An Introduction to mixture models

by

Franck Picard

Research Report No. 7
March 2007

Statistics for Systems Biology Group
Jouy-en-Josas/Paris/Evry, France
`http://genome.jouy.inra.fr/ssb/`

An introduction to mixture models

Franck Picard

Laboratoire Statistique et Génome
UMR CNRS 8071 - INRA 1152 - Univ. d'Evry, France

March 26, 2007

The purpose of cluster analysis is to determine the inner structure of clustered data when no information other than the observed values is available. Interest in clustering has increased due to the emergence of new domains of application, such as astronomy, biology, physics and social sciences. Most clustering done in practice is based largely on heuristic or distance-based procedures, such as hierarchical agglomerative clustering or iterative relocation procedures. These methods present two major advantages: their construction is intuitive and the associated computational time is reasonable. Nevertheless their lack of statistical basis appears to be a limitation for their use, since classical questions in clustering such as the number of clusters, can hardly be theoretically handled by heuristic procedures.

Clustering methods based on probability models offer a principal alternative to heuristic-based algorithms. In this context the data are viewed as coming from a mixture of probability distributions, each representing a different cluster. In addition to clustering purposes, finite mixtures of distributions have been applied to a wide variety of statistical problems such as discriminant analysis, image analysis and survival analysis. To this extent finite mixture models have continued to receive increasing attention from both theoretical and practical points of view.

In a mixture model based approach to clustering the data are assumed to have arisen from a mixture of an initially specified number of populations in different proportions. Let us note $Y = \{Y_1, \ldots, Y_n\}$ a random sample of size n, where Y_t is a q-dimensional random vector with probability density function $f(y_t)$ on $\mathbb{R}^q$, and y_t its realization. We suppose that the density of

email: picard@genopole.cnrs.fr

Y_t can be written in the form

$$f(y_t) = \sum_{p=1}^{P} \pi_p f_p(y_t),$$

where $f_p(y_t)$ is a component density of the mixture, and π_p the weight of population p (with the constraints $0 \leq \pi_p \leq 1$ and $\sum_p \pi_p = 1$). In many applications the component densities are assumed to belong to some parametric family. In this case, they are specified as $f(y_t; \theta_p)$, where θ_p is the unknown vector of parameters of the postulated form for the p^{th} component of the mixture. Let $\psi = (\pi_1, \dots, \pi_{P-1}, \theta_1, \dots, \theta_P)$ denote the vector containing all the unknwon parameters of the mixture. Section 1 will be devoted to the formulation of mixture models in the parametric context.

Since we are interested in clustering it appears that one information is missing regarding the observed sample: the assignment of data points to the different clusters. A new random variable is introduced and noted Z_{tp} that equals 1 if data point y_t belongs to population p, and 0 otherwise. We suppose that variables $\{Z_1, \dots, Z_n\}$ are independent (with $Z_t = \{Z_{t1}, \dots, Z_{tP}\}$) and that the conditional density of Y_t given $\{Z_{tp} = 1\}$ is $f(y_t; \theta_p)$. Therefore variables Z_{tp} can be viewed as categorial variables that indicate the labelling of the data points. Thus Z_t is assumed to be distributed according to a multinomial distribution consisting of one draw on p categories with probabilities $\pi_1, \dots, \pi_P$:

$$\{Z_{t1}, \dots, Z_{tP}\} \sim \mathcal{M}(1; \pi_1, \dots, \pi_P).$$

In terms of clustering, the p^{th} mixing proportion can be viewed as the prior probability that one data point belongs to population p. The posterior probability of Z_{tp} given the observed value of y_t will be central for clustering purposes:

$$\tau_{tp} = \Pr\{Z_{tp} = 1 | Y_t = y_t\} = \frac{\pi_p f(y_t; \theta_p)}{\sum_{\ell=1}^{P} \pi_\ell f(y_t; \theta_\ell)}.$$

In order to formalize the incomplete data structure of mixture models, let $X = (Y, Z)$ denote the complete data vector, whose only component being observed is Y. This reformulation clearly shows that mixture models can be viewed as a particular example of models with hidden structure such as hidden Markov models or models with censored data.

If the label of each data point was observed, the estimation of the mixture parameters would be straightforward since the parameters of each density component $f(y_t;\theta_p)$ could be estimated only via the data points from population p. Nevertheless the categorial variables are hidden, and the estimation can only be based on the observed data Y. The main reason for the important work on estimation methodology for mixtures is that explicit formulas for parameter estimates are not available in a closed form, leading to the need for iterative estimation procedures. Fitting mixture distributions can be handled by a wide variety of techniques, such as graphical methods, the method of moments, maximum likelihood and Bayesian approaches. It has only been since 30 years that considerable advances have been made in the fitting of mixture models, especially via the maximum likelihood method, thanks to the publication of Dempster et al. (1977) and to the introduction of the EM algorithm.

The purpose of the EM algorithm is the iterative computation of maximum likelihood estimators when observations can be viewed as incomplete data. The basic idea of the EM algorithm is to associate a complete data model to the incomplete structure that is observed in order to simplify the computation of maximum likelihood estimates. Similarly, a complete data likelihood is associated to the complete data model. The EM algorithm exploits the simpler MLE computation of the complete data likelihood to optimize the observed data likelihood. Section 2 is devoted to the general description of the EM algorithm and to its general properties. Despite a wide range of successful applications and the important work on its properties, the EM algorithm presents two intrinsic limitations: it appears to be slow to converge and as many iterative procedures, is sensitive to the initialization step. This has lead to the development of modified versions of the EM algorithm, which will be detailed in section 2.

Once the mixture model has been specified and its parameters have been estimated, one central question remains: "How many clusters?". Mixture models present a main advantage compared with heuristic cluster algorithms in which there is no established method to determine the number of clusters. With the underlying probability model, the problem of choosing the number of components can be reformulated as a statistical model choice problem. Testing for the number of components in a mixture appears to be difficult since the classical likelihood ratio test does not hold for mixtures. On the contrary, criteria based on penalized likelihood, such as the Bayesian Information Criterion (BIC) have been successfully applied to mixture models. Nevertheless, it appears that those criteria do not consider the specific objective of mixture models in the clustering context. This has lead to the construction of classification-based criteria. These criteria will be discussed in Section 3.

1. Mixture models in the parametric context

1.1 *Definition of the model*

Let $Y = \{Y_1, \dots, Y_n\}$ denote a random sample of size n where Y_t is a vector of $\mathbb{R}^q$, y_t its realization and $f(y_t)$ its density function. In the mixture model context the density of Y_t is supposed to be a mixture of P parametric densities such that:

$$f(y_t;\psi) = \sum_{p=1}^{P} \pi_p f(y_t;\theta_p), \tag{1}$$

with the constraint $\sum_{p=1}^{P} \pi_p = 1$, P being fixed. Coefficients π_p can be viewed as the weights of the p^{th} component of the mixture, which is characterized by parameter θ_p. $\psi = (\pi_1, \dots, \pi_{P-1}, \theta_1, \dots, \theta_P)$ denotes the vector of parameters of the model.

Mixture models are reformulated as an incomplete data problem since the assignment of the observed data is unknown. If we note $X_t = \{Y_t, Z_t\}$ the complete data vector whose only component being observed is Y_t, its density function is then:

$$g(x_t;\psi) = \prod_{p=1}^{P} [\pi_p f(y_t;\theta_p)]^{z_{tp}}. \tag{2}$$

1.2 *Clustering via mixture models*

When mixture models are used in the clustering context, the aim is to provide a partition of the data into P groups, with P being fixed. The populations' weights are interpreted as *prior* probabilities of belonging to a given population. $\Pr\{Z_{tp} = 1\} = \pi_p$ represents the probability to assign one data point to population p when the only available information about the data are the weights of each group.

In the complete data specification the clustering procedure aims at recovering the associated label variables $z_1, \dots, z_n$ having observed $y_1, \dots, y_n$. After the mixture model has been fitted and its parameter ψ has been estimated, a probabilistic clustering of the observations is provided in terms of their *posterior* probabilities of component membership:

$$\hat{\tau}_{tp} = \Pr_{\hat{\psi}}\{Z_{tp} = 1 | Y_t = y_t\} = \frac{\hat{\pi}_p f(y_t;\hat{\theta}_p)}{\sum_{\ell=1}^{P} \hat{\pi}_\ell f(y_t;\hat{\theta}_\ell)}.$$

Probabilities $\hat{\tau}_{t1}, \ldots, \hat{\tau}_{tP}$ are the estimated probabilities that data point y_t belongs to the first, second, ..., P^{th} component of the mixture.

Instead of fuzzy classification results each data point can be assignated to a particular population with the maximum *a posteriori* rule (MAP):

$$\hat{z}_{tp} = \begin{cases} 1 & \text{if } p = \underset{\ell}{\text{Argmax}}\,\{\hat{\tau}_{t\ell}\}, \\ \\ 0 & \text{otherwise.} \end{cases}$$

2. Fitting mixture models via the EM algorithm

The estimation of the parameters of a mixture can be handled by a variety of techniques from graphical to Bayesian methods (see Titterington et al. (1985) for an exhaustive review of those methods). Nevertheless the maximum likelihood method has focused many attentions, mainly due to the existence of an associated statistical theory. Given a sample of n independent observations from a mixture defined in 1.1, the likelihood function is:

$$\mathcal{L}(y; \psi) = \prod_{t=1}^{n} \left\{ \sum_{p=1}^{P} \pi_p f(y_t; \theta_p) \right\}.$$

The particularity of mixture models is that the maximization of the likelihood defined above with respect to ψ is not straightforward and requires iterative procedures. The EM algorithm has become the method of choice for estimating the parameters of a mixture model, since its formulation leads to straighforward estimators.

2.1 *General presentation of the EM algorithm*

In the incomplete data formulation of mixture models let us note $\mathcal{X}$ the complete data sample space from which x arises, $\mathcal{Y}$ the observed sample space and $\mathcal{Z}$ the hidden sample space. It follows that $\mathcal{X} = \mathcal{Y} \times \mathcal{Z}$, and $x = (y, z)$. The density of the observed data X can be written:

$$g(x; \psi) = f(y; \psi) k(z|y; \psi),$$

where $f(y; \psi)$ is the density of the observed data and $k(z|y; \psi)$ is the conditional density of the missing observations given the data. This leads to the definition of different likelihoods: the observed/incomplete-data likelihood $\mathcal{L}(y; \psi)$ and the unobserved/complete-data likelihood $\mathcal{L}^c(x; \psi)$. These likelihoods are linked with the relationship:

$$\log \mathcal{L}^c(x; \psi) = \log \mathcal{L}(y; \psi) + \log k(z|y; \psi),$$

with

$$\log \mathcal{L}^c(x;\psi) = \sum_{t=1}^{n} \log g(x_t;\psi),$$

and

$$\log k(z|y;\psi) = \sum_{t=1}^{n}\sum_{p=1}^{P} z_{tp} \log \mathbb{E}\left\{Z_{tp}|Y_t = y_t\right\}.$$

Since the hidden variables are not observed, the EM machinery consists of the indirect optimization of the incomplete-data likelihood *via* the iterative optimization of the conditional expectation of the complete-data likelihood using the current fit for ψ. If we note $\psi^{(h)}$ the value of the parameter at iteration h, it follows that:

$$\log \mathcal{L}(y;\psi) = Q(\psi;\psi^{(h)}) - H(\psi;\psi^{(h)}), \qquad (3)$$

with conventions:

$$\begin{aligned} Q(\psi;\psi^{(h)}) &= \mathbb{E}_{\psi^{(h)}}\left\{\log \mathcal{L}^c(X;\psi)|Y\right\}, \\ H(\psi;\psi^{(h)}) &= \mathbb{E}_{\psi^{(h)}}\left\{\log k(Z|Y;\psi)|Y\right\}, \end{aligned}$$

where $\mathbb{E}_{\psi^{(h)}}\left\{\cdot\right\}$ denotes the expectation operator, taking the current fit $\psi^{(h)}$ for ψ.

The EM algorithm consits of two steps:

- E-step: calculate $Q(\psi;\psi^{(h)})$,
- M-step: choose $\psi^{(h+1)} = \underset{\psi}{\text{Argmax}}\left\{Q(\psi;\psi^{(h)})\right\}$.

The E- and M- steps are repeated alternatively until the difference $|\psi^{(h+1)} - \psi^{(h)}|$ changes by an arbitrarily small amount. Note that another stopping rule could be the difference of log-likelihoods between two steps, $|\log \mathcal{L}(y;\psi^{(h+1)}) - \log \mathcal{L}(y;\psi^{(h)})|$. However if the log-likelihood is "flat" with respect to ψ this difference can be stable whereas parameter $\psi^{(h)}$ keeps changing.

The key property of the EM algorithm established by Dempster et al. (1977) is that the incomplete data log-likelihood increases after each iteration of the algorithm. The proof of this theorem is based on the definition of the M-step that ensures

$$Q(\psi;\psi^{(h+1)}) \geq Q(\psi;\psi^{(h)}),$$

while the application of the Jensen inequality gives

$$H(\psi;\psi^{(h+1)}) \leq H(\psi;\psi^{(h)}).$$

Put together and considering relation 2.1, these inequalities ensure the monotonicity of the likelihood sequence:

$$\log \mathcal{L}(y;\psi^{(h+1)}) \geq \log \mathcal{L}(y;\psi^{(h)}).$$

This inequality proves that the EM sequence of likelihoods must converge if the likelihood is bounded above.

2.2 *Formulation of the EM algorithm for mixture models*

When applied to the special case of mixture models the log-likelihoods are written in the form:

$$\begin{aligned}
\log \mathcal{L}(y;\psi) &= \sum_{t=1}^{n} \log f(y_t;\psi) = \sum_{t=1}^{n} \log \left\{ \sum_{p=1}^{P} \pi_p f(y_t;\theta_p) \right\} \\
\log \mathcal{L}^c(x;\psi) &= \sum_{t=1}^{n} \log g(x_t;\psi) = \sum_{t=1}^{n} \sum_{p=1}^{P} z_{tp} \log \left\{ \pi_p f(y_t;\theta_p) \right\}
\end{aligned}$$

Since the complete data log-likelihood is linear in the unobservable data z_{tp} the E-step only requires the computation of the conditional expectation of the missing information given the observed data y_t, using the current fit $\psi^{(h)}$ for ψ. It gives

$$Q(\psi;\psi^{(h)}) = \sum_{t=1}^{n} \sum_{p=1}^{P} \mathbb{E}_{\psi^{(h)}} \left\{ Z_{tp} | Y_t = y_t \right\} \log \left\{ \pi_p f(y_t;\theta_p) \right\},$$

with

$$\mathbb{E}_{\psi^{(h)}} \left\{ Z_{tp} | Y_t = y_t \right\} = \Pr_{\psi^{(h)}} \left\{ Z_{tp} = 1 | Y_t = y_t \right\} = \tau_{tp}^{(h)},$$

and

$$\tau_{tp}^{(h)} = \frac{\pi_p^{(h-1)} f(y_t;\theta_p^{(h-1)})}{\sum_{\ell=1}^{P} \pi_\ell^{(h-1)} f(y_t;\theta_\ell^{(h-1)})}.$$

Then

$$Q(\psi;\psi^{(h)}) = \sum_{t=1}^{n} \sum_{p=1}^{P} \tau_{tp}^{(h)} \log \left\{ \pi_p f(y_t;\theta_p) \right\}.$$

The M-step requires the global maximization of $Q(\psi;\psi^{(h)})$ with respect to ψ to give an updated estimate $\psi^{(h+1)}$.

For finite mixture models, the estimation of the mixing proportions is done via constrained maximization of the incomplete-data log-likelihood which gives:

$$\hat{\pi}_p^{(h+1)} = \frac{\sum_{t=1}^n \tau_{tp}^{(h)}}{n}.$$

This estimator has a natural interpretation: it summarizes the contribution of each data point y_t to the p^{th} component of the mixture via its *posterior* probability of membership. As for the updating of θ, it is obtained as an appropriate root of

$$\sum_{t=1}^{n}\sum_{p=1}^{P} \tau_{tp}^{(h)} \frac{\partial \log f(y_t;\theta_p)}{\partial \theta} = 0.$$

2.3 *Information matrix using the EM algorithm*

Once the parameters of the mixture have been estimated via maximum likelihood, a natural question is to assess the standard errors of the estimator $\hat{\psi}$. This can be done with the evaluation of the expected information matrix

$$\mathcal{I}(\psi) = \mathbb{E}_Y \left\{ \frac{-\partial^2}{\partial\psi\partial\psi^T} \log \mathcal{L}(Y;\psi) \right\},$$

with $\log \mathcal{L}(y;\psi)$ being the incomplete-data likelihood calculated on the available observations, and $\mathbb{E}_Y\{\cdot\}$ designating the expectation operator with respect to the random variable Y.

In practice this quantity is often estimated by the observed information matrix calculated at $\hat{\psi}$, $I(\hat{\psi}, y)$, with the relationship

$$\mathcal{I}(\psi) = \mathbb{E}_Y\left\{I(\psi;Y)\right\}.$$

Efron and Hinkley (1978) have provided a justification for this approximation. Since the data Y are considered as incomplete within the EM framework, $I(\psi;Y)$ will be denoted as the incomplete-data observed information matrix.

The use of the EM algorithm is often motivated by the analytic form of the observed-data likelihood, whose gradient or curvature matrices are difficult to derive analytically (which is typically the case for mixture models). As the estimation problem has been solved using the missing-data framework of EM, the derivation of the information matrix $I(\psi; y)$ can be simplified using the missing information principle introduced by Woodbury (1971).

Missing information principle

If we consider the formulation of mixtures as a missing-data problem, we define the complete-data observed information matrix based on the complete-data log-likelihood:

$$I^c(\psi; x) = \frac{-\partial^2}{\partial\psi\partial\psi^T} \log \mathcal{L}^c(x; \psi).$$

Since the incomplete data and the complete data likelihood are linked by definition:

$$\log \mathcal{L}(y; \psi) = \log \mathcal{L}^c(x; \psi) - \log k(z|y; \psi),$$

on differentiating both sides twice with respect to ψ, we have

$$I(\psi; y) = I^c(\psi; x) - I^m(\psi, z),$$

where

$$I^m(\psi, z) = \frac{-\partial^2}{\partial\psi\partial\psi^T} \log k(z|y; \psi)$$

is the missing-data observed information matrix. This term can be viewed as the "missing information", the consequence of having observed only y and not z.

Since the complete-data are not fully observed, we take the conditional expectation of both sides over Y that yields to:

$$I(\psi; y) = \mathbb{E}_{X|Y}\left\{I^c(\psi; X)\right\} - \mathbb{E}_{Z|Y}\left\{I^m(\psi, Z)\right\} \tag{4}$$

Then the problem is to formulate the conditional expectations of $I^c(\psi; x)$ and $I^m(\psi, z)$ in directly computable terms within the EM framework.

Extracting the observed information matrix in terms of the complete-data likelihood

Let us introduce the score notation such that:

$$\begin{aligned} S(y; \psi) &= \frac{\partial}{\partial\psi} \log \mathcal{L}(y; \psi), \\ S^c(x; \psi) &= \frac{\partial}{\partial\psi} \log \mathcal{L}^c(x; \psi). \end{aligned}$$

Louis (1982) gives a formulation of the missing information matrix, in the form:

$$\mathbb{E}_{Z|Y}\left\{I^m(\psi, Z)\right\} = \mathbb{E}_{X|Y}\left\{S^c(X;\psi)S^c(X;\psi)^T\right\} - S(y;\psi)S(y;\psi)^T,$$

meaning that the all the conditional expectations calculated in 4 can be computed in the EM algorithm only using the conditional expectation of the gradient and curvature of the complete-data likelihood.

Since $S(y;\psi) = 0$ for $\psi = \hat{\psi}$, Formula 4 is restated as:

$$I(\hat{\psi}; y) = \mathbb{E}_{X|Y}\left\{I^c(\psi; X)\right\}\big|_{\psi=\hat{\psi}} - \mathbb{E}_{X|Y}\left\{S^c(X;\psi)S^c(X;\psi)^T\right\}\big|_{\psi=\hat{\psi}}.$$

Hence the observed information matrix of the initial incomplete-data problem can be computed as the conditional moments of the gradient and curvature matrix of the complete-data likelihood introduced in the EM framework.

2.4 *Convergence properties of the EM algorithm*

It has been seen in previous sections that the EM algorithm generates a sequence $\left(\psi^{(h)}\right)_{h\geq 0}$ which increases the incomplete data log-likelihood at each iteration. The convergence of this EM-generated sequence has been studied by many authors, such as Dempster et al. (1977) and Wu (1983). Under some regularity conditions of the model, Wu (1983) shows the convergence of the sequence $\psi^{(h)}$ to a stationary point of the incomplete-data likelihood. The convergence of the EM algorithm to a local maximum of the incomplete data likelihood has also been established by Wu (1983) under restrictive hypothesis, that have been released by Delyon et al. (1999). One important theorem is provided by Wu (1983):

Suppose that $Q(\psi, \Phi)$ is continuous in both ψ and Φ, then all the limit points of any instance $\{\psi^{(h)}\}$ of the EM algorithm are stationary points of $\mathcal{L}(\psi)$ and $\mathcal{L}(\psi^{(h)})$ converges monotonically to some value $\mathcal{L}^*$ for some stationary point ψ^*.

Moreover in many practical situations $\mathcal{L}^*$ will be a local maximum. In general if the likelihood has several stationary points the convergence of an EM sequence to a local/global maximum or to a saddle point will depend on the choice of the starting value $\psi^{(0)}$, unless the likelihood is unimodal.

2.5 *Modified versions of the EM algorithm*

Despite appealing features, the EM algorithm presents some well documented shortcomings: the resulting estimate $\hat{\psi}$ can strongly depend on the starting position $\psi^{(0)}$, the rate of convergence can be slow and it can provide a saddle point of the likelihood function rather than a local maximum. For these reasons several authors have proposed modified versions of the EM algorithm: deterministic improvements (Louis (1982), Meilijson (1989), Green (1990) for instance), and stochastic modifications (Broniatowski et al. (1983) Celeux and Dielbolt (1985) Wei and Tanner (1990), Delyon et al. (1999)).

Broniatowski et al. (1983) proposed a Stochastic EM algorithm (SEM) which provides an attractive alternative to EM. The motivation of the simulation step (S-step) is based on the Stochastic Imputation Principle, where the purpose of the S-step is to fill-in for the missing data z with a single draw from $k(z|y;\psi^{(h)})$. This imputation of z is based on all the available amount of information about ψ and provides a pseudo complete sample. More precisely the current *posterior* probabilities $\tau_{tp}^{(h)}$ are used in the S-step wherein a single draw from distribution $\mathcal{M}_P(1;\tau_{t1}^{(h)},\ldots,\tau_{tP}^{(h)})$ is used to assign each observation to one of the component of the mixture. The deterministic M-Step and the stochastic S-Step generate a Markov Chain $\psi^{(h)}$ which converges to a stationary distribution under mild conditions. In pratice a number of iterations is required as a burn in period to allow $\psi^{(h)}$ to approach its stationary regime. In mixture models 100-200 iterations are often used for burn in.

This stochastic step can be viewed as a random perturbation of the sequence $\psi^{(h)}$ generated by EM. This perturbation prevents the algorithm from staying near an unstable fixed point of EM, and prevents stable fixed points corresponding to insignificant local maxima of the likelihood. The Stochastic EM algorithm provides an interesting alternative to the limitations of EM, concerning local maxima and starting values.

Other stochastic versions of the EM algorithm have been proposed, among them, the Stochastic Annealing EM algorithm (SAEM, Celeux and Dielbolt (1992)) which is a modification of SEM, the Monte Carlo EM (Wei and Tanner (1990)), which replaces analytic computation of the conditional expectation of the complete-data log-likelihood by a Monte Carlo approximation, and a stochastic approximation of EM (Delyon et al. (1999)). Nevertheless, empirical studies from Dias and Wedel (2004) and Biernacki et al. (2003) suggest the practical use of SEM in the context of mixture models, for its simplicity of implementation compared with Monte Carlo-based improvements, for its quick rate of convergence, and for its property to avoid spurious local maximizers.

3. Choosing the number of clusters via model selection criteria

Choosing the number of clusters is often the first question that is asked by/to the analyst. Two approaches can be considered to answer this question. The first one can be to fix this number and to propose different classifications. Since every clustering method (heuristically or model-based) can be run for a fixed number of groups, this strategy can be applied to any method. Nevertheless, the question can be to score different classifications with different numbers of clusters. In the model-based context, the choice of the number of clusters can be formulated as a model selection problem, and it can be performed with a penalized criterion, such as:

$$\log \mathcal{L}_P(y; \hat{\psi}) - \beta pen(P),$$

with $\log \mathcal{L}_P(y; \hat{\psi})$ being the observed data log-likelihood for a mixture with P clusters, calculated at $\psi = \hat{\psi}$, β a positive constant and $pen(P)$ an increasing function with respect to the number of clusters.

3.1 *Bayesian approaches for model selection*

As previously described in the context of segmentation methods (**??**), the purpose of model selection is to select a candidate model m_i among a finite collection of models $\{m_1, \dots, m_\ell\}$, in order to estimate function f from which the data $Y = \{Y_1, \dots, Y_n\}$ are drawn. Each model is characterized by a density g_{m_i} whose parameters ψ_i are of dimension ν_i.

In the Bayesian context, ψ_i and m_i are viewed as random variables with *prior* distributions noted $\Pr\{m_i\}$ and $\Pr\{\psi_i | m_i\}$ for ψ_i when model m_i is fixed. This formulation is flexible since additional information can be modelled through *prior* distributions, and if no information is available a non-informative prior can be used. The Bayesian Information Criterion (BIC) developed by Schwartz (1978) aims at selecting the model which maximizes the posterior probability $\Pr\{m_i | Y\}$. Using the Bayes formula:

$$\Pr\{m_i | Y\} = \frac{\Pr\{Y | m_i\} \Pr\{m_i\}}{\Pr\{Y\}},$$

and considering the case where the prior distribution $\Pr\{m_i\}$ is non informative, the search for the best model only requires the computation of distribution $\Pr\{Y | m_i\}$ which is the integrated likelihood of the data for model

m_i. This distribution can be approximated using the Laplace approximation method (see Lebarbier and Mary-Huard (2004) for more details), which yields to the following penalized criterion:

$$BIC_i = -2\Pr\{Y|m_i\} \simeq -2\log g_{m_i}(Y, \hat{\psi}_i) + \nu_i \times \log(n),$$

where $\hat{\psi}_i$ is the maximum likelihood estimator of ψ_i. The BIC is used to assess a score to each model m_i and the selected model is such that:

$$\hat{m}_{BIC} = \underset{i}{\mathrm{Argmax}}\, BIC_i.$$

Interestingly regularity conditions for BIC do not hold for mixture models, since the estimates of some mixing proportions can be on the boundary of the parameter space. Nevertheless there is considerable practical support for its use in this context (see Fraley and Raftery (1998) for instance). Other approaches have been considered for Bayesian model selection (see Kass and Raftery (1995) for a complete review on Bayes Factors for instance). Nevertheless the BIC has focused much attention, for its simplicity of implementation and for its statistical properties. Gassiat and Dacunha-Castelle (1997) have shown that the use of BIC leads to a consistent estimator of the number of clusters.

3.2 *Strategy-oriented criteria*

Other criteria have been defined for the special case of mixture models. They can be based on Bayesian methods, on the entropy function of the mixture, or on information theory. The reader is referred to McLachlan and Peel (2000) for a complete review on the construction of those criteria. Empirical comparisons of those criteria have been extensively used to determine the "best" criterion. As noted by Biernacki et al. (2000), the use of the BIC can lead to an overestimation of the number of clusters regardless the clusters separation. Moreover estimating the "true" numbers of clusters, which is the objective of the BIC, is not necessarily suitable in a practical context. For these reasons, Biernacki et al. (2000) propose a new criterion, the Integrated Classification Criterion (ICL) that considers the clustering objective of mixture models. In this paragraph we present the main steps of the construction of ICL.

In a mixture model context, the integrated likelihood is noted $f(y|m_P)$

for a model m with P clusters. It is calculated such that:

$$f(y|m_P) = \int_{\Psi_P} f(y|m_P, \psi) h(\psi|m_P) d\psi,$$

with

$$f(y|m_P, \psi) = \prod_{t=1}^{n} f(y_t|m_P, \psi),$$

Ψ_P being the parameter space of model m_P, and $h(\psi|m_P)$ a non-informative prior distribution on ψ. Instead of considering the incomplete-data integrated likelihood for which the BIC approximation is not valid, the authors suggest to use the complete-data integrated likelihood or integrated classification likelihood:

$$f(y, z|m_P) = \int_{\Psi_P} f(y, z|m_P, \psi) h(\psi|m_P) d\psi,$$

with

$$f(y, z|m_P, \psi) = \prod_{t=1}^{n} \prod_{p=1}^{P} \left\{ \pi_p f(y_t; \theta_p) \right\}^{z_{tp}}.$$

Then the idea is to isolate the contribution of the missing data z by conditioning on z, and it follows that:

$$f(y, z|m_P) = f(y|z, m_P) f(z|m_P),$$

provided that $h(\psi|m_P) = h(\theta|m_P) h(\pi|m_P)$.

The authors emphasize that the BIC approximation is valid for the term $f(y|z, m_P)$, such that:

$$\log f(y|z, m_P) \simeq \max_{\theta} \log f(y|z, m_P, \theta) - \frac{\lambda_P}{2} \log(n),$$

where λ_P is the number of free components in θ. Note that the parameter θ which maximizes $\log f(y|z, m_P, \theta)$ is not the maximum likelihood estimator. Nevertheless, the authors propose to use the maximum likelihood estimator as an approximation.

As for term $f(z|m_P)$ it can be directly calculated using a Dirichlet prior $\mathcal{D}(\delta, \ldots, \delta)$ on proportion parameters. It follows that:

$$f(z|m_P) = \int \pi_1^{n_1} \ldots, \pi_P^{n_P} \frac{\Gamma(P\delta)}{\Gamma(\delta)^P} 1\!\!1_{\sum_p \pi_p = 1} d\pi,$$

with n_p being the number of data points belonging to cluster p. Then parameter δ is fixed at $1/2$ which corresponds to the Jeffreys non-informative distribution for proportion parameters.

The last steps of the construction of ICL consists in replacing the missing data z which are unknown by the recovered label variables $\tilde{z}$ using a MAP rule. Then an approximation of $f(z|m_P)$ is given when n is large. It follows that:

$$ICL(m_P) = \max_{\psi} \log f(y, \tilde{z}|m_P, \psi) - \frac{\nu_P}{2} \log(n),$$

with ν_P the number of free parameters for model m_P. Therefore the ICL criterion is an " *la BIC*" approximation of the completed log-likelihood or classification log-likelihood. Since this criterion considers the classification results to score each model it has been shown to lead to a more sensible partitioning of the data, compared with BIC.

The performance of ICL have been tested based on real and simulated data sets. Compared with BIC, ICL tends to select a lower number of clusters which provides good clustering results in real situations, compared with BIC which tends to select a too overly high number of clusters. When the data are simulated, ICL tends to select a lower number of clusters if the groups are not well separated, contrary to BIC which finds the true number of classes. From a theoretical point of view, no result has yet been demonstrated for the properties of ICL.

References

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**, 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.

Broniatowski, M., Celeux, G. and Dielbolt, J. (1983). Reconnaissance de mlanges de densits par un algorithme d'apprentissage probabiliste. *Data analysis and informatics* **3**, 359–374.

Celeux, G. and Dielbolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistic Quaterly* **2**, 73–82.

Celeux, G. and Dielbolt, J. (1992). A stochastic approximation type EM algorithm for the mixture problem. *Stichastics and Stochastic reports* **41**, 119–134.

Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics* **27**, 94–28.

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

Dias, J. and Wedel, M. (2004). An empirical comparison of EM, SEM and MCMC performances for problematic Gaussian mixture likelihoods. *Statistics and Computing* **14**, 323–332.

Efron, B. and Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. *Biometrika* **65**, 457–487.

Fraley, C. and Raftery, A. (1998). How many clusters ? *The Computer Journal* **41**, 578–587.

Gassiat, E. and Dacunha-Castelle, D. (1997). Estimation of the number of components in a mixture. *Bernoulli* **3**, 279–299.

Green, P. (1990). On the use of the EM algorithm for penalized likelihood estimation. *JRSS-B* **52**, 443–452.

Kass, E. and Raftery, A. (1995). Bayes factors. *Journal of the American statistical Association* **90**, 773–795.

Lebarbier, E. and Mary-Huard, T. (2004). Le critre BIC: fondements thoriques et interprtation. Technical Report 5315, INRIA.

Louis, T. A. (1982). Finding the observed information matrix when using the EM-algorithm. *JRSS-B* **44**, 226–233.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley Inter-Science.

Meilijson, I. (1989). A fast improvement to the EM algorithm in its own terms. *JRSS-B* **51**, 127–138.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461–464.

Titterington, D., Smith, A. and Makov, U. (1985). *Statistical analysis of finite mixture distributions.* Wiley.

Wei, G. and Tanner, M. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *JASA* **82**, 528–550.

Woodbury, M. (1971). Discussion of paper by Hartley and Hocking. *Biometrics* **27**, 808–817.

Wu, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics* **11**, 95–103.