A LASSO-penalized BIC for mixture model selection

Sakyajit Bhattacharya · Paul D. McNicholas

Received: 27 November 2012 / Revised: 10 June 2013 / Accepted: 17 October 2013 / Published online: 8 November 2013
© Springer-Verlag Berlin Heidelberg 2013

Abstract The efficacy of family-based approaches to mixture model-based clustering and classification depends on the selection of parsimonious models. Current wisdom suggests the Bayesian information criterion (BIC) for mixture model selection. However, the BIC has well-known limitations, including a tendency to overestimate the number of components as well as a proclivity for underestimating, often drastically, the number of components in higher dimensions. While the former problem might be soluble by merging components, the latter is impossible to mitigate in clustering and classification applications. In this paper, a LASSO-penalized BIC (LPBIC) is introduced to overcome this problem. This approach is illustrated based on applications of extensions of mixtures of factor analyzers, where the LPBIC is used to select both the number of components and the number of latent factors. The LPBIC is shown to match or outperform the BIC in several situations.

Keywords BIC · LASSO · Mixture models · Model-based clustering · Model selection

Mathematics Subject Classification 62F99 · 62H30 · 62H25

This work was supported by a Collaborative Research and Development grant from the Natural Sciences and Engineering Research Council of Canada, a Grant-in-Aid from Compusense Inc., a Collaborative Research grant from the Ontario Centres of Excellence, and an Early Researcher Award from the Ontario Ministry of Research and Innovation.

S. Bhattacharya · P. D. McNicholas
Department of Mathematics and Statistics, University of Guelph, Guelph, ON N1G 2W1, Canada
e-mail: pmcnicho@uoguelph.ca
1 Introduction

Consider \( n \) realizations \( x = (x_1, x_2, \ldots, x_n) \) of a \( p \)-dimensional random variable \( X \) that follows a \( G \)-component finite Gaussian mixture model. The likelihood is given by

\[
\mathcal{L}(\hat{\theta} \mid x) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g \phi(x_i \mid \mu_g, \Sigma_g),
\]

where \( \pi_g > 0 \), with \( \sum_{g=1}^{G} \pi_g = 1 \), are the mixing proportions, \( \phi(x \mid \mu_g, \Sigma_g) \) is the multivariate Gaussian density with mean \( \mu_g \) and covariance matrix \( \Sigma_g \), and we denote the model parameters using \( \hat{\theta} = (\pi_1, \ldots, \pi_G, \mu_1, \ldots, \mu_G, \Sigma_1, \ldots, \Sigma_G) \).

A model-based clustering approach assumes that each component or some combination thereof corresponds to a cluster. When fitting the model in (1), the main task is to decide the number of components \( G \). Titterington et al. (1985), McLachlan and Basford (1988), and McLachlan and Peel (2002) extensively review mixture models with a focus on Gaussian mixture models. Fraley and Raftery (2002) present a review of work on Gaussian mixtures with a focus on clustering, discriminant analysis, and density estimation. They discuss a family of Gaussian mixture models that arises from the imposition of constraints upon an eigen-decomposition of the component covariance structure. The family of mixture models they discuss, known as MCLUST, is actually a subset of the Gaussian parsimonious clustering models (GPCMs) of Celeux and Govaert (1995). When using the MCLUST models, one must choose the appropriate member of the family, i.e., the covariance structure, in addition to deciding the number of components \( G \).

Ghahramani and Hinton (1997) introduce a mixture of factor analyzers model, which is further developed by Tipping and Bishop (1999) and McLachlan and Peel (2000). Through foisting constraints on the covariance structure, McNicholas and Murphy (2008, 2010) develop mixtures of factor analyzers into a family of parsimonious Gaussian mixture models (PGMMs). In addition to selecting the member of the family (i.e., the covariance structure) and the number of components, one must now also select the number of latent factors. Further complicating the model selection problem here is the fact that PGMMs are often applied to high-dimensional data. McNicholas et al. (2010) explain why the PGMMs are particularly suited to the analysis of high-dimensional data; amongst the most salient points is the fact that, unlike families such as MCLUST, the number of covariance parameters is linear in data dimensionality for every member of the PGMM family.

There are a number of well-known methods to select the best mixture model but the Bayesian information criterion (BIC; Schwartz 1978) remains by far and away the most popular. We have

\[
\text{BIC} = 2 \log \mathcal{L}(\hat{\theta} \mid x) - \rho \log n,
\]

where \( \hat{\theta} \) is the maximum likelihood estimate (MLE) of \( \theta \), \( \mathcal{L} \) is the likelihood, \( \rho \) is the number of free parameters, and \( n \) is the number of observations. For a family of mixture models, the model having the maximum BIC is selected. Some other
well-known model selection criteria are the Akaike information criteria (AIC; Akaike 1974), deviance information criteria (Zhu and Carlin 2000), consistent AIC (CAIC), CAICFCorrected (Bozdogan 1987), and the more recent information complexity criterion (ICOMP; Bozdogan 2000). The last is a consistent information criterion, using the number of estimated free parameters in the penalty term and thus enforcing a stricter penalty than the BIC. However, the BIC remains by far the most popular criterion for mixture model selection within the literature. The use of the BIC is theoretically justified by a number of authors, e.g., Kass and Wasserman (1995), Kass and Raftery (1995), and Keirbin (2000). In particular, the BIC has some useful asymptotic properties, e.g., the criterion consistently chooses the correct model under an increasing number of observations (Shibata 1986).

Nevertheless, the BIC is not without drawbacks. The criterion is derived using a Laplace approximation and its precision is influenced by the specific form of the prior density of the parameters as well as the correlation structure between observations. Recently, Clyde et al. (2007) rectified the problems of the marginal distribution of the parameter caused by the Laplace approximation. In addition, Fraley and Raftery (2007) propose a Bayesian regularization for Gaussian mixtures. Their method assumes predefined priors that lead to a modified version of the BIC, using posterior modes instead of the MLEs of the parameters. The resulting method avoids degeneracies, singularities, and the problem of flat priors. However, another more serious problem has not been addressed, i.e., the problem of high-dimensional cases.

The penalty term in the BIC is $\rho \log n$, cf. (2). Therefore, in a high-dimensional setting where $p \gg n$, the penalty term dominates the likelihood and so the BIC is prone to underfitting. Parametric estimation for high-dimensional cases has been studied by a number of authors, mostly within the linear regression set-up. The celebrated LASSO method (Tibshirani 1996) is perhaps the most popular among them. This method minimizes the residual sum of squares under the constraint that the sum of the absolute values of the regression coefficients is less than some constant, leading to sparse solutions of the coefficients and thus an interpretable model. Different variations of the LASSO have since been proposed, depending on the nature of the regression and the asymptotic behaviour. These include the adaptive LASSO (Zou 2006), the fused LASSO (Tibshirani et al. 2005), and the graphical LASSO (Friedman et al. 2008). Fan and Li (2001) provide a theoretical discussion of variable selection via a non-concave penalized likelihood procedure where the LASSO is a special case. They also propose that a good penalized estimation should satisfy the oracle properties, i.e., it should be consistent and the estimates should be asymptotically Gaussian.

Following the idea of Fan and Li (2001), Khalili and Chen (2007) propose the use of the penalized likelihood in finite mixtures of regression models, where the penalty is non-concave and the LASSO is a special case. They also devise a method of selecting the tuning parameter as well as conditions under which the estimation procedure would satisfy the oracle properties. Their method is especially suitable for finite mixtures of regression models, though no new model selection criterion is proposed. It should also be noted that the theoretical results regarding the asymptotic properties are somehow strange, because the authors use the same tuning parameter when comparing two different estimates for a fixed cluster. Chen and Chen (2008) propose an extended BIC for regression in high-dimensional settings. The extended
BIC assumes a prior inversely proportional to the size of the assumed model instead of a flat prior. The criterion is consistent and computationally cheap. Interestingly, the authors do not propose any penalized likelihood; instead they maximize the natural likelihood, thus using the conventional estimation procedure. The above estimation procedures, though interesting and useful, are mainly for regression-type problems and are not applicable to mixture model-based clustering and classification. Also, as the authors rightly point out, the approach is computationally infeasible if $p \gg n$. Nevertheless, useful extensions are possible. Herein, we draw upon some mathematical results from Fan and Li (2001) and Khalili and Chen (2007), especially on the issues of the choice of penalty and consistency.

The use of penalized likelihood in mixture model-based clustering was proposed by Pan and Shen (2007), where a LASSO-type penalty is applied to the likelihood. From there, they go on to propose a modified BIC that would be well-suited for high-dimensional settings. The limitation of their method is that this criterion works only for a common, diagonal component covariance matrix. Furthermore, the authors do not study the asymptotic properties, which are important in the sense that the classical LASSO method can be inconsistent (cf. Zou 2006). An ideal criterion should be analytically derivable from the penalized likelihood, work well for an arbitrary model, and have some good asymptotic properties. The work presented herein attempts to address these requirements by proposing a LASSO-penalized BIC (LPBIC) for model selection within high-dimensional settings for the PGMM family.

While deriving the MLE of the unknown parameters, we use a penalized likelihood approach. In particular, instead of maximizing the likelihood $\mathcal{L}(\theta \mid x)$, we maximize the penalized log-likelihood

$$\log \mathcal{L}(\theta \mid x) - \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} \varphi(\mu_{gj}).$$

We use a LASSO-like penalty for $\varphi(\mu_{gj})$. In particular, $\varphi(\mu_{gj}) = n\lambda_n |\mu_{gj}|$, where $\mu_{gj}$ is the $j$th element in $\mu_g$ and $\lambda_n$ is the tuning parameter that depends on $n$. Though a LASSO penalty is used here, other types of non-concave penalties can also be suitable. For example, one might use the hard thresholding (HARD) penalty $\varphi(\mu_{gj}) = [\lambda_n^2 - (\sqrt{n} \mu_{gj} - \lambda_n)^2 I(\sqrt{n} \mu_{gj} < \lambda_n)]$ or the smoothly clipped absolute deviation (SCAD) penalty, as discussed by Fan and Li (2001). One problem with using such an $L_1$-norm penalty is that the oracle properties might not be fully satisfied: the estimation can be consistent but not asymptotically normal. HARD or SCAD penalties satisfy both these properties and these issues are discussed in more detail in Sect. 3. However, we still prefer the LASSO-type penalty because it is computationally easier due to its convexity.

From this penalized likelihood, we derive a model selection criterion. We use a modified alternating expectation-conditional maximization (AECM) algorithm (Meng and Dyk 1997) to estimate the parameters in the PGMM models. We show that our LPBIC generally outperforms the BIC for the PGMM family in high-dimensional settings.

The remainder of this paper is laid out as follows. In Sect. 2, we discuss parameter estimation under the penalized likelihood approach and derive an LPBIC. The asymp-
totic properties of the LPBIC are discussed (Sect. 3) and we illustrate our approach on real and simulated data (Sect. 4). The real data considered exhibit the ‘small $n$, large $p$’ property and our data analysis results are compared with the BIC. The paper concludes with a discussion (Sect. 5). The mathematical derivation of the LPBIC and its asymptotic properties are described in appendices.

2 Method

Again, suppose we observe $x = (x_1, x_2, \ldots, x_n)$ with $f(x \mid \vartheta) = \sum_{g=1}^{G} \pi_g \phi(x \mid \mu_g, \Sigma_g)$, where $\phi(x \mid \mu_g, \Sigma_g)$ is the multivariate Gaussian density with mean $\mu_g$ and covariance matrix $\Sigma_g$. Now, instead of maximizing the likelihood $\mathcal{L}(\vartheta \mid x)$, we maximize the penalized log-likelihood

$$\log \mathcal{L}_{\text{pen}}(\vartheta \mid x) = \log \mathcal{L}(\vartheta \mid x) - n\lambda_n \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} |\mu_{gj}|,$$

where $\mu_g$ and $\lambda_n$ are defined as before. Hereafter, we denote $\varphi(\mu) = \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} \varphi(\mu_{gj})$ and so

$$\log \mathcal{L}_{\text{pen}}(\vartheta \mid x) = \log \mathcal{L}(\vartheta \mid x) - \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p} \varphi(\mu_{gj}) = \log \mathcal{L}(\vartheta \mid x) - \varphi(\mu).$$

Before going into details of parameter estimation, we make two assumptions. First, we observe that the penalty function is non-concave and singular at the origin; it does not have a second derivative at 0. We locally approximate the penalty by a quadratic function as first suggested by Heiser (1995). The parameters are estimated by successive iterations. Suppose $\mu^{(m)}$ is the estimate of $\mu$ after $m$ iterations. The penalty can be locally approximated as

$$\varphi(\mu) \approx n\lambda_n \sum_{g=1}^{G} \pi_g \sum_{j=1}^{p_g} \left[ |\mu_{gj}^{(m)}| + \frac{1}{2} \frac{\mathbf{1}[\mu_{gj}^{(m)}]}{\mu_{gj}^{(m)}} \left( \mu_{gj}^{(m)} - \mu_{gj}^{(m)^2} \right) \right],$$

where $p_g$ is the number of non-zero elements in $\mu_g$. The approximation of the sum of the absolute values of the mean components as above is exactly equal to the majorization of the absolute value of the mean components by quadratic majorization. A number of authors, including Fan and Li (2001) and Leeuw (1994), have used this majorization approach for maximizing the penalized likelihood. Moreover, Lange et al. (2000) prove the majorizing property of such functions—i.e., the penalized likelihood increases at each iteration of the expectation-maximization (EM) algorithm (Dempster et al. 1977)—in the context of various optimization transfer algorithms including the EM. Recently, Deun et al. (2011) use this approach for maximizing the penalized log-likelihood in the context of a sparse principal components analysis approach for different penalty functions, including the LASSO, group LASSO, elitist LASSO, and ridge regression. This is an example of a majorization-minimization approach, which replaces functions that are complicated to minimize by surrogate functions that are
easier to minimize (cf. Hunter and Lange 2004). These surrogate functions lie above the original function, and minimizing them is equivalent to minimizing the original function.

We also assume that the marginal distribution of the mixing proportions \( \pi = (\pi_1, \pi_2, \ldots, \pi_G) \) is uniform on the simplex and that \( \mu_g \sim \mathcal{N}(\tilde{\mu}_g, I(\tilde{\mu}_g)^{-1}) \), for \( g = 1, 2, \ldots, G \), where \( \tilde{\mu}_g \) is the MLE derived by maximizing the penalized likelihood \( \mathcal{L}_{\text{pen}} \) and \( I(\tilde{\mu}_g) \) is the unit information matrix at \( \tilde{\mu}_g \).

Define \( z_i = (z_{i1}, \ldots, z_{iG}) \) to denote the component membership of the \( i \)th observation, so that \( z_{ig} = 1 \) if \( x_i \) belongs to the \( g \)th component and \( z_{ig} = 0 \) otherwise.

To estimate the parameters, we use the AECM algorithm composed of two stages. In the first stage, when estimating \( \pi_g \) and \( \mu_g, z_1, \ldots, z_n \) is treated as the missing data.

The expected complete-data log-likelihood is

\[
Q(\pi, \mu) = \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \log \pi_g + \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig} \log \{ \phi(x_i \mid \mu_g, \Sigma_g) \} - \phi(\mu),
\]

where \( \hat{z}_{ig} = \hat{\pi}_g \phi(x_i \mid \hat{\mu}_g, \hat{\Sigma}_g) / \sum_{j=1}^{G} \hat{\pi}_j \phi(x_i \mid \hat{\mu}_j, \hat{\Sigma}_j) \). The M-step maximizes \( Q \) to update the parameter estimates \( \pi_g \) and \( \mu_g \). The estimation of \( \pi_g \) is complicated and has a complex analytic form. However, we have observed that, in practical applications, the analytical estimate is equivalent to the estimate derived by the EM algorithm. Hence, in our analyses (Sect. 4), \( \pi_g \) can be estimated via

\[
\hat{\pi}_g = \frac{\sum_{i=1}^{n} \hat{z}_{ig}}{n}.
\]

For the mean parameters,

\[
\frac{\partial Q}{\partial \mu_g} = \hat{\Sigma}_g^{-1} \sum_{i=1}^{n} \hat{z}_{ig} (x_i - \mu_g) - n \lambda_n \hat{\pi}_g \text{sign}(\mu_g).
\]

By equating the above expression to 0, we get

\[
\tilde{\mu}_{gj} = \hat{\mu}_{gj} - \lambda_n (\hat{\Sigma}_g \text{sign}(\hat{\mu}_g))_j = 0,
\]

where \( \tilde{\mu}_{gj} = \sum_{i=1}^{n} \hat{z}_{ig} x_{ig} / \sum_{i=1}^{n} \hat{z}_{ig} \) is the update of \( \mu_{gj} \) if no penalty term is involved. Also, imposing a LASSO penalty on the mean components implies that the estimate of \( \mu_{gj} \) for a non-penalized case, i.e. \( \tilde{\mu}_{gj} \), is shrunken towards zero, leading to a sparse solution. Therefore, the new estimate is either zero or a value lying between zero and the original estimate \( \tilde{\mu}_{gj} \). In other words, either \( \hat{\mu}_{gj} = 0 \) or \( \text{sign}(\tilde{\mu}_{gj}) = \text{sign}(\tilde{\mu}_{gj}) \).

Thus, if \( \tilde{\mu}_{gj} \neq 0 \), then

\[
| \tilde{\mu}_{gj} - \hat{\mu}_{gj} | < | \tilde{\mu}_{gj} |.
\]

Using (5) in (6) and the sparsity described above, we get \( \lambda_n < | \tilde{\mu}_{gj} | / (\hat{\Sigma}_g \text{sign}(\tilde{\mu}_g))_j \) and \( \text{sign}(\tilde{\mu}_{gj}) = \text{sign}((\hat{\Sigma}_g \text{sign}(\tilde{\mu}_g))_j) \), because \( \lambda_n \) is non-negative. Thus, (5) and the above arguments lead to an estimate of \( \mu_{gj} \).
\[ \hat{\mu}_{gj} = \text{sign}(\hat{\mu}_{gj})\left[|\hat{\mu}_{gj}| - \lambda_n|\left(\hat{\Sigma}_g\text{sign}(\hat{\mu}_g)\right)_j|\right]_+, \]

where, for any \( \alpha, \alpha_+ = \alpha \) if \( \alpha > 0 \) and \( \alpha_+ = 0 \) otherwise. \( \hat{\mu}_{gj} \) is a shrunk estimate of \( \mu_{gj} \) in the sense that \( \hat{\mu}_{gj} = 0 \) if \( \lambda_n \geq |\hat{\mu}_{gj}|/|\left(\hat{\Sigma}_g\text{sign}(\hat{\mu}_g)\right)_j| \). Otherwise, \( \hat{\mu}_{gj} \) is obtained by shrinking the usual EM estimate \( \hat{\mu}_{gj} \) by the amount \( \lambda_n|\left(\hat{\Sigma}_g\text{sign}(\mu_{gj})\right)_j| \) towards 0.

At the second stage of the AECM algorithm, we take the missing data as the group labels \( z_1, \ldots, z_n \) and the unobserved latent factors \( u_1, \ldots, u_n \) to estimate the covariance matrix under the PGMM set-up. The component covariance matrices \( \Sigma_1, \ldots, \Sigma_G \) are updated as usual, see McNicholas and Murphy (2008, 2010) for details. The first stage, where the \( \mu_g \) and \( \pi_g \) are estimated based on the complete-data \( (x, z) \), and the second stage, where the constituent parts of the \( \Sigma_g \) are estimated based on the complete-data \( (x, z, u) \), are iterated until convergence. Extensive details on an AECM algorithm for fitting the members of the PGMM family are given by McLachlan and Peel (2000) and McNicholas et al. (2010).

To derive a model selection criterion from the penalized log-likelihood, we maximize (3). Using (4), the second term of (3) becomes

\[ \frac{\lambda_n}{G} \sum_{g=1}^{G} \sum_{j=1}^{p_g} \left[ |\hat{\mu}_{gj}| + \frac{1}{2} \frac{\text{sign}(\mu_g)}{\hat{\mu}_{gj}} (\mu_{gj}^2 - \hat{\mu}_{gj}^2) \right], \]

where \( p_g \) is the number of non-zero mean components in class \( g \).

Here, we make an assumption that, for a given model, the mixture components are chosen independently so that the parameters for any two clusters are independent. Hence, using the weak law of large numbers with the BIC-type approximation to \( \log L(\theta \mid x) \), the penalized BIC is

\[ \text{LPBIC} = 2 \log L(\hat{\theta} \mid x) - \tilde{\rho} \log n \]

\[ -2n\lambda_n \sum_{g=1}^{G} \sum_{j=1}^{p_g} \left[ |\hat{\mu}_{gj}| + \frac{1}{|\hat{\mu}_{gj}|} (I(\hat{\mu}_g)^{-1})_{jj} - \text{sign}(\hat{\mu}_{gj}) \right], \]

where \( \tilde{\rho} \) is the number of estimated parameters that are non-zero. Intuitively, the LPBIC further penalizes the traditional BIC by both the absolute mean and absolute coefficient of variation of the parameters. The derivation is discussed in detail in Appendix A.

### 3 Asymptotic properties

#### 3.1 Properties

The consistency of a model selection criterion is closely related to the asymptotic identifiability of the model. In general, a model \( \mathcal{F} \) with the parameter set \( \theta \) is called identifiable if, for any two different sets of parameters \( \theta_1 \) and \( \theta_2 \),
We assume that our model satisfies the asymptotic identifiability condition. In the context of mixture models, a criterion is consistent if it can correctly select the number of components and the true set of parameters. If the true parameter set \( \theta_0 \) is decomposed as \( (\theta_{01}, \theta_{02}) \) such that \( \theta_{02} \) contains only the zero elements, and if any estimated parameter \( \hat{\theta} \) that is sufficiently close to \( \theta_0 \) is likewise decomposed as \( (\hat{\theta}_1, \hat{\theta}_2) \), then in order to satisfy consistency we should have \( P(\hat{\theta}_2 = 0) \rightarrow 1 \) as \( n \rightarrow \infty \) and \( \hat{\theta}_1 \rightarrow \theta_{01} \) in probability. Thus, the criterion should choose just as it would if the true number of clusters and the true parameters were known. Based on this idea, we study the consistency of the LPBIC with the help of the following assumptions:

1. Let \( p = O(n^\alpha) \) and \( \lambda_n = o(\log n/n) \). Define an estimate \( \hat{\theta} \) of \( \theta \) such that \( \| \hat{\theta} - \theta_0 \| = O(n^\kappa) \) for \( \kappa > -\infty \).
2. Let \( \theta = (\theta_1, \theta_2, \ldots, \theta_\nu) \). Then, there exist finite real numbers \( M_1 \) and \( M_2 \) (possibly depending on \( \kappa \)) such that
   \[
   \sup_j \left| \frac{\partial \log L(\theta | x)}{\partial \theta_j} \right| \leq M_1(x) \quad \text{and} \quad \sup_{j,k} \left| \frac{\partial^2 \log L(\theta | x)}{\partial \theta_j \partial \theta_k} \right| \leq M_2(x).
   \]
3. \( I(\theta) \) is positive-definite for all \( \theta \).

Then, under Assumptions 1 to 3, and assuming that the asymptotic identifiability condition is satisfied, we state the following theorem. The proof is given in Appendix B.

**Theorem 1** If \( \kappa < \min \{0, (\alpha - 1)/2\} \), then the LPBIC chooses the number of components and the set of parameters just as it would if \( \theta_0 \) were known as \( n \rightarrow \infty \). In other words, under the condition \( \kappa < \min \{0, (\alpha - 1)/2\} \), if there exists an estimate \( \hat{\theta} \) such that \( \| \hat{\theta} - \theta_0 \| = O(n^\kappa) \) and \( \text{LPBIC}(\hat{\theta}) \geq \text{LPBIC}(\theta) \) for all \( \theta \) such that \( \| \theta - \theta_0 \| = O(n^\kappa) \), then

a. \( P(\hat{\theta}_2 = 0) \rightarrow 1 \) as \( n \rightarrow \infty \), and
b. \( \hat{\theta}_1 \rightarrow \theta_{01} \) in probability as \( n \rightarrow \infty \).

We prove only Part a with some of the arguments proposed by Khalili and Chen (2007) for a false match rate setting. The method is modified for mixture models with high-dimensional set-up. Part b of the theorem can be proven by the method described in Fan and Li (2001). To prove Part b, we need \( \sqrt{n}\lambda_n \rightarrow 0 \) as \( n \rightarrow \infty \), which is satisfied by Assumption 1. This is particularly important because LASSO-type penalties do not satisfy the oracle property, i.e., they do not ensure the existence of a \( \sqrt{n} \)-consistent MLE of \( \theta \) that satisfies Part a and Part b. This is because the existence of a \( \sqrt{n} \)-consistent MLE requires that \( \sqrt{n}\lambda_n \rightarrow \infty \) and the consistency of \( \hat{\theta}_1 \) requires that \( \sqrt{n}\lambda_n \rightarrow 0 \). Hence, under a tighter assumption, we show that if such an estimator exists, then it satisfies consistency. However, other non-concave penalties, such as SCAD or HARD, can satisfy the oracle property with proper choice of the tuning parameter.
3.2 Choice of $\lambda_n$

Generally, the tuning parameters are chosen by cross-validation (Stone 1974) or generalized cross-validation (Craven and Wahaba 1979). We should remember that $\lambda_n$ depends on $n$. To satisfy the asymptotic properties, we require $\lambda_n = o(\log n/n)$. Khalili and Chen (2007) derive a component-wise deviance-based generalized cross validation with the above conditions in order to estimate $\lambda_n$. The method, though originally used in regression, also serves well for mixture models. The present paper takes the working sequence $\lambda_n = 1/p$ (as $p = O(n^\alpha)$ and $\alpha \geq 1$) and studies the behaviour of the LPBIC. The methods proposed by Khalili and Chen (2007), modified for a mixture model, are also considered and provide a range for the values of $\lambda_n$. It is observed that for moderately large $n$ ($n \geq 50$), $\lambda_n = 1/p$ falls into that range. For our data analysis (Sect. 4), we study the behaviour of the LPBIC for different values of $\lambda_n$ within that range. For illustration, though, a single $\lambda_n$ is chosen because the behaviour of the LPBIC is uniform over different $\lambda_n$ values within that range.

4 Data analysis

4.1 Overview

We analyze two high-dimensional data sets—one simulated and one real—and compare the results using the LPBIC to those with the BIC for the PGMM family. Although run as cluster analyses, the true group memberships are known in each case and we use the adjusted Rand index (ARI: Rand 1971; Hubert and Arabie 1985) to reflect classification agreement. A value of 1 indicates perfect agreement and a value of 0 would be expected under random classification.

4.2 Simulated data

We generate a simulated $p$-dimensional Gaussian data set consisting of three groups. We set $\mu_1 = -5.5\mathbf{1}$, $\Sigma_1$ isotropic; $\mu_2 = 2\mathbf{1}$, $\Sigma_2$ diagonal; and $\mu_3 = 3\mathbf{1}$, $\Sigma_3$ full. More specifically, the diagonal elements of $\Sigma_1$ are taken to be 0.5 and the diagonal elements of $\Sigma_2$ are randomly taken from a uniform distribution (0,1). $\Sigma_3$ is taken to be the covariance matrix of an AR(1) series with autocorrelation 0.9 and $\sigma = 1$. So, the $(i, j)$th element of the matrix is $0.9|i-j|$. 

Simulations are run in two steps. In the first step, we run a single simulation for each different $p$ and fixed $n$, with $n_1 = 40$, $n_2 = 30$, $n_3 = 30$. Here, we compare the performance of the LPBIC with the traditional BIC. In the second step, we run 25 simulations for $p = 500$ and vary $n$ as well as the relative size of the clusters within each $n$. In this scenario, the performance of the LPBIC is compared with other well-known model selection criteria such as the BIC, AIC, and CAIC.

In the first step, we run simulations for $p \in \{100, 250, 500\}$. LPBIC values are observed for each member of the PGMM family for $G = 1, \ldots, 4$ and $q = 1, 2, 3$, where $G$ is the number of components and $q$ is the number of latent factors. The results (Table 1) show that the LPBIC consistently chooses $G = 3$ as $p$ gets larger.
Table 1 Best models chosen by the LPBIC and BIC, respectively, for high-dimensional simulated data

|   | LPBIC | BIC |
|---|-------|-----|
|   | G     | q   | Model | ARI | G | q | Model | ARI |
| $p = 100$ | 3 | 3 | CUC | 0.88 | 3 | 3 | CUC | 0.86 |
| $p = 250$ | 3 | 2 | CUC | 0.82 | 2 | 1 | CCC | 0.62 |
| $p = 500$ | 3 | 3 | CUC | 0.97 | 2 | 1 | CCC | 0.49 |

![Fig. 1](image-url) Performance of the LPBIC and BIC for 25 simulations for $n = 100$, $n_1 = 40$, $n_2 = 30$, and $n_3 = 30$. The left plot shows the selection of number of components. The right plot shows the ARIs of the models selected.

but that the BIC fails in higher dimensions, choosing a $G = 2$ component model. The models in the table (i.e., CUC, CCC) are members, which correspond to covariance structures, of the PGMM family as described by McNicholas and Murphy (2008, 2010). The associated ARI values (Table 1) confirm that the models selected by the LPBIC capture the underlying group structure better than those chosen by the BIC, especially in higher dimensions.

The effect of increasing dimension on the performance of the BIC is clear: the BIC chooses fewer mixture components and latent factors as well as a more parsimonious covariance structure. However, the LPBIC chooses the same number of components and the same covariance structure each time, and the number of factors does not decrease with $p$.

In the second step, we generate 25 simulations of the $p = 500$ dimensional data for $n = 100$, 200, 250. For each $n$, three choices of relative cluster size are considered: 4:3:3, 3:4:3, and 3:3:4. Each choice assigns more weight to one cluster and equally partitions the rest of the weight to the other two clusters. The weights are so chosen to study how the performance of different model selection criteria is affected by varying the importance of different component covariance structures.

We study the behaviour of the BIC, LPBIC, AIC, and CAIC for selecting $G$ and for clustering performance (i.e., ARI). For $n = 100$ and $n_1 = 40$, $n_2 = 30$, and $n_3 = 30$, the results (Fig. 1) show the LPBIC correctly chooses the number of components ($G = 3$) 23 times but the BIC selects $G = 3$ only four times out of 25. As expected,
the BIC tends to choose too few components. The ARIs for models selected using the LPBIC are higher than those selected using the BIC. Out of 25 simulations, the ARI for the LPBIC is higher than for the BIC in 21 cases, illustrating generally superior clustering performance. The performance of the LPBIC, BIC, AIC, and CAIC are compared for different $n$ and varying cluster size in Fig. 2, which shows that the LPBIC usually has the best performance in selecting the right number of clusters. However, CAIC sometimes outperforms the LPBIC in terms of ARI. Also, the performance of the BIC improves and the performance of the LPBIC, CAIC, and BIC become more comparable as $n$ increases. However, for $n = 200$, both the LPBIC and CAIC demonstrate slight drawbacks in selecting the number of components when greater weight is put on the component with full covariance structure. This is because as more weight is put on the third component, the possibility of clusters overlapping increases. Because the LPBIC only penalizes the means and not the covariance matrices, it can mistakenly merge two overlapping clusters. This is a drawback of the LPBIC that we discuss further in Sect. 5.

4.3 Leukaemia data

Golub et al. (1999) presents data on two forms of acute leukaemia: acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML). Affymetrix arrays were used to collect measurements for 7,129 genes on 72 tissues (47 ALL and 25 AML). McLachlan et al. (2002) reduce the data set as follows:

1. Genes with expression falling outside the interval $(100, 16,000)$ are removed.
2. Genes with expression satisfying max/min $\leq 5$ or max-min $\leq 500$ are removed.

McNicholas and Murphy (2010) further reduce the number of genes to 2,030 by applying the select-genes software (cf. McLachlan et al. 2002). We analyze these
2,030 genes using 20 different random starts for the initial $\hat{z}_{ig}$. We run our approach for $G \in \{1, 2, 3\}$ and $q = 1, \ldots, 6$.

A summary of the models selected by the LPBIC and the BIC is given in Table 2. The BIC chooses a CCU model with $G = 1$ component and $q = 2$ factors. The LPBIC chooses a CUC model with $G = 2$ components and $q = 1$ factors. The ARI of the model chosen using the LPBIC (0.47) is greater than for the model chosen using the BIC (0.29). The model selected using the LPBIC misclassifies eleven of the 72 samples (Table 3).

### Table 2

Comparison of the performance of the LPBIC and BIC for PGMM model selection for the leukaemia data

|       | Value  | $G$ | $q$ | Model | ARI |
|-------|--------|-----|-----|-------|-----|
| BIC   | $-400,394$ | 1   | 2   | CCU   | 0.29 |
| LPBIC | $-391,023$ | 2   | 1   | CUC   | 0.47 |

### Table 3

Classification table for the best model chosen by the LPBIC for the leukaemia data

|       | 1 | 2 |
|-------|---|---|
| ALL   | 39 | 3 |
| AML   | 8  | 22|

## 5 Discussion

The paper proposes an LPBIC through a penalized likelihood-based approach in the context of model selection for a family of Gaussian mixture models, i.e., the PGMM family. The approach is mainly intended for high-dimensional settings, where the BIC has some unattractive problems due to an ‘exploding’ penalty term. Our LPBIC approach does not use the total number of independent parameters to be estimated in its penalty term but, rather, the total number of non-zero cluster centroid parameters to be estimated. This has some advantages. Because the likelihood is penalized by a tuning parameter, many of the elements of the component means become 0, thereby reducing the number of independent estimable parameters. The loss of information due to penalizing the likelihood is somehow compensated for by both the absolute mean and absolute coefficient of variation of the mean parameters.

The choice of tuning parameters is an important aspect in this scenario because no theoretical results exist to specify the best choice. Recently, Wang et al. (2007, 2009) proposed some interesting mathematical methods for choosing the tuning parameters without requiring cross-validation. However, their method is most suitable to low-dimensional settings. Herein, we followed an approach close to the one proposed by Fan and Li (2001), though careful modifications have been taken to preserve the asymptotic properties, accounting for the nature of the data.

Our method seems to consistently choose the correct number of clusters for high-dimensional data, as shown through the analysis of real and simulated data. Our analyses suggest that the LPBIC is an improvement over the BIC in high-dimensional settings. However, the oracle property is lost because the LASSO may fail to simultaneously satisfy consistency, sparsity, and asymptotic normality. That said, the LASSO
has some computational advantages because of convexity and, hence, it is preferred over alternative non-concave penalties.

Of course, the LPBIC is not without its issues. One problem arises by locally approximating the penalty function: if an estimator is shrunken, it stays at 0. Another arises if the initial domain of the estimates does not contain the posterior mode, or even if the posterior mode lies at the boundary of the domain. This second problem, which will lead to failure, is a general problem with the EM algorithm. Another drawback is that the penalty is solely a function of the cluster centroids (i.e., $G_p$ parameters). Hence, clusters that are well separated will likely be identified, but overlapping clusters can be incorrectly merged. Adding the component variances and covariances, which define each clusters’ shape, compactness, and volume can lead to greater parsimony and possibly help with overlapping cluster identifiability. In the MCLUST scenario, not penalizing the covariance matrices could lead to serious consequences as most of the parameters are in the covariance matrices. However, this is not a very serious problem in the PGMM set-up as the number of covariance parameters is linear in data dimensionality for every member of the PGMM family.

Future work will focus on the use of penalties that lead to consistent model selection criteria. We are particularly interested in the group LASSO and adaptive LASSO, which lead to the oracle properties, and we want to apply this approach to other families, such as MCLUST. We shall also study the penalization of the variance parameters as it will generate greater parsimony.

**Acknowledgments** This work was presented at the Statistical Society of Canada annual meeting in Guelph, Canada (June 2012) and at the MBC2 meeting in Catania, Italy (September 2012). The authors wish to thank the numerous people who provided helpful comments and feedback at these meetings. The authors are grateful to two anonymous reviewers for their very helpful comments on this work.

**Appendix A: Derivation of the LPBIC**

To derive the LPBIC, we closely follow the derivation of the usual BIC. We aim to maximize (3). Using (4), the second term becomes

$$n\lambda_n \sum_{g=1}^{G} \int \pi_g \sum_{j=1}^{p_g} |\mu_{gj}| d\pi_g = n \frac{\lambda_n}{G} \sum_{g=1}^{G} \sum_{j=1}^{p_g} \left[ |\hat{\mu}_{gj}| + \frac{1}{2} \frac{\text{sign}(\hat{\mu}_{gj})}{\hat{\mu}_{gj}} (\mu^2_{gj} - \hat{\mu}^2_{gj}) \right],$$

where $p_g$ is the number of non-zero mean components in class $g$. Under the assumption made in Sect. 2, $\mu_g$ is at most $p_g$ dependent, and the weak law of large numbers holds. In a large-$p$ setting, $\sum_{g=1}^{G} p_g$ is a large number and so $\sum_{g=1}^{G} \sum_{j=1}^{p_g} (\mu^2_{gj} - \hat{\mu}^2_{gj}) / \sum_{g=1}^{G} p_g \xrightarrow{P} \sum_{g=1}^{G} \sum_{j=1}^{p_g} (I(\hat{\mu}_g)^{-1})_{jj} / \sum_{g=1}^{G} p_g$. Thus, the second term becomes

$$n\lambda_n \sum_{k=1}^{G} \sum_{j=1}^{p_k} \left( |\hat{\mu}_{kj}| + \frac{(I(\hat{\mu}_g)^{-1})_{jj}}{|\hat{\mu}_{kj}|} \right).$$
The first term, using Taylor’s expansion, is

\[ \int \exp \left[ \log \mathcal{L}(\theta \mid x) \mathcal{G}(\theta) \right] d\theta = \int \exp \left[ \log \mathcal{L}(\hat{\theta} \mid x) \mathcal{G}(\hat{\theta}) + (\theta - \hat{\theta}) \frac{\partial \log \mathcal{L}(\theta \mid x) \mathcal{G}(\theta)}{\partial \theta} - \frac{1}{2} (\theta - \hat{\theta})^T \mathcal{H}(\theta - \hat{\theta}) \right] d\theta, \]

where \( \mathcal{H} \) is the second derivative matrix of \( \log \mathcal{L}(\theta \mid x) \mathcal{G}(\theta) \). Because \( \hat{\theta} \) is derived maximizing the penalized likelihood, the second term within the integral becomes \( (\theta - \hat{\theta}) \partial \mathcal{G}(\mu) / \partial \theta \), where \( \mathcal{G}(\mu) \) is the LASSO penalty function. Using (4), the mean-value theorem, and the fact that the \( \hat{\theta} \) values are close to \( \bar{\theta} \), the second term within the integral is \( n \lambda_n / G \sum_{g=1}^{G} \sum_{j=1}^{p_g} \text{sign}(\mu_{gj}) \).

The third term within the integral similarly becomes \( 1/2 (\bar{\theta} - \hat{\theta})^T \mathcal{H}_{\hat{\theta}}(\bar{\theta} - \hat{\theta}) \), where \( \bar{\theta} \) is the set of non-zero parameters and \( \hat{\theta} \) is their estimate. Using Laplace approximation on \( \mathcal{H} \) and applying the weak law of large numbers, as in the usual BIC, we arrive at \( \log \mathcal{L}(\hat{\theta} \mid x) - 1/2 \tilde{\rho} \log n \), where \( \tilde{\rho} = \dim(\hat{\theta}) \). This, combined with the second term of (3), gives (7).

**Appendix B: Proof of the asymptotic property of the LPBIC**

First, suppose the true number of clusters \( G \) is known with the corresponding parameter \( \bar{\theta} \). Let the true parameter be \( \bar{\theta}_0 \). Let \( \hat{\theta} \) be an arbitrary estimate of \( \bar{\theta} \). Let \( \tilde{\rho}_0 \) and \( \tilde{\rho}_1 \) be the corresponding numbers of non-zero parameters, and \( \lambda_n^0 \) and \( \lambda_n^1 \) be the corresponding tuning parameters. We first prove that, for an arbitrary estimate \( \hat{\theta} \) satisfying \( || \hat{\theta} - \bar{\theta}_0 || = \mathcal{O}(n^\kappa) \), \( \text{LPBIC}(\hat{\theta}_1, \hat{\theta}_2) - \text{LPBIC}(\hat{\theta}_1, 0) \leq 0 \) as \( n \to \infty \). We note that

\[ \text{LPBIC}(\hat{\theta}_1, \hat{\theta}_2) - \text{LPBIC}(\hat{\theta}_1, 0) = 2l(\hat{\theta}_1, \hat{\theta}_2 \mid x) - 2l(\hat{\theta}_1, 0 \mid x) - \left[ \Lambda(\hat{\theta}_1, \hat{\theta}_2) - \Lambda(\hat{\theta}_1, 0) \right], \]

where \( l = \log \mathcal{L} \) and \( \Lambda \) is the penalty part of the LPBIC. Using the mean-value theorem,

\[ l(\hat{\theta}_1, \hat{\theta}_2 \mid x) - l(\hat{\theta}_1, 0 \mid x) = \left[ \frac{\partial l(\hat{\theta}_1, \xi)}{\partial \hat{\theta}_2} \right] \hat{\theta}_2, \]

where \( || \xi || \leq || \hat{\theta}_2 || = \mathcal{O}(n^\kappa) \). Also,

\[ \left| \left| \frac{\partial l(\hat{\theta}_1, \xi)}{\partial \hat{\theta}_2} - \frac{\partial l(\hat{\theta}_0, 0)}{\partial \hat{\theta}_2} \right| \right| \leq \left| \left| \frac{\partial l(\hat{\theta}_1, \xi)}{\partial \hat{\theta}_2} - \frac{\partial l(\hat{\theta}_1, 0)}{\partial \hat{\theta}_2} \right| \right| + \left| \left| \frac{\partial l(\hat{\theta}_1, 0)}{\partial \hat{\theta}_2} - \frac{\partial l(\hat{\theta}_0, 0)}{\partial \hat{\theta}_2} \right| \right| \leq \sum_{i=1}^{n} M_2(z_i) \left[ ||\xi|| + ||\hat{\theta}_1 - \bar{\theta}_0|| \right] = \left[ ||\xi|| + ||\hat{\theta}_1 - \bar{\theta}_0|| \right] \mathcal{O}(n) = \mathcal{O}(n^{\kappa+1}) \]
from Assumption 2. Also, from the last part of the first line of (8), which is of order \( \mathcal{O}(n^{\kappa+1}) \), we can conclude that \( \partial l(\hat{\theta}_0, \theta)/\partial \theta_2 \) is of order \( \mathcal{O}(n^{\kappa+1}) \), as is \( \partial l(\hat{\theta}_1, \hat{\theta}_2)/\partial \theta_2 \). Therefore, from these order assessments, we conclude that \( l(\hat{\theta}_1, \hat{\theta}_2) - l(\hat{\theta}_1, 0) = \mathcal{O}(n^{\kappa+1}) \sum_{g=1}^G \sum_{j=p_g+1}^p \hat{\mu}_{gj} \), where \( p_g \) is defined as in (4).

For the \( \Lambda(\hat{\theta}_1, \hat{\theta}_2) - \Lambda(\hat{\theta}_1, 0) \) part, note that

\[
\frac{2n\lambda_n}{G} \sum_{g=1}^G \sum_{j=1}^{p_g} \left[ |\hat{\mu}_{gj}| + \frac{(I(\hat{\mu}_{gj})^{-1})_{jj}}{|\hat{\mu}_{gj}|} - \text{sign}(\hat{\mu}_{gj}) \right] = \mathcal{O}(n^{\alpha+1}) \lambda_n
\]

because the summation part is some constant times \( p = \mathcal{O}(n^\alpha) \), using Assumption 1. We also have \( (\hat{\rho}_1 - \bar{\rho}_0) \log n = \sum_{g=1}^G (p - p_g) \log n = \mathcal{O}(n^\alpha) \log n \). Hence,

\[
\text{LPBIC}(\hat{\theta}_1, \hat{\theta}_2) - \text{LPBIC}(\hat{\theta}_1, 0) = \mathcal{O}(n^{\kappa+1}) \sum_{g=1}^G \sum_{j=p_g+1}^p \hat{\mu}_{gj} - \mathcal{O}(n^\alpha) \log n - (\lambda_n^{(1)} - \lambda_n^{(0)}) \mathcal{O}(n^{\alpha+1}).
\]

The first term of the above expression is \( \mathcal{O}(n^{\kappa+1}) \sum_{g=1}^G \sum_{j=p_g+1}^p \hat{\mu}_{gj} = \mathcal{O}(n^{2\kappa+1}) \). Using Assumption 1, i.e., that \( \lambda_n = o(\log n/n) \), and by order comparison, we can conclude that the leading terms in the above expression are \( \mathcal{O}(n^{2\kappa+1}) \) and \( \mathcal{O}(n^\alpha) \log n \). Because \( \alpha > 2\kappa + 1 \). LPBIC(\( \hat{\theta}_1, \hat{\theta}_2 \)) - LPBIC(\( \hat{\theta}_1, 0 \)) \leq 0 as \( n \to \infty \).

Now, let \( \hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2) \) be an estimate of \( \theta \) such that \( \hat{\theta}_1, 0 \) is a maximizer of LPBIC(\( \hat{\theta}_1, 0 \)) satisfying \( || \hat{\theta} - \theta_0 || = \mathcal{O}(n^\kappa) \). It suffices to show that in the neighbourhood \( || \hat{\theta} - \theta_0 || = \mathcal{O}(n^\kappa) \), LPBIC(\( \hat{\theta}_1, \hat{\theta}_2 \)) - LPBIC(\( \hat{\theta}_1, 0 \)) < 0 with probability tending to 1 as \( n \to \infty \). We note that

\[
\text{LPBIC}(\hat{\theta}_1, \hat{\theta}_2) - \text{LPBIC}(\hat{\theta}_1, 0) = [\text{LPBIC}(\hat{\theta}_1, \hat{\theta}_2) - \text{LPBIC}(\hat{\theta}_1, 0)] + [\text{LPBIC}(\hat{\theta}_1, 0) - \text{LPBIC}(\hat{\theta}_1, 0)],
\]

where \( \text{LPBIC}(\hat{\theta}_1, \hat{\theta}_2) - \text{LPBIC}(\hat{\theta}_1, 0) \leq 0 \) with probability tending to 1 (by the previous result) and \( \text{LPBIC}(\hat{\theta}_1, 0) - \text{LPBIC}(\hat{\theta}_1, 0) \leq 0 \) with probability tending to 1 because \( (\hat{\theta}_1, 0) \) is a maximizer of LPBIC(\( \hat{\theta}_1, 0 \)). Thus, \( (\hat{\theta}_1, 0) \) maximizes LPBIC(\( \hat{\theta}_1, \hat{\theta}_2 \)) with probability tending to 1 as \( n \to \infty \). Hence, we conclude that \( \text{P}(\hat{\theta}_2 = 0) \to 1 \) as \( n \to \infty \), which completes the proof.

The case of unknown clusters can be similarly proved. If the estimated number of components is \( G_1 \) and the true number is \( G \), then the estimated parameter corresponding to \( G_1 \) is, say, \( \hat{\theta} \). We can again decompose \( \hat{\theta} \) as \( (\hat{\theta}_1, \hat{\theta}_2) \) and similarly show that \( \hat{\theta}_2 \to 0 \) in probability. Here, \( \hat{\theta}_1 \) is comprised of the clusters belonging to \( \hat{\theta}_0 \).

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6): 716–723
Bozdogan H (1987) Model selection and Akaike’s information criterion (AIC): the general theory and its analytical extensions. Psychometrika 52:345–370
Bozdogan H (2000) Akaikeas information criterion and recent developments in informational complexity. Invited paper in the Special Issue of Journal of Mathematical Psychology on Methods for Model Selection Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recognit 28:781–793
Chen J, Chen Z (2008) Extended Bayesian information criteria for model selection with large model spaces. Biometrika 95(3):759–771
Clyde M, Berger J, Bullard F, Ford E, Jeffrey W, Luo R, Paulo R, Lorenzo T (2007) Current challenges in Bayesian model choice. Statistical Challeng Modern Astron IV 371:224–240
Craven P, Wahaba G (1979) Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer Math 31:377–403
De Leeuw J (1994) Information systems and data analysis. In: Bock HH, Lenski W, Richter MM (eds) Block relaxation algorithms in statistics. Springer, Berlin, pp 308–325
Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Ser B 39(1):1–38
Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360
Fraley C, Raftery A (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97:611–631
Fraley C, Raftery AE (2007) Bayesian regularization for normal mixture estimation and model-based clustering. J Classif 24:155–181
Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441
Ghahramani Z, Hinton GE (1997) The EM algorithm for factor analyzers. Tech. Rep. CRG-TR-96-1, University of Toronto, Toronto
Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537
Heiser WJ (1995) Recent advances in descriptive multivariate analysis. In: Krzanowski WJ (ed) Convergent computation by iterative majorization: theory and applications in multidimensional data analysis. Oxford University Press, Oxford, pp 157–189
Hubert L, Arabie P (1985) Comparing partitions. J Classif 2:193–218
Hunter DL, Lange K (2004) A tutorial on MM algorithms. Am Stat 58(1):30–37
Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90(430):773–795
Kass RE, Wasserman L (1995) A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. J Am Stat Assoc 90(431):928–934
Keirbin C (2000) Consistent estimation of the order of mixture models. Sankhya Indian J Stat Ser A 62(1):49–66
Khalili A, Chen J (2007) Variable selection in finite mixture of regression models. J Am Stat Assoc 102(479):1025–1038
Lange K, Hunter DR, Yang I (2000) Optimization transfer using surrogate objective functions. J Comput Graph Stat 9:1–20
McLachlan G, Basford K (1988) Mixture models: inference and applications to clustering. Marcel Dekker Inc, New York
McLachlan G, Peel D (2002) Finite mixture model. Wiley, New York
McLachlan GJ, Peel D (2000) Mixtures of factor analyzers. In: Proceedings of the seventh international conference on machine learning. Morgan Kaufmann, San Francisco, pp 599–606
McLachlan GJ, Bean R, Peel D (2002) A mixture model-based approach to the clustering of microarray expression data. Bioinformatics 18:412–422
McNicholas PD, Murphy TB (2008) Parsimonious Gaussian mixture models. Stat Comput 18:285–296
McNicholas PD, Murphy TB (2010) Model-based clustering of microarray expression data via latent Gaussian mixture models. Bioinformatics 26(21):2705–2712
McNicholas PD, Murphy TB, McDaid AF, Frost D (2010) Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. Comput Stat Data Anal 54(3):711–723
Meng XL, van Dyk D (1997) The EM algorithm—an old folk song sung to a fast new tune (with discussion). J R Stat Soc Ser B 59:511–567
Pan W, Shen X (2007) Penalized model-based clustering with application to variable selection. J Mach Learn Res 8:1145–1164

Rand WM (1971) Objective criteria for the evaluation of clustering methods. J Am Stat Assoc 66:846–850

Schwartz G (1978) Estimating the dimensions of a model. Ann Stat 6:461–464

Shibata R (1986) Consistency of model selection and parameter estimation. J Appl Probab 23:127–141

Stone M (1974) Cross-validatory choice and assessment of statistical predictions (with discussion). J R Stat Soc Ser B 36:111–147

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288

Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K (2005) Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B 67:91–108

Tipping TE, Bishop CM (1999) Mixtures of probabilistic principal component analysers. Neural Comput 11(2):443–482

Titterington DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, Chichester

Van Deun K, Wilderjans TF, Van Den Berg RA, Antoniadis A, Van Mechelen I (2011) A flexible framework for sparse simultaneous component based data integration. BMC Bioinforma 12:1971–1987

Wang H, Runze L, Tsai CL (2007) Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika 94(3):553–568

Wang H, Bo L, Ling C (2009) Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc Ser B 71(3):671–683

Zhu L, Carlin BP (2000) Comparing hierarchical models for spatio-temporally misaligned data using the deviance information criterion. Stat Med 19(17–18):2265–2278

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429