}$$

This estimate is unbiased and its variation coefficient is given by

$$\mathrm{c}_v[\hat{f}(\mathcal{D})] = \frac{\sqrt{\mathrm{Var}[\hat{f}(\mathcal{D})]}}{\mathrm{E}[\hat{f}(\mathcal{D})]} = \sqrt{\frac{1}{S}\left(\sum_{\mathbf{z}^u \in \mathcal{Z}^u} \frac{f^2(\mathbf{z}^u|\mathcal{D})}{I_{\mathcal{D}}(\mathbf{z}^u)} - 1 \right)}. \tag{1.95}$$

In order to approximate the ideal importance function $I_{\mathcal{D}}^*(\mathbf{z}^u)$, *i.e.* this one minimizing the variance and defined by

$$I_{\mathcal{D}}^*(\mathbf{z}^u) = f(\mathbf{z}^u|\mathcal{D}) = \int_{\Theta} f(\mathbf{z}^u|\mathcal{D}; \boldsymbol{\theta}) f(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}, \tag{1.96}$$

Biernacki *et al.* [2011] propose to make use of the following "Bayesian" instrumental distribution

$$I_{\mathcal{D}}(\mathbf{z}^u) = \frac{1}{R \# \mathcal{P}(\mathbf{z}^l)} \sum_{r=1}^{R} \sum_{\rho \in \mathcal{P}(\mathbf{z}^l)} f(\mathbf{z}^u|\mathcal{D}; \rho(\boldsymbol{\theta}^{(r)})), \tag{1.97}$$

where the set $\mathcal{P}(\mathbf{z}^l)$ denotes all label permutations of $\boldsymbol{\theta}$ on the set $\{1, \ldots, K\} \backslash \{k : z_{ik} = z_{ik}^l\}$ of label permutations not already fixed[5] by $\mathbf{z}^l$ and where $\{\boldsymbol{\theta}^{(r)}\}$ are chosen to be independent realisations of $f(\boldsymbol{\theta}|\mathcal{D})$. The sum over all label permutations $\mathcal{P}(\mathbf{z}^l)$ provides an importance density which is labelling invariant, like the ideal one[6]. Moreover, independence of $\{\boldsymbol{\theta}^{(r)}\}$, although not necessary for ensuring the validity of the unbiasedness of the estimator (1.94) and the variation coefficient (1.95), is recommended for a good estimation of (1.96) from the strong law of large numbers. In practice, a Gibbs sampler can be used[7] and the derived criterion will be called ILbayes (IL for Integrated Likelihood). Note that ILbayes is depending on both S and R. Note also that, in practice,

[5] If no label permutation if known ($n^l = 0$), then $\mathcal{P}(\mathbf{z}^l)$ contains all $K!$ label permutations on {1,...,K}. It can be huge for moderate to large values of K and thus (1.97) can be intractable.

[6] Because the prior distribution is symmetric in the components of the mixture, the posterior distribution is invariant under a permutation of the component labels (see for instance McLachlan and Peel [2000], Chap. 4). This lack of identifiability of $\boldsymbol{\theta}$ corresponds to the so-called *label switching* problem.

[7] An iteration of a possible Gibbs sampler for the latent class model is the following (see for instance Biernacki *et al.* [2011]) with priors defined in (1.91): $\boldsymbol{\pi}|\mathbf{z} \sim \mathsf{D}(\frac{1}{2} + n_1, \ldots, \frac{1}{2} + n_K)$, $\boldsymbol{\alpha}_k^j|\mathbf{x}, \mathbf{z} \sim \mathsf{D}(\frac{1}{2} + n_k^{j1}, \ldots, \frac{1}{2} + n_k^{jm_j})$ and, for $i = n^l + 1, \ldots, n$, by $\mathbf{z}_i^u|\mathbf{x}_i, \mathbf{z}_i^l; \boldsymbol{\theta} \sim \mathsf{M}(t_{i1}(\boldsymbol{\theta}), \ldots, t_{iK}(\boldsymbol{\theta}))$.

calculating ILbayes for values of $K > 6$ can be unreachable because of the factorial term involved in (1.97).

In order to illustrate the BIC and the ILbayes behaviour, we consider observations described by six variables ($d = 6$) with numbers of levels $m_1 = \ldots = m_4 = 3$ and $m_5 = m_6 = 4$ and a four component mixture ($K = 4$) with equal mixing proportions, $\boldsymbol{\pi} = (0.25, 0.25, 0.25, 0.25)$. The parameter $\boldsymbol{\alpha}$ is chosen to get a low cluster overlapping, about 11% of error rate, which corresponds to 15% of the worst error rate equal to 0.75. Detail of parameter value is given in Biernacki *et al.* [2011]. Figure 1.11 displays a data sample on the first two axes of a correspondence analysis. 20 samples are generated for three different sample sizes $n \in \{320, 1\,600, 3\,200\}$. For each sample, the EM algorithm has been run 10 times with random initial parameters (uniform distribution on the parameter space) for a sequence of 1 000 iterations and the best run is retained as being the maximum likelihood estimate. The mean of the retained number of mixture components with BIC and ILbayes criteria is displayed on Table 1.4. We notice that ILbayes performs better than BIC.

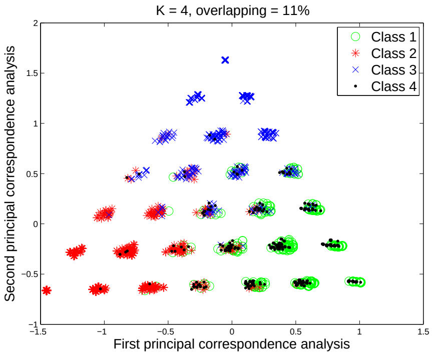


Figure 1.11: A sample ($n = 1\,600$) arising from $K = 4$ mixture situation for low overlapping. It is displayed on the first plane of a correspondence analysis and an i.i.d. uniform noise on $[0, 0.01]$ has been added on both axes for each point in order to clarify the visualisation.

n	320	1 600	3 200
BIC	3.0	3.5	4.0
ILbayes	3.4	4.0	4.0

Table 1.4: Mean of the chosen number of groups for BIC and ILbayes criteria when $K = 4$ for the latent class model. ILbayes is performed with $R = 50$ and $S = 100$.

1.4 Model selection in (semi-)supervised classification

1.4.1 Need to select a model

In the (semi-)supervised setting, usually the number of components is known and model selection essentially addresses model structure complexity and also variable selection. Model structure complexity corresponds for instance to particular constraints on the Gaussian matrices in the Gaussian mixture case. However, notice that variable subsets were not considered as possible models in the previous density estimation context (Section 1.3).

The reason for choosing a model in the (semi-)supervised classification setting is again the universal bias/variance trade-off. Nevertheless, this trade-off has primarily to be obtained on the discriminant rule $r_{\boldsymbol{\theta}}$, rule given by the MAP of $\mathbf{t}(\boldsymbol{\theta})$, instead of the density value $f_{\boldsymbol{\theta}}$. We recall also the notation of the theoretical error rate $e(r_{\boldsymbol{\theta}})$ associated to the rule $r_{\boldsymbol{\theta}}$. Denoting by r the optimal MAP rule obtained from the true (unknown) distribution f, we define also

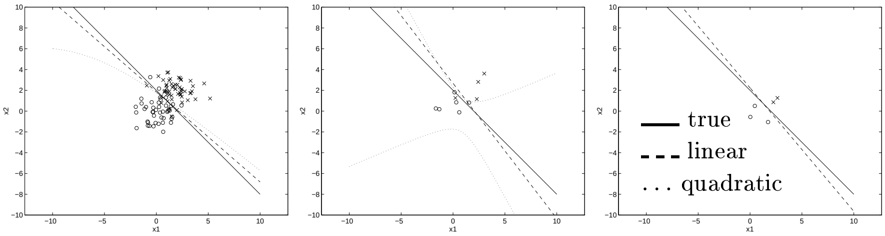
$$\boldsymbol{\theta}^*_{\mathbf{m}} = \arg\min_{\boldsymbol{\theta}\in\Theta_{\mathbf{m}}} e(r_{\boldsymbol{\theta}}) - e(r) \tag{1.98}$$

the best parameter associated to the model $\mathcal{S}_{\mathbf{m}}$ with regards to the best discriminant rule r. We then have the simple but important following decomposition, $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$ denoting as before the MLE:

$$\begin{aligned} e(r_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) - e(r) &= \Big\{e(r_{\boldsymbol{\theta}^*_{\mathbf{m}}}) - e(r)\Big\} + \Big\{e(r_{\hat{\boldsymbol{\theta}}_{\mathbf{m}}}) - e(r_{\boldsymbol{\theta}^*_{\mathbf{m}}})\Big\} & (1.99)\\ &= \Big\{\text{bias}_{\mathbf{m}}\Big\} + \Big\{\text{variance}_{\mathbf{m}}\Big\}. & (1.100)\end{aligned}$$

We notice thus that this bias/variance trade-off differs from this one produced in the density estimation context (see Equation (1.50)). Consequently, the best models in the density setting could be different from the best ones in the semi-supervised setting. Ripley [1996] (p. 27) illustrates for instance a situation where well-separated components has definitively not the same effect for density estimation and for discrimination. The question which is then addressed in this section is to propose specific model choice criteria taking fully into account the discriminant purpose. Such criteria will involve naturally error $e(r_{\boldsymbol{\theta}})$ and also conditional probabilities $\mathbf{t}(\boldsymbol{\theta})$.

Figure 1.12 illustrates influence of the model and of the sample size on the estimated discriminant rule (obtained by the plug-in method) in a supervised setting. We observe that the less is the sample size, the furthest the complex quadratic borderline is from the true simple linear borderline. In addition, when the sample size is too low ($n = 5$), the quadratic borderline is no more available since the estimate is singular. It corresponds to the limit case of an infinite variance situation. We see also that the simple linear estimate borderline has less variance than the quadratic one around the true borderline. This dependence on the model structure would be similarly illustrated in the semi-supervised setting.



$n = 100$ $n = 10$ $n = 5$

Figure 1.12: Illustration of the variance of estimates in the supervised classification setting: influence of the sample size on both the estimated spherical linear and general quadratic borderlines when the true borderline is spherical linear.

In a semi-supervised setting now, we illustrate the importance of selecting a subset of variables. We consider data simulated according to a design where *all* variables contribute to discrimination but with less and less information. This matter of fact causes an increase in the classification error rate. The experimental setting corresponds to $K = 2$ groups of same proportions ($\pi_1 = \pi_2 = 0.5$) and the class-conditional distributions are Gaussian distributions in dimension $d = 50$ with $\mathbf{X}_1|\mathbf{z}_{11} = 1 \sim \mathsf{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{X}_1|\mathbf{z}_{12} = 1 \sim \mathsf{N}(\boldsymbol{\mu}, \mathbf{I})$ with $\mu_j = \frac{1}{j}\ \forall j \in \{1, \dots, 50\}$. Thus, variables provide less and less discriminant information. The order in which variables are selected from 1 to 50 is assumed to be known. With the true model all the variables will be selected, but the less informative variables will dramatically increase the classifier variance. We consider 100 data sets with $n^l = 100$ label data and $n^u = 1\ 000$ unlabelled data. The optimal and the actual error rates, associated respectively to rules $r_{\hat{\boldsymbol{\theta}}}$ and $r_{\boldsymbol{\theta}}$, are evaluated through a test sample of size 50 000. The apparent error rate of $r_{\hat{\boldsymbol{\theta}}}$ is evaluated on the learning set. See more details on error rates in the next section. All error rates are shown Figure 1.13 and we can see that the optimal and apparent error rates decrease as the number of selected variables increases, while the actual error rate on the test sample decreases and then increases.

1.4.2 Error rates-based criteria

The aim of (semi-)supervised is to provide a discriminant rule with the minimum error rate. Ideally, it corresponds thus to retain the model where the associated rule $\hat{r} = r_{\hat{\boldsymbol{\theta}}}$ obtained from $\mathcal{D}$ leads to the less error in average. It

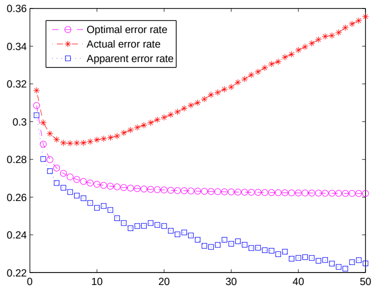


Figure 1.13: Variable selection for simulated data in the semi-supervised context: error rates according to the number of selected variables.

corresponds to the criterion e expressed by:

$$e = \mathbb{E}_{\mathcal{D}}[e(\hat{r})] = 1 - \mathbb{E}_{\mathcal{D}(\mathbf{X}'_1,\mathbf{Z}'_1)}[Z'_{1\hat{r}(\mathbf{X}'_1)}], \tag{1.101}$$

$(\mathbf{X}'_1, \mathbf{Z}'_1)$ being a rv independent of $(\mathbf{X}_1, \mathbf{Z}_1)$ but with identical pdf. Several classical estimates of e exist. The most simple of them if the *apparent* error rate $\hat{e}^{\text{a}}$ defined by

$$\hat{e}^{\text{a}} = 1 - \frac{1}{n^l}\sum_{i=1}^{n^l} Z_{i\hat{r}(\mathbf{X}_i)}. \tag{1.102}$$

It is a consistent estimate of e but it is well-known to have an optimistic bias, it means an underestimation of the error rate in average with $\mathbb{E}_{\mathcal{D}}[\hat{e}^{\text{a}}] \leq e$, since the same sample is used to learn and also to test the rule.

The so-called *partition* error rate $\hat{e}^{\text{p}}_{\{1\}}$ estimate is more relevant because it divides the whole data set into two different subsamples. The first one (the training or the learning sample), denoted by $\mathcal{D}^{\{1\}} = (\mathcal{D}^{\{1\}}_l, \mathcal{D}^{\{1\}}_u)$, is composed by a labelled subset $\mathcal{D}^{\{1\}}_l$ and an unlabelled subset $\mathcal{D}^{\{1\}}_u$. It is used for *learning* the discriminant rule, denoted by $\hat{r}^{\{1\}}$. Then, the second subsample (the testing or test sample) is composed by all the remaining *labelled* data $\bar{\mathcal{D}}^{\{1\}}_l = \mathcal{D}_l \backslash \mathcal{D}^{\{1\}}_l$ and is used for *testing* the rule $\hat{r}^{\{1\}}$ (note that the unlabelled data of $\mathcal{D}_u \backslash \mathcal{D}^{\{1\}}_u$ are thus discarded). Evaluating error $\hat{r}^{\{1\}}$ is finally given by

$$\hat{e}^{\text{p}}_{\{1\}} = 1 - \frac{1}{\#\bar{\mathcal{D}}^{\{1\}}_l}\sum_{\mathbf{x}_i \in \bar{\mathcal{D}}_l} Z_{i\hat{r}^{\{1\}}(\mathbf{X}_i)}. \tag{1.103}$$

Note that a proper use of this partition estimate in a semi-supervised setting is to remove the same proportion of labelled and unlabelled data from the training sample. It produces an unbiased estimate of $\mathbb{E}_{\mathcal{D}^{\{1\}}}[e(\hat{r}^{\{1\}})]$ since we can easily verify that

$$\mathbb{E}_{\mathcal{D}}[\hat{e}^{\text{p}}_{\{1\}}] = 1 - \mathbb{E}_{\mathcal{D}^{\{1\}}(\mathbf{X}'_1,\mathbf{Z}'_1)}[Z'_{1\hat{r}^{\{1\}}(\mathbf{X}'_1)}] = \mathbb{E}_{\mathcal{D}^{\{1\}}}[e(\hat{r}^{\{1\}})]. \tag{1.104}$$

We immediately notice that $\mathbb{E}_{\mathcal{D}^{\{1\}}}[e(\hat{r}^{\{1\}})] \simeq e$ only if the learning set $\mathcal{D}^{\{1\}}$ is large enough. Thus, it make sense to select quite small testing sets for increasing the sample size of the learning set. The limit is of course a unique individual, so $\#\mathcal{D}_l^{\{1\}} = 1$. However, restricted excessively the size of the learning set could provide an estimate $\hat{e}^{\{1\}}$ with too large variance.

The principle of V-fold *cross-validation* can then be applied for restricted this variance while preserving a small testing data set. It consists in splitting at random $\mathcal{D}_u$ and $\mathcal{D}_l$ in V blocks of (approximately) equal sizes $\{\mathcal{D}_\ell^{\{1\}}, \ldots, \mathcal{D}_\ell^{\{V\}}\}$ and $\{\mathcal{D}_u^{\{1\}}, \ldots, \mathcal{D}_u^{\{V\}}\}$, respectively, and then to compute the following error estimate

$$\hat{e}^{\text{cv}} = \frac{1}{V}\sum_{v=1}^{V} \hat{e}^{\text{p}}_{\{v\}}. \tag{1.105}$$

Random variables $\hat{e}^{\text{p}}_{\{1\}}, \ldots, \hat{e}^{\text{p}}_{\{V\}}$ having the same distribution but being non-independent, we can verify that it remains an unbiased estimate of $\mathbb{E}_{\mathcal{D}^{\{1\}}}[e(\hat{r}^{\{1\}})]$ since

$$\mathbb{E}_{\mathcal{D}}[\hat{e}^{\text{cv}}] = \mathbb{E}_{\mathcal{D}}[\hat{e}^{\text{p}}_{\{1\}}] = \mathbb{E}_{\mathcal{D}^{\{1\}}}[e(\hat{r}^{\{1\}})], \tag{1.106}$$

while its variance is less than this one of $\hat{e}^{\text{p}}_{\{1\}}$ since

$$\mathbb{V}_{\mathcal{D}}[\hat{e}^{\text{cv}}] = \frac{1}{V^2}\mathbb{V}_{\mathcal{D}}\left[\sum_{v=1}^{V} \hat{e}^{\text{p}}_{\{v\}}\right] < \mathbb{V}_{\mathcal{D}^{\{1\}}}[\hat{e}^{\text{p}}_{\{1\}}]. \tag{1.107}$$

This last inequality is the consequence that two rv Y_1 and Y_2 of same distribution verify $\mathbb{V}[Y_1+Y_2] = 2\mathbb{V}[Y_1] + 2\mathsf{Cov}[Y_1, Y_2]$ and that also $\mathsf{Cov}[Y_1, Y_2] < \mathbb{V}[Y_1]$ if no functional relationship exists between both rv.

The main competitors to $\hat{e}^{\text{cv}}$ are the *Jacknife* estimate Tukey [1958] and also the *bootstrap* estimate Efron [1983]. However, the V-fold cross-validation criterion leads to good results with a low cost of implementation.

Nevertheless, resampling methods like the V-fold cross-validation criterion has two important drawbacks. Firstly, the choice of V may affect the model selection. Secondly, computing V discriminant rules can be time consuming, especially in the semi-supervised setting where unlabelled data require to use an algorithm like EM each time. In the supervised context, this problem vanishes sometimes, as in the Gaussian case where a closed-form updated formula for the discriminant rule is available (Biernacki and Govaert [1999] Appendix A).

1.4.3 A predictive deviance criterion

BEC: A Bayesian entropic criterion

A good approximation of the conditional distribution $f(\mathbf{z}^l|\mathbf{x})$ is expected to produce a good classifier (see Equation (1.9)). Consequently, it makes sense to choose a generative classification model $\mathcal{S}_{\mathbf{m}}$ that gives the largest conditional integrated likelihood $f(\mathbf{z}^l|\mathbf{x},\mathbf{m})$. In this Bayesian perspective, the BEC criterion to be maximized is a BIC-like approximation of $\ln f(\mathbf{z}^l|\mathbf{x},\mathbf{m})$:

$$\text{BEC} = \ln f(\mathcal{D};\hat{\boldsymbol{\theta}}_{\mathcal{D}}) - \ln f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathbf{x}}), \tag{1.108}$$

where $\hat{\boldsymbol{\theta}}_{\mathbf{x}}$ is the MLE of $\hat{\boldsymbol{\theta}}$ derived from $\mathbf{x}$ with the model $\mathcal{S}$. The computational cost of the BEC criterion is approximately twice as large as the computational cost of AIC or BIC, since both $\hat{\boldsymbol{\theta}}_{\mathcal{D}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{x}}$ have to be estimated through an EM algorithm, but it nevertheless remains significantly cheaper than cross-validation.

From a theoretical point of view, if the sampling distribution belongs to a single model of the model collection, this model will be asymptotically selected by BEC (Bouchard and Celeux [2006]). However, when there are several nested true models, BEC can select arbitrarily complex models among them.

From a practical point of view, BEC has been proved to behave better than AIC and BIC for many classification problems, though it often selects more complex generative classifiers than the cross-validated error rate criterion (Bouchard and Celeux [2006]).

AICcond: A predictive deviance criterion

A specific criterion for selecting a classifier in the semi-supervised setting has been proposed by Vandewalle *et al.* [2013]. This criterion is designed to select a generative model that has good classification performances and a low computational cost. It can be seen as a penalized BEC criterion and also as a predictive version of the AIC criterion.

In the frequentist perspective view, when seeking to select a generative classifier with good prediction performances, one particularly interesting quantity is the predictive deviance of the classification model, which is related to the conditional likelihood of the model knowing the predictors. Similarly to the AIC criterion genesis, the aim is to find the model that minimizes an expected Kullback-Leibler divergence. In our case both distributions involved in this divergence are the estimated conditional distribution of $\mathbf{Z}^l|\mathbf{x}$ and the true conditional distribution:

$$2\mathbb{E}_{\mathcal{D}\mathcal{D}'}[\ln f(\mathbf{z}^{l'}|\mathbf{x}') - \ln f(\mathbf{z}^{l'}|\mathbf{x}';\hat{\boldsymbol{\theta}}_{\mathcal{D}})], \tag{1.109}$$

with $\mathcal{D}$ and $\mathcal{D}'$ two independent samples. Since the first term does not depend on the model, it is equivalent to finding the model that maximizes:

$$E_{cond} = 2\mathbb{E}_{\mathcal{D}\mathcal{D}'} \ln f(\mathbf{z}^{l'}|\mathbf{x}';\hat{\boldsymbol{\theta}}_{\mathcal{D}}). \tag{1.110}$$

Proposition 1 in Vandewalle *et al.* [2013] provides the following estimate of E_{cond} under the hypothesis that there is a true model $\mathcal{S}$, that n^l is a realization

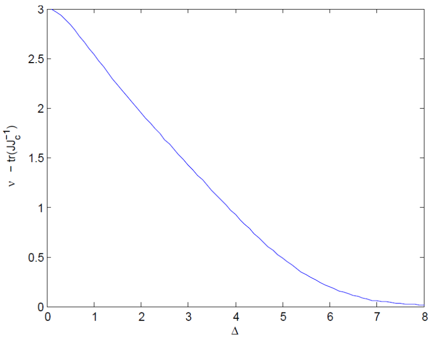


Figure 1.14: Value of the penalty according to the class separation.

of the rv $N^l \sim \mathsf{B}(n,\beta)$, the binomial distribution of parameters n and $\beta \in [0,1]$ (thus $N^l/n \overset{a.s.}{\to} \beta$ when $n \to \infty$), and also that standard regularity conditions hold [Jennrich, 1969; Amemiya, 1973; White, 1981]:

$$E_{cond} = 2[\ln f(\mathcal{D};\hat{\boldsymbol{\theta}}_{\mathcal{D}}) - \ln f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathbf{x}})] - [D - \text{trace}(\mathbf{J}\mathbf{J}_{\beta}^{-1})] + O_p(\sqrt{n}), \quad (1.111)$$

$\mathbf{J}$ and $\mathbf{J}_\beta$ are respectively the Fisher information matrices for unlabelled and partially-labelled data evaluated at the true parameter value $\boldsymbol{\theta}^*$ and already defined in 1.24.

Equation (1.111) exhibits a specific penalty $[D - \text{trace}(\mathbf{J}\mathbf{J}_{\beta}^{-1})]$, which depends on the class overlap and can be related to the number of so-called *predictive parameters* present in the generative model. Indeed, when groups are well-separated, $\mathbf{J} \approx \mathbf{J}_c$ and consequently $\mathbf{J} \approx \mathbf{J}_\beta$ so that $D - \text{trace}(\mathbf{J}\mathbf{J}_{\beta}^{-1}) \approx 0$. Moreover, the more the groups overlap, the larger the value of $D - \text{trace}(\mathbf{J}\mathbf{J}_{\beta}^{-1})$. This claim can be made precise in particular Gaussian situations (see Vandewalle [2009]) and we illustrate it in the following example.

Suppose that data are generated according to the homoscedastic distribution $\mathbf{X}_1|Z_{11} = 1 \sim \mathsf{N}(0,1)$, $\mathbf{X}_1|Z_{12} = 2 \sim \mathsf{N}(\Delta,1)$ and $\pi_1 = \pi_2 = 0.5$. In this case it is possible to compute the penalty. Figure 1.14 displays the value of the penalty according to Δ for a heteroscedastic Gaussian model in the supervised setting ($\beta = 1$). The penalty is largest when the classes are not separated. It is important to note that when $\Delta = 0$ the penalty is equal to the number of parameters involved in the quadratic logistic regression, which corresponds to the predictive expression of the previous Gaussian model.

However, the penalty $D - \text{trace}(\mathbf{J}\mathbf{J}_{\beta}^{-1})$ is difficult to derive, because in a mixture framework the information matrices will need to be computed. For this reason, Vandewalle *et al.* [2013] provide a simple means of approximating it, under the same previous hypotheses:

$$[D - \text{trace}(\mathbf{J}\mathbf{J}_{\beta}^{-1})] = 2(\ln f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathbf{x}}) - \ln f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathcal{D}})) + O_p(\sqrt{n}). \quad (1.112)$$

This gives the following expression for E_{cond}:

$$E_{cond} = 2\ln f(\mathbf{z}^l|\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathcal{D}}) - 4[\ln f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathbf{x}}) - \ln f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathcal{D}})] + O_p(\sqrt{n}), \qquad (1.113)$$

which finally leads to the criterion, to be maximized,

$$\mathrm{AIC}_{cond} = \ln f(\mathbf{z}^l|\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathcal{D}}) - 2\ln\frac{f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathbf{x}})}{f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathcal{D}})}. \qquad (1.114)$$

The approximation error centred at zero involved in AIC_{cond} is relatively high (of order $O_p(\sqrt{n})$) as for AIC. However, note that AIC_{cond} is different from the usual AIC criterion in the predictive setting, even in the absence of additional unlabelled data. In addition, AIC_{cond} can be viewed as an overpenalized BEC criterion, since it can be written

$$\mathrm{AIC}_{cond} = \mathrm{BEC} - \ln\frac{f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathbf{x}})}{f(\mathbf{x};\hat{\boldsymbol{\theta}}_{\mathcal{D}})}. \qquad (1.115)$$

The additional penalty is expected to prevent the appearance of a plateau when considering true nested models, since Vandewalle *et al.* [2013] proved that in case of two nested models $\mathcal{S}_1$ and $\mathcal{S}_2$, with $\mathcal{S}_1 \subset \mathcal{S}_2$, then

$$\mathbb{E}_{\mathcal{D}}[\mathrm{AIC}_{cond_1}] - \mathbb{E}_{\mathcal{D}}[\mathrm{AIC}_{cond_2}] > 0, \qquad (1.116)$$

if the number of data points is large enough and AIC_{cond_k} denoting the value of the AIC_{cond} criterion obtained with the model $\mathcal{S}_k$. Thus, AIC_{cond} tends to prefer the less complex model among two nested true models. Moreover, like BEC, AIC_{cond} selects the right model when there is only one as proved also in Vandewalle *et al.* [2013].

To illustrate the AIC_{cond} behaviour, we retrieve the variable selection example described at the end of Section 1.4.1 but with more values of n^u and n^l. For this experiment, the performances of the cross-validation criterion $\hat{e}^{\mathrm{cv}}$ for $V \in \{1,3\}$ (denoted by $\hat{e}^{\mathrm{cv}}_V$), of BEC and of AIC_{cond} criteria are compared. The results are summarized in tables 1.5 and 1.6, where NbVar* denotes the optimal number of variables derived from the actual error rate function and Err* the corresponding error rate. Those tables show that AIC_{cond} performs the best, since it selects on average the number of variables closest to the optimal number of variables (Table 1.5) and produces a low classification error rate (Table 1.6). Moreover, it has the lowest standard deviations. Cross-validation also produces good results in both settings, while BEC behaves poorly because it selects too many variables. This experiment shows that for nested reliable models, AIC_{cond} leads to the selection of a parsimonious model with good prediction performances, in contrast to BEC.

(n^l, n^u)	BEC	AIC$_{cond}$	$\hat{e}_3^{\mathrm{cv}}$	$\hat{e}_{10}^{\mathrm{cv}}$	NbVar*
(100, 1 000)	17.5 (12.6)	**9.2** (7.8)	10.7 (10.3)	10.0 (9.5)	6 (3.6)
(1 000, 10 000)	33.8 (30.6)	**22.0** (17.8)	21.1 (18.5)	21.4 (25.5)	23 (6.2)

Table 1.5: Variable selection for simulated data: Average number of selected variables for each criterion (best criterion in bold and standard deviations in brackets).

(n^l, n^u)	BEC	AIC$_{cond}$	$\hat{e}_3^{\mathrm{cv}}$	$\hat{e}_{10}^{\mathrm{cv}}$	Err*
(100, 1 000)	30.42 (2.21)	29.75 (1.10)	**29.70** (1.23)	29.82 (1.00)	28.55 (0.54)
(1 000, 10 000)	27.18 (0.34)	**27.17** (0.21)	**27.17** (0.29)	27.21 (0.27)	27.03 (0.12)

Table 1.6: Variable selection for simulated data: Error rate (%) for the different criteria (best criterion in bold and standard deviations in brackets).

1.5 Model selection in clustering

1.5.1 Need to select a model

In the model-based clustering context, the model set involved is potentially very large because it includes the model structure (Gaussian covariance matrices for instance), the number of groups and also the set of discriminant variables[8]. In addition, it is the situation where the data set is the smallest because since it is only composed of data positions $\mathbf{x}$. Finally, in comparison to the density estimation context and to the (semi-)supervised context, the clustering setting is the most difficult for two reasons: variety of models and poor data information.

In the model-based clustering setting, the bias/variance trade-off can be expressed in the following manner. We note $\mathrm{err}(\mathbf{z}_1, \mathbf{z}_2) \geq 0$ a distance-like measure between two partitions $\mathbf{z}_1$ and $\mathbf{z}_2$. When the number of groups in each partition is identical, it can be the classical empirical error rate. When the number of groups differs, it can be for instance the Rand criterion defined in Rand [1971]. We also define, with $\mathbf{z}(\boldsymbol{\theta})$ the MAP derived from $\boldsymbol{\theta}$,

$$\boldsymbol{\theta}^*_{\mathbf{m}} = \arg \min_{\boldsymbol{\theta} \in \Theta_{\mathbf{m}}} \mathrm{err}(\mathbf{z}, \mathbf{z}(\boldsymbol{\theta})) \qquad (1.117)$$

the best parameter associated to the model $\mathcal{S}_{\mathbf{m}}$ with regards to the true partition $\mathbf{z}$. We then have the simple but important following decomposition,

[8] In Maugis *et al.* [2009], variable selection in the Gaussian model-based setting is expressed as a model selection problem. They model differently three kinds of variables: variables interesting for the clustering, variables redundant for the clustering and variables uninteresting for the clustering. Then model/variable selection relies on a BIC criterion for instance (see Chap. **??**, Section **??**).

$\hat{\boldsymbol{\theta}}_{\mathbf{m}} = \hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{m}}$ denoting as before the MLE:

$$\begin{aligned}
&\mathrm{err}(\mathbf{z}, \mathbf{z}(\hat{\boldsymbol{\theta}}_{\mathbf{m}})) \\
&= \Big\{\mathrm{err}(\mathbf{z}, \mathbf{z}(\boldsymbol{\theta}^*_{\mathbf{m}})) - \mathrm{err}(\mathbf{z}, \mathbf{z})\Big\} + \Big\{\mathrm{err}(\mathbf{z}, \mathbf{z}(\hat{\boldsymbol{\theta}}_{\mathbf{m}})) - \mathrm{err}(\mathbf{z}, \mathbf{z}(\boldsymbol{\theta}^*_{\mathbf{m}}))\Big\} \qquad (1.118) \\
&= \Big\{\mathrm{bias}_{\mathbf{m}}\Big\} + \Big\{\mathrm{variance}_{\mathbf{m}}\Big\}. \qquad (1.119)
\end{aligned}$$

We notice again that this bias/variance trade-off differs from the one produced in the density estimation context (see Equation (1.50)). Consequently, the best models in the density setting could be different from the best ones in the clustering setting. In particular, it can be much more dramatic to make a mistake on the number of groups in clustering than in density estimation. Thus, similarly to the (semi-)supervised situation, the question to be addressed in this section is to propose specific model choice criteria taking fully into account the partitioning purpose. Such criteria will involve naturally entropy terms $\xi(\boldsymbol{\theta}, \mathbf{t}(\boldsymbol{\theta}))$ and also conditional probabilities $\mathbf{t}(\boldsymbol{\theta})$.

In order to illustrate the variance effect on the accuracy estimate of the partition, we retrieve the example given in Section 1.3.1 but we display now in Table 1.7 the empirical error estimate err of the partition instead of the Kullback-Leibler divergence. Again, we see that the partition accuracy decreases with the model complexity, revealing the effect of the variance. We note also that the variance decreases with the sample size.

n	$\mathbf{m}$	$\mathrm{err}(\mathbf{z}, \hat{\mathbf{z}}_{\mathbf{m}})$
40	spherical	0.0967
	general	0.1100
200	spherical	0.0840
	general	0.0872

Table 1.7: Effect of the variance of $\hat{\mathbf{z}}_{\mathbf{m}}$ on the partition estimation quality. $\hat{\mathbf{z}}_{\mathbf{m}}$ denotes the partition obtained from the MAP of the estimated parameter $\hat{\boldsymbol{\theta}}_{\mathbf{m}}$.

1.5.2 Partition-based criteria

Criteria *not using* the likelihood term

Some criteria propose to retain the model leading to the best group separability. It is the case of the so-called PC criterion (*Partition Coefficient*) of Bezdeck [1981] which sums the square of all conditional probabilities $\mathbf{t}(\hat{\boldsymbol{\theta}})$. There is also the so-called MIR criterion (*Minimum Information Ratio*) of Windham and Cutler [1992], and its variants, involving a ratio of the complete-data Fisher information matrix $\mathbf{J}_c(\hat{\boldsymbol{\theta}})$ and of the observed-data Fisher information matrix

$\mathbf{J}(\hat{\boldsymbol{\theta}})$. This ratio gives a measure of the ability of the data set to be partitioned with the model. Generally, these criteria have poor theoretical justification and are also difficulty to apply for distinguishing $K = 1$ (no structure) from $K > 1$. To overcome this drawback, there is a need to aggregate a measure of the model adequacy to the measure of partitioning ability. The log-likelihood value can reach this task as we now show.

Criteria *using* the likehood term

The entropy term $\xi(\boldsymbol{\theta}; \mathbf{t}(\boldsymbol{\theta}))$ measures the groups overlap: a small value indicates poor overlap between groups whereas a large value corresponds to strong overlap. The following fundamental relationship between the log-likelihood and the entropy is given by Hathaway [1986]:

$$\ell(\boldsymbol{\theta}; \mathbf{x}) = \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}(\boldsymbol{\theta})) + \xi(\boldsymbol{\theta}; \mathbf{t}(\boldsymbol{\theta})). \tag{1.120}$$

The NEC criterion (*Normalized Entropy Criterion*) of Celeux and Soromenho [1996] and Biernacki *et al.* [1999] is established from this link. It is expressed as a normalization of the entropy by two log-likelihood terms:

$$\mathrm{NEC}_K = \begin{cases} \dfrac{\xi_K}{\ell_K - \ell_1} & \text{if } K > 1 \\ 1 & \text{if } K = 1 \end{cases} \tag{1.121}$$

with $\ell_k = \ell(\hat{\boldsymbol{\theta}}_k; \mathcal{D})$ and $\xi_k = \xi(\hat{\boldsymbol{\theta}}_k; \hat{\mathbf{t}}_k)$ where $\hat{\boldsymbol{\theta}}_k$ is the MLE for k groups and $\hat{\mathbf{t}}_k = \mathbf{t}(\hat{\boldsymbol{\theta}}_k)$. It has to be noticed that $\hat{\boldsymbol{\theta}}_K$ and $\hat{\boldsymbol{\theta}}_1$ must be obtained with the same constraints on the parameters (for instance, in the Gaussian case, a spherical model for both numbers of groups). We retain then the model $\mathcal{S}_{\mathbf{m}}$ with the lowest $\mathrm{NEC}_{\mathbf{m}}$ value. The NEC value itself appears to be meaningful since the partitioning evidence is associated to NEC values less than 1.

Another approach has been proposed to merge the log-likelihood and an entropic term. The retained criterion in Biernacki and Govaert [1997] is simply the complete-data log-likelihood $\ell(\hat{\boldsymbol{\theta}}; \mathbf{x}, \hat{\mathbf{z}})$, $\hat{\mathbf{z}}$ being the MAP of $\hat{\boldsymbol{\theta}}$. It corresponds to the so-called CL criterion (*Completed Likelihood*):

$$\mathrm{CL} = \ell(\hat{\boldsymbol{\theta}}; \mathbf{x}, \hat{\mathbf{z}}) = \ell(\hat{\boldsymbol{\theta}}; \mathbf{x}) - \xi(\hat{\boldsymbol{\theta}}; \hat{\mathbf{z}}). \tag{1.122}$$

The retained model is this one leading to the largest CL value. This criterion can be seen as the maximum log-likelihood value combined with an entropic penalty term indicating the group overlapping. It is thus quite different from the AIC or the BIC criteria for which the penalty term is related to the model complexity. This entropic term corresponds also to minus the logarithm of the conditional probability of the partition $\hat{\mathbf{z}}$, since we can write $f(\hat{\mathbf{z}}|\mathbf{x}; \hat{\boldsymbol{\theta}}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \hat{t}_{ik}^{\hat{z}_{ik}}$. Thus, the quantity $\xi(\boldsymbol{\theta}; \mathbf{z})$ measures a dissimilarity between the conditional probabilities $\mathbf{t}$ and the partition $\mathbf{z}$ which is the closest from a certain point of view.

We have to note that both NEC and CL show a certain ability to select K but fail to select other kinds of models like the Gaussian structure on the covariance matrices. It seems to lack a penalty term involving the model complexity.

1.5.3 The Integrated Completed Likelihood criterion

Integrated completed likelihood for model selection

We remember that in the clustering context, observed data are restricted to $\mathcal{D} = \mathbf{x}$. In a Bayesian context, model selection was thus relying on the calculus of the *observed-data* integrated likelihood $f(\mathbf{x}|\mathbf{m})$ given in (1.76) in Section 1.3.3. If complete data $(\mathbf{x}, \mathbf{z})$ were known, model selection would be similarly performed by retaining the model $\mathcal{S}_\mathbf{m}$ maximizing the *complete-data* integrated likelihood $f(\mathbf{x}, \mathbf{z}|\mathbf{m})$ expressed in (1.90). The following straightforward relationship exists between the integrated complete-data and observed-data likelihoods:

$$\ln f(\mathbf{x}, \mathbf{z}|\mathbf{m}) = \ln f(\mathbf{x}|\mathbf{m}) + \ln f(\mathbf{z}|\mathbf{x}, \mathbf{m}). \tag{1.123}$$

Thus, as already noticed in Biernacki *et al.* [2011], the *complete-data* integrated likelihood can be interpreted as the classical integrated likelihood penalized by a measure of the cluster overlap expressed through $f(\mathbf{z}|\mathbf{x}, \mathbf{m})$. It means that it tends to realize a compromise between the adequacy of the model to the data measured by $\ln f(\mathbf{x}|\mathbf{m})$ and the evidence of data partitioning measured by $\ln f(\mathbf{z}|\mathbf{x}, \mathbf{m})$. For instance, highly overlapping mixture components typically lead to a low value of $f(\mathbf{z}|\mathbf{x}, \mathbf{m})$ and consequently dos not favour a high value of $f(\mathbf{x}, \mathbf{z}|\mathbf{m})$. However, the partition $\mathbf{z}$ being hidden in clustering, Biernacki *et al.* [2011] propose to replace it by its MAP estimate $\hat{\mathbf{z}}_\mathbf{m}$ associated to the MLE $\hat{\boldsymbol{\theta}}_\mathbf{m} = \hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{m}}$. Then, it gives the so-called ICL (*Integrated Completed Likelihood*) criterion which retains the model $\mathcal{S}_\mathbf{m}$ associated to its maximum value[9]:

$$\mathrm{ICL}_\mathbf{m} = \ln f(\mathbf{x}, \hat{\mathbf{z}}_\mathbf{m}|\mathbf{m}). \tag{1.125}$$

The question we now address is how to practically calculate ICL and also to identify its properties.

Asymptotic approximation

Biernacki *et al.* [2000] propose to proceed in two steps for approximating the previous ICL criterion. First, they use a BIC-like approximation of the *complete-data* integrated likelihood:

$$\ln f(\mathbf{x}, \mathbf{z}|\mathbf{m}) = \ln f(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{z},\mathbf{m}}|\mathbf{m}) - \frac{D_\mathbf{m}}{2} \ln n + O_p(1), \tag{1.126}$$

[9] Another definition of ICL is also used sometimes, with $\hat{\mathbf{t}} = \mathbf{t}(\hat{\boldsymbol{\theta}})$:

$$\mathrm{ICL}_\mathbf{m} = \ln f(\mathbf{x}, \hat{\mathbf{t}}_\mathbf{m}|\mathbf{m}). \tag{1.124}$$

where $\hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{z},\mathbf{m}}$ denotes the MLE associated to complete data $(\mathbf{x},\mathbf{z})$ with model $\mathcal{S}_{\mathbf{m}}$. But, in case of the right model $\mathcal{S}_{\mathbf{m}}$, we have both $\hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{z}} \overset{a.s.}{\to} \boldsymbol{\theta}^*$ and $\hat{\boldsymbol{\theta}} \overset{a.s.}{\to} \boldsymbol{\theta}^*$, $\hat{\boldsymbol{\theta}}$ still denoting the MLE associated to $\mathbf{x}$ and also index $\mathbf{m}$ being omitted. Thus, for n large enough, we can make the approximation $\hat{\boldsymbol{\theta}}_{\mathbf{x},\mathbf{z}} \approx \hat{\boldsymbol{\theta}}$. Then, we replace the missing cluster indicators $\mathbf{z}$ by their MAP values $\hat{\mathbf{z}}$ associated to the MLE $\hat{\boldsymbol{\theta}}$. It finally leads to the so-called ICLbic criterion[10]

$$\text{ICLbic} = \ln f(\mathbf{x},\hat{\mathbf{z}};\hat{\boldsymbol{\theta}}) - \frac{D}{2}\ln n. \qquad (1.128)$$

Remark that the so-called AWE criterion (*Approximate Weight of Evidence*) also proposed in a Bayesian context by Banfield and Raftery [1993] is very similar to ICLbic. However, it uses the complete-data estimate $\hat{\boldsymbol{\theta}}_c$ defined in (1.29) and it penalizes more strongly the number of parameters.

By some simple algebra, The ICLbic criterion can also be viewed either as a partition complexity (measured by an entropy-like term) penalized version of the BIC criterion or as a model complexity penalized version of the CL criterion:

$$\begin{aligned} \text{ICLbic} &= \text{BIC} - \xi(\hat{\boldsymbol{\theta}};\hat{\mathbf{z}}) && (1.129)\\ &= \text{CL} - \frac{D}{2}\ln n. && (1.130) \end{aligned}$$

Robustness of ICL to model misspecification

This trade-off between the model adequacy (log-likelihood), the model complexity (number of parameters) and the partitioning evidence (entropy) provides robustness properties for the ICL/ICLbic criterion as we now illustrate. We consider experiments from a bivariate mixture of a uniform and a Gaussian cluster. One of the 50 simulated data sets of size $n = 200$ is displayed in Figure 1.15 and the mixture characteristics are as follows:

- non-Gaussian component: $\pi_1 = 0.5$, $f_1(\mathbf{x}_1) = 0.25\ \mathbf{1}_{[-1,1]}(x^1)\ \mathbf{1}_{[-1,1]}(x^2)$ where $\mathbf{1}_{[-1,1]}$ denotes the indicator function in the interval $[-1,1]$;
- Gaussian component: $\pi_2 = 0.5$, $\boldsymbol{\mu}_2 = (3.3,0)'$, $\boldsymbol{\Sigma}_2 = \mathbf{I}$.

When running the EM algorithm, only the most simple spherical model is considered and K is varying from one to five. Percentage of times K is chosen is displayed in Table 1.8. In this case BIC has a disappointing behaviour. This example highlights a well-known tendency of this criterion: when the

[10] The following other definition is also widely used:

$$\text{ICLbic} = \ln f(\mathbf{x},\hat{\mathbf{t}};\hat{\boldsymbol{\theta}}) - \frac{D}{2}\ln n. \qquad (1.127)$$

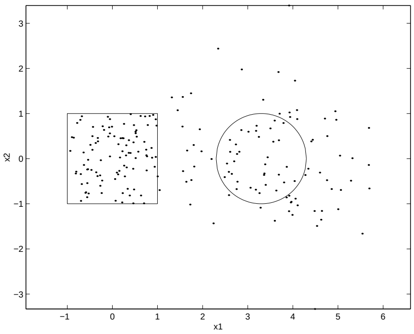


Figure 1.15: A uniform and a Gaussian component.

clustering model at hand (here a Gaussian mixture model) does not fit well the data, BIC tends to overestimate the number of components. On the contrary, ICLbic includes an entropic term $\xi(\hat{\boldsymbol{\theta}};\hat{\mathbf{z}})$ which penalizes overlapping groups and which balances the lack of fit of the data in the model at hand . Thus, ICL is expected to be more robust to violations of the model specifications than BIC, as it appears in this experiment.

K	1	2	3	4	5
BIC	.	**60**	.	32	8
ICLbic	.	**100**	.	.	.

Table 1.8: Non-Gaussian component samples: percentage of times K is chosen with the spherical Gaussian model.

Question on the consistency of ICL

A counterpart of this robustness of ICL/ICLbic is that it is not consistent for the number of components if their overlap is two high. Indeed, ICLbic tends to underestimate the true number of components in this situation, even asymptotically. We illustrate this fact from both a theoretical and a practical point of view in the simple situation where two components are really present.

We note $\delta_n = n(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2^{*p})'\mathbf{J}(\boldsymbol{\theta}_2^*)(\boldsymbol{\theta}_2^* - \boldsymbol{\theta}_2^{*p})$ with $\mathbf{J}(\boldsymbol{\theta}_2^*)$ the Fisher matrix for a data unit calculated with the true parameter $\boldsymbol{\theta}_2^*$ (see Equation (1.24)) and $\boldsymbol{\theta}_2^{*p}$ its projected value on the parameter subspace associated to the one component case. Moreover, denoting by $\chi_a^2(b)$ a rv with the non-central χ^2 distribution with a degrees of freedom and non-centrality parameter b, we define $\mu_n = \mathbb{E}[\chi^2_{\Delta D}(\delta_n)] = \Delta D + \delta_n$, $\sigma_n^2 = \mathbb{V}[\chi^2_{\Delta D}(\delta_n)] = 2(\Delta D + \delta_n)$, $\Delta D = D_2 - D_1$, $\Delta\xi = \xi(\hat{\boldsymbol{\theta}}_2;\hat{\mathbf{z}}(\hat{\boldsymbol{\theta}}_2)) - \xi(\hat{\boldsymbol{\theta}}_1;\hat{\mathbf{z}}(\hat{\boldsymbol{\theta}}_1))$ with $\hat{\mathbf{z}}(\hat{\boldsymbol{\theta}}_K)$ the MAP partition obtained from

$\mathbf{t}(\hat{\boldsymbol{\theta}}_K)$ and finally $\Delta\ell = \ell(\hat{\boldsymbol{\theta}}_2;\mathcal{D}) - \ell(\hat{\boldsymbol{\theta}}_1;\mathcal{D})$. The probability of choosing the wrong model (one group instead two groups) by ICLbic is given by

$$p(\text{ICLbic}_2 < \text{ICLbic}_1) = p(2\Delta\ell < \Delta D\ln n{+}2\Delta\xi) \leq p(2\Delta\ell < \Delta D\ln n{+}2n\ln 2), \tag{1.131}$$

the last inequality being implied by $\Delta\xi < n\ln 2$ (the entropy of two components is higher than that for one component). Noting now that $2\Delta\ell$ and $\chi^2_{\Delta D}(\delta_n)$ have asymptotically the same distribution, then the probability of choosing the wrong model by ICLbic is asymptotically less than

$$p(\chi^2_{\Delta D}(\delta_n) < \Delta D\ln n + 2n\ln 2) \leq p(|\mu_n - \chi^2_{\Delta D}(\delta_n)| > \mu_n - \Delta D\ln n - 2n\ln 2). \tag{1.132}$$

Finally, the Chebishev inequality gives

$$p(\chi^2_{\Delta D}(\delta_n) < \Delta D\ln n + 2n\ln 2) \leq \frac{\sigma_n^2}{(\mu_n - \Delta D\ln n - 2n\ln 2)^2} \overset{n\to\infty}{\longrightarrow} 0, \tag{1.133}$$

provided that $\mu_n - \Delta D\ln n - 2n\ln 2 > 0$, thus provided that the two components are sufficiently separated since μ_n is a measure of the overlapping. In addition, noting that μ_n and $-2n\ln 2$ are of same order with n, then the IClbic consistency is not guaranteed for a quite large degree of overlapping, even asymptotically.

We now numerically illustrate the fact that ICLbic can be inconsistent, even asymptotically, if components are not well-separated. We draw 100 samples of sizes $n = 100, 400, 700, 1\ 000$ from a univariate Gaussian mixture with same proportions, with unit variances and with a distance between the two centres successively equal to $\Delta\mu = 2.9, 3.0, 3.1, 3.2, 3.3$. The EM algorithm is then run with a model with one and two components on all 100 samples and for all values of n and $\Delta\mu$. Table 1.9 displays the percentage of times the right number of components (two) is chosen by ICLbic and by BIC. We clearly identify a threshold around $\Delta\mu = 3.0$ where ICLbic switches from non consistency towards consistency.

$\Delta\mu$	2.9		3.0		3.1		3.2		3.3	
n	BIC	ICL	BIC	ICL	BIC	ICL	BIC	ICL	BIC	ICL
100	94	23	96	31	97	44	95	45	97	60
400	100	9	100	21	100	48	100	70	100	85
700	100	8	100	15	100	39	100	72	100	96
1 000	100	6	100	16	100	56	100	75	100	91

Table 1.9: Percentage of times two components is chosen as a function of their overlapping .

ICL with a new contrast point of view

Alternatively, Baudry [2012] considers that ICLbic is a criterion relying on the (fuzzy) complete-data log-likelihood $\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}(\boldsymbol{\theta}))$, instead of the log-likelihood $\ell(\boldsymbol{\theta}; \mathbf{x})$. From Equation (1.120), it can be rewritten as the following penalized log-likelihood:

$$\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}(\boldsymbol{\theta})) = \ell(\boldsymbol{\theta}; \mathbf{x}) - \xi(\boldsymbol{\theta}; \mathbf{t}(\boldsymbol{\theta})). \tag{1.134}$$

This author proposes the following new ICLbic-like criterion

$$\widetilde{\text{ICLbic}} = \ell(\tilde{\boldsymbol{\theta}}; \mathbf{x}, \mathbf{t}(\tilde{\boldsymbol{\theta}})) - \frac{D}{2} \ln n, \tag{1.135}$$

where

$$\tilde{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta} \in \Theta} \ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}(\boldsymbol{\theta})). \tag{1.136}$$

Thus $\widetilde{\text{ICLbic}}$ is here a penalized contrast with a BIC-like penalty. It no longer involves any entropic penalty because here entropy is a part of the contrast itself. This criterion is then proved to be consistent (only) from this new contrast point of view. It appears that the ICLbic and $\widetilde{\text{ICLbic}}$ criteria are very close both by their expressions and by their numerical behaviour. In addition, since $\tilde{\boldsymbol{\theta}}$ is more difficult to obtain that the MLE $\hat{\boldsymbol{\theta}}$, ICLbic could be preferred.

Note that Baudry [2012] also proposes to use the slope heuristics to obtain a data-driven penalty associated to the contrast $\ell(\boldsymbol{\theta}; \mathbf{x}, \mathbf{t}(\boldsymbol{\theta}))$.

Combining ICL and BIC

Baudry *et al.* [2010] proposed to combine BIC and ICL in the following manner for obtaining the model flexibility given by BIC while preserving the clustering evidence given by ICL. Firstly, they choose the number of *components* by BIC. Secondly, they merge the more overlapped components in order to obtain the number of *groups* initially proposed by ICL. Finally, a mixture of mixture is obtained: a group may be composed by several components. Other strategies of combinations are possible by looking directly at the entropy value.

Combining ICL and an external partition

Baudry *et al.* [2012a] assumed that an *external* partition $\mathbf{y}$ with J groups is known and proposed to use it to reveal an (unknown) *internal* partition $\mathbf{z}$ with K groups. Noting $n_{jk} = \#\{i : y_{ij} = 1 \text{ and } z_{ik} = 1\}$ the elements of the contingency table cross-tabulating $\mathbf{y}$ and $\mathbf{z}$, and noting also $n_{.k} = \sum_{j=1}^{J} n_{jk}$, they derived the so-called SICL criterion (*Supervised Integrated Completed Likelihood*) expressed by

$$\text{SICL} = \text{ICL} + \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk} \ln \frac{n_{jk}}{n_{.k}}. \tag{1.137}$$

The last additional term quantifies the strength of the link between both partitions, making a subtle trade-off between model adequacy, evidence of partitioning $\mathbf{z}$ and also accordance between partitions $\mathbf{y}$ and $\mathbf{z}$.

Exact ICL criterion for the latent class model

We have seen in Section 1.3.3 that conjugate Jeffreys non informative prior distributions are available for all the parameters of the latent class model. Thus, using the associated closed-form of the integrated complete-data likelihood given in (1.92) and then replacing the missing labels $\mathbf{z}$ by $\hat{\mathbf{z}}$ in $\ln f(\mathbf{x}, \mathbf{z})$, we obtain the following non-asymptotic expression for the ICL criterion:

$$\begin{aligned}\mathrm{ICL} = \ln f(\mathbf{x}, \hat{\mathbf{z}}) = &\\ &\sum_{k=1}^{K}\sum_{j=1}^{d}\left\{\sum_{h=1}^{m_j}\ln\Gamma\left(\hat{n}_k^{jh}+\tfrac{1}{2}\right)-\ln\Gamma(\hat{n}_k+\tfrac{m_j}{2})\right\}-\ln\Gamma(n+\tfrac{K}{2})+\ln\Gamma(\tfrac{K}{2})\\ &+K\sum_{j=1}^{d}\left\{\ln\Gamma(\tfrac{m_j}{2})-m_j\ln\Gamma(\tfrac{1}{2})\right\}+\sum_{k=1}^{K}\ln\Gamma(\hat{n}_k+\tfrac{1}{2})-K\ln\Gamma(\tfrac{1}{2}), \quad (1.138)\end{aligned}$$

where $\hat{n}_k = \#\{i : \hat{z}_{ik} = 1\}$ and $\hat{n}_k^{jh} = \#\{i : \hat{z}_{ik} = 1, x_i^{jh} = 1\}$.

In order to illustrate the ICL and the ICLbic behaviour, we consider observations described by six variables ($d = 6$) with numbers of levels $m_1 = \ldots = m_4 = 3$ and $m_5 = m_6 = 4$ and a two component mixture ($K = 2$) with unbalanced mixing proportions $\boldsymbol{\pi} = (0.3, 0.7)$. The parameter $\boldsymbol{\alpha}$ is chosen to get successively a low cluster overlapping (about 5% of error rate), a middle overlapping (about 10% of error rate) and a high overlapping (about 20% of error rate), to be compared to the worst error rate equal to 30%. Detail of parameter values is given in Biernacki *et al.* [2011]. Figure 1.16 displays a data sample on the first two axes of a correspondence analysis. 20 samples are generated for three different sample sizes $n \in \{320, 1\,600, 3\,200\}$. For each sample, the EM algorithm has been run 10 times with random initial parameters (uniform distribution on the parameter space) for a sequence of 1 000 iterations. The mean of the retained number of mixture components with ICL and ICLbic criteria is displayed on Table 1.10. We notice that ICL has ability to detect structures with lower sample sizes than ICLbic. In addition, we notice again that ICL/ICLbic are not consistent when the overlapping is too high.

1.6 Experiments on real data sets

In this section, we illustrate the behaviour of numerous criteria described in the previous three sections on various real data sets. It gathers the three settings of density estimation, semi-supervised classification and clustering. At the same

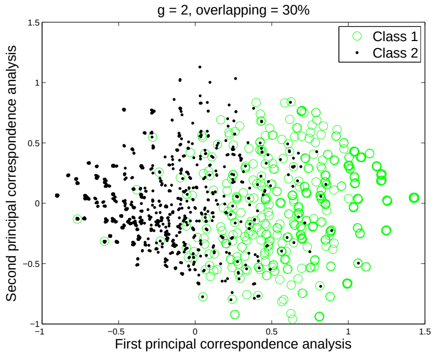


Figure 1.16: A sample ($n = 1\,600$) arising from a $K = 2$ mixture situation for medium overlapping. It is displayed on the first plane of a correspondence analysis and an i.i.d. uniform noise on $[0, 0.01]$ has been added on both axes for each point in order to clarify the visualisation.

n	320			1 600			3 200		
Overlap (%)	5	10	20	5	10	20	5	10	20
ICLbic	2.0	1.5	1.0	2.0	2.0	1.0	2.0	2.0	1.0
ICL	2.0	1.9	1.0	2.0	2.0	1.0	2.0	2.0	1.0

Table 1.10: Mean of the chosen number of groups for ICL and ICLbic criteria when $K = 2$ for the latent class model.

time, it is the opportunity to discover their use with mixture models dedicated to particular kinds of data: interval data, rank data, mixed data…

1.6.1 BIC: extra-solar planets

In numerous fields, the collected data are available only in grouped form, *i.e.* their exact position inside a given subset, or bin, is unknown. Grouped data may occur systematically when a measurement instrument has finite resolution but it may also occur intentionally when real-valued variables are quantized to simplify data collection. In the context of Gaussian mixtures, some features has already been studied for such data. In particular, McLachlan and Jones [1988] and Cadez *et al.* [2002] adapted the EM algorithm in order to reach the MLE for both univariate and multivariate normal mixtures. Since the bin dimension is a crucial feature for grouped data, Cadez *et al.* [2002] performed also some simulation experiments to observe the effect of the bin dimension on the MLE of the mixture parameter in the case of a two-component bivariate Gaussian mixture. They note that increasing the bin dimension obviously decreases the quality of the MLE although substantial differences between both MLE of grouped and individual data are obtained only with quite wide bins. But, as far as we know, the effect of the bin dimension on model selection problems has not yet been studied. Thus, the aim of this experiment is to study the influence of data precision on the BIC behaviour for selecting a model, in particular here the number of components in a Gaussian mixture.

We consider extra-solar planets from single planetary systems for which both mass and eccentricity are not exactly known at the date of June 25 2004. Data are obtained from the Paris Observatory[11]. Mass (measured in Jupiters, one Jupiter mass corresponding to 318 Earths), eccentricity and the associated uncertainty for both variables are given for the 10 concerned planets in Table 1.11. Figure 1.17(a) displays this data set and it shows that uncertainty is often very high.

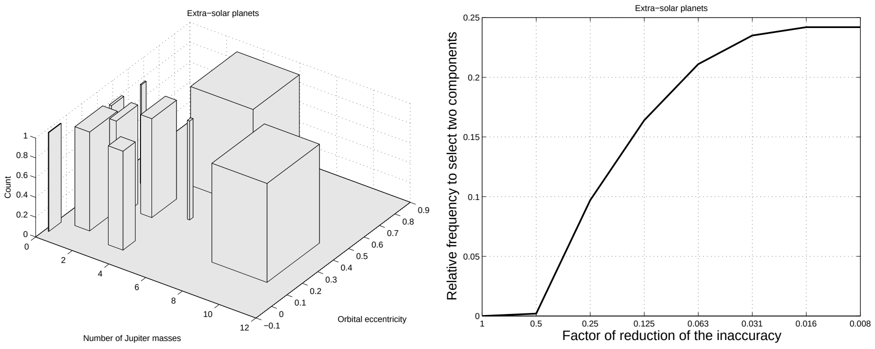
Name of the planetary system	Jupiter Mass		Eccentricity	
HD 76700	0.197	± 0.017	0.00	± 0.04
HD 217107	1.28	± 0.4	0.14	± 0.09
HD 195019	3.43	± 0.4	0.05	± 0.04
HD 52265	1.13	± 0.06	0.29	± 0.04
HD 73526	3.0	± 0.3	0.34	± 0.08
HR 810	1.94	± 0.18	0.24	± 0.07
HD 210277	1.24	± 0.03	0.450	± 0.015
HD 2039	4.85	± 1.7	0.68	± 0.15
Gl 614	4.74	± 0.06	0.338	± 0.011
HD 30177	9.17	± 1.5	0.30	± 0.17

Table 1.11: Extra-solar planets from single planetary systems for which both mass and eccentricity are not exactly known at the date of June 25 2004 (source: Extra-solar Planets Catalog of the Paris Observatory at `http://www.obspm.fr/encycl/cat1.html`).

Retaining the homoscedastic diagonal model with free mixing proportions, the EM algorithm is launched on the extra-solar data set for one and two components. In this situation, the BIC criterion selects only one component.

However, in the future, we can reasonably expect a reduction of uncertainty by the evolution of the measurement instruments. Thus, we propose to study the influence of decreasing uncertainty on the number of components (between 1 and 2) selected by the BIC criterion. To this end, we artificially decrease the bin dimensions of both mass and eccentricity by multiplying each side of all rectangles of uncertainty successively by factors 0.5^u where $u = 1, \ldots, 7$. Obviously, we do not know where to place the narrower rectangles inside the rectangles of the initial data set. Consequently, for each $u = 1, \ldots, 7$, 1 000

[11] `http://www.obspm.fr/encycl/cat1.html`



(a) (b)

Figure 1.17: Extra-solar planets: (a) initial data, (b) frequency to select two components by BIC for each uncertainty decreasing factor 0.5^u ($u = 0, \ldots, 7$).

data sets are generated in the following manner: for each of the 10 planets, the associated uncertainty rectangle is uniformly drawn inside the initial uncertainty rectangle. Then, the EM algorithm is run again for the 1 000 × 7 artificial data sets. Figure 1.17(b) displays the relative frequency of choosing two components by BIC among 1 000 replications for each 0.5^u value of the decreasing factor ($u = 1, \ldots, 7$). Note that the selected number of components for the initial data set is also available in this figure: it corresponds to a factor $0.5^0 = 1$.

We remark that, when uncertainty decreases, the frequency of choosing two components regularly increases. It becomes stable at about 0.24 from a factor equal to 0.5^6. From an astronomic point of view, the probability of having two components will increase when the accuracy will become better. For instance, dividing uncertainty by 4 (it means multiplying by a factor $0.5^2 = 0.25$ on the figure) may lead to a new data set with probability of around 0.1 (i.e. 10%) that BIC discovers two components. If uncertainty completely disappears in the future (so all data are exactly known), then the probability of having an individual data set with two components by the BIC criterion is about 0.24 (*i.e.* approximately a quarter), the frequency value obtained with the very small bin dimensions 0.5^6 and 0.5^7.

1.6.2 AIC$_{cond}$/BIC/AIC/BEC/$\hat{e}^{\text{cv}}$: benchmark data sets

We compare now the behaviour of the previous semi-supervised classification specific criteria (BEC, AIC$_{cond}$, $\hat{e}^{\text{cv}}$) to general density estimation criteria (AIC, BIC) on some real data sets. Results are extracted from Vandewalle *et al.* [2013]. In each case, the RMIXMOD[12] software has been used. We consider

[12] http://www.mixmod.org/ and http://cran.r-project.org/web/packages/Rmixmod/

benchmark data sets from the UCI database repository[13] and Pattern Recognition data sets[14]. Performances of criteria for selecting a Gaussian model are compared among the six following constraints on *homoscedastic* covariance matrices: spherical (with equal or free volume), diagonal (*idem*) and general (*idem*). Features of the data sets are summarized in Table 1.12. If a test set is provided, its predictors are used to learn the parameters of the classification models in the semi-supervised setting and its labels are used to compute the error rate. Otherwise, 100 random splits of n^u unlabelled data and n^l labelled data are generated. Table 1.13 shows that AIC$_{cond}$, BEC and cross-validation have a similar behaviour and outperform BIC and AIC, as is the case for the Parkinson and Pima data sets.

Dataset	n	d	K	Test set	n^u	n^l
Crab	200	5	4	no	150	50
Iris	150	4	3	no	100	50
Parkinson	195	22	2	no	95	100
Pima	532	7	2	yes	332	200
Wine	178	13	3	no	89	89

Table 1.12: Variable parameter selection for benchmark data sets: Experimental setting.

	BIC	AIC	BEC	AIC$_{cond}$	$\hat{e}_3^{\text{cv}}$	$\hat{e}_{10}^{\text{cv}}$
Crab	**6.63**	6.75	6.80	6.77	7.81	7.78
Iris	2.98	2.98	**2.91**	**2.91**	3.25	3.21
Parkinson	26.45	30.68	15.43	**15.16**	18.20	16.38
Pima	25.00	25.00	**19.58**	**19.58**	22.53	**19.58**
Wine	3.24	**1.17**	1.45	1.47	1.73	1.70

Table 1.13: Variable parameter selection for benchmark data sets: error rate of each criterion on UCI data sets (the criterion producing the lowest error rate is shown in bold).

1.6.3 AIC$_{cond}/\hat{e}_V^{\text{cv}}$: textile data set

We now consider a three-class problem extracted from Vandewalle *et al.* [2013]. The RMIXMOD software has been used. The data are the near infra red (NIR) spectra of different manufactured textile materials. The three-class NIR data set contains 223 NIR spectra of manufactured textiles of various compositions. The classification problem is to recover the physical characterisation of the textiles, which can take three values Devos *et al.* [2009]. The data were naturally separated into a learning sample (132 textiles) and a test sample (91 textiles)

[13] http://archive.ics.uci.edu/ml/

[14] http://www.stats.ox.ac.uk/pib/PRNN/

with the labels of the test sample initially unknown. The NIR spectra were measured on an XDS rapid content analyzer instrument in reflectance mode in the range $1100-2500$ *nm* at 0.5 *nm* apparent resolution (2 800 data points per spectrum). Standard Gaussian models are too complex for this data set, since the number of variables is too large. Parsimonious high-dimensional Gaussian models can be used Jacques *et al.* [2010], although the large number of tuning parameters make these unattractive in the semi-supervised setting.

A variable pre-selection step is performed, based on the analysis of variance (ANOVA) Toher *et al.* [2005]. For each variable an ANOVA is performed with respect to the class membership of the data, and the F statistic is plotted according to the variable number in Figure 1.18. This preprocessing step searches for the most discriminant variables, taking into account its ordered nature. As remarked in Toher *et al.* [2005], this method is competitive with wavelets for NIR data. It can be seen that the F statistic presents 20 peaks, each variable corresponding to a peak yielding more information than its neighbours. These 20 variables are chosen and sorted in decreasing order of F statistic. The model selection problem is then equivalent to choosing the right number of variables among those 20 ordered variables. In this setting, a general quadratic Gaussian model is used. Error rates with respect to the number of selected variables are presented in Table 1.14. As expected, this error rate computed on the test sample decreases and then increases according to the number of selected variables. The optimal number of variables is 13 and 14, which produces an error rate of 7.69%, which is in accordance with the error rates produced by other methods on these data (8.8% with SVM Devos *et al.* [2009]). The selection criteria $\hat{e}_3^{\text{cv}}$, $\hat{e}_{10}^{\text{cv}}$, BEC and AIC$_{cond}$ are compared in a semi-supervised setting, where the test sample is used as an unlabelled sample to improve the classification function. Table 1.15 shows that the three criteria produce good results, AIC$_{cond}$ and BEC performing the best.

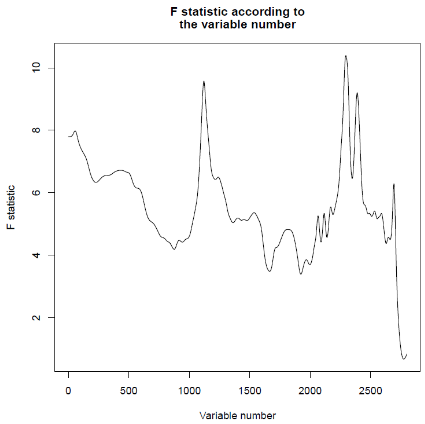


Figure 1.18: F statistic according to the variable number.

Nb of variables	1	2	3	4	5	6	7	8	9	10
Error rate (%)	64.84	59.34	26.37	27.47	28.57	19.78	24.18	20.88	18.68	18.68
Nb of variables	11	12	13	14	15	16	17	18	19	20
Error rate (%)	18.68	12.09	7.69	7.69	9.89	7.69	10.99	10.99	18.68	20.88

Table 1.14: Error rate according to the number of selected variables.

Criterion	Number of variables	Error rate (%)
AIC$_{cond}$	14	7.69
BEC	13	7.69
$\hat{e}^{\mathrm{cv}}_{10}$	15	9.89
$\hat{e}^{\mathrm{cv}}_{3}$	10	18.68

Table 1.15: Number of selected variables and resulting error rate according to the criterion.

1.6.4 BIC: social comparison theory

The following data set has been provided by Dr Hans Kuyper who is a researcher at the Faculty of Behavioural and Social Sciences at the University of Groningen (The Netherlands). His research domain is “social comparison theory”. It is known that most persons compare themselves with others, in order to evaluate themselves, to get positive feelings, or to improve themselves. More specifically, his interest goes to the question of knowing along which dimensions persons prefer to compare themselves, given a free choice situation. It is original since in most research there is no free choice, as the comparison dimension is part of the experimental design. The subject of the present research topic, therefore, is “preference for comparison dimensions”.

All his research is in secondary education. The present data were collected in third classes (US grade 9), when most students were 15 years. The data were collected with a questionnaire, during regular school time. The social comparison items were one part of the questionnaire. The tasks in the questionnaire had to be suitable for students of all ability levels. The Dutch system of secondary education is highly tracked (one of the most tracked systems in the world). In the social comparison part of the questionnaire were several subtopics. This part started with a few remarks about comparing with others, for instance that it is quite normal to do such thing. The second social comparison question was as follows: “Which things do you prefer to compare with other children of your age? Put a 1 in front of what you prefer to compare most, a 2 in front of what you prefer next, and so on. More than 3 is not necessary, but is allowed”. We offered 13 “objects” $\mathcal{O}_j$ ($j = 1, \ldots, 13$), *i.e.* aspects or dimensions from which the students could choose: $\mathcal{O}_1$) “your popularity”, $\mathcal{O}_2$) “how well you do in sports”, $\mathcal{O}_3$) “your appearance”, $\mathcal{O}_4$) “how much money you can spend”, $\mathcal{O}_5$) “how you are feeling”, $\mathcal{O}_6$) “your parents”, $\mathcal{O}_7$) “your clothes”, $\mathcal{O}_8$) “your grades at school", $\mathcal{O}_9$) "how well you can express your opinions", $\mathcal{O}_{10}$) "your hobby?s", $\mathcal{O}_{11}$) "how "courageous" you are", $\mathcal{O}_{12}$) "how smart you are", $\mathcal{O}_{13}$) "the kind of friends you have". These topics were assumed (and partly known) to be important dimensions for this age group. As the questionnaire had to be suitable for students of all ability levels, except the lowest levels, it has been decided to ask only partial rank orders, *i.e.* the highest three ranks. Finally, the final data set if composed by $n = 1\ 567$ students with only one *ranking* variable ($d = 1$) for which the space $\mathcal{X}$ corresponds to the permutation space of size 13! ("!" stands for factorial). In addition, 85% of students provided only partial ranks, for instance only the first three objects they preferred. Among the 15% of full ranking data, note also that 20% of them contain tie situations. Finally, this data set is thus very partial.

We use the model proposed for partial ranking data in Biernacki and Jacques [2013] and Jacques and Biernacki [2014]. It corresponds to a mixture of a specific distribution for rank data parameterized by $\boldsymbol{\alpha}_k = (\boldsymbol{\mu}_k, \lambda_k)$, $\boldsymbol{\mu}_k$ being the rank modal value of this distribution and $\lambda_k \in [0.5, 1]$ being its so-called precision parameter. When $\lambda_k = 0.5$, it gives the uniform distribution; when $\lambda_k = 1$, it gives the Dirac distribution on $\boldsymbol{\mu}_k$. This model is implemented in the RANKCLUSTER[15] R package of Jacques *et al.* [2014] with a specific SEM-Gibbs. The command line for running this package on this data set for $K = 1, \ldots, 5$ is the following:

```
R> res=rankclust(x,13,1:5).
```

It provides the BIC values given in Figure 1.19. Note that confidence intervals for BIC are given since the log-likelihood is intractable for this model and so has been estimated (see Jacques and Biernacki [2014] for more details). We note that a clear hesitation between one and two groups appear, certainly due to the high degree of missing data (partial rankings and ties).

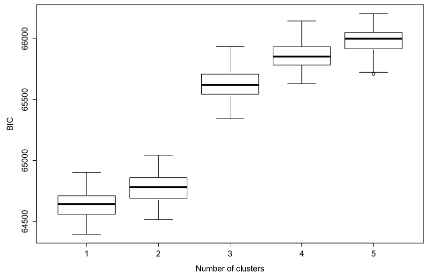


Figure 1.19: BIC value, and its associated confidence interval, for different number of groups on the social comparison theory data set.

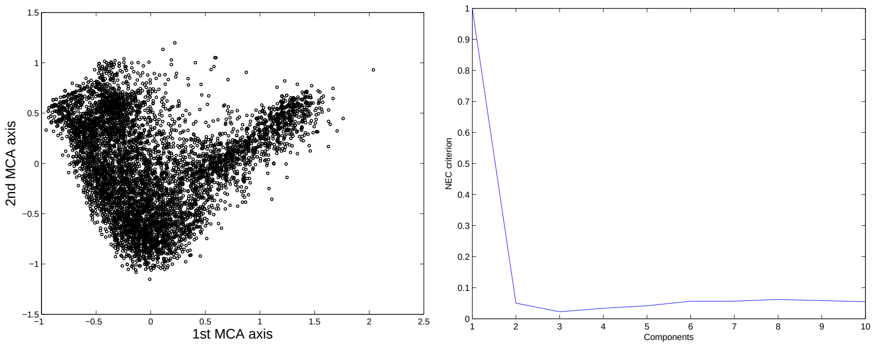
[15] http://cran.r-project.org/web/packages/Rankcluster/index.html

The estimated parameter of the dispersion for the one-group case is $\hat{\lambda}_1 \approx 0.65$. It indicates that the component distribution is quite uniform, thus denoting no particular preference between objects in the data set.

For the two-groups case, a large group ($\hat{\pi}_1 \approx 0.93$) and a small group ($\hat{\pi}_1 \approx 0.07$) are present. The first one corresponds again to a very flat distribution $\hat{\lambda}_1 \approx 0.65$, thus similar to the first group obtained in the previous one group case. The second group is more interesting since it exhibits a more tight distribution ($\hat{\lambda}_2 \approx 0.8$) which was probably masked by the previous one-group case. This group is potentially interesting for the researcher in social sciences and it can be described in depth by its meaningful parameter of preferences $\hat{\boldsymbol{\mu}}_2$ for further studies.

1.6.5 NEC: marketing data

We consider the marketing data set described in Hastie *et al.* [2001] concerning the $d = 13$ demographic attributes (nominal and ordinal variables) of $n = 6\,876$ shopping mall customers in the San Francisco Bay (it corresponds to the *complete* data observations among 8 993 observations). Here are examples of attributes with the corresponding levels between brackets: SEX (1. Male, 2. Female), MARITAL STATUS (1. Married, 2. Living together, not married, 3. Divorced or separated, 4. Widowed, 5. Single, never married), AGE (1. 14 thru 17, 2. 18 thru 24, 3. 25 thru 34, 4. 35 thru 44, 5. 45 thru 54, 6. 55 thru 64, 7. 65 and Over), *etc.* Data are displayed Figure 1.20(a) on the first two multiple correspondence analysis axes.



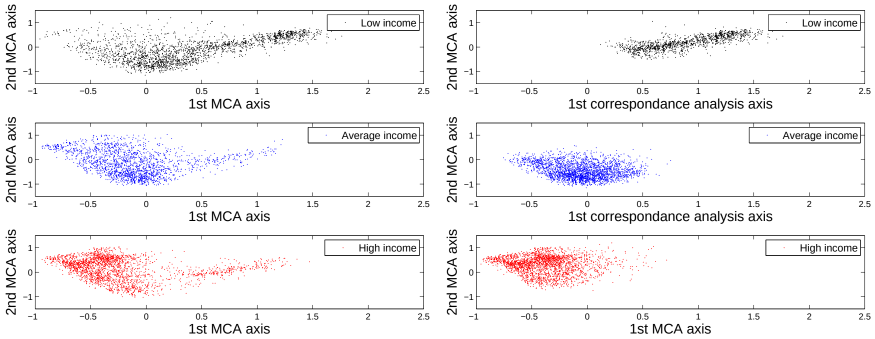
(a) (b)

Figure 1.20: Marketing data set: (a) data on the first two multiple correspondence analysis axes, (b) the NEC values for several numbers of groups.

We use the RMIXMOD package to search for a hidden structure in this data set. The following command line in R runs an EM algorithm with $K \in \{1, \ldots, 10\}$ and the NEC criterion for selecting the number of groups:

```
R> out = mixmodCluster(x, nbcluster = 1:10, criterion = "NEC").
```

The NEC criterion values are given in Figure 1.20(b) and it appears that $K = 3$ components are selected. There exists a possible true partitioning of this data set which corresponds to the following three groups of annual income of households (personal income if single), as displayed in Figure 1.21(a): less that 19 999$ (group of "low income"), between 20 000$ and 39 999$ (group of "average income"), more than 40 000$ (group of "high income"). We see in Figure 1.21(b) that the three group estimated partition is highly correlated to this true partitioning.

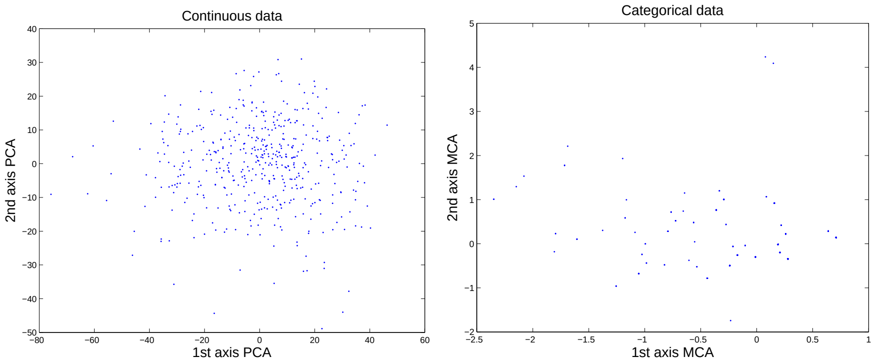


(a) (b)

Figure 1.21: Marketing data set: (a) *true* underlying partition, (b) *estimated* partition.

1.6.6 ICL: prostate cancer data

Hunt and Jorgensen [1999] (see also McLachlan and Peel [2000] p. 139–142) considered the clustering of patients on the basis of petrial variates alone for the prostate cancer clinical trial data of Byar and Green [1980] which is reproduced in Andrews and Herzberg [1985] p. 261–274. This data set was obtained from a randomized clinical trial comparing four treatments for $n = 506$ patients with prostatic cancer grouped on clinical criteria into two Stages 3 and 4 of the disease. As reported by Byar and Green [1980], Stage 3 represents local extension of the disease without evidence of distance metastasis, while Stage 4 represents distant metastasis as evidenced by elevated acid phosphatase, X-ray evidence, or both. Twelve pre-trial variates were measured on each patient, composed by eight continuous variables (age, weight, systolic blood pressure, diastolic blood pressure, serum haemoglobin, size of primary tumour, index of tumour stage and histolic grade, serum prostatic acid phosphatase) and four categorical variables with various numbers of levels (performance rating, cardiovascular disease history, electrocardiogram code, bone metastases). The skewed variables "size of primary tumour" and "serum prostatic acid phosphatase" were transformed by using a square root and a logarithm transformation, respectively. Observations that had missing values in any of the twelve pretreatment covariates were omitted from further analysis, leaving $n = 475$ out of the original 506 observations available. Figure 1.22(a) and (b) displays continuous and categorical data, respectively, on the first two factorial axes. It seems difficult to distinguish groups on these axes.



(a) (b)

Figure 1.22: Prostate cancer data: (a) continuous data on the first two principal component analysis axes, (b) categorical data on the first two multiple correspondence analysis axes.

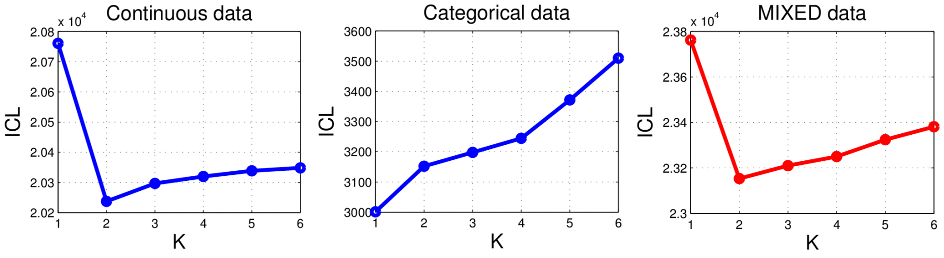
We propose to perform three different clustering procedures: a first one on only continuous variables with the diagonal Gaussian model, a second one on only categorical variables with the multivariate multinomial latent class model and a last one with all variables (mixed case) with the so-called Gaussian-multinomial model. This model assumes that continuous and categorical variables are mutually independent conditionally to the group membership while the conditional continuous variable distribution is diagonal Gaussian and while the continuous categorical variable distribution is multivariate multinomial with independence. Thus, the corresponding component pdf can be written

$$f(\mathbf{x}_1; \boldsymbol{\alpha}_k) = f(\mathbf{x}_1; \boldsymbol{\alpha}_k^{cont}) \cdot f(\mathbf{x}_1; \boldsymbol{\alpha}_k^{cat}) \tag{1.139}$$

where $\boldsymbol{\alpha}_k = (\boldsymbol{\alpha}_k^{cont}, \boldsymbol{\alpha}_k^{cat})$, $\boldsymbol{\alpha}_k^{cont} = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is the Gaussian parameter with $\boldsymbol{\Sigma}_k$ diagonal and where $\boldsymbol{\alpha}_k^{cat}$ is the multivariate multinomial parameter. This particular model is implemented in the RMIXMOD software and the command line to launch it for $K \in \{1, \ldots, 6\}$, selected through the ICLbic criterion, is the following:

```
R> out = mixmodCluster(x, nbCluster = 1:6,
   + dataType = "composite", criterion = "ICL").
```

The RMIXMOD software is also run for the pure continuous and the pure categorical cases with the same number of components and with the same criterion. Results of the corresponding ICL values are displayed in Figure 1.23(a)(b)(c), each sub figure corresponding to a particular data situation. We note that only the continuous and the mixed cases allow to choose a two-group structure by ICLbic.



(a) (b) (c)

Figure 1.23: Prostate cancer data: ICLbic values with (a) continuous data only, (b) categorical data only, (c) mixed continuous and categorical.

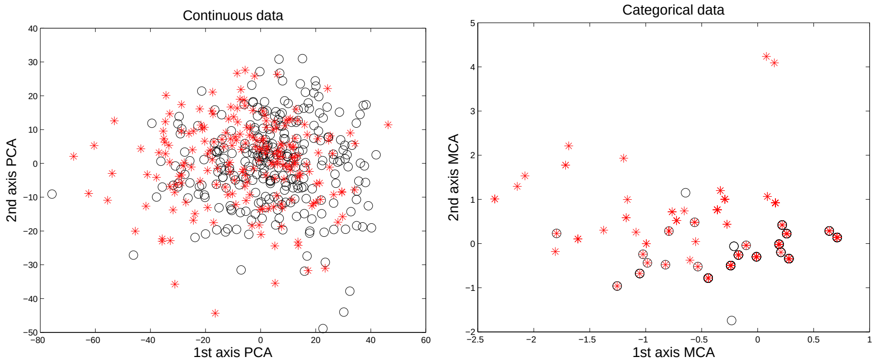
The two group estimated partition for the continuous, categorical and mixed cases is also given in Table 1.16 in comparison to the true partition in Stage 3 and Stage 4. It appears that categorical data alone are not able to provide a relevant partitioning of data. However, associated with continuous data (mixed case) they allow to improve slightly the partition estimated by the continuous variables alone. It indicates thus that categorical variables contain some partitioning information also. Figure 1.24(a) and (b) displays this mixed case estimated partition for continuous and categorical data, respectively, on the first two factorial axes.

Variables	Continuous		Categorical		Mixed	
Error (%)	9.46		47.16		8.63	
True \ estimated group	1	2	1	2	1	2
Stage 3	247	26	142	131	252	21
Stage 4	19	183	120	82	20	182

Table 1.16: Prostate cancer data: classification error rate and missclassification table for the three kinds of variables.

1.6.7 BIC: density estimation in the steel industry

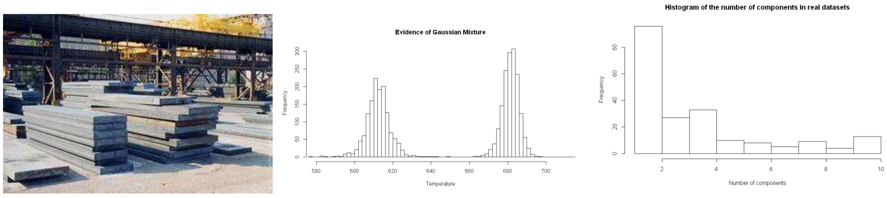
The work of Thery *et al.* [2014] takes place in the steel industry context, with a quality oriented objective. The purpose is to understand and to prevent



(a) (b)

Figure 1.24: Prostate cancer data with the too group partition estimated in the mixed case: (a) continuous data on the first two principal component analysis axes, (b) categorical data on the first two multiple correspondence analysis axes.

quality problems on finished products, knowing the whole process. The correlations between involved features can be strong because many parameters of the whole process are highly correlated (physical laws, process rules, *etc.*). A quality parameter (confidential) is considered as a response variable y and 205 variables from the whole process are measured to explain it. It is then a regression problem with the goal to explain y from these 205 variables. However, some of these industrial variables are naturally highly correlated. For instance, denoting by ρ the linear correlation coefficient between two variables, the width and the weight of a steel slab (see an illustration of a slab in Figure 1.25(a)) gives $|\rho| = 0.905$, the temperature before and after some tool gives $|\rho| = 0.983$, the roughness of both faces of the product gives $|\rho| = 0.919$, *etc.* Consequently, performing directly a regression on y with such covariates would lead to very unstable estimates. For this reason, Thery *et al.* [2014] developed a specific method which identifies intra linear regressions which are present between the 205 variables in order to obtain an uncorrelated variable subset. This procedure relies on a whole generative process, thus it is needed to have a density estimation of all potentially uncorrelated variables. To this end, the density of each variable is estimated by a univariate Gaussian mixture, each related number of components being selected by a BIC criterion. The RMIXMOD packages is used to perform these estimations. Thus, each variable being replicated 3 000 times, we have 205 univariate data sets $\mathbf{x}$ of identical size $n = 3\ 000$. An example of one of this variable (temperature) is displayed by its histogram in Figure 1.25(b). Figure 1.25(c) gives also the distribution of the number of components estimated for all the 205 data sets. We note that the flexibility of Gaussian mixtures allows to obtain quite parsimonious densities since the estimated value of K remains quite moderate.



(a) (b) (c)

Figure 1.25: Steel industry: (a) a steel slab, (b) Example of a non-Gaussian real variable easily modeled by a Gaussian mixture, (c) distribution of the number of components found for each covariate.

1.6.8 BIC: partitioning communes of Wallonia

This illustration is extracted from Thomas *et al.* [2008]. The purpose is to classify the $n = 262$ communes of Wallonia (made up of urban, suburban, periurban and rural areas) in terms of so-called $d = 2$ fractals at a local level. By definition, a fractal is a rough or fragmented geometric shape that can be subdivided into parts, each of which is (at least approximately) a smaller copy of the whole. Fractals are generally self-similar and independent of scale. The use of fractals in urban analysis was mainly developed in the 1990s. The first fractal variable is associated to built-up surfaces and the second one to their perimeters.

In many situations, practitioners decide to perform a clustering procedure on a one to one transformation $\mathbf{g}(\mathbf{x}) = (g(x_i^j), i = 1, \ldots, n\ j = 1, \ldots, d)$ of the initial data set instead of on the initial data set $\mathbf{x}$ itself. The reasons are generally either that the new data set $\mathbf{g}(\mathbf{x})$ "seems to have a better specific mixture shape" than $\mathbf{x}$, or that its unit has a particular meaning for the practitioner. Typically, standard transformations are $g(x_i^j) = x_i^j$ (identity), $g(x_i^j) = \exp(x_i^j)$ or $g(x_i^j) = \ln(x_i^j)$. The second transformation expresses data in the same units as fractals indices, which is a traditional quantity for many geographers. This may be a sufficient reason to consider such a transformation. However, to avoid the difficult task of proposing and justifying a particular transformation, the practitioner may use the statistical framework to choose one of the suggested transformations automatically. We describe this interesting and innovative feature below.

If the new sample $\mathbf{g}(\mathbf{x})$ arises from a mixture model $f(\cdot; \boldsymbol{\theta})$ then the initial sample $\mathbf{x}$ arises from another distribution $f_{\mathbf{g}}(\cdot; \boldsymbol{\theta})$ which is a transformation of $f(\mathbf{x}; \boldsymbol{\theta})$. Consequently, it is possible to interpret any transformation $\mathbf{g}$ as another kind of model $\mathcal{S}$ and to employ the BIC criterion to select this transformation. Denoting by $\mathbf{H_g}$ the Jacobian of the transformation $\mathbf{g}$, and by $\hat{\boldsymbol{\theta}}_{\mathbf{g}}$ the MLE obtained with $\mathbf{g}(\mathbf{x})$, we retain the transformation $\mathbf{g}$ leading to the largest of the following BIC expressions:

$$\mathrm{BIC}_{\mathbf{g}} = \ell(\hat{\boldsymbol{\theta}}_{\mathbf{g}}; \mathbf{g}(\mathbf{x})) - \frac{D}{2} \ln n + \ln |\mathbf{H}_{\mathbf{g}}|. \tag{1.140}$$

The 262 communes can now be classified with a $K = 6$ component Gaussian mixture (the number of components is here imposed by the geographer), with the three previous standard transformations **g** (identity, exponential, logarithm) and with all 28 Gaussian of Celeux and Govaert [1995]. A model is thus composed by the couple transformation and constraints on covariance matrices/mixing proportions, leading so to $3 \times 28 = 84$ models in competition. The BIC criterion retains the simplest model (spherical with equal mixing proportions) and also the exponential transformation. As said before, such a transformation was expected by geographers. The partitioning result is illustrated in Figure 1.26(a). The map reveals strong effects of contiguity: communes close to each other look alike in terms of fractal dimensions. Groups are, however, spread out all over the region. The six groups lead to the following geographical interpretation, with in brackets the three communes which are closest to the centre of each group (Mahalanobis distance):

- **Group 1** Peri-urban I and small cities (Brugelette, Heron, Nandrin);

- **Group 2** Rural I: compact isolated hamlets (Lierneux, Havelange, Merbes-le-C);

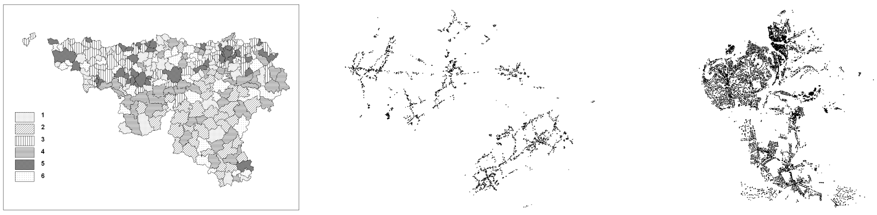
- **Group 3** Peri-urban II and eastern (Hainaut) part (Pepinster, Saint-Georges, Blegny);

- **Group 4** Rural II: hamlets with a linear structure (Erquelinnes, Baelen, Rendeux);

- **Group 5** Urban, thus homogeneous, fully urbanised communes (Ottignies, Châtelet, Chaudfontaine);

- **Group 6** Rural III: rural communes with hamlets and one (small) city centre (Gesves, Jalhay, Ciney).

Figure 1.26(b) and (c) respectively display the map of a commune of Group 1 and a commune of Group 5, revealing high differences between both structures. In addition, we show that fractal indices partition the region into sub-areas that do not correspond to "natural landscapes" but result from the history of urbanisation. Urban sprawl seems to affect most communes, even the remotest villages: traditional (compact, ribbon, *etc.*) villages are transformed into more complex and heterogeneous shapes.



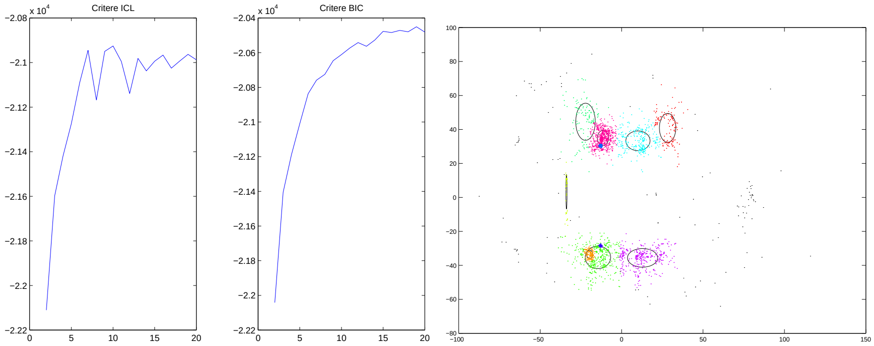
(a) (b) (c)

Figure 1.26: Communes of Wallonia: (a) the estimated six component partitioning, (b) Héron commune map as an example of Group 1, (c) Chaudfontaine commune map as an example of Group 5.

1.6.9 ICLbic/BIC: acoustic emission control

This example is extracted from Biernacki *et al.* [2000]. It is concerned with flaws detection on a pressurized vessel by acoustic emission. During a pressurization control, the vessel sounds (the *events*) are located on its surface. The first step of the flaw detection procedure consists of grouping those events in homogeneous groups. Data at hand are $n = 2\,061$ event locations in a rectangle of $\mathbb{R}^2$ representing the vessel (so, $d = 2$).

In this setting, a Gaussian mixture model with equal proportions, diagonal variance matrices with different volumes appears to be relevant. Moreover, the uniform background noise is taken into account with a uniform distribution on the rectangle where the sounds are located. It is worth noting that adding such a uniform distribution in the mixture is straightforward and simply leads to consider the proportion of the uniform component as an additional parameter (see for instance Banfield and Raftery [1993]).



(a) (b) (c)

Figure 1.27: Acoustic emission control: (a) ICLbic values, (b) BIC values, (c) the ten-cluster partition retained by ICL.

For this industrial example, the problem is to find a relevant number of mixture components leading to a clear grouping of the sound locations. Figure 1.27(a) and (b) displays the values of ICLbic and BIC, respectively, when K is varying from 2 to 20. BIC increases almost monotonically with K and does not provide evidence for any K value. On the contrary, ICLbic gives a preference for the ten-cluster partition which is depicted in Figure 1.27(c) by the iso-density of each of the ten components. In particular, it seems that the ten-cluster partition selected ICLbic captures the high density regions appearing in this data set.

1.6.10 ICLbic/ICL/BIC/ILbayes: a seabird data set

This example is extracted from Biernacki *et al.* [2011]. Puffins are pelagic seabirds from the family Procellaridae. A data set of $n = 153$ puffins divided into three subspecies *dichrous* (84 birds), *lherminieri* (34 birds) and *subalaris* (35 birds) is considered [Bretagnolle, 2007]. These birds are described by the five plumage and external morphological characters displayed in Table 1.17. Figure 1.28 (a) displays the birds on the first correspondence analysis plan.

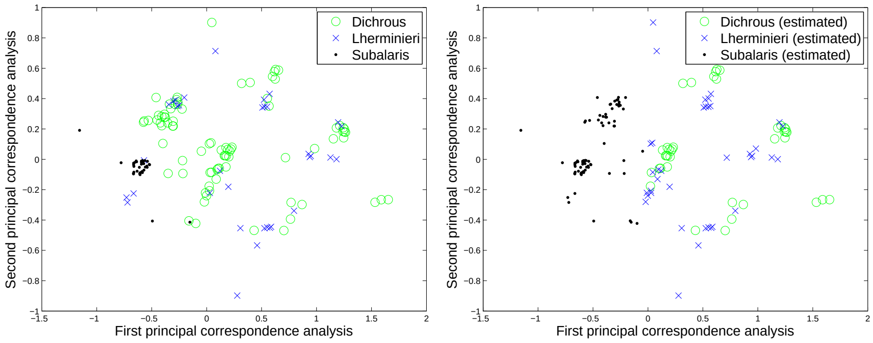
	levels				
variables	1	2	3	4	5
gender	male	female			
eyebrows^a	none	....		very pronounced	
collar^a	none	....			continuous
sub-caudal	white	black	black & white	black & WHITE	BLACK & white
border^a	none	......	many		

^{a} using a paper pattern

Table 1.17: Details of plumage and external morphological characters for the seabird data set.

For a number of groups varying from $K = 1$ to 6, asymptotic criteria BIC and ICLbic and non-asymptotic criteria ILbayes and ICL are computed. Table 1.18 displays values of all of them for each number of components. It appears that only non-asymptotic criteria ICL and ICLbayes select three groups, whereas asymptotic criteria select less groups: one for ICLbic and two for BIC. The estimated three-group partition, where labels are chosen to ensure the minimum error rate with the true partition, is given in Figure 1.28 (b). It has to be compared with the true partition given in Figure 1.28 (a). It leads to 55 misclassified birds (35.95% of birds), a rand criterion value of 0.6121 and a corrected rand criterion value of 0.1896 (Rand [1971]).

However, it has to be noticed that the ICL values for one, two and three groups are quite similar. It seems to point out that there are little differences between the birds, and that it could be hazardous to discriminate the subspecies with the available variables. Moreover, it appears that ICLbic and



(a) (b)

Figure 1.28: Seabird data set on the first two correspondence analysis axes: (a) with the *true partition* and (b) with the *EM estimated partition*. An i.i.d. uniform noise on $[0, 0.1]$ has be added on both axes for each individual in order to improve visualisation.

	$\hat{K}$					
criteria	1	2	3	4	5	6
ICLbic	**-714.03**	-727.33	-741.37	-774.01	-802.47	-830.83
ICL	-712.08	-712.57	**-711.81**	-727.44	-737.46	-741.79
BIC	-714.03	**-711.14**	-729.97	-754.58	-784.49	-814.61
ILbayes	-712.08	-693.41	**-692.88**	-694.01	*-695.21*	*-696.00*

Table 1.18: Value of ICL, ICLbic, BIC and ILbayes (with $R = 50$ and $S = 1\ 000$) criteria for different number of groups on the seabird data set. Boldface indicates maximum value for each criterion. Italic indicates an upper bound value for ILbayes (see detail in Biernacki *et al.* [2011]).

ICL do not behave the same since ICLbic has a marked preference for the one component solution (no clustering). BIC favours the two-group solution, but the no-cluster solution cannot be completely discarded. On the contrary, ILbayes clearly rejects the no clustering solution and favours three groups, emphasizing again the potentially high difference between the two types of criteria of ICL-type and of BIC-type for revealing structures in data sets.

1.7 Future methodological challenges

We identify two main challenges for model selection in mixtures: the increasing number of proposed models and the increasing volume of data (individuals and/or variables). In addition, both problems are not totally unrelated.

The increasing number of models

The number of models is expected to have a linear-like increase because new ones are regularly proposed for dealing with particular situations. In addition, some models can be combined, like the Gaussian structure and the number of components, implying this time a multiplying-like increase of models. But an exponential-like increase of models is also possible as soon as discrete parameters are involved in models. It is the case for instance in variable selection or also in the categorical case in Marbac *et al.* [2013].

Having a huge model set $\mathcal{M}$ than implies two important consequences. First, from a computational point of view, the whole model set cannot be exhaustively browsed. Thus, some specific strategies have to be performed for obtaining efficient trajectories inside $\mathcal{M}$. For instance, stochastic chains on $\mathcal{M}$ can be a candidate strategy, as the seminal work on the reversible jump of Green [1995]. See also a particular Gibbs strategy in Marbac *et al.* [2013] and Thery *et al.* [2014] where the chain is guided by the BIC value.

The second consequence of having a very large $\mathcal{M}$ is about the criteria validity. Indeed, asymptotic criteria like AIC, BIC or ICLbic are defined relatively to a given error order which, when the number of models highly increases, may be too crude for making accurate distinction between some of them. Note that when the number of models grows, the set of "close" models, hence poorly indistinguishable models, is expected to grow also. A solution for dealing with this phenomenon in the Bayesian context is either to implement non-asymptotic criteria, or to define a non-uniform prior $f(\mathbf{m})$ on $\mathcal{M}$. For instance, in Thery *et al.* [2014], a *hierarchical* uniform distribution has been put on a particular decomposition of $\mathcal{S}_{\mathbf{m}}$, resulting in a higher penalty for more complex models while preserving a non-informative approach. In the frequentist setting, the heuristics slope has also to be adapted for large $\mathcal{M}$. For instance, Meynet and Maugis-Rabusseau [2012] give some proposal for variable selection in the model-based clustering framework.

The increasing volume of data

The "Big Data" era implies an increasing number of individuals and/or variables. From the model selection point of view, it may increase a lot the computation time, in particular in mixtures where EM-like algorithms are quite slow. Simultaneously, a larger volume of data encourage to try a larger model set $\mathcal{M}$, as testing a much larger upper bound for the number of groups. Indeed, we expect to discover finer structures when the data set grows!

Possible solutions are sampling strategies. However, the risk of them is to miss some fine structures in data. Thus, some specific researchs could be needed to overcome this difficulty.

Bibliography

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

Aitkin, M., Anderson, D. and Hinde, J. [1981]. Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society (series B)*, **47**(1), 67–75.

Aitkin, M. and Rubin, D. B. [1985]. Estimation and Hypothesis Testing in Finite Mixture Models. *Journal of the Royal Statistical Society, Series B*, **47**, 67–75.

Akaike, H. [1973]. Information Theory as an Extension of the Maximum Likelihood Principle. In B. Petrov, F. Csaki (editors), *Second International Symposium on Information Theory*, 267–281. Budapest, Akademiai Kiado.

Akaike, H. [1974]. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716–723.

Amemiya, T. [1973]. Regression analysis when the dependent variables is truncated normal. *Econometrica*, **41**, 997–1016.

Andrews, D. F. and Herzberg, A. M. [1985]. *Data: A Collection of Problems from Many. Fields for the Student and Research Worker*. Springer-Verlag.

Arlot, S. and Celisse, A. [2010]. A survey of cross-validation procedures for model selection. *Statistics Surveys*, **4**, 40–79.

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

Baudry, J., Raftery, A., Celeux, G., Lo, K. and Gottardo, R. [2010]. Combining Mixture Components for Clustering. *Journal of Computational and Graphical Statistics*, **9**(2), 332–353.

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., Cardoso, M., Celeux, G., Amorim, M.-J. and Sousa Ferreira, A. [2012a]. Enhancing the selection of a model-based clustering with external qualitative variables. *Rapport de recherche RR-8124*, INRIA. URL `http://hal.inria.fr/hal-00747387`.

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

Benaglia, T., Chauveau, D. and Hunter, D. [2011]. *Bandwidth Selection in an EM-like algorithm for nonparametric multivariate mixtures*, chapter Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger. World Scientific Publishing Co.

Besag, J. [1986]. On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society B*, **48**, 259–302.

Bezdeck, J. [1981]. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum, New York.

Biecek, P., Szczurek, E., Vingron, M. and Tiuryn, J. [2012]. The R package bgmm: Mixture modeling with uncertain knowledge. *Journal of Statistical Software*, **47**(3).

Biernacki, C. [1997]. *Choix de modèles en classification*. Phd. thesis, UTC Compiègne.

Biernacki, C., Celeux, G. and Govaert, G. [1999]. An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model. *Pattern Recognition Letters*, **20**(3), 267–272.

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. [2003]. Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models. *Computational Statistics and Data Analysis*, **41**, 561–575.

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.

Biernacki, C. and Govaert, G. [1997]. Using the Classification Likelihood to Choose the Number of Clusters. *Computing Science and Statistics*, **29**(2), 451–457.

Biernacki, C. and Govaert, G. [1999]. Choosing Models in Model-based Clustering and Discriminant Analysis. *Journal of Statistical Computation and Simulation*, **64**, 49–71.

Biernacki, C. and Jacques, J. [2013]. A generative model for rank data based on insertion sort algorithm. *Computational Statistics and Data Analysis*, **58**, 162–176.

Biernacki, C. and Jacques, J. [2015]. Model-Based Clustering of Multivariate Ordinal Data Relying on a Stochastic Binary Search Algorithm. *Statistics and Computing*. In press, URL `https://hal.inria.fr/hal-01052447`.

Biernacki, C. and Lourme, A. [2013]. Gaussian Parsimonious Clustering Models Scale Invariant and Stable by Projection. *Statistics and Computing*, **23**(5).

Binder, S. [1978]. Bayesian cluster analysis. *Biometrika*, **65**, 31–38.

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

Bock, H. [1981]. Statistical Testing and Evaluation Methods in Cluster Analysis. In *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions*, 116–146. Calcutta.

Bordes, L., Chauveau, D. and Vandekerkhove, P. [2007]. A stochastic EM algorithm for a semiparametric mixture model. *Computational Statistics and Data Analysis*, **51**(11), 5429–5443.

Bouchard, G. and Celeux, G. [2006]. Selection of Generative Models in Classification. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **28**(4), 544–554.

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

Bozdogan, H. [1981]. *Multi-Sample Cluster Analysis and Approaches to Validity Studies in Clustering Individuals.* Thèse de doctorat, Department of Mathematics, University of Illinois, Chicago, IL 60680.

Bozdogan, H. [1983]. Determining the number of clusters in the standart multivariate normal mixture model using model-selection criteria. *Technical report no. uic/dqm/a83-1*, Quantitative Methods Department, University of Illinois, Chicago, IL 60680. ARO contract DAAGL90820K-0155.

Bretagnolle, V. [2007]. Personal communication, source: Museum.

Bryant, P. and Williamson, J. [1978]. Asymptotic Behaviour of Classification Maximum Likelihood Estimates. *Biometrika*, **65**, 273–281.

Byar, D. and Green, S. [1980]. The choice of treatment for cancer patients based on covariate information: Application to prostate cancer. *Bulletin du Cancer*, **67**, 477–490.

Cadez, I., Smyth, P., McLachlan, G. and McLaren, C. [2002]. Maximum Likelihood Estimation of Mixture Densities for Binned and Truncated Multivariate Data. *Machine Learning*, **47**(1), 7–34.

Casella, G., Robert, C. and Wells, M. [2000]. Mixture models, latent variables and partitioned importance sampling. *Technical Report 2000-03*, CREST, INSEE, Paris.

Celeux, G., Chrétien, S., Forbes, F. and Mkhadri, A. [2001]. A Component-wise EM Algorithm for Mixtures. *Journal of Computational and Graphical Statistics*, **10**, 699–712.

Celeux, G. and Diebolt, J. [1985]. The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, **2**(1), 73–92.

Celeux, G. and Diebolt, J. [1990]. Une version de type recuit simulé de l'algorithme EM. *Notes aux Comptes Rendus de l'Académie des Sciences*, **310**, 119–124.

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. [1992]. A Classification EM Algorithm and two Stochastic Versions. *Computational Statistics & Data Analysis*, **14**, 315–332.

Celeux, G. and Govaert, G. [1993]. Comparison of the Mixture and the Classification Maximum Likelihood in Cluster Analysis. *Journal of Statistical Computation and Simulation*, **47**, 127–146.

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

Celeux, G. and Soromenho, G. [1996]. An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model. *Journal of Classification*, **13**, 195–212.

Chib, S. [1995]. Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**(432), 1313–1321.

Coleman, D. A. and Woodruff, D. L. [2000]. Cluster Analysis for Large Datasets: An Effective Algorithm for Maximizing the Mixture Likelihood. *Journal of Computational and Graphical Statistics*, **9**(4), 672–688.

Cramér, H. [1946]. *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.

Cutler, A. and Windham, M. P. [1993]. Information-Based Validity Functionals for Mixture Analysis. In K. H. Bozdogan (editor), *Proceedings of the first US-Japan Conference on the Frontiers of Statistical Modeling*, 149–170. Amsterdam.

Day, N. E. [1969]. Estimating the components of a mixture of normal distributions. *Biometrika*, **56**, 463–474.

Dempster, A. P., Laird, N. M. and Rubin, D. B. [1977]. Maximum Likelihood from Incomplete Data (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

Devos, O., Ruckebusch, C., Durand, A., Duponchel, L. and Huvenne, J. [2009]. Support vector machines (SVM) in near infrared (NIR) spectroscopy: Focus on parameters optimization and model interpretation. *Chemometrics and Intelligent Laboratory Systems*, **96**(1), 27–33.

Efron, B. [1983]. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association*, **78**, 316–331.

Everitt, B. [1981]. A Monte-Carlo investigation of the likelihood ratio test for the number of components in a mixture of normal distributions. *Multivariate Behavioral Research*, **16**, 171–180.

Ferguson, T. S. [1982]. An inconsistent Maximum Likelihood Estimate. *Journal of the American Statistical Association*, **77**(380), 831–834.

Flury, B. [1997]. *A first course in multivariate statistics.* Springer, New York.

Fraley, C. and Raftery, A. E. [2002]. Model-based Clustering, Discriminant Analysis and Density Estimation. *Journal of the American Statistical Association*, **97**, 611–631.

Friedman, H. and Rubin, J. [1967]. On some Invariant Criteria for Grouping Data. *Journal of the American Statistical Association*, **62**, 1159–1178.

Frühwirth-Schnatter, S. [2006]. *Finite Mixture and Markov Switching Models.* Springer Series in Statistics.

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

Gower, J. C. [1974]. Maximal predictive classification. *Biometrics*, **30**, 643–654.

Green, P. [1995]. Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *Biometrika*, **82**(4), 711–732.

Hastie, T. and Tibshirani, R. [1996]. Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 155–176.

Hastie, T., Tibshirani, R. and Friedman, J. [2001]. *The Elements of Statistical Learning.* Springer.

Hathaway, R. J. [1986]. Another Interpretation of the EM Algorithm for Mixture Distributions. *Statistics and Probability Letters*, **4**, 53–56.

Hunt, L. and Jorgensen, M. [1999]. Mixture Model Clustering: a Brief Introduction to the MULTIMIX Program. *Australian and New Zealand Journal of Statistics*, **41**(2), 153–171.

Jacques, J. and Biernacki, C. [2014]. Model-based clustering for multivariate partial ranking data. *Journal of Statistical and Planning Inference*, **149**, 201–217.

Jacques, J., Bouveyron, C., Girard, S., Devos, O., Duponchel, L. and Ruckebusch, C. [2010]. Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *Journal of Chemometrics*, **24**(11-12), 719–727.

Jacques, J., Grimonprez, Q. and Biernacki, C. [2014]. Rankcluster: An R Package for clustering multivariate partial ranking. *The R Journal*, **6**(1), 101–110.

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

Jennrich, R. I. [1969]. Asymptotic properties of non linear least square estimators. *Annals of Mathematical Statistics*, **40**, 633–643.

Kass, R. E. and Raftery, A. E. [1995]. Bayes Factors and Model Uncertainty. *Journal of the American Statistical Association*, **90**, 773–795.

Kass, R. E. and Wasserman, L. [1995]. A reference Bayesian Test for Nested Hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, **90**, 928–934.

Keribin, C. [2000]. Consistent estimation of the order of mixture models. *Sankhya, Series A*, **62**(1), 49–66.

Kiefer, J. and Wolfowitz, J. [1956]. Consistency of the maximum likelihood estimates in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics*, **27**, 887–906.

Lebarbier, E. and Mary-Huard, T. [2004]. Le critère BIC : fondements théoriques et interprétation. *Research Report RR-5315*, INRIA. URL http://hal.inria.fr/inria-00070685.

Lehmann, E. [1983]. *Theory of Point Estimation*. Wiley, New-York.

Leroux, B. G. [1992]. Consistent estimation of a mixing proportion. *Annals of Statistics*, **20**, 1350–1360.

Liu, C. and Rubin, D. B. [1994]. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **81**, 633–648.

Liu, C. and Sun, D. X. [1997]. Acceleration of EM Algorithm for Mixtures Models using ECME. In *ASA Proceedings of The Stat. Comp. Session*, **5**, 109–114.

Marbac, M., Biernacki, C. and Vandewalle, V. [2013]. Model-based clustering for conditionally correlated categorical data. *Travaux universitaires RR-8232*, INRIA. URL http://hal.inria.fr/hal-00787757.

Marbac, M., Biernacki, C. and Vandewalle, V. [2014]. Model-based clustering of Gaussian copulas for mixed data. URL http://hal.archives-ouvertes.fr/hal-00987760.

Marden, J. [1995]. *Analyzing and modeling rank data*, volume 64 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, London.

Mardia, K. S. and Jupp, P. E. [2000]. *Directional Statistics.* Wiley.

Marin, J.-M., Mergersen, K. and Robert, C. P. [2005]. *Bayesian modelling and inference on mixture of distributions.* Elsevier B. V., Handbbok of Statistics, Vol. 25.

Markatou, M., Basu, A. and Lindsay, B. G. [1998]. Weighted Likelihood Equations with Bootstrap Root Search. *Journal of the American Statistical Association*, **93**(442), 740–750.

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

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. [1987]. On Bootstraping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *Applied Statistics*, **36**, 318–324.

McLachlan, G. and Jones, P. [1988]. Finite Mixture Models to Grouped and Truncated Data via the EM algorithm. *Biometrics*, **44**, 571–578.

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

McLachlan, G. J. [1992]. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York.

McLachlan, G. J. and Krishnam, T. [1997]. *The EM algorithm and Extensions.* Wiley, New York.

McNicholas, P. and Browne, R. [2013]. Discussion of 'How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification'. *Journal of the Royal Statistical Society: Series C*, **62**(3), 352–353.

Meila, M. and Heckerman, D. [2001]. An Experimental Comparison of Model-Based Clustering Methods. *Machine Learning*, **42**, 9–29.

Mendenhall, W. and Hader, R. J. [1958]. Estimation of parameters of mixed exponentially distributed failure time distributions from censored life test data. *Biometrika*, **45**, 504–520.

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, D. J. and Browning, J. [2003]. A Mixture Model and EM-Based Algorithm for Class Discovery, Robust Classification, and Outlier Rejection in Mixed Labeled/Unlabeled Data Sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **25**(11), 1468–1483.

Monfrini, E. [2003]. Unicité dans la méthode des moments pour les mélanges de deux distributions normales. *C. R. Acad. Sci. Paris*, **336**(Série I), 89–94.

Murata, N., Yoshizawa, S. and Amari, S. [1991]. A criterion for determining the number of parameters in an artificial neural network model. In T. Hohonen, K. M?kisara, O. Simula, J. Kangas (editors), *Artificial Neural Networks. Proceesings of ICANN-91*, **1**, 9–14. Amsterdam: North Holland.

Murata, N., Yoshizawa, S. and Amari, S. [1993]. Learning curves, model selection and complexity of neural networks. In *NIPS5*, 607–614.

Murata, N., Yoshizawa, S. and Amari, S. [1994]. Network information criterion – determining the number of hidden units for artificial neural network models. *IEEE Transactions on Neural Networks*, **5**, 865–872.

Nadif, M. and Govaert, G. [1998]. Clustering for binary data and mixture models: Choice of the model. *Applied Stochastic Models and Data Analysis*, **13**, 269–278.

Neyman, J. and Scott, E. [1948]. Consistent Estimators Based on Partially Consistent Observations. *Econometrica*, **16**, 1–32.

Nowicki, K. and Snijders, T. A. B. [2001]. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, **96**, 1077–1087.

Pearson, K. [1894]. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society*, **185**, 71–110.

Pilla, R. S. and Lindsay, B. G. [2001]. Alternative EM methods for Nonparametric Finite Mixture Models. *Biometrika*, **88**, 535–550.

Raftery, A. E. [1995]. Bayesian Model Selection in Social Research (with discussion). In P. V. Marsden (editor), *Sociological Methodology*, 111–195. Cambridge, Mass.: Blackwells.

Rand, W. M. [1971]. Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, **66**, 846–850.

Rao, C. R. [1948]. The Utilization of Multiple Measurements in Problems of Biological Classification (with discussion). *Journal of the Royal Statistical Society, Series B*, **10**, 159–203.

Redner, R. and Walker, H. [1984]. Mixture densities, Maximum likelihood and the EM algorithm. *SIAM Review*, **26**(2), 195–239.

Ripley, B. D. [1996]. *Neural Networks and Pattern Recognition.* Cambridge University Press, New York.

Rissanen, J. [1989]. *Stochastic Complexity in Statistical Inquiry.* World Scientific, Teaneck, New Jersey.

Robert, C. P. [1994]. *The Bayesian Choice: a Decision-Theoretic Motivation.* New York: Springer-Verlag.

Robert, C. P. [2001]. *The Bayesian Choice.* Springer Verlag, second edition, New York.

Roeder, K. and Wasserman, L. [1997]. Practical Bayesian Density Estimation Using Mixtures of Normals. *Journal of the American Statistical Association*, **92**(439), 894–902.

Schwarz, G. [1978]. Estimating the Dimension of a Model. *Annals of Statistics*, **6**, 461–464.

Scott, A. and Symons, M. [1971]. Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387–397.

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

Small, C. G., Wang, J. and Yang, Z. [2000]. Eliminating Multiple Root Problems in Estimation. *Statistical Science*, **15**(4), 313–341.

Smyth, P. [2000]. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing*, **10**, 63–72.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. [2002]. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, **64**(4), 583–639.

Stefanski, L. A. and Carroll, R. J. [1987]. Conditional scores and optimal scores for generalized linear measurement-error models. *Biometrika*, **74**(4), 703–716.

Stone, M. [1977]. An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of Royal Statistical Society, Series B*, **39**, 44–47.

Tarone, R. D. and Gruenhage, G. [1975]. A note on the uniqueness of roots of the likelihood equations for vector-valued parameters. *Journal of the American Statistical Association*, **70**, 903–904.

Thery, C., Biernacki, C. and Loridant, G. [2014]. Model-Based Variable Decorrelation in Linear Regression.

Thomas, I., Frankhauser, P. and Biernacki, C. [2008]. The morphology of built-up landscapes in Wallonia (Belgium): a classification using fractal indices. *Landscape and Urban Planning*, **84**, 99–115.

Tierney, L. and Kadane, J. B. [1986]. Accurate approximations for posterior moments and marginal distributions. *Journal of the American Statistical Association*, **81**, 82–86.

Toher, D., Downey, G. and Murphy, T. B. [2005]. Gaussian mixture models for the classification of high-dimensional vibrational spectroscopy data. *Young Statisticians Meeting.*

Tukey, J. [1958]. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, **29**, 164.

Ueda, N. and Nakano, R. [1998]. Deterministic Annealing EM Algorithm. *Neural Networks*, **11**, 271–282.

Vandewalle, V. [2009]. *Estimation et sélection en classification semi-supervisée.* Thèse de doctorat, Université de Lille 1.

Vandewalle, V., Biernacki, C., Celeux, G. and Govaert, G. [2013]. A predictive deviance criterion for selecting a generative model in semi-supervised classification. *Computational Statistics and Data Analysis*, **64**, 220–236.

Wald, H. [1949]. Note on the Consistency of the Maximum Likelihood Estimate. *Annals of the Mathematical Statistics*, **20**, 595–601.

Ward, J. [1963]. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, **58**, 236–244.

White, H. [1981]. Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, **76**(374), 419–433.

White, H. [1982]. Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, **50**(1), 1–25.

Windham, M. and Cutler, A. [1992]. Information Ratios for Validating Cluster Analysis. *Journal of the American Statistical Association*, **87**, 1188–1192.

Wolfe, J. H. [1971]. A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Multinormal Distributions. *Technical Bulletin STB 72-2*, US Naval Personnel Research Activity, San Diego, California.