Exact and Monte Carlo Calculations of Integrated Likelihoods for the Latent Class Model

C. Biernacki[a,*], G. Celeux[b], G. Govaert[c]

[a] *CNRS & Université de Lille 1, Villeneuve d'Ascq, France*
[b] *INRIA, Orsay, France*
[c] *CNRS & Université de Technologie de Compiègne, Compiègne, France*

Abstract

The latent class model or multivariate multinomial mixture is a powerful approach for clustering categorical data. It uses a conditional independence assumption given the latent class to which a statistical unit is belonging. In this paper, we exploit the fact that a fully Bayesian analysis with Jeffreys non informative prior distributions does not involve technical difficulty to propose an exact expression of the integrated *complete-data* likelihood, which is known as being a meaningful model selection criterion in a clustering perspective. Similarly, a Monte Carlo approximation of the integrated *observed-data* likelihood can be obtained in two steps: An exact integration over the parameters is followed by an approximation of the sum over all possible partitions through an importance sampling strategy. Then, the exact and the approximate criteria experimentally compete respectively with their standard asymptotic BIC approximations for choosing the number of mixture components. Numerical experiments on simulated data and a biological example highlight that asymptotic criteria are usually dramatically more conservative than the non-asymptotic presented criteria, not only for moderate sample sizes as expected but also for quite large sample sizes. This research highlights that asymptotic standard criteria could often fail to select some interesting structures present in the data.

*Corresponding author. Tel. +33 3 20 43 68 76, Fax. +33 3 20 43 43 02.
Email addresses: christophe.biernacki@math.univ-lille1.fr (C. Biernacki), gilles.celeux@math.u-psud.fr (G. Celeux), gerard.govaert@utc.fr (G. Govaert)

Key words: Categorical data, Bayesian model selection, Jeffreys conjugate prior, importance sampling, EM algorithm, Gibbs sampler

1. Introduction

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 clusters. It has been proved to be successful in many practical situations (see for instance Aitkin et al., 1981).

In this paper, we consider the problem of choosing a relevant latent class model. In the Gaussian mixture context, the BIC criterion (Schwarz, 1978) 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 behavior in practical situations. And, any criterion related to the asymptotic BIC approximation may suffer this limitation. In this paper, we take profit from the possibility to avoid asymptotic approximation of integrated likelihoods to propose alternative non-asymptotic criteria.

Actually, a conjugate Jeffreys non informative prior distribution is available for the latent class model parameters (contrary to what happens for Gaussian mixture models) and integrating the complete-data likelihood leads to a closed form formula. Thus, the integrated *complete-data* likelihood proposed in Biernacki et al. (2000) as a Bayesian *clustering* criterion can be exactly and easily computed without needing any BIC approximation. Moreover, the integrated *observed-data* likelihood, more commonly named *marginal likelihood* (see for instance Frühwirth-Schnatter, 2006), can be non asymptotically approximated through two steps: An exact integration of the complete data distribution over

the parameters is followed by an approximation of the sum over all possible partitions to get the marginal distribution of the observed data. This approximation involves a Bayesian importance sampling strategy. The Bayesian instrumental distribution is derived in a natural way using the fact that Bayesian inference is efficiently implemented through a Gibbs sampler thanks to conjugate properties.

The main purpose of this paper is to present those *non-asymptotic* Bayesian (latent class) model selection criteria and to compare them with their *asymptotic* versions. Second, it gives the opportunity to highlight the important difference between the *complete-data* and *observed-data* criteria.

The paper is organised as follows. In Section 2, the standard latent class model is described; furthermore maximum likelihood (ML) and non informative Bayesian inferences are briefly sketched. The exact integrated *complete-data* likelihood and the approximate integrated *observed-data* likelihood are respectively described in Section 3 and Section 4. Numerical experiments on both simulated and real data sets for selecting a relevant number of mixture components are presented in Section 5. A discussion section ends the paper by summarising the pros and cons of each evaluated strategy in order to help practitioners to make their choice. It gives also some possible extensions of this work.

2. The latent class model

2.1. The model

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. In the standard latent class model, data are supposed to arise independently from a mixture of g multivariate multinomial distributions with probability distribution function (pdf)

$$\mathrm{p}(\mathbf{x}_i; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k \mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_k) \tag{1}$$

with

$$\mathrm{p}(\mathbf{x}_i;\boldsymbol{\alpha}_k) = \prod_{j=1}^{d}\prod_{h=1}^{m_j}(\alpha_k^{jh})^{x_i^{jh}}, \tag{2}$$

where $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ is denoting the vector parameter of the latent class model to be estimated, with $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ the vector of mixing proportions of the g latent clusters, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ 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. The latent class model is assuming that the variables are *conditionally independent* knowing the latent clusters.

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

2.2. Maximum likelihood inference

Since the latent class structure is a mixture model, the EM algorithm (Dempster et al., 1977; McLachlan and Krishnan, 1997) is a preferred tool to derive the ML estimates of these model parameters (see McLachlan and Peel, 2000). The observed-data log-likelihood of the model is

$$L(\boldsymbol{\theta};\mathbf{x}) = \sum_{i=1}^{n} \ln\left(\sum_{k=1}^{g}\pi_k \prod_{j=1}^{d}\prod_{h=1}^{m_j}(\alpha_k^{jh})^{x_i^{jh}}\right). \tag{3}$$

Noting the unknown indicator vectors of the g clusters by $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ with $\mathbf{z}_i = (z_{i1}, \ldots, z_{ig})$ where $z_{ik} = 1$ if $\mathbf{x}_i$ arose from cluster k, $z_{ik} = 0$ otherwise, the complete-data log-likelihood is

$$L(\boldsymbol{\theta};\mathbf{x},\mathbf{z}) = \sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} \ln\left(\pi_k \prod_{j=1}^{d}\prod_{h=1}^{m_j}(\alpha_k^{jh})^{x_i^{jh}}\right). \tag{4}$$

From this complete-data log-likelihood, the equations of the EM algorithm are easily derived and this algorithm is as follows from an initial position $\boldsymbol{\theta}^0 = (\boldsymbol{\pi}^0, \boldsymbol{\alpha}^0)$.

- E step: Calculation of the conditional probability $t_{ik}(\boldsymbol{\theta}^r)$ that $\mathbf{x}_i$ arose from cluster k ($i = 1, \ldots, n; k = 1, \ldots, g$), r denoting the iteration index,

$$t_{ik}(\boldsymbol{\theta}^r) = \frac{\pi_k^r \mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_k^r)}{\sum_{\ell=1}^g \pi_\ell^r \mathrm{p}(\mathbf{x}_i; \boldsymbol{\alpha}_\ell^r)}. \tag{5}$$

- M step: Updating of the mixture parameter estimates,

$$\pi_k^{r+1} = \frac{\sum_i t_{ik}(\boldsymbol{\theta}^r)}{n} \quad \text{and} \quad (\alpha_k^{jh})^{r+1} = \frac{\sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^r) x_i^{jh}}{\sum_{i=1}^n t_{ik}(\boldsymbol{\theta}^r)}. \tag{6}$$

2.3. Bayesian inference

Since the Jeffreys non informative prior distribution for a multinomial distribution $\mathcal{M}_g(\mathrm{p}_1, \ldots, \mathrm{p}_g)$ is a conjugate Dirichlet distribution $\mathcal{D}_g(\frac{1}{2}, \ldots, \frac{1}{2})$, a fully non informative Bayesian analysis is possible for latent class models contrary to the Gaussian mixture model situation (see for instance Marin et al., 2005). Thus, using the prior distribution $\mathcal{D}_g(\frac{1}{2}, \ldots, \frac{1}{2})$ for the mixing weights, and noting $n_k = \#\{i : z_{ik} = 1\}$, the full conditional distribution of $\boldsymbol{\pi}$ is given by

$$\mathrm{p}(\boldsymbol{\pi}|\mathbf{z}) = \mathcal{D}_g(\tfrac{1}{2} + n_1, \ldots, \tfrac{1}{2} + n_g). \tag{7}$$

In a similar way, using the prior distribution $\mathcal{D}_{m_j}(\frac{1}{2}, \ldots, \frac{1}{2})$ for $\boldsymbol{\alpha}_k^j = (\alpha_k^{j1}, \ldots, \alpha_k^{jm_j})$, with $k = 1, \ldots, g$ and $j = 1, \ldots, d$, the full conditional distribution for $\boldsymbol{\alpha}_k^j$ is, noting $n_k^{jh} = \#\{i : z_{ik} = 1, x_i^{jh} = 1\}$,

$$\mathrm{p}(\boldsymbol{\alpha}_k^j|\mathbf{x}, \mathbf{z}) = \mathcal{D}_{m_j}(\tfrac{1}{2} + n_k^{j1}, \ldots, \tfrac{1}{2} + n_k^{jm_j}). \tag{8}$$

Finally, since the conditional probabilities of the indicator vectors $\mathbf{z}_i$ are given, for $i = 1, \ldots, n$, by

$$\mathrm{p}(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{M}_g(t_{i1}(\boldsymbol{\theta}), \ldots, t_{ig}(\boldsymbol{\theta})), \tag{9}$$

the Gibbs sampling implementation of the fully non informative Bayesian inference is straightforwardly deduced from those formulas and is not detailed

further here (see for instance Frühwirth-Schnatter, 2006, Section 9.5.3.). In addition, since $\mathbf{z}$ is discrete and finite, the convergence of the chain on $\boldsymbol{\theta}$ towards the stationary distribution $\mathrm{p}(\boldsymbol{\theta}|\mathbf{x})$ is geometric (see Robert, 2007, Subsection 6.3.3 for instance).

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. In order to deal with this problem, some authors as Stephens (2000) or Celeux et al. (2000) apply a clustering-like method to possibly change the component labels of the simulated values for $\boldsymbol{\theta}$.

3. The exact integrated complete-data likelihood

Defined in a Bayesian perspective, the integrated complete-data likelihood of a mixture is defined by

$$\mathrm{p}(\mathbf{x}, \mathbf{z}) = \int_{\Theta} \mathrm{p}(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \mathrm{p}(\boldsymbol{\theta}) d\boldsymbol{\theta}, \tag{10}$$

Θ being the whole unconstrained parameter space and $\mathrm{p}(\boldsymbol{\theta})$ being the prior distribution of the model parameter $\boldsymbol{\theta}$ on Θ. A BIC-like approximation can be used:

$$\ln \mathrm{p}(\mathbf{x}, \mathbf{z}) = \ln \mathrm{p}(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}) - \frac{\nu}{2} \ln n + O_{\mathrm{p}}(1) \tag{11}$$

where ν is the number of parameters to be estimated and where $\hat{\boldsymbol{\theta}}$ corresponds to the ML of $\boldsymbol{\theta}$ obtained from the observed data $\mathbf{x}$ (since $\hat{\boldsymbol{\theta}}$ and the ML estimate of $\boldsymbol{\theta}$ obtained from the complete data $(\mathbf{x}, \mathbf{z})$ are both consistent). Replacing the missing cluster indicators $\mathbf{z}$ by their Maximum A Posteriori (MAP) values $\hat{\mathbf{z}}$ for $\hat{\boldsymbol{\theta}}$ defined by

$$\hat{z}_{ik} = \begin{cases} 1 & \text{if } \arg\max_{\ell} t_{i\ell}(\hat{\boldsymbol{\theta}}) = k \\ 0 & \text{otherwise,} \end{cases} \tag{12}$$

Biernacki et al. (2000) obtained the following ICLbic criterion

$$\mathrm{ICLbic} = \ln \mathrm{p}(\mathbf{x}, \hat{\mathbf{z}}; \hat{\boldsymbol{\theta}}) - \frac{\nu}{2} \ln n. \tag{13}$$

This criterion aims favoring mixture situations giving rise to a partitioning of the data with the greatest evidence and, as a consequence, it appears to be robust against model misspecification (see Biernacki et al., 2000, and the experiments in the present paper).

Fortunately, in the context of multivariate multinomial distributions and in the non informative setting, there is no need to use such an asymptotic approximation because conjugate Jeffreys non informative prior distributions are available for all the parameters. Thus, the integrated complete-data likelihood (10) is closed form as shown hereunder.

Jeffreys non informative Dirichlet prior distributions for the mixing proportions and the latent class parameters are

$$\mathrm{p}(\boldsymbol{\pi}) = \mathcal{D}_g(\tfrac{1}{2},\ldots,\tfrac{1}{2}) \quad \text{and} \quad \mathrm{p}(\boldsymbol{\alpha}_k^j) = \mathcal{D}_{m_j}(\tfrac{1}{2},\ldots,\tfrac{1}{2}). \tag{14}$$

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, 2007, Subsection 3.3.3), that

$$\mathrm{p}(\mathbf{x},\mathbf{z}) = \frac{\Gamma(\frac{g}{2})}{\Gamma(\frac{1}{2})^g} \frac{\prod_{k=1}^g \Gamma(n_k+\frac{1}{2})}{\Gamma(n+\frac{g}{2})} \prod_{k=1}^g \prod_{j=1}^d \frac{\Gamma(\frac{m_j}{2})}{\Gamma(\frac{1}{2})^{m_j}} \frac{\prod_{h=1}^{m_j} \Gamma\left(n_k^{jh}+\frac{1}{2}\right)}{\Gamma(n_k+\frac{m_j}{2})}. \tag{15}$$

Replacing the missing labels $\mathbf{z}$ by $\hat{\mathbf{z}}$ in $\ln \mathrm{p}(\mathbf{x},\mathbf{z})$, *i.e.* mimicking the previously described ICLbic criterion principle, the so-called ICL criterion is defined as follows:

$$\begin{aligned}\mathrm{ICL} = \ln \mathrm{p}(\mathbf{x},\hat{\mathbf{z}}) = \\ \sum_{k=1}^g \sum_{j=1}^d \left\{ \sum_{h=1}^{m_j} \ln \Gamma\left(\hat{n}_k^{jh}+\tfrac{1}{2}\right) - \ln \Gamma(\hat{n}_k+\tfrac{m_j}{2}) \right\} - \ln \Gamma(n+\tfrac{g}{2}) + \ln \Gamma(\tfrac{g}{2}) \\ + g \sum_{j=1}^d \left\{ \ln \Gamma(\tfrac{m_j}{2}) - m_j \ln \Gamma(\tfrac{1}{2}) \right\} + \sum_{k=1}^g \ln \Gamma(\hat{n}_k+\tfrac{1}{2}) - g \ln \Gamma(\tfrac{1}{2}),\end{aligned} \tag{16}$$

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

4. The approximate integrated observed-data likelihood

The integrated observed-data likelihood (or integrated likelihood in short) is

$$\mathrm{p}(\mathbf{x}) = \int_{\Theta} \mathrm{p}(\mathbf{x};\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{17}$$

A standard asymptotic approximation is given by

$$\ln \mathrm{p}(\mathbf{x}) = \ln \mathrm{p}(\mathbf{x};\hat{\boldsymbol{\theta}}) - \frac{\nu}{2}\ln n + O_{\mathrm{p}}(1), \tag{18}$$

where $\hat{\boldsymbol{\theta}}$ is the ML estimator previously defined in Section 3, and leads to the BIC criterion (Schwarz, 1978)

$$\mathrm{BIC} = \ln \mathrm{p}(\mathbf{x};\hat{\boldsymbol{\theta}}) - \frac{\nu}{2}\ln n. \tag{19}$$

For much mixture models, (17) is difficult to calculate and simulation-based approaches or approximations based on density ratios can be involved (see for instance Frühwirth-Schnatter, 2006, Chapter 5 or Frühwirth-Schnatter, 2004).

4.1. An approximate computation by importance sampling

Usually numerical approximations rely on Monte Carlo integration over $\boldsymbol{\theta}$, for instance by invoking an importance sampling strategy. Alternatively, we use the fact that the model on complete data $(\mathbf{x}, \mathbf{z})$ is conjugate to obtaining the marginal over $\boldsymbol{\theta}$ in closed-form and then to performing a sum over $\mathbf{z}$.

Thus denoting by $\mathcal{Z}$ all possible combinations of labels $\mathbf{z}$, Equation (17) can be written (see Frühwirth-Schnatter, 2006, p.140)

$$\mathrm{p}(\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} \mathrm{p}(\mathbf{x}, \mathbf{z}), \tag{20}$$

where the integrated likelihood $\mathrm{p}(\mathbf{x})$ is explicit since the integrated complete-data likelihood $\mathrm{p}(\mathbf{x}, \mathbf{z})$ can be exactly calculated for the latent class model (see the previous section).

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

$$\hat{\mathrm{p}}(\mathbf{x}) = \frac{1}{S}\sum_{s=1}^{S} \frac{\mathrm{p}(\mathbf{x}, \mathbf{z}^s)}{I_{\mathbf{x}}(\mathbf{z}^s)}. \tag{21}$$

This estimate is unbiased and its variation coefficient is given by

$$\mathrm{c}_v[\hat{\mathrm{p}}(\mathbf{x})] = \frac{\sqrt{\mathrm{Var}[\hat{\mathrm{p}}(\mathbf{x})]}}{\mathrm{E}[\hat{\mathrm{p}}(\mathbf{x})]} = \sqrt{\frac{1}{S}\left(\sum_{\mathbf{z}\in\mathcal{Z}} \frac{\mathrm{p}^2(\mathbf{z}|\mathbf{x})}{I_{\mathbf{x}}(\mathbf{z})} - 1\right)}. \tag{22}$$

In order to approximate the ideal importance function $I^*_{\mathbf{x}}(\mathbf{z})$, *i.e.* this one minimising the variance and defined by

$$I^*_{\mathbf{x}}(\mathbf{z}) = \mathrm{p}(\mathbf{z}|\mathbf{x}) = \int_{\Theta} \mathrm{p}(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta}|\mathbf{x})d\boldsymbol{\theta}, \tag{23}$$

we propose to make use of the following "Bayesian" instrumental distribution[1]

$$I_{\mathbf{x}}(\mathbf{z}) = \frac{1}{Rg!}\sum_{r=1}^{R}\sum_{q=1}^{g!} \mathrm{p}(\mathbf{z}|\mathbf{x};\rho_q(\boldsymbol{\theta}^r)), \tag{24}$$

where the set $\{\rho_1(\boldsymbol{\theta}), \ldots, \rho_{g!}(\boldsymbol{\theta})\}$ denotes all $g!$ distinct label permutations of $\boldsymbol{\theta}$ and where $\{\boldsymbol{\theta}^r\}$ are chosen to be independent realisations of $\mathrm{p}(\boldsymbol{\theta}|\mathbf{x})$. The sum over all label permutations provides an importance density which is labelling invariant, like the ideal one. Moreover, independence of $\{\boldsymbol{\theta}^r\}$, although not necessary for ensuring the validity of the unbiasedness of the estimator (21) and the variation coefficient (22), is recommended for a good estimation of (23) from the strong law of large numbers. In practice, a Gibbs sampler can be used and the derived criterion will be called ILbayes (IL for Integrated Likelihood). Note that ILbayes is depending on both S and R.

Remark. As previously noticed, the support of the importance sampling function $I_{\mathbf{x}}(\mathbf{z})$ needs to include the support of $\mathrm{p}(\mathbf{z}|\mathbf{x})$ in order to avoid infinite

[1]Note that if evaluation of (17) was performed first by a sum over $\mathbf{z}$ and then by an importance sampling procedure over $\boldsymbol{\theta}$, the importance sampling function could be written $I_{\mathbf{x}}(\boldsymbol{\theta}) = (Rg!)^{-1}\sum_{r=1}^{R}\sum_{q=1}^{g!}\mathrm{p}(\boldsymbol{\theta}|\mathbf{x};\rho_q(\mathbf{z}^r))$, where the $\mathbf{z}^r$s are obtained from a Gibbs sampler and where ρ_q is now a permutation of $\mathbf{z}$.

variance in (22). For sufficiently large R, the "Bayesian" strategy ensures this property. Note that, although the naive uniform importance sampling function $I_{\mathbf{x}}(\mathbf{z}) = 1/\#\{\boldsymbol{z} : \boldsymbol{z} \in \mathcal{Z}\}$ verifies this property too, it does not encompass the target distribution in its high-density region. Thus, this latter is expected to be an inappropriate strategy[2].

In Biernacki et al. (2008), we consider an importance sampling function which was ignoring the label switching problem. This choice was criticized by a reviewer, who suggested to use the importance function (24). As he guessed, it appears that this choice leads to dramatically reduce the variability in the model choice.

4.2. An upper bound for avoiding to compute ILbayes for large values of g

It appears that the functional evaluation of $I_{\mathbf{x}}(\mathbf{z})$ in (24) is very expensive in practice as soon as $g = 5$ or $g = 6$ because of the sum over ρ_q. In order avoiding to systematically evaluate $I_{\mathbf{x}}(\mathbf{z})$ for "high" g values we propose to first compute the following lower bound $I^{inf}_{\mathbf{x}}(\mathbf{z})$ of $I_{\mathbf{x}}(\mathbf{z})$:

$$I^{inf}_{\mathbf{x}}(\mathbf{z}^s) = \frac{1}{Rg!} \sum_{r=1}^{R} \mathrm{p}(\mathbf{z}^s|\mathbf{x}; \rho_{q_s}(\boldsymbol{\theta}^r)) \leq I_{\mathbf{x}}(\mathbf{z}^s), \tag{25}$$

where $\mathbf{z}^s$ arises from $I_{\mathbf{x}}(\mathbf{z})$ and ρ_{q_s} denotes the permutation associated to the component in $I_{\mathbf{x}}(\mathbf{z})$ (among the $Rg!$ components) having generated $\mathbf{z}^s$. We choose this particular permutation because it is expected to provide a tight lower bound. Using $I^{inf}_{\mathbf{x}}(\mathbf{z}^s)$ instead of $I_{\mathbf{x}}(\mathbf{z}^s)$ in (21) leads to an upper bound $\mathrm{ILbayes}_{sup}$ of ILbayes. Thus, when the unknown number of components g has to be selected in the set $\{1, \ldots, g_{sup}\}$, there is no need to calculate the integral ILbayes(g) for $g \geq 5$ as soon as $\mathrm{ILbayes}_{sup}(g) < \max_{g' \in \{1,\ldots,4\}} \mathrm{ILbayes}(g')$. This ad hoc trick saves huge computation time in numerical experiments of Section 5 ($g_{sup} = 6$) since it avoids computing exactly ILbayes in about 98.8% and 99.7% of situations for $g = 5$ and $g = 6$ respectively.

[2]For datasets of Section 5, this strategy leads to prefer one-cluster solutions for all cases in which multiple clusters are present (results not reported in the paper are in Biernacki et al. (2008)).

4.3. Link between ICL and the integrated likelihood

The following straightforward relationship exists between the integrated complete-data and observed-data likelihoods:

$$\ln \mathrm{p}(\mathbf{x}, \hat{\mathbf{z}}) = \ln \mathrm{p}(\mathbf{x}) + \ln \mathrm{p}(\hat{\mathbf{z}}|\mathbf{x}). \tag{26}$$

Thus, as already noticed in Biernacki et al. (2000), the ICL criterion defined in (16) can be interpreted as the integrated likelihood penalized by a measure of the cluster overlap. It means that ICL tends to realize a compromise between the adequacy of the model to the data measured by $\ln \mathrm{p}(\mathbf{x})$ and the evidence of data partitioning measured by $\ln \mathrm{p}(\hat{\mathbf{z}}|\mathbf{x})$. For instance, highly overlapping mixture components leads typically to a low value of $\mathrm{p}(\hat{\mathbf{z}}|\mathbf{x})$ and consequently does not favor a high value of ICL.

5. Numerical experiments

We now compare experimentally the behaviour of the four criteria ICL, ICLbic, ILbayes and BIC in order to highlight main practical differences between asymptotic/non asymptotic and complete/observed data strategies. First, we distinguish two different situations for simulated data: A situation where the data arose from one of the mixtures in competition and a situation where the latent class model did not give rise to the data. Then, we treat an example on a real data set.

Note that, throughout this section, the upper bound $\mathrm{ILbayes}_{sup}$ is used for each situation involving $g \in \{5, 6\}$.

5.1. Simulated data: Well specified model

We compare here non-asymptotic *vs.* asymptotic criteria: First the pair ICL/ICLbic, then the pair ILbayes/BIC. We will compare more specifically complete-data (ICL-type) *vs.* observed-data (IL-type) criteria in Subsection 5.2.

5.1.1. Design of experiments.

Observations are described by six variables ($d = 6$) with numbers of levels $m_1 = \ldots = m_4 = 3$ and $m_5 = m_6 = 4$. Two different numbers of mixture components are considered: A two component mixture ($g = 2$) with unbalanced mixing proportions, $\boldsymbol{\pi} = (0.3\ 0.7)'$, and a four component mixture ($g = 4$) with equal mixing proportions, $\boldsymbol{\pi} = (0.25\ 0.25\ 0.25\ 0.25)'$. In each situation, three values of the parameter $\boldsymbol{\alpha}$ are chosen to get a low, a moderate and a high cluster overlapping, respectively defined as 15%, 30% and 60% of the worst possible error rate (situation where $\boldsymbol{\alpha}_k^{jh} = 1/m_j$). For the previous structures associated to $g = 2$ and $g = 4$, this worst error rate is 0.30 and 0.75 respectively. More precisely, the chosen structure for $\boldsymbol{\alpha}$ is expressed by

$$\alpha_k^{jh} = \begin{cases} \frac{1}{m_j} + (1-\delta)\frac{m_j-1}{m_j} & \text{if } h = \Big[(k-1) \bmod m_j\Big] + 1 \\ \frac{\left(1-\frac{1}{m_j}-(1-\delta)\frac{m_j-1}{m_j}\right)}{m_j-1} & \text{otherwise,} \end{cases} \tag{27}$$

where $0 \leq \delta \leq 1$ allows to fit mixture parameters with the required overlapping: $\delta = 0$ corresponds to the minimum overlap because $\alpha_k^{jh} = 0$ or 1, whereas $\delta = 1$ corresponds to the maximum overlap because $\alpha_k^{jh} = 1/m_j$. Since the overlap is a continuous and non decreasing function of δ, the value $\boldsymbol{\alpha}$ associated to a given overlap is easily derived from a numerical procedure. Table 1 provides computed values of δ for each situation. In addition, Figure 1 displays a data sample for $g = 2$ and $g = 4$ on the first two axes of a correspondence analysis.

[Table 1 about here.]

[Figure 1 about here.]

5.1.2. Results for the ICL criteria.

For each parameter structure, 20 samples are generated for three different sample sizes $n \in \{320, 1\,600, 3\,200\}$. For each sample and for a number of mixture components varying from $g = 1$ to 6, 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 all criteria is displayed on Tables 2 and 3 respectively for $g = 2$ and $g = 4$.

[Table 2 about here.]

[Table 3 about here.]

As expected, it appears that ICL and ICLbic behave the same for large sample sizes. Sometimes, asymptotic behaviour of both criteria is reached for small sample sizes (low and high overlap situations). However, when asymptotic behaviour is reached only for larger sample sizes (typically for moderate overlap situations), ICL converges far faster than ICLbic towards this limit. We also notice than, before reaching its asymptotic behaviour, ICLbic is much more conservative than ICL since it detects less components than ICL. Thus, ICL can be preferred to ICLbic since it behaves better and is not really more complex to compute.

5.1.3. Results for the IL criteria.

The same samples and experimental conditions that were previously defined are used. In addition the Gibbs sampler, initialised at random from a uniform distribution on the parameter space, generates a sequence of 11 000 parameters, the first 1 000 draws corresponding to the *burn-in* period. The R values $\boldsymbol{\theta}^r$ are selected in the remaining sequence of size 10 000 every $10\,000/R$ draws. Since values $R = 50$ and $R = 100$ are retained, it guarantees that the selected draws are quasi independent. Indeed, a value of $\boldsymbol{\theta}^r$ is selected in the remaining sequence of size 10 000 every 100 draws when $R = 100$, and every 200 draws when $R = 50$.

From Tables 2 and 3, it appears that variability of the ILbayes criterion is not really significant for R and S values in the selected range. Moreover, as for ICL comparisons, the ILbayes criterion reaches its asymptotic behaviour far faster than BIC. This advantage is particularly marked when the overlap is high, so when the data structure is harder to guess. Thus, it illustrates again

the interest of non-asymptotic approximations of the integrated likelihood for the latent class model. However, contrary to ICL criterion, the ILbayes (and ILbayes$_{sup}$) criterion is more computational demanding than its asymptotic version BIC. This time difference is expressed in Table 4. A modified MATLAB version of the MIXMOD software (Biernacki et al., 2006) was used on a Dell PowerEdge2950 equipped with two Quad-Core Intel(R) Xeon(R) CPU X5460 3.16GHz and with 31.5 GB memory, and using FedoraCore10 OS. It appears that ILbayes$_{sup}$ computing cost (due mainly to the Gibbs run) stands roughly between 1.5 and 3.3 times the computing cost of BIC (due mainly to the 10 EM runs), depending on R, S, $\hat{g}$ and n values. This ratio reaches values between 1.6 and 21.6 for ILbayes when $\hat{g} \in \{3, 4\}$ and values between 3 and 90 for ILbayes when $\hat{g} = 5$. We see here the marked computing advantage of ILbayes$_{sup}$ on ILbayes for some $\hat{g}$ values. In particular, for $n = 3200$, $\hat{g} = 5$, $R = 100$, $S = 1000$, ILbayes$_{sup}$ needs 2.28 minutes of computing time whereas ILbayes requires about one hour.

[Table 4 about here.]

5.2. Simulated data: Misspecified model

In this subsection, we focus on the difference which could occur in practice between ICL and IL criteria. From Tables 2 and 3, it is apparent that ICL criteria have a tendency to underestimate the right number of components. This tendency appears more marked in the high overlap case where ICL always underestimates the right number of clusters, even for large n. In this case, the entropic penalty term in ICL is high and actually there is no evidence for data partitioning (see the right column in Figure 1).

The realistic case where the data generator does not obey the variables conditional independence assumption is now considered. It will allow us to highlight the possible interest of ICL in a cluster analysis context.

5.2.1. Design of experiments.

Two well separated components (about 0.07 error rate) are considered in a situation where the conditional independence assumption is not true. Data

have been generated with the following procedure:

1. Firstly, a sample of size n is drawn from a two component Gaussian mixture in $\mathbb{R}^6$ with mixing proportions $\boldsymbol{\pi} = (0.3\ 0.7)'$, with centers $\boldsymbol{\mu}_1 = (-2\ \ 2\ -2\ -2\ -2\ -2)'$ and $\boldsymbol{\mu}_2 = (2\ -2\ \ 2\ \ 2\ \ 2\ \ 2)'$ and with variance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{DAD}'$ where

$$\mathbf{A} = 10 \times \begin{bmatrix} 4 & 0 & \mathbf{0}_4' \\ 0 & 2 & \mathbf{0}_4' \\ \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{I}_4 \end{bmatrix} \quad \text{and} \quad \mathbf{D} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & \mathbf{0}_4' \\ 1/\sqrt{2} & -1/\sqrt{2} & \mathbf{0}_4' \\ \mathbf{0}_4 & \mathbf{0}_4 & \mathbf{I}_4 \end{bmatrix}. \quad (28)$$

   The four-variate identity matrix is denoted by $\mathbf{I}_4$ and $\mathbf{0}_4$ denotes the four-variate zero vector. It is to be noticed that conditional independence between axes 1 and 2 is broken since they are correlated for both mixture components.

2. Then, $\mathbb{R}^6$ is discretized in the following manner in order to obtain categorical data: (1) axes 1 to 4 are divided into three levels $]-\infty, -2[$, $[-2, 2[$ and $[2, \infty[$, (2) axes 5 and 6 are divided into four levels $]-\infty, -1[$, $[-1, 0[$, $[0, 1[$ and $[1, \infty[$. Thus, the same dimension space and number of levels per variable than in the simulated data of Section 5.1 is retrieved.

The other experimental conditions are similar to these ones considered in Section 5.1, excepted that four different sample sizes are retained ($n \in \{320, 1\,600, 3\,200, 16\,000\}$).

5.2.2. Results.

Mean of the estimated g values is displayed on Table 5 for all criteria. It clearly appears that ICL and ICLbic favor two clusters for most of sample sizes whereas BIC and ILbayes prefer a higher number of components when the sample size significantly increases. It illustrates the robustness of ICL criteria already noticed in the Gaussian situation by Biernacki et al. (2000) where ICLbic was able to select well separated clusters even when the model was misspecified. On the contrary, the IL criteria (BIC and ILbayes) are focused on detecting latent classes providing a good fit of the mixture with the data without considering the cluster overlap.

[Table 5 about here.]

5.3. A biological data set

5.3.1. The data.

Puffins are pelagic seabirds from the family Procellaridae. A data set of 153 puffins divided into three subspecies *dichrous* (84 birds), *lherminieri* (34 birds) and *subalaris* (35 birds) is considered (Bretagnolle, 2007)[3]. These birds are described by the five plumage and external morphological characters displayed in Table 6. Figure 2 (a) displays the birds on the first correspondence analysis plan.

[Table 6 about here.]

[Figure 2 about here.]

5.3.2. Results for ICL criteria.

For number of clusters varying from $g = 1$ to 6, EM is run 10 times at random (uniform distribution on the parameter space) for 1 000 iterations and the run providing the largest likelihood is considered as the ML estimate. Table 7 displays values of all criteria for each number of components. It appears that only ICL selects three clusters. The corresponding estimated partition, where labels are chosen to ensure the minimum error rate with the true partition, is given in Figure 2 (b). It has to be compared with the true partition given in Figure 2 (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). The confusion table between the two partitions in given in Table 8.

[Table 7 about here.]

[Table 8 about here.]

[3]Data can be obtained by contacting Vincent Bretagnolle, Centre d'Etudes Biologiques de Chizé, Villiers en Bois, 79360, Beauvoir sur Niort, France (tel.: 33 5.49.09.78.17, email: breta@cebc.cnrs.fr).

On another hand, it has to be noticed that the ICL values for one, two and three clusters are quite similar. It seems to point out that there is little differences between the birds, and that it could be hazardous to discriminate the sub-species with the available variables. Moreover, it appears that ICLbic and ICL do not behave the same since ICLbic has a marked preference for the one component solution (no clustering).

5.3.3. Results for IL criteria.

Experiments are now focused on BIC and ILbayes criteria. The implemented Gibbs sampler is the same as with the simulated data sets. For $R \in \{50, 100\}$ and $S \in \{1\,000, 10\,000, 100\,000\}$, ILbayes and ILbayes$_{sup}$ are computed respectively for $g \in \{1, \ldots, 4\}$ and $g \in \{5, 6\}$ and associated values are displayed in Table 7. We note again that the values of R and S have not too much impact on the chosen number of clusters by ILbayes.

BIC favors the two-cluster solution, but the no-cluster solution cannot be completely discarded (Table 9 gives the associated confusion table between the true partition and the two-cluster solution). On the contrary, ILbayes clearly rejects the no clustering solution and favors 3 clusters, emphasizing again the potentially high difference between the two types of criteria for revealing structures in datasets.

[Table 9 about here.]

6. Discussion

In this paper, we exploit the fact that the Jeffreys non informative prior distribution of the parameters of the multivariate multinomial mixture model is a conjugate distribution. It implies that the integrated complete-data likelihood can be expressed explicitly. Moreover, it helps to derive a non-asymptotic approximation of the integrated observed-data likelihood. A simple and efficient numerical procedure to get such a non-asymptotic approximation is proposed.

Monte Carlo numerical experiments for selecting the number of clusters in a latent class model highlight the interest of using exact or approximate non-

asymptotic criteria instead of standard asymptotic criteria as ICLbic or BIC. In particular, they illustrate the fact that asymptotic criteria may fail to detect interesting structures in the data for small sample sizes. More precisely, the pros and cons of each criterion can be summarized as following:

- From a time computing point of view, ICL and ICLbic being absolutely equivalent, the ICL criterion can definitively be preferred to the ICLbic criterion.

- Concerning the ILbayes and BIC criteria, ILbayes leads to better results than BIC even for quite moderate values of R and S. Consequently, ILbayes appears as a challenging criterion to BIC for a moderate increase of the CPU times to be paid as long as the upper bound ILbayes_{sup} is useful for "high" values of g. Otherwise, ILbayes can be difficult to evaluate for high g and its use cannot be recommended.

- On another hand, this paper underlines the possible interest of using the integrated complete-data likelihood criterion rather than the integrated observed-data likelihood criterion. The first one explicitly favors models leading to well separated clusters. This feature implies some robustness against model misspecification, as the violation of the conditional independence assumption of the latent class model. It appears that ICL is preferable when the clustering model is misspecified while implying that some power is lost in model selection. Since in practical situations, the clustering model is often misspecified, ICL could be recommended when clustering is the main purpose of the analysis.

From the encouraging results obtained for non-asymptotic criteria in this latent class model context, it is now challenging to decline such criteria in other model-based situations. It includes for instance the possibility to design such criteria to variants on the latent class model considering constrained parameters to get more parsimonious models (see Celeux and Govaert, 1991).

Finally, alternative criteria such as the deviance information criterion (DIC)

(Spiegelhalter et al. , 2002), which can be viewed as a Bayesian analogue of AIC with a similar justification, could be considered in the latent class analysis context. In particular, the criterion DIC4 proposed in Celeux et al. (2006) in the context of missing data models could be an alternative to ICL since both criteria are based on the completed likelihood. However as noted in Frühwirth-Schnatter (2006) Section 4.4.2, AIC tends to select too many components even for a correctly specified mixture. Thus it is doubtful that DIC-like criteria could outperform ILbayes and ICL criteria for latent class models. But it remains an interesting avenue for future research.

Acknowledgments. Authors are indebted to a reviewer for his comments and suggestions which help to greatly improve this paper. In particular, he suggested to consider the importance sampling function we finally used to approximate the marginal likelihood of a model.

References

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

Biernacki, C., Celeux, G., 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., Govaert, G. Langrognet, F., 2006. Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. Computational Statistics and Data Analysis 52 (2), 587–600.

Biernacki, C., Celeux, G., Govaert, G., 2008. Exact and Monte Carlo calculations of integrated likelihoods for the latent class model. Tech. Rep. RR6609, INRIA.

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

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

Celeux G, Forbes F, Robert CP, Titterington DM. Deviance information criteria for missing data models. Bayesian Analysis. 2006;1:651–674.

Celeux, G., Govaert, G., 1991. Clustering criteria for discrete data and latent class models. Journal of Classification 8 (2), 157–176.

Celeux, G., Hurn, M., Robert, C. P., 2000. Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association 95, 957–970.

Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1–38.

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

Frühwirth-Schnatter, S., 2004. Estimating marginal likelihoods for mixture and markov switching models using bridge sampling techniques. The Econometrics Journal 7, 143–167.

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

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

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

McLachlan, G. J., Krishnan, K., 1997. The EM Algorithm. Wiley, New York.

McLachlan, G. J., Peel, D., 2000. Finite Mixture Models. Wiley, New York.

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

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

Robert, C. P., 2007. The Bayesian Choice. Springer Verlag, second edition, New York.

Schwarz, G., 1978. Estimating the number of components in a finite mixture model. Annals of Statistics 6, 461–464.

Spiegelhalter, D., Best, N., Carlin, B., and van der Linde, A. 2002. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B, 64: 583–639. (With discussion).

Stephens, M. A., 2000. Dealing with label-switching in mixture models. Journal of the Royal Statistical Society series B 62, 795–809.

Table 1: Error rate and corresponding value of δ for each parameter structure. The reference overlap case (denoted by "maximum"), corresponding to the worst possible error rate, is also given.

		$g = 2$		$g = 4$	
overlap	% of max.	error rate	δ	error rate	δ
low	15	0.0450	0.4713	0.1125	0.4770
moderate	30	0.0900	0.5822	0.2250	0.6097
high	60	0.1800	0.7313	0.4500	0.7900
maximum	100	0.3000	1.0000	0.7500	1.0000

Table 2: Mean of the chosen number of groups for each criterion when $g = 2$ (well specified model situation).

			n								
			320			1 600			3 200		
			overlap (%)			overlap (%)			overlap (%)		
Criterion	R	S	15	30	60	15	30	60	15	30	60
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
BIC	-	-	2.0	2.0	1.0	2.0	2.0	2.0	2.0	2.0	2.0
ILbayes	50	100	2.2	2.2	2.0	2.0	2.0	2.0	2.0	2.0	2.0
		1 000	2.1	2.1	1.9	2.0	2.0	2.0	2.0	2.0	2.0
	100	100	2.0	2.1	2.0	2.0	2.0	2.0	2.0	2.0	2.0
		1 000	2.1	2.1	1.9	2.0	2.1	2.0	2.1	2.0	2.0

Table 3: Mean of the chosen number of groups for each criterion when $g = 4$ (well specified model situation).

			n								
			320			1 600			3 200		
			overlap (%)			overlap (%)			overlap (%)		
Criterion	R	S	15	30	60	15	30	60	15	30	60
ICLbic	-	-	3.0	1.0	1.0	3.0	1.1	1.0	3.0	1.0	1.0
ICL	-	-	3.1	1.5	1.0	3.0	1.6	1.0	3.0	2.2	1.0
BIC	-	-	3.0	2.2	1.0	3.5	3.0	1.1	4.0	3.0	1.5
ILbayes	50	100	3.4	3.0	1.1	4.0	3.1	1.8	4.0	3.6	2.5
		1 000	3.5	3.0	1.2	4.0	3.2	1.9	4.0	3.8	2.5
	100	100	3.4	3.1	1.4	4.0	3.2	1.8	4.0	3.8	2.5
		1 000	3.4	3.0	1.4	4.0	3.2	1.9	4.0	3.7	2.5

Table 4: Computer time for a unique sample (in seconds). The EM time (10 runs) or the Gibbs time has been added when a criterion used its results. A modified MATLAB version of the MIXMOD software (Biernacki et al., 2006) was used on a Dell PowerEdge2950 equipped with two Quad-Core Intel(R) Xeon(R) CPU X5460 3.16GHz and with 31.5 GB memory, and using FedoraCore10 OS.

			n								
			320			1 600			3 200		
			$\hat{g}$			$\hat{g}$			$\hat{g}$		
Criterion	R	S	3	4	5	3	4	5	3	4	5
ICLbic, ICL, BIC	-	-	13	17	20	22	29	35	27	35	42
ILbayes	50	100	20	28	60	42	67	159	69	116	301
		1 000	29	66	246	81	228	994	143	432	1941
	100	100	21	33	85	46	86	257	77	151	491
		1 000	39	106	457	120	403	1920	219	756	3743
ILbayes$_{sup}$	50	100	19	24	31	38	50	61	62	81	100
		1 000	22	28	37	47	60	72	78	100	121
	100	100	19	24	31	39	50	62	63	83	102
		1 000	24	29	39	53	67	80	91	115	137

Table 5: Mean of the chosen number of clusters when the conditional independence assumption is not verified ($g = 2$).

			n			
Criterion	R	S	320	1 600	3 200	16 000
ICLbic	-	-	1.5	2.0	2.0	2.0
ICL	-	-	1.8	2.0	2.0	2.0
BIC	-	-	2.0	2.1	3.0	4.0
ILbayes	50	100	2.3	3.0	3.1	4.0
		1 000	2.2	3.0	3.2	4.0
	100	100	2.3	3.0	3.1	4.0
		1 000	2.6	3.0	3.1	4.0

Table 6: Details of plumage and external morphological characters for the seabird data set.

	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 7: Value of ICL, ICLbic, BIC and ILbayes criteria for different number of clusters on the seabird data set. Boldface indicates maximum value for each criterion. Italic indicates ILbayes$_{sup}$ value.

			$\hat{g}$					
criteria	R	S	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	50	1 000	-712.08	-693.41	**-692.88**	-694.01	*-695.21*	*-696.00*
		10 000	-712.08	**-693.10**	-693.42	-693.83	*-694.18*	*-695.17*
		100 000	-712.08	-693.11	**-692.91**	-693.85	*-693.74*	*-692.61*
	100	1 000	-712.08	-693.16	**-692.15**	-693.36	*-694.04*	*-694.75*
		10 000	-712.08	-693.14	**-692.59**	-693.61	*-694.17*	*-693.63*
		100 000	-712.08	-693.14	**-692.58**	-693.48	*-693.24*	*-693.72*

Table 8: Confusion table between the true partition $\mathbf{z}$ and the *three clusters* partition $\hat{\mathbf{z}}$ estimated from the EM solution.

	$\hat{\mathbf{z}}$		
$\mathbf{z}$	*dichrous*	*lherminieri*	*subalaris*
dichrous	39	14	31
lherminieri	0	24	10
subalaris	0	0	35

Table 9: Confusion table between the true partition $\mathbf{z}$ and the *two clusters* partition $\hat{\mathbf{z}}$ estimated from the EM solution.

	$\hat{\mathbf{z}}$	
$\mathbf{z}$	group 1	group 2
dichrous	36	48
lherminieri	12	22
subalaris	35	0

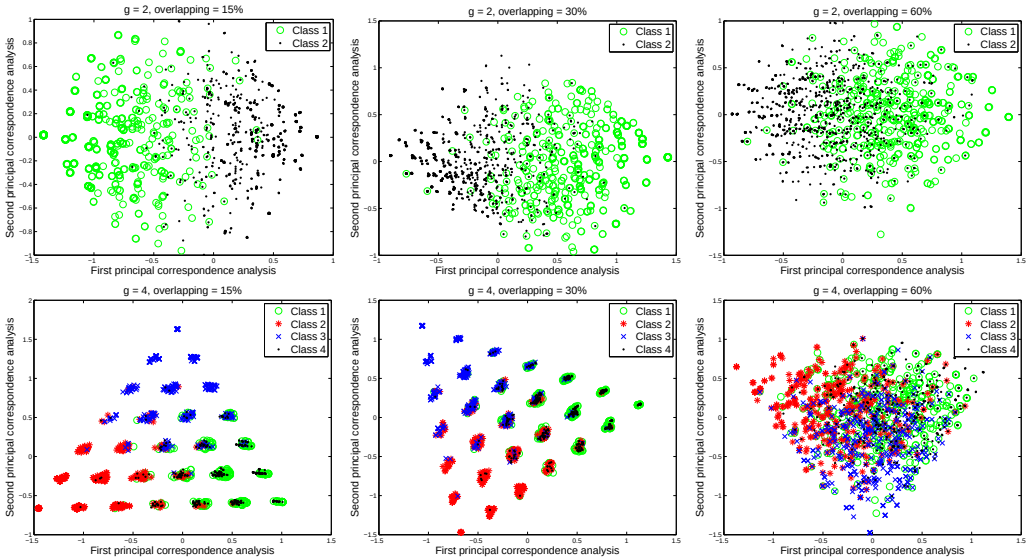
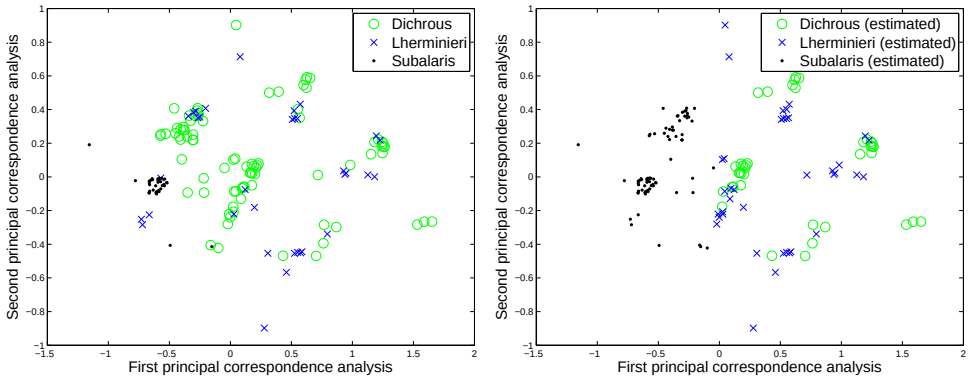


Figure 1: A sample ($n = 1\,600$) arising from $g = 2$ (top) and $g = 4$ (bottom) mixture situation for low, moderate and high overlapping. It is displayed on the first plane of a correspondence analysis and an i.i.d. uniform noise on $[0, 0.01]$ has been added on both axes for each point in order to clarify the visualisation.



(a) (b)

Figure 2: Seabirds 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.