Statistica Sinica **27** (2017), 147-169
doi:http://dx.doi.org/10.5705/ss.2014.105

MODEL SELECTION FOR GAUSSIAN MIXTURE MODELS

Tao Huang, Heng Peng and Kun Zhang

Shanghai University of Finance and Economics,
Hong Kong Baptist University and
Carnegie Mellon University & Max Planck Institute for Intelligent Systems

Abstract: This paper is concerned with an important issue in finite mixture modeling, the selection of the number of mixing components. A new penalized likelihood method is proposed for finite multivariate Gaussian mixture models, and it is shown to be consistent in determining the number of components. A modified EM algorithm is developed to simultaneously select the number of components and estimate the mixing probabilities and the unknown parameters of Gaussian distributions. Simulations and a data analysis are presented to illustrate the performance of the proposed method.

Key words and phrases: EM algorithm, Gaussian mixture models, model selection, penalized likelihood.

1. Introduction

The finite mixture model is a flexible and powerful way to model data that stem from multiple populations and is heterogeneous, such as data from pattern recognition, computer vision, image analysis, and machine learning. Gaussian mixture model is an important mixture model family, and it is well known that any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities (Lindsay (1995); McLachlan and Peel (2000)). However, as demonstrated by Chen (1995), when the number of components is unknown, the optimal convergence rate of the estimate of a finite mixture model is slower than the optimal convergence rate when it is known. Recently, Nguyen (2013) and Ho and Nguyen (2015) suggested the use of Wasserstein distance to systematically investigate the identifiability problem and the optimal rates of estimating convergence for the parameters of multiple types in finite mixtures without constraint conditions. In particular, they pointed out that the finite multivariate Gaussian mixture model is not second-order identifiable, and its optimal estimating convergence rate is unusually slow when the model is over-fitted. In practice, with too many components, the mixture model may over-fit the data and yield poor interpretations, while with too few components, the mixture model may not be flexible enough to approximate the underlying data structure. Thus the selection of the number of components is not only of theoretical interest, but of value in practical applications.

Most conventional methods for determining the order of a finite mixture model are based on the likelihood function and some information criteria, such as AIC and BIC. Leroux (1992) investigated the properties of AIC and BIC and showed that these criteria do not underestimate the true number of components. Roeder and Wasserman (1997) showed the consistency of BIC when a normal mixture model is used to estimate a density function nonparametrically. Using the locally conic parameterization method developed by Dacunha-Castelle and Gassiat (1997), Keribin (2000) investigated the consistency of the maximum penalized likelihood estimator for an appropriate penalization sequence. Another class of methods is based on the distance measured between the fitted model and the nonparametric estimate of the population distribution, such as the penalized minimum-distance method (Chen and Kalbfleisch (1996)), the Kullback-Leibler distance method (James, Priebe and Marchette (2001)) and the Hellinger distance method (Woo and Sriram (2006)). To avoid the irregularity of the likelihood function for the finite mixture model when the number of components is unknown, Ray and Lindsay (2008) suggested the use of a quadratic risk-based approach to select the number of components. These methods are all based on the complete model search algorithm and, as a consequence, the computation burden is heavy. To improve computational efficiency, Chen and Khalili (2008) proposed a penalized likelihood method with the SCAD penalty (Fan and Li (2001)) for mixtures of univariate location distributions. They applied SCAD to penalize the differences of location parameters, which can then merge some subpopulations by shrinking such differences to zero. However, similar to most conventional order selection methods, their penalized likelihood method can be only used for one-dimensional location mixture models. Moreover, it is not always reasonable to merge Gaussian components with the same mean but different variance. Bunea et al. (2010) studied sparse density estimation via ℓ_1 penalization (SPADES). They assumed that the densities of true mixture components come from a large known candidate density pool, and then selected the mixture components by penalizing the mixing weights. According to their conditions on the densities of true mixture components, SPADES can be effective only if the local distances of true mixture components are quite large in comparison to their variances or covariance matrices.

Bayesian approaches have also been used to find a suitable number of components of the finite mixture model. For instance, Corduneanu and Bishop (2001) and Bishop (2006) applied the variational inference method to determine the number of components. Moreover, with suitable priors on the parameters, the maximum a posteriori (MAP) estimator can be used for model selection. In particular, Ormoneit and Tresp (1998) and Zivkovic and van der Heijden (2004) put the Dirichlet prior on the mixing probabilities, and Brand (1999) applied the entropic prior on the same parameters to favor models with small entropy. The MAP estimator then drives the mixing probabilities associated with unnecessary components toward extinction. Based on an improper Dirichlet prior, Figueiredo and Jain (2002) suggested using the minimum message length criterion to determine the number of components, and further proposed an efficient algorithm for learning a finite mixture from multivariate data. Although these Bayesian approaches preform well in practice, in general their theoretical justifications are still missing. The main challenge is that the objective function does not change continuously and thus encounter a sudden drop when a component is eliminated, as *zero* is not in the support region of the prior distribution for the mixing probabilities, such as the Dirichlet prior.

We propose a new penalized likelihood method for estimating finite Gaussian mixture models. Intuitively, if some of the mixing probabilities are shrunk to zero, the corresponding components are eliminated and a suitable number of components is retained. By doing this, we can deal with multivariate Gaussian mixture models, and do not need to assume different mean vectors or the same covariance matrix for different components. We propose to penalize the logarithm of mixing probabilities. Our proposed penalized likelihood method is significantly different from various Bayesian methods in the objective function and theoretical properties. When a component is eliminated, i.e., the mixing weight of that component is shrunk to zero, the objective function of our proposed method changes continuously. This enables us to fully investigate the statistical properties of the proposed method, and especially the consistency of the model selection.

The rest of the paper is organized as follows. In Section 2, we propose a new penalized likelihood method for finite multivariate Gaussian mixture models, and describe a modified EM algorithm to simultaneously select the number of components and estimate the unknown parameters. In Section 3, we derive asymptotic properties of the estimated number of components. In Section 4, simulation studies are presented to illustrate the performance of the proposed method. Some discussions are given in Section 5. Proofs are relegated to the Appendix.

2. Gaussian Mixture Model Selection

2.1. Penalized likelihood method

The density of a d-dimensional random variable $\mathbf{x}$ can be approximated by a weighted sum of some Gaussian densities

$$f(\mathbf{x}) = \sum_{m=1}^{M} \pi_m \phi(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \tag{2.1}$$

where $\phi(\mathbf{x};\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)$ is a Gaussian density with mean vector $\boldsymbol{\mu}_m$ and covariance matrix $\boldsymbol{\Sigma}_m$, and π_m, $m=1,\ldots,M$, are the positive mixing probabilities that satisfy $\sum_{m=1}^{M}\pi_m=1$. For identifiability of the Gaussian mixture model (GMM), let M be the smallest integer such that $\pi_m>0$ for $1\le m\le M$, and $(\boldsymbol{\mu}_a,\boldsymbol{\Sigma}_a)\ne(\boldsymbol{\mu}_b,\boldsymbol{\Sigma}_b)$ for $1\le a\ne b\le M$. Given the number of components M, the complete set of parameters of the GMM, $\boldsymbol{\theta}=\{\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\mu}_M,\boldsymbol{\Sigma}_M,\pi_1,\ldots,\pi_M\}$, can be conveniently estimated by maximum likelihood via the EM algorithm. To avoid overfitting and underfitting, an important issue is to determine the number of components M.

Intuitively, if some of the mixing probabilities are shrunk to zero, the corresponding components are eliminated and a number of components is smaller retained. Denote by y_{im} the indicator variable if the ith observation arises from the mth component; the conditional expected complete data log-likelihood function (McLachlan and Peel (2000)) is

$$\begin{aligned}\ell(\boldsymbol{\theta})&=\log\prod_{i=1}^{n}f(\mathbf{x}_i;\boldsymbol{\theta})\\&=\mathrm{E}\left\{\sum_{i=1}^{n}\sum_{m=1}^{M}y_{im}\left[\log\pi_m+\log\phi(\mathbf{x}_i;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)\right]\,\middle|\,\mathbf{x}_i,i=1,\ldots,n\right\}\\&=\sum_{i=1}^{n}\sum_{m=1}^{M}h_{im}\log\pi_m+\sum_{i=1}^{n}\sum_{m=1}^{M}h_{im}\log\phi(\mathbf{x}_i;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m),\end{aligned}\tag{2.2}$$

where h_{im} is the posterior probability that the ith observation belongs to the mth component. This expression contains $\log\pi_m$, whose gradient grows very fast when π_m is close to zero. The L_p types of penalties may not be able to set insignificant π_m to zero.

We give a simple illustration of how the likelihood function changes when a mixing probability approaches zero. In particular, a data set of 1,000 points was randomly generated from a single bivariate Gaussian distribution. A GMM with two components, $f(\mathbf{x})=\pi_1\phi(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)+(1-\pi_1)\phi(\mathbf{x};\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, was used to fit the data. The learned two Gaussian components are depicted in Figure 1(a), and $\hat{\pi}_1$ is 0.227. For each fixed π_1, we optimized all other parameters $\{(\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i),i=1,2\}$ by maximizing the likelihood function. Figure 1(b) shows that the minimized negative log-likelihood function changes almost linearly with $\log(\pi_1)$ when π_1 is close to zero, albeit with some small upticks. Thus the derivative of the log-likelihood function with respect to π_1 is approximately proportional to $1/\pi_1$ when π_1 is close to zero, and it would dominate the derivative of $\frac{\partial}{\partial\pi}\max_{\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\Sigma_1,\Sigma_2}\ell(\boldsymbol{\theta})$.

Thus (2.2) suggests that we need to consider penalizing $\log\pi_m$ to achieve the sparsity of $\boldsymbol{\pi}=\{\pi_1,\ldots,\pi_M\}$. We choose to penalize $\log((\epsilon+\pi)/\epsilon)=\log(\epsilon+$

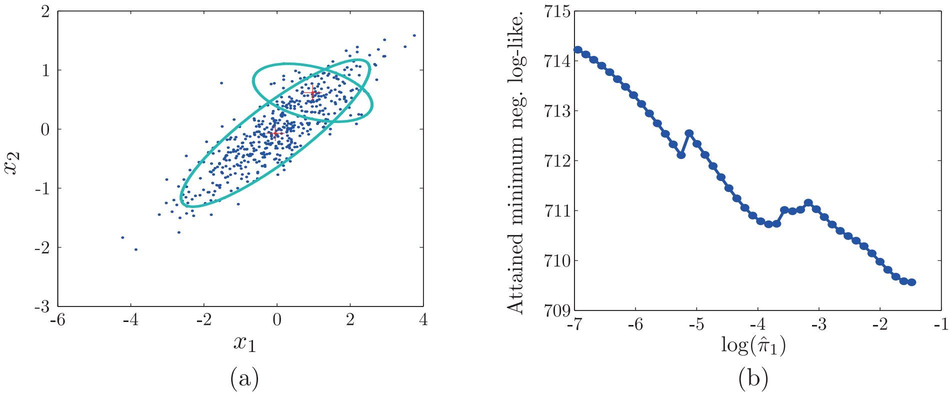


Figure 1. An illustration on the behavior of the negative log-likelihood function when a mixing probability is close to zero. (a) A simulated data set and a learned two-component GMM model. (b) The minimized negative log-likelihood as a function of $\log \pi_1$.

$\pi) - \log(\epsilon)$, where ϵ is a very small positive number, say 10^{-6} or $o(n^{-1/2}\log^{-1} n)$, as suggested in the proof of Theorem 1 and 2. Here $\log(\epsilon + \pi) - \log(\epsilon)$ is a monotonically increasing function of π, and it is shrunk to zero when the mixing probability π goes to zero. We propose the penalized log-likelihood function

$$\ell_P(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - n\lambda D_f \sum_{m=1}^{M} \left[\log(\epsilon + \pi_m) - \log(\epsilon)\right], \tag{2.3}$$

where $\ell(\boldsymbol{\theta})$ is the log-likelihood function, λ is a tuning parameter, and D_f is the number of free parameters for each component. For a GMM with arbitrary covariance matrices, each component has $D_f = 1+d+d(d+1)/2 = d^2/2+3d/2+1$ number of free parameters. Although D_f is a constant and can be removed from (2.3), it simplifies the search range of λ in numerical study and hence is kept.

The objective function of our proposed penalized likelihood method is similar to that derived, with Dirichlet prior, from the Bayesian point of view. In the mathematical sense, such Bayesian methods cannot shrink the mixing probabilities to zero exactly since the objective function is not continuous when some of mixing probabilities shrink to zero. As discussed by Fan and Li (2001), such discontinuity poses challenges to investigate the statistical properties of related penalization or Bayesian methods.

Fan and Li (2001) suggested that a good penalty function should yield an estimator with three properties: unbiasedness, sparsity, and continuity. It is obvious that $\log((\epsilon + \pi_m)/\epsilon)$ would over penalize large π_m and yield a biased estimator. Hence, we also consider the penalized log-likelihood function

$$\ell_P(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - n\lambda D_f \sum_{m=1}^{M} \left[\log(\epsilon + p_\lambda(\pi_m)) - \log(\epsilon)\right]. \tag{2.4}$$

Here $p_\lambda(\pi)$ is the SCAD penalty function proposed by Fan and Li (2001) that is conveniently characterized through its derivative:

$$p'_\lambda(\pi) = I(\pi \leq \lambda) + \frac{(a\lambda - \pi)_+}{(a-1)\lambda} I(\pi > \lambda),$$

for some $a > 2$ and $\pi > 0$. For a relatively large π_m and $\pi_m > a\lambda$, $p_\lambda(\pi_m)$ is a constant, and the estimator of this π_m is expected to be unbiased.

2.2. Modified EM algorithm

We propose a modified EM algorithm to maximize (2.3) and (2.4) iteratively in two steps. First to maximize (2.3), by combining (2.2) and (2.3), the expected penalized log-likelihood function is

$$\sum_{i=1}^{n}\sum_{m=1}^{M} h_{im}\log\pi_m + \sum_{i=1}^{n}\sum_{m=1}^{M} h_{im}\log\phi(\mathbf{x}_i; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) - n\lambda D_f \sum_{m=1}^{M}\left[\log(\epsilon+\pi_m) - \log(\epsilon)\right]. \tag{2.5}$$

In the E step, we are given the current estimate, $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\mu}}_1^0, \hat{\boldsymbol{\Sigma}}_1^0, \ldots, \hat{\boldsymbol{\mu}}_M^0, \hat{\boldsymbol{\Sigma}}_M^0, \hat{\pi}_1^0, \ldots, \hat{\pi}_M^0)$, and calculate the posterior probability

$$h_{im} = \frac{\hat{\pi}_m^0 \phi(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_m^0, \hat{\boldsymbol{\Sigma}}_m^0)}{\sum_{m=1}^{M} \hat{\pi}_m^0 \phi(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_m^0, \hat{\boldsymbol{\Sigma}}_m^0)}.$$

In the M step, we update $\boldsymbol{\theta} = \{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M, \pi_1, \ldots, \pi_M\}$ by maximizing the expected penalized log-likelihood function (2.5). We can update $\{\pi_1, \ldots, \pi_M\}$ and $\{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M\}$ separately, as they are not intermingled in (2.5). To obtain an estimate for $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$, we aim to solve the set of equations

$$\frac{\partial}{\partial \pi_m}\left[\sum_{i=1}^{n}\sum_{m=1}^{M} h_{im}\log\pi_m - n\lambda D_f \sum_{m=1}^{M}\log(\epsilon+\pi_m) - \beta(\sum_{m=1}^{M}\pi_m - 1)\right] = 0. \tag{2.6}$$

Given that ϵ is close to zero, ideally such that $1/\pi_m \approx 1/(\pi_m+\epsilon)$ for any π_m, by making use of $\sum_{i=1}^{n}\sum_{m=1}^{M} h_{im} = n$ and straightforward calculations, we obtain $\beta = n(1 - M\lambda D_f)$. Since $\pi_m, m = 1, \ldots, M$, are always nonnegative, we have

$$\hat{\pi}_m^1 = \max\left\{0, \frac{1}{1 - M\lambda D_f}\Big[\frac{1}{n}\sum_{i=1}^{n} h_{im} - \lambda D_f\Big]\right\}. \tag{2.7}$$

Some $\hat{\pi}_m^1$ may be shrunk to zero and, subsequently, the constraint $\sum_{m=1}^{M}\hat{\pi}_m^1 = 1$ may not be satisfied. However, this neither decreases the likelihood function nor affects the estimate of the posterior probability h_{im} in the E-step or the update of π_m in the M-step. We need only to normalize $\hat{\pi}$ by enforcing $\sum_{m=1}^{M} \hat{\pi}_m = 1$ after the EM algorithm converges.

The update equations on $\{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_M\}$ are the same as those of the standard EM algorithm for GMM (McLachlan and Peel (2000)). Specifically, we update $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ as

$$\hat{\boldsymbol{\mu}}_m^1 = \frac{\sum_{i=1}^{n} h_{im}\mathbf{x}_i}{\sum_{i=1}^{n} h_{im}}, \qquad \hat{\boldsymbol{\Sigma}}_m^1 = \frac{\sum_{i=1}^{n} h_{im}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_m^1)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_m^1)^T}{\sum_{i=1}^{n} h_{im}}.$$

In summary the proposed modified EM algorithm works as follows. It starts with a pre-specified large number of components, and whenever a mixing probability is shrunk to zero by (2.7), the corresponding component is deleted, and thus fewer components are retained for the remaining EM iterations. Here we abuse the notation M for the number of components at beginning of each EM iteration, and through the updating process, M becomes smaller. For a given EM iteration step, it is possible that zero, one, or more than one component are deleted.

The modified EM algorithm for maximizing (2.4) is similar to the one for (2.3), the only difference is in the M step for maximizing $\boldsymbol{\pi}$. Given the current estimate $(\hat{\pi}_1^0, \ldots, \hat{\pi}_M^0)$ for $\boldsymbol{\pi}$, to solve

$$\frac{\partial}{\partial \pi_m}\left[\sum_{i=1}^{n}\sum_{m=1}^{M} h_{im}\log\pi_m - n\lambda D_f \sum_{m=1}^{M}\log(\epsilon + p_\lambda(\pi_m)) - \beta(\sum_{m=1}^{M}\pi_m - 1)\right] = 0,$$

we substitute $\log(\epsilon + p_\lambda(\hat{\pi}_m))$ by its linear approximation $\log(\epsilon + p_\lambda(\hat{\pi}_m^0)) + [(p'_\lambda(\hat{\pi}_m^0))/(\epsilon + p_\lambda(\hat{\pi}_m^0))](\hat{\pi}_m - \hat{\pi}_m^0)$. By $\sum_{i=1}^{n}\sum_{m=1}^{M} h_{im} = n$, we first update the value of β by

$$\beta = n - n\lambda D_f \sum_{m=1}^{M} \frac{p'_\lambda(\hat{\pi}_m^0)\hat{\pi}_m^0}{\epsilon + p_\lambda(\hat{\pi}_m^0)}.$$

Then by straightforward calculations, π_m can be updated as

$$\hat{\pi}_m^1 = \frac{1}{T_m}\sum_{i=1}^{n} h_{mi}, \tag{2.8}$$

where

$$T_m = n - n\lambda D_f \sum_{m=1}^{M} \frac{p'_\lambda(\hat{\pi}_m^0)\hat{\pi}_m^0}{\epsilon + p_\lambda(\hat{\pi}_m^0)} + n\lambda D_f \frac{p'_\lambda(\hat{\pi}_m^0)}{\epsilon + p_\lambda(\hat{\pi}_m^0)}.$$

In the numerical study, (2.8) is seldom exactly zero. To avoid possible numerical instability, if an updated $\hat{\pi}_m^1$ is smaller than a pre-specified small threshold, we set it to zero and remove the corresponding component from the mixture model. Because of the consistency of the proposed penalized likelihood method, we can set this threshold value as small as possible, though a smaller threshold value increases the modified EM algorithm's iteration steps and computation time. In our numerical studies, we set this threshold to be 10^{-4}, smaller than the smallest mixing probabilities in the simulation examples.

2.3. Selection of tuning parameters

To obtain the final estimate of the mixture model by maximizing (2.3) or (2.4), one needs to select the tuning parameters λ (for both) and a (for maximizing (2.4)). Our simulation studies show that the numerical results are not sensitive to the selection of a. Following the suggestion of Fan and Li (2001), we set $a = 3.7$. For the standard LASSO (Tibshirani (1996)) and SCAD penalized regressions, there are many methods to select λ, such as generalized cross-validation (GCV) and BIC (Fan and Li (2001); Wang, Li and Tsai (2007)). Here we define a BIC value

$$\mathrm{BIC}(\lambda) = \sum_{i=1}^{n} \log\Big\{ \sum_{m=1}^{\hat{M}} \pi_m \phi(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\Sigma}}_m) \Big\} - \frac{1}{2}\hat{M} D_f \log n,$$

and take $\hat{\lambda} = \arg\max_{\lambda} \mathrm{BIC}(\lambda)$, where $\hat{M}$ is the estimate of the number of components and $\hat{\boldsymbol{\mu}}_m$ and $\hat{\boldsymbol{\Sigma}}_m$ are the estimates of $\boldsymbol{\mu}_m$ and $\boldsymbol{\Sigma}_m$ for a given λ.

3. Asymptotic Properties

It is possible to extend our method to more generalized mixture models, but we only show the consistency of model selection for Gaussian mixture models.

We need some conditions to derive asymptotic properties.

P1: $\|\mu_i\| \le C_1, \|\boldsymbol{\Sigma}_i\| \le C_2, i = 1, \ldots, M$, where C_1 and C_2 are large enough constants.

P2: $\min_{i,k}\{\alpha_k(\boldsymbol{\Sigma}_i), k = 1, \ldots, d, i = 1, \ldots, M\} \ge C_3$, where $\alpha_k(\boldsymbol{\Sigma}_i)$ is the kth eigenvalue of $\boldsymbol{\Sigma}_i$ and C_3 is a positive constant.

Remark 1. If the number of components is known, say K, the maximum likelihood function, or the penalized maximum likelihood function, has $K!$ equivalent solutions corresponding to the $K!$ ways of assigning K sets of parameters to K components. The identifiability problem is an important issue if we wish to interpret the estimated parameter values for a selected model, but our main focus is to determine the model order and find a good density estimate with the finite mixture model. The identifiability problem is irrelevant as all equivalent solutions yields the same estimates of the order and the density function.

Remark 2. Compared to the conditions in Dacunha-Castelle and Gassiat (1997, 1999), conditions (P1) and (P2) are slightly stronger. Without lose of generality, we assume that the parameters of the mixture model are in a bounded compact space. This is not only for mathematical conveniences, but also to ensure the identifiability and to avoid the ill-posedness problems of the finite mixture model, as discussed in Bishop (2006). These conditions are also practically reasonable for our modified EM algorithm, as discussed by Figueiredo and Jain (2002).

Conditions (P1) and (P2) extend that of Hathaway (1985),

$$\Omega_c = \{\Psi \in \Omega : \frac{\sigma_h^2}{\sigma_i^2} \geq C > 0, 1 \leq h \neq i \leq g\},$$

where Ω denotes the unconstrained parameter space, g is the number of components for the initial univariate Gaussian mixture model, and σ_i^2 and σ_h^2 are variances of the Gaussian components in the model. With this condition, singularities in the likelihood function are avoided, which occur when the mean of a component is set equal to any observed value and the variance goes to zero. Hathaway (1985) also proposed a constrained parameter space for the multivariate case,

$$\Omega_c = \{\Psi \in \Omega : \text{all eigenvalues of} \quad \Sigma_h \Sigma_i^{-1} \geq C > 0, 1 \leq h \neq i \leq g\}.$$

Our proposed conditions (P1) and (P2) are stronger than such a requirement, and hence undesirable properties of multivariate Gaussian mixture models can be avoided. In fact, the conditions (P1) and (P2) naturally hold when the variances of Gaussian mixture components are the same.

Theorem 1. *Under conditions* (P1) *and* (P2), *if* $\sqrt{n}\lambda \to \infty$, $\lambda \to 0$ *and* $\epsilon = o(1/\sqrt{n})$, *there exists a local maximizer* $(\theta, \boldsymbol{\beta})$ *of* ℓ_P, *given in Appendix* (A.3), *such that* $\theta = O_p(1/\sqrt{n})$. *For such a local maximizer, the estimated number of components* $\hat{q}_n \to q$ *with probability tending to one.*

Theorem 2. *Under conditions* (P1) *and* (P2), *if* $\sqrt{n}\lambda \to C$ *and* $\epsilon = o(1/\sqrt{n}\log n)$ *where* C *is a constant, there exists a local maximizer* $(\theta, \boldsymbol{\beta})$ *of* ℓ_P, *given in Appendix* (A.2), *such that* $\theta = O_p(1/\sqrt{n})$. *For such a local maximizer, the estimated number of components* $\hat{q}_n \to q$ *with probability tending to one.*

Our method is rather general as we do not impose conditions on the difference of mean vectors or assume a common covariance for different mixture components. In practice it is easier to select an appropriate tuning parameter for (A.3) than for (A.2) to guarantee the consistency of the final model selection and estimation. In particular, the proposed BIC method always selects a reasonable tuning parameter. Let Component$_\lambda$ denote the number of components selected by (A.3) using the tuning parameter λ, and λ_{BIC} be the tuning parameter λ selected by the proposed BIC method in Section 2.3.

Table 1. Parameter estimation with standard deviation ($M = 10$) for Example 1.

Component		Mixing Probability	Mean		Covariance (eigenvalue)	
	True	0.3333	-1	1	2	0.2
1	(2.3)	0.3342(0.0201)	-0.9911(0.0861)	1.0169(0.1375)	2.0034(0.3022)	0.1981(0.0264)
	(2.4)	0.3356(0.0187)	-1.0022(0.0875)	1.0007(0.1428)	2.0205(0.2769)	0.1973(0.0265)
	True	0.3333	1	1	2	0.2
2	(2.3)	0.3317(0.0196)	1.0151(0.0845)	0.9849(0.1318)	1.9794(0.2837)	0.1977(0.0303)
	(2.4)	0.3321(0.0193)	1.0108(0.0790)	0.9904(0.1253)	1.9825(0.2980)	0.1957(0.0292)
	True	0.3333	0	-1.4142	2	0.2
3	(2.3)	0.3341(0.0171)	0.0019(0.1324)	-1.4112(0.0405)	1.9722(0.2425)	0.1973(0.0258)
	(2.4)	0.3322(0.0159)	0.0014(0.1449)	-1.4103(0.0404)	1.9505(0.2424)	0.1978(0.0267)

Theorem 3. *Under conditions* (P1) *and* (P2), $\Pr(\text{Component}_{\lambda_{BIC}} = q) \to 1$.

The proofs of Theorems 1, 2 and 3 are given in the Supplementary Materials for the paper.

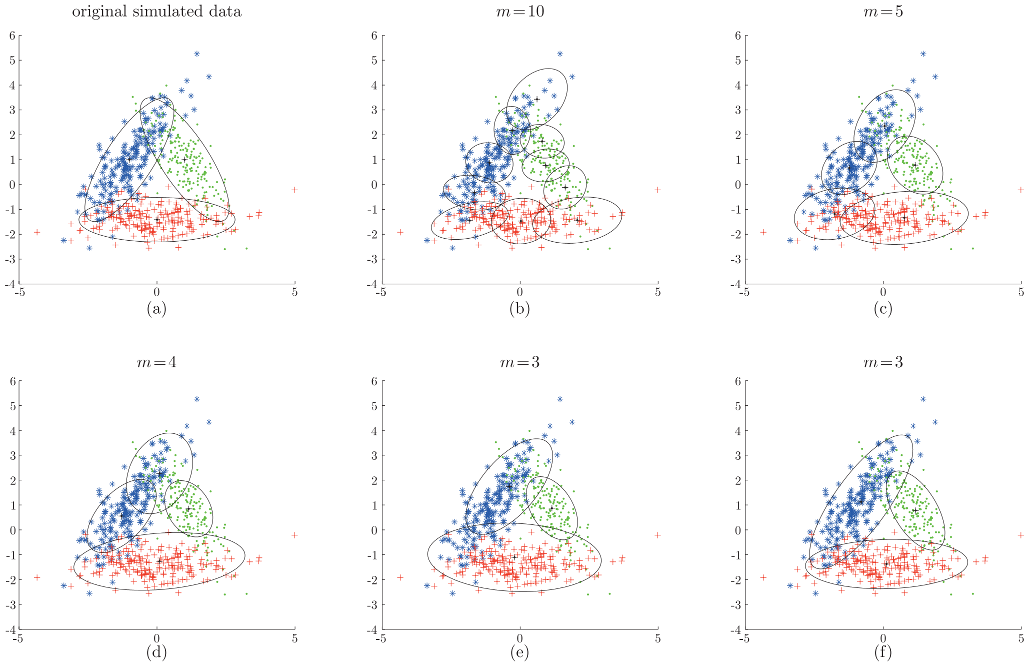
4. Numerical Studies

Example 1. We generated 600 observations from a three-component bivariate normal mixture with mixing probabilities $\pi_1 = \pi_2 = \pi_3 = 1/3$, mean vectors $\boldsymbol{\mu}_1 = [-1, 1]^T$, $\boldsymbol{\mu}_2 = [1, 1]^T$, $\boldsymbol{\mu}_3 = \left[0, -\sqrt{2}\right]^T$, and covariance matrices

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.65 & 0.7794 \\ 0.7794 & 1.55 \end{bmatrix}, \ \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.65 & -0.7794 \\ -0.7794 & 1.55 \end{bmatrix}, \ \boldsymbol{\Sigma}_3 = \begin{bmatrix} 2 & 0 \\ 0 & 0.2 \end{bmatrix}.$$

These three components are obtained by rotating and shifting a common Gaussian density $\mathcal{N}(0, \mathrm{diag}(2, 0.2))$.

We ran our proposed penalized likelihood methods (2.3) and (2.4) for 300 times. The initial maximum number of components M was set to be 10 or 50, the initial value for the modified EM algorithm was estimated by K-means clustering with $K = 10$ or 50, and the tuning parameter λ was selected by our proposed BIC method. Figure 2 shows the evolution of the modified EM algorithm for (2.4), with the maximum number of components as 10. We compare our proposed methods with the traditional AIC and BIC methods. Figure 3(a−c) shows the histograms of the estimated numbers of component. Our proposed methods do better in identifying the correct number of components than do the AIC and BIC methods. The proposed methods always correctly estimate the number of components regardless of the initial maximum number of components. Figure 3(d) depicts the evolution of the penalized log-likelihood function (2.3) for the simulated data set in Figure 2(a) in one run.



(a) (b) (c) (d) (e) (f)

Figure 2. One typical run of Example 1. (a) a simulated data set. (b) initialization for $M = 10$ components, (c-e) three intermediate estimates for $M = 6, 5, 4$, respectively, (f) the final estimate for $M = 3$.

Table 2. Parameter estimation with standard deviation ($M = 50$) for Example 1.

Component		Mixing Probability	Mean		Covariance (eigenvalue)	
1	True	0.3333	-1	1	2	0.2
	(2.3)	0.3342(0.0201)	-1.0080(0.0881)	0.9854(0.1372)	1.9603(0.2857)	0.1974(0.0296)
	(2.4)	0.3320(0.0190)	-1.0017 (0.0859)	0.9985(0.1389)	1.9604(0.2830)	0.1960 (0.0286)
2	True	0.3333	1	1	2	0.2
	(2.3)	0.3347(0.0170)	0.9879(0.0885)	1.0166(0.1385)	1.9531(0.2701)	0.1981(0.0283)
	(2.4)	0.3345 (0.0182)	0.9987(0.0896)	1.0044(0.1402)	1.9661(0.2460)	0.1971 (0.0248)
3	True	0.3333	0	-1.4142	2	0.2
	(2.3)	0.3329(0.0198)	0.0210(0.1329)	-1.4105(0.0344)	1.9717(0.2505)	0.1975(0.0265)
	(2.4)	0.3334(0.0164)	0.0117(0.1302)	-1.4116(0.0372)	1.9736(0.2769)	0.1998(0.0281)

When the number of components is correctly identified, we summarize the estimation of the unknown parameters of Gaussian distributions and the mixing probabilities in Tables 1 and 2 with different initial maximum number of components. For the covariance matrix, we used eigenvalues since the three components have the same shape as $\mathcal{N}(0, \text{diag}(2, 0.2))$. Tables 1 and 2 show that the modified EM algorithm gives accurate estimates for parameters and mixing probabilities. The final estimate of these parameters is robust to the initialization of the max-

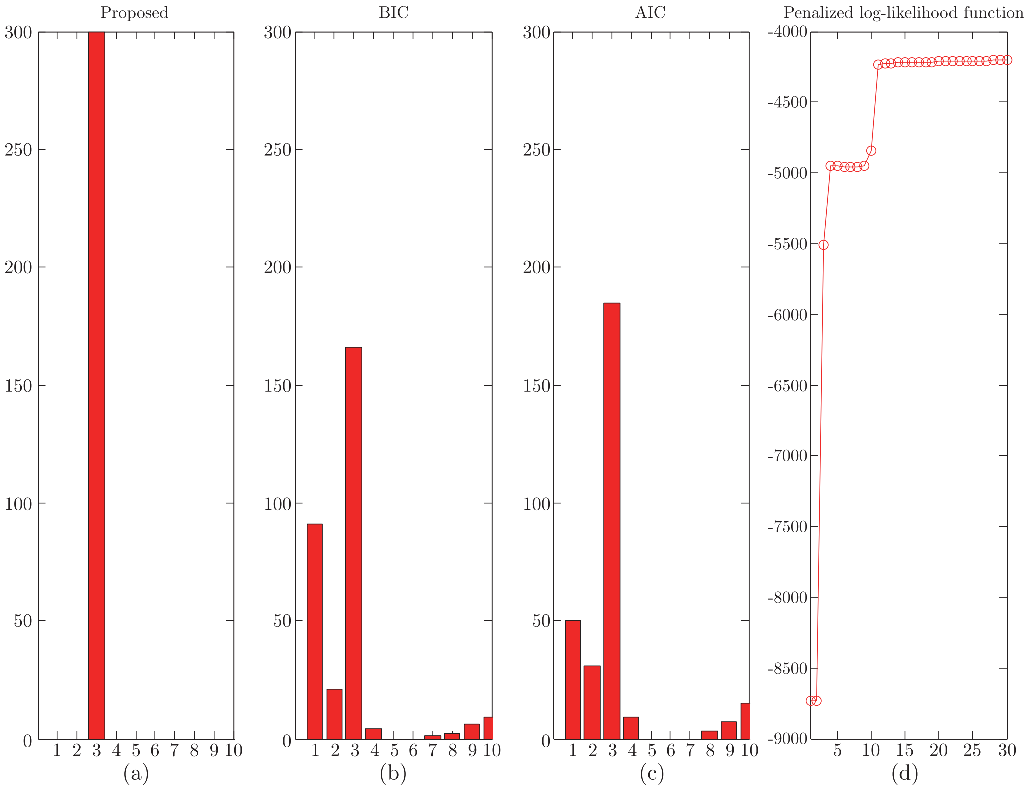


Figure 3. Histogram of estimated numbers of components for Example 1. (a) the proposed method (2.3), (b) BIC, (c) AIC. (d) The penalized log likelihood function for a typical run.

imum number of components.

Example 2. We considered a situation where the mixture components overlap and may have the same means but different covariance matrices. Neither of the proposed methods of Chen and Khalili (2008) and Bunea et al. (2010) is expected to work well here, as some components have the same mean. Specifically, we generated 1,000 samples with mixing probabilities $\pi_1 = \pi_2 = \pi_3 = 0.3$, $\pi_4 = 0.1$, mean vectors $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = [-2, -2]^T$, $\boldsymbol{\mu}_3 = [2, 0]^T$, $\boldsymbol{\mu}_4 = [1, -4]^T$, and covariance matrices

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.1 & 0 \\ 0 & 0.2 \end{bmatrix}, \; \boldsymbol{\Sigma}_2 = \begin{bmatrix} 2 & 2 \\ 2 & 7 \end{bmatrix}, \boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.5 & 0 \\ 0 & 4 \end{bmatrix}, \; \boldsymbol{\Sigma}_4 = \begin{bmatrix} 0.125 & 0 \\ 0 & 0.125 \end{bmatrix}.$$

We ran our proposed methods 300 times. The maximum number of components M was set to be 10 or 50, the initial value for the modified EM algorithm was estimated by K-means clustering, and the tuning parameter λ was selected

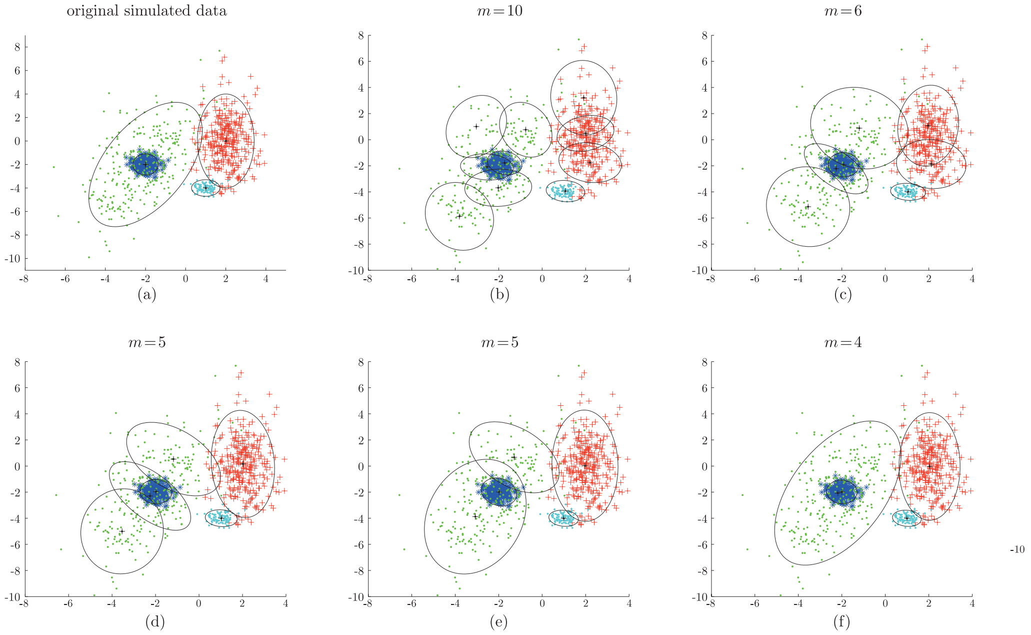


Figure 4. One typical run of Example 2. (a) a simulated data set. (b) initialization for $M = 10$ components, (c-e) three intermediate estimates for $M = 7, 6, 5$, respectively, (f) the final estimate for $M = 4$.

Table 3. Parameter estimation with standard deviation ($M = 10$) for Example 2.

Component		Mixing Probability	Mean		Covariance (eigenvalue)	
	True	0.3	-2	-2	0.1	0.2
1	(2.3)	0.3022(0.0093)	-2.0010(0.0216)	-1.9989(0.0291)	0.0979(0.0114)	0.2010 (0.0242)
	(2.4)	0.3009(0.0095)	-1.9995(0.0206)	-1.9975(0.0319)	0.0990(0.0119)	0.2003(0.0226)
	True	0.3	-2	-2	1.2984	7.7016
2	(2.3)	0.2995(0.0112)	-1.9989(0.1133)	-1.9963(0.1837)	1.2864(0.1407)	7.7219(0.7301)
	(2.4)	0.3017(0.0118)	-1.9995(0.1202)	-2.0049(0.1811)	1.2926(0.1343)	7.5856(0.7301)
	True	0.3	2	0	0.5	4
3	(2.3)	0.3019(0.0083)	1.9943(0.0483)	0.0001(0.1294)	0.4986(0.0529)	3.9951(0.3496)
	(2.4)	0.3012(0.0087)	1.9995(0.0511)	-0.0001(0.1244)	0.4963(0.0544)	3.9998(0.3911)
	True	0.1	1	-4	0.125	0.125
4	(2.3)	0.0964(0.0038)	1.0005(0.0373)	-3.9966(0.0394)	0.1143(0.0245)	0.1339(0.0252)
	(2.4)	0.0962(0.0047)	0.9993(0.0394)	-4.0013(0.0417)	0.1167(0.0259)	0.1317(0.0278)

by our proposed BIC method. Figure 4 shows the evolution of the modified EM algorithm for (2.3) with the initial maximum number of components as 10 for one simulated data set. Figure 5 shows that our method always identifies the number of components correctly and performs much better than AIC and BIC methods. Tables 3 and 4 show that the modified EM algorithm gives accurate estimates for both parameters and mixing probabilities. The final estimates

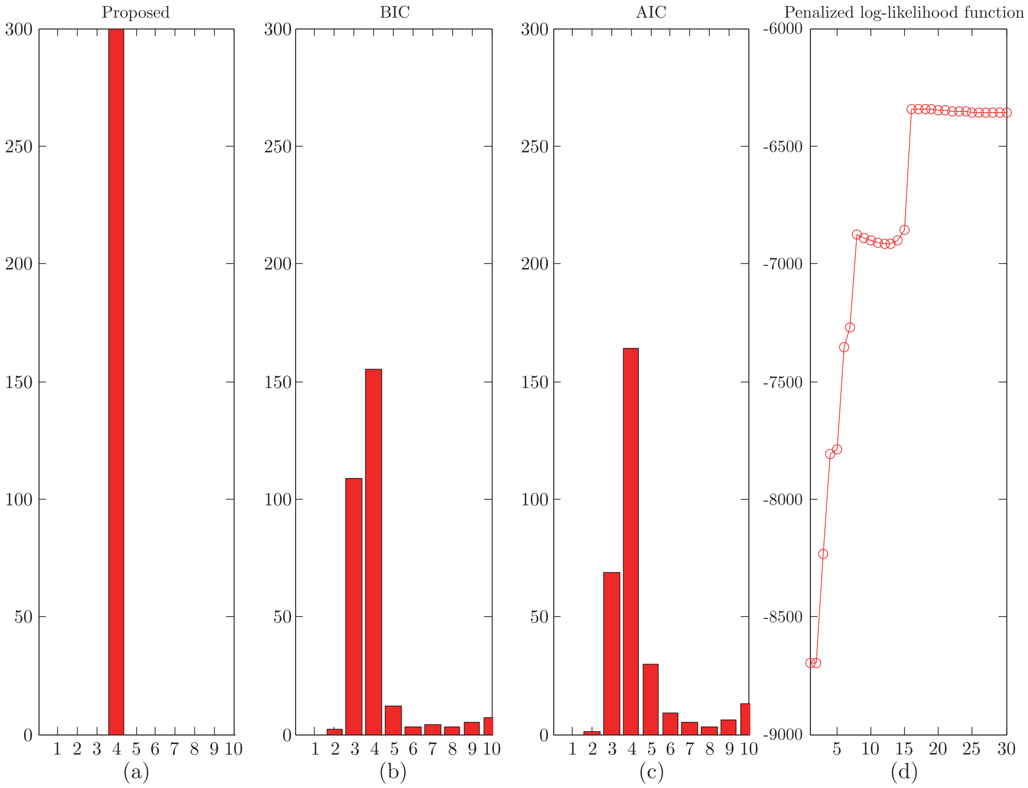


Figure 5. Histogram of estimated numbers of components for Example 2. (a) the proposed method (2.3), (b) BIC, (c) AIC. (d) The penalized log likelihood function for one typical run.

Table 4. Parameter estimation with standard deviation ($M = 50$) for Example 2.

Component		Mixing Probability	Mean		Covariance (eigenvalue)	
	True	0.3	-2	-2	0.1	0.2
1	(2.3)	0.3016(0.0107)	-1.9982(0.0223)	-1.9998(0.0312)	0.0986(0.0110)	0.2034 (0.0241)
	(2.4)	0.3009(0.0095)	-1.9986(0.0218)	-1.9978(0.0320)	0.0993(0.0110)	0.2010(0.0238)
	True	0.3	-2	-2	1.2984	7.7016
2	(2.3)	0.3002(0.0128)	-2.0040(0.1086)	-2.0052(0.1819)	1.2823(0.1386)	7.6696(0.7538)
	(2.4)	0.3017(0.0115)	-1.9988(0.1173)	-2.0116(0.1811)	1.2757(0.1318)	7.6734(0.7476)
	True	0.3	2	0	0.5	4
3	(2.3)	0.3015(0.0083)	1.9986(0.0500)	0.0054(0.1365)	0.4998(0.0531)	3.9951(0.3651)
	(2.4)	0.3012(0.0084)	2.0015(0.0505)	0.0102(0.1268)	0.4915(0.0524)	3.9751(0.3770)
	True	0.1	1	-4	0.125	0.125
4	(2.3)	0.0966(0.0044)	0.9983(0.0408)	-4.0019(0.0431)	0.1150(0.0251)	0.1327(0.0258)
	(2.4)	0.0962(0.0050)	1.0011(0.0402)	-4.0019(0.0425)	0.1154(0.0256)	0.1313(0.0254)

of these parameters are robust to the initialization of the maximum number of components.

Table 5. Frequency of correct order selection (%) for Example 3.

M	20	40	60	80	100	120	140	160	180	200	
$k = 2, n = 50, m = 25$	89	84	88	89	88	88	92	85	90	87	
M	100	150	200	250	300	350	400	450	500	550	600
$k = 5, n = 300, m = 150$	98	98	97	99	98	98	100	100	97	100	98

Example 3. Bunea et al. (2010) studied sparse density estimation via ℓ_1 penalization (SPADES), and, under some regularity conditions, they showed their proposed method can recover the true number of components in the finite mixture models with high probability. For comparison, we considered a univariate Gaussian mixture model with the true number of components $k = 2$ and $k = 5$. The mixture components were chosen at random from a large pool of M Gaussian distributions $\mathcal{N}(aj, 1), 1 \leq j \leq M$, where for $k = 2$, a was 4, and for $k = 5$, a was 5. The mixing probabilities are all equal to $1/k$. The maximum size M of the candidate pool was $M = 200$ for $k = 2$ and $M = 600$ for $k = 5$. Similar to Bunea et al. (2010), our results were based on $T = 100$ simulations.

To compare the accuracy of order selection, we considered the above two finite mixture model settings with the sample size $n = 50$ for $k = 2$, and $n = 300$ for $k = 5$. We let the size of the candidate Gaussian pool M change from 20 to 200 for $k = 2$, and M change from 100 to 600 for $k = 5$. The initial value for the modified EM algorithm was estimated by K-means clustering with 25 and 150 components for $k = 2$ and $k = 5$, respectively. Bunea et al. (2010) used a larger initial order for SPADES. Table 5 reports the frequencies of selecting the correct order base on 100 simulations. For $k = 2$ and $n = 50$, our result is slightly better than the result shown by Figure 2 in Bunea et al. (2010). For $k = 5$ and $n = 300$, our results are clearly better than theirs.

To evaluate the effect of the mean distance between mixture components on the accuracy of order estimate, we changed a from 0 to 5 for the setting $k = 2$, $n = 100$ and $M = 25$. We considered two initial estimates for our modified EM algorithm. One was by the K-means clustering with 25 components, and the other was a Gaussian candidate pool $\mathcal{N}(aj, 1)$ with $M = 25$, $\pi_j = 1/M$, and $1 \leq j \leq M$. As shown in Figure 6, the percentage of times that the estimated order was the true order was similar for these two initial estimates. Compared to Figure 3 of Bunea et al. (2010), Figure 6 shows that, when the distance between the means is relatively small, our method is more adaptive in detecting the true number of mixture components.

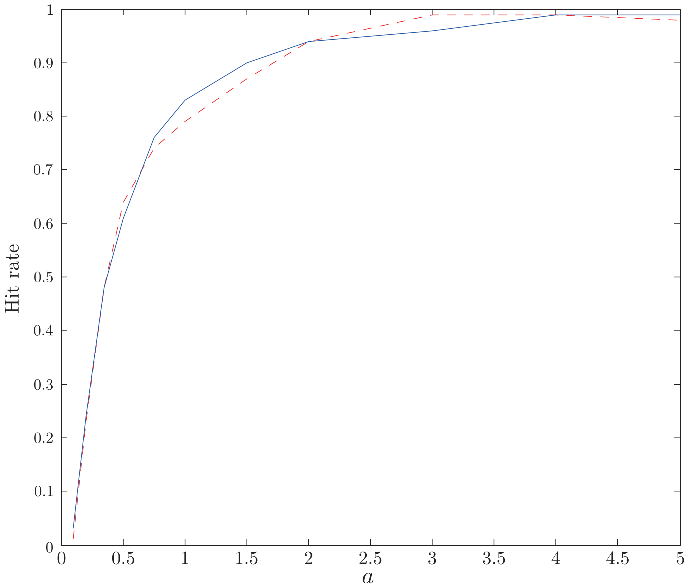


Figure 6. The percentage of times the estimated order $\hat{k}$ is the real order k for different values of a under the setting $k = 2$, $n = 100$ and $M = 25$. The solid line: $\mathcal{N}(aj, 1), j = 1, 2, \ldots, 25$ with the weight 1/25 as the initial estimate. The dash line: K-means estimate with $K = 25$ as the initial estimate.

5. Conclusions and Discussions

In this paper, we proposed a penalized likelihood method for multivariate finite Gaussian mixture models that integrates model selection and parameter estimation. The method involves light computations and is attractive when there are many possible candidate models. Under mild conditions, our proposed method can select the number of components consistently. Although we mainly focused on Gaussian mixture models, we believe our method can be extended to more generalized mixture models, such as the mixture of factor analyzers (Ghahramani and Hinton (1997)).

Our proposed modified EM algorithm gradually discards insignificant components, and does not generate new components or split any large components. If necessary, for complex problems, one can perform the split-and-merge operations (Ueda et al. (1999)) after certain EM iterations to improve the final results. We only show the convergence of our algorithm through simulations, and further theoretical investigation is needed. Classical acceleration methods, such as Louis' method, Quasi-Newton method and the Hybrid method (McLachlan and Peel (2000)), may be used to improve the convergence rate of our algorithm.

Another practical issue is the choice of the tuning parameter λ for the penalized likelihood function. We propose a BIC selection method, and simulation results show that it works well. Moreover, our simulation results show that the final estimate is quite robust to the initial number of components, given that is reasonably large.

We proposed two penalized likelihood functions at (2.3) and (2.4). Although the numerical results obtained by these two penalized likelihood functions are similar they likely have different theoretical properties. We have shown the consistency of model selection and tuning parameter selection by maximizing (2.4) and the proposed BIC method under mild conditions. We have also shown the consistency of model selection by maximizing (2.3), but the conditions are somewhat restrictive. In particular, the consistency of the proposed BIC method for (2.3) needs further investigations.

An ongoing work is to investigate how to extend our method to the mixture of factor analyzers (Ghahramani and Hinton (1997)), to integrate clustering and dimensionality reduction.

Supplementary Materials

Simulation Example 4, Data Analysis and Proofs of Theorem 2 and 3.

Acknowledgements

Tao Huang's research was supported in part by Shanghai Pujiang Program (13PJC048), Program for Changjiang Scholars and Innovative Research Team in SUFE (PCSIRT-IRT13077), and the State Key Program in the Major Research Plan of NSF of China Grant (No. 91546202). Heng Peng's research was supported part by CERG grants by the Research Grants Council of Hong Kong (No. HKBU 202012 and HKBU 12302615), FRG grants from Hong Kong Baptist University (No. FRG2/12-13/077 and FRG2/14-15/064) and NSF of CHINA Grant (No. 10871054). Kun Zhang was supported in part by the Research Grants Council of Hong Kong under the General Research (No. LU342213). Dr. Peng is the corresponding author of this paper.

The authors would like to thank the Editor, an associate editor, and the referees for their helpful comments and suggestions that have helped us substantially improve the quality of the paper. We are very grateful to a referee who pointed out a flaw in a proof of an earlier version of the paper.

Appendix. Proof of Theorem 1

For the Gaussian mixture model, we assume that there are q Gaussian mixture components and $q \leq M$. Without loss of generality, assume that $\pi_i = 0$ for $i = 1, \ldots, M-q$, $\pi_i = \pi_l^0$ and $\phi(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \phi(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0)$ for $i = M-q+1, \ldots, M, l =$

$1, \ldots, q$. Then by the idea of locally conic models (Dacunha-Castelle and Gassiat (1997, 1999)), the density function of the Gaussian mixture model can be rewritten as

$$f(\mathbf{x}, \boldsymbol{\theta}) = f(\mathbf{x}, \theta, \boldsymbol{\beta}) = \sum_{i=1}^{M-q} \alpha_i \theta \cdot \phi(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \sum_{l=1}^{q} (\pi_l^0 + \rho_l \theta) \cdot \phi(\boldsymbol{\mu}_l^0 + \theta \delta_\mu^l, \boldsymbol{\Sigma}_l^0 + \theta \delta_\Sigma^l),$$

where $\boldsymbol{\beta} = (\alpha_1, \ldots, \alpha_{M-q}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{M-q}, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_{M-q}, \delta_\mu^1, \ldots, \delta_\mu^q, \delta_\Sigma^1, \ldots, \delta_\Sigma^q, \rho_1, \ldots, \rho_q)$, and $(\pi_1, \ldots, \pi_M)$ in (2.1) can be defined as $\pi_i = \alpha_i \theta$, $i = 1, \ldots, M-q$ and $\pi_i = \pi_l^0 + \rho_l \theta$, $i = M-q+1, \ldots, M$, $l = 1, \ldots, q$. By the restrictions imposed on the $\boldsymbol{\beta}$

$$\begin{aligned} &\alpha_i \geq 0,\ \mu_i \in \mathbf{R}^d, \text{and } \boldsymbol{\Sigma}_i \in \mathbf{R}^{d \times d}, i = 1, \ldots, M-q, \\ &\delta_\mu^l \in \mathbf{R}^d,\ \delta_\Sigma^l \in \mathbf{R}^{d \times d}, \text{and } \rho \in \mathbf{R},\ l = 1, \ldots, q, \qquad\qquad (\mathrm{A.1}) \\ &\sum_{i=1}^{M-q} \alpha_i + \sum_{l=1}^{q} \rho_l = 0, \text{and } \sum_{i=1}^{M-q} \alpha_i^2 + \sum_{l=1}^{q} \rho_l^2 + \sum_{l=1}^{q} \|\delta_\mu^l\|^2 + \sum_{l=1}^{q} \|\delta_\Sigma^l\|^2 = 1 \end{aligned}$$

and, by permutation, such a parametrization is locally conic and identifiable.

After the parametrization, the penalized likelihood functions (2.3) and (2.4) can, respectively, be rewritten as

$$\begin{aligned} \ell_P(\boldsymbol{\theta}) &= \ell(\boldsymbol{\theta}) - n\lambda D_f \sum_{m=1}^{M} [\log(\epsilon + \pi_m)) - \log(\epsilon)] \\ &\hat{=} \ell_P(\theta, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i, \theta, \boldsymbol{\beta}) - n\lambda D_f \sum_{m=1}^{M} [\log(\epsilon + \pi_m) - \log(\epsilon)], \qquad (\mathrm{A.2}) \end{aligned}$$

$$\begin{aligned} \ell_P(\boldsymbol{\theta}) &= \ell(\boldsymbol{\theta}) - n\lambda D_f \sum_{m=1}^{M} [\log(\epsilon + p_\lambda(\pi_m)) - \log(\epsilon)] \\ &\hat{=} \ell_P(\theta, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i, \theta, \boldsymbol{\beta}) - n\lambda D_f \sum_{m=1}^{M} [\log(\epsilon + p_\lambda(\pi_m)) - \log(\epsilon)] .(\mathrm{A.3}) \end{aligned}$$

For the key ideas of the proof of Theorem 1, first assume that the true Gaussian mixture density is $g_0 = \sum_{l=1}^{q} \pi_l^0 \phi(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0)$, and then define $\mathcal{D}$ as the subset of functions of the form

$$\begin{aligned} \sum_{l=1}^{q} \pi_l^0 \sum_{i=1}^{d} \frac{\delta_{\mu_i}^l D_i^1 \phi(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0)}{g_0} &+ \sum_{l=1}^{q} \pi_l^0 \sum_{i \geq j=1}^{d} \frac{\delta_{\Sigma_{i,j}}^l D_{i,j}^1 \phi(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0)}{g_0} \\ &+ \sum_{l=1}^{q} \rho_l \frac{\phi(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0)}{g_0} + \sum_{i=1}^{M-q} \alpha_i \frac{\phi(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{g_0}, \end{aligned}$$

where D_i^1 is the derivative of $\phi(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ for the ith component of $\boldsymbol{\mu}_l$, $D_{i,j}^1$ is the derivative of $\phi(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ for the (i, j) component of $\boldsymbol{\Sigma}_l$. Functions in $\mathcal{D}$, $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, \ldots, M-q$ and $(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0), l = 1, \ldots, q$ satisfy conditions P1 and P2; for any $\boldsymbol{\Sigma}_i, i = 1, \ldots, M-q$, there exists a $\boldsymbol{\Sigma}_l^0, 1 \leq \ell \leq q$ such that $\boldsymbol{\Sigma}_i \leq (1+\kappa)\boldsymbol{\Sigma}_l^0$ where $0 \leq \kappa < 1$.

Proposition 1. *$\mathcal{D}$ is a Donsker class.*

Proof. The functions in $\mathcal{D}$ can be decomposed as

$$I_1 = \sum_{l=1}^{q} \pi_l^0 \sum_{i=1}^{d} \frac{\delta_{\mu_i}^l D_i^1 \phi(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0)}{g_0} + \sum_{l=1}^{q} \pi_l^0 \sum_{i \geq j=1}^{d} \frac{\delta_{\Sigma_{i,j}}^l D_{i,j}^1 \phi(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0)}{g_0} + \sum_{l=1}^{q} \rho_l \frac{\phi(\boldsymbol{\mu}_l^0, \boldsymbol{\Sigma}_l^0)}{g_0},$$

$$I_2 = \sum_{i=1}^{M-q} \alpha_i \frac{\phi(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)}{g_0}.$$

Hence $\mathcal{D} = \{f + g : f \in \mathcal{D}_1, g \in \mathcal{D}_2\}$ where $\mathcal{D}_1$ is the function set with form I_1 indexed by δ_* and ρ_* and $\mathcal{D}_2$ is the function set with form I_2 indexed by α_*, $\boldsymbol{\mu}_*$ and $\boldsymbol{\Sigma}_*$.

Here I_1 is a linear combination of given functions, and the linear combination coefficients should satisfy (A.1). Hences as shown by Example 19.17 in van der Vaart (1998) and under conditions P1 and P2, the class of the functions in $\mathcal{D}_1$ is a Donsker class.

On the other hand, for I_2, under conditions P1 and P2, $\boldsymbol{\mu}_i, i = 1, \ldots, M-q$, are bounded and $\boldsymbol{\Sigma}_i, i = 1, \ldots, M-q$, are positive with bounded eigenvalues. Then under the conditions P1 and P2, for each $1 \leq l \leq q$ define $\mathcal{F}_{il} = \{\alpha_i \phi(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)/g_0 : (\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \in \Omega_l\}$, where $\Omega_l = \{\alpha, \boldsymbol{\mu}, \boldsymbol{\Sigma} : -1 \leq \alpha \leq 1, \boldsymbol{\Sigma} \leq (1+\kappa)\boldsymbol{\Sigma}_l^0\}$ and $0 \leq \kappa < 1$. Through some cumbersome calculations involving matrix operations and matrix derivatives, we can show that functions $\alpha_i \phi(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)/g_0$ in $\mathcal{F}_{il}$ satisfy a Lipschitz condition, and that the $L_2(g_0)$ norm of the Lipschitz coefficient is bounded. Therefore for each $1 \leq l \leq q$, the class of functions in $\mathcal{F}_{il}$ is a Donsker class by Theorem 19.5 and Example 19.7 in van der Vaart (1998). Hence, under the conditions P1 and P2, $\mathcal{F}_i = \bigcup_{l=1}^q \mathcal{F}_{il} = \{\alpha_i \phi(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)/g_0 : (\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \in \bigcup_{l=1}^q \Omega_l\}$ is a Donsker class. So by Example 19.20 in van der Vaart (1998), $\mathcal{D}_2 = \{I_2 : I_2 = \sum_{i=1}^{M-q} h_i,\ h_i \in \mathcal{F}_i\}$ is a Donsker class.

Finally as shown by Example 19.20 in van der Vaart (1998), the class of functions in $\mathcal{D} = \{f + g : f \in \mathcal{D}_1, g \in \mathcal{D}_2\}$ is a Donsker class.

Proof of Theorem 1. To prove the theorem, we first show that there exists a maximizer $(\theta, \boldsymbol{\beta})$ such that $\theta = O_p(1/\sqrt{n})$. In fact, it is sufficient to show that, for a large constant C, $\ell(\theta, \boldsymbol{\beta}) < \ell(0, \boldsymbol{\beta})$ where $\theta = C/\sqrt{n}$ and $\boldsymbol{\beta}$ is in a local compact area $\Omega = \{\boldsymbol{\beta} : (\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \in \bigcup_{l=1}^q \Omega_l,\ i = 1, \ldots, M-q\}$ where Ω_l is defined in the proof of Proposition 1. If $\theta = C/\sqrt{n}$, then

$$\begin{aligned}
&\ell_p(\theta, \boldsymbol{\beta}) - \ell_p(0, \boldsymbol{\beta})\\
&\le \sum_{i=1}^n \{\log f(\mathbf{x}_i, \theta, \boldsymbol{\beta}) - \log g_0(\mathbf{x}_i)\}\\
&\quad -n\lambda D_f \sum_{m=M-q+1}^{M} [\log(\epsilon + p_\lambda(\pi_m)) - \log(\epsilon + p_\lambda(\pi^0_{m-M+q}))] \hat{=} I_1 + I_2.
\end{aligned}$$

For I_2, because of $\theta = C/\sqrt{n}$ and by the restriction condition on $\rho_l, l = 1, \ldots, q$, we have $|\pi_m - \pi^0_{m-M+q}| \le C/\sqrt{n}$ when $m > M - q$. By the property of the penalty function, we then have

$$\begin{aligned}
|I_2| &= |-n\lambda D_f \sum_{m=M-q+1}^{M} [\log(\epsilon + p_\lambda(\pi_m)) - \log(\epsilon + p_\lambda(\pi^0_{m-M+q}))]|\\
&= |-n\lambda D_f \sum_{m=M-q+1}^{M} [\log(\epsilon + a\lambda) - \log(\epsilon + a\lambda)]| = 0.
\end{aligned}$$

For I_1, we have

$$\begin{aligned}
I_1 = &\sum_{i=1}^n \frac{f(\mathbf{x}_i, \theta, \boldsymbol{\beta}) - g_0(\mathbf{x}_i)}{g_0(\mathbf{x}_i)} - \frac{1}{2}\sum_{i=1}^n \left(\frac{f(\mathbf{x}_i, \theta, \boldsymbol{\beta}) - g_0(\mathbf{x}_i)}{g_0(\mathbf{x}_i)}\right)^2\\
&+ \frac{1}{3}\sum_{i=1}^n U_i \left(\frac{f(\mathbf{x}_i, \theta, \boldsymbol{\beta}) - g_0(\mathbf{x}_i)}{g_0(\mathbf{x}_i)}\right)^3
\end{aligned}$$

if $\theta = C/\sqrt{n}$, where $|U_i| \le 1$. Expanding $f(\mathbf{x}, \theta, \boldsymbol{\beta})$ up to the second order gives

$$f(\mathbf{x}, \theta, \boldsymbol{\beta}) = g_0(\mathbf{x}) + \theta \cdot f'(\mathbf{x}, 0, \boldsymbol{\beta}) + \frac{\theta^2}{2} \cdot f''(\mathbf{x}, \theta^*, \boldsymbol{\beta}),$$

for some $\theta^* \le \theta$.

Noticing $\theta = C/\sqrt{n}$, $\mathrm{E}f'/g_0 = 0$, $\mathrm{E}f''/g_0 = 0$, by conditions P1 and P2, and Proposition 1 for the class $\mathcal{D}$, we have

$$I_1 = \left\{\sum_{i=1}^n \theta \frac{f'(\mathbf{x}_i, 0, \boldsymbol{\beta})}{g_0(\mathbf{x}_i)} - \frac{1}{2}\sum_{i=1}^n \theta^2 \left(\frac{f'(\mathbf{x}_i, 0, \boldsymbol{\beta})}{g_0(\mathbf{x}_i)}\right)^2\right\}(1 + o_p(1)).$$

Since $(1/\sqrt{n})\sum_{i=1}^n f'(\mathbf{x}_i, 0, \boldsymbol{\beta})/g_0(\mathbf{x}_i)$ converges uniformly in distribution to a Gaussian process by Proposition 1 and $\sum_{i=1}^n (f'(\mathbf{x}_i, 0, \boldsymbol{\beta})/g_0(\mathbf{x}_i))^2$ is of order $O_p(n)$ by the Law of Large Numbers, we have

$$I_1 = \frac{C}{\sqrt{n}} \cdot O_P(\sqrt{n}) - \frac{C^2}{n} \cdot O_p(n).$$

When C is large enough, the second term of I_1 dominates other terms in the penalized likelihood ratio function. Then we have $\ell_p(\theta, \boldsymbol{\beta}) - \ell_p(0, \boldsymbol{\beta}) < 0$ with probability tending to one. Hence there exists a maximizer $(\theta, \boldsymbol{\beta})$ with probability tending to one such that $\theta = O_p(1/\sqrt{n})$.

Next we show that there exists a maximizer $(\hat{\theta}, \hat{\boldsymbol{\beta}})$ satisfying $\hat{\theta} = O_p(1/\sqrt{n})$ such that $\hat{q} = q$ or $\hat{\pi}_m = 0, m = 1, \ldots, M-q$. In fact, when $\hat{\theta} = O_p(1/\sqrt{n})$, by the restriction condition on α_i, we have $\hat{\pi}_m = O_p(1/\sqrt{n})$, $m = 1, \ldots, M-q$. A Lagrange multiplier β is taken into account for the constraint $\sum_{m=1}^{M} \hat{\pi}_m = 1$. Then it is sufficient to show that

$$\frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \hat{\pi}_m} < 0 \quad \text{for} \quad \hat{\pi}_m < \varepsilon_n \tag{A.4}$$

with probability tending to one for the maximizer $(\theta, \boldsymbol{\beta})$, where $\varepsilon_n = Cn^{-1/2}$, $m \leq M-q$, and $\ell^*(\boldsymbol{\theta}) = \ell_p(\boldsymbol{\theta}) - \beta(\sum_{m=1}^{M} \pi_m - 1)$. To show that (A.4) holds, we consider the partial derivatives for $\hat{\pi}_m, m > M-q$ and have

$$\frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \hat{\pi}_m} = \sum_{i=1}^{n} \frac{\phi_m(\mu_m, \boldsymbol{\Sigma}_m)}{\sum_{i=1}^{M} \hat{\pi}_i \phi_i(\mu_i, \boldsymbol{\Sigma}_i)} - n\lambda D_f \frac{1}{\epsilon + \hat{\pi}_m} - \beta = 0. \tag{A.5}$$

The first term in (A.5) is of order $O_p(n)$ by the Law of Large Numbers. Given $m > M-q$ and $\theta = O_p(1/\sqrt{n})$, we have $\hat{\pi}_m = \pi^0_{m-M+q} + O_p(1/\sqrt{n}) > (1/2) \cdot \min(\pi^0_1, \ldots, \pi^0_q)$; then the second term is $O_p(n\lambda) = o_p(n)$, and moreover $\beta = O_p(n)$.

Next, consider

$$\frac{\partial \ell^*(\boldsymbol{\theta})}{\partial \hat{\pi}_m} = \sum_{i=1}^{n} \frac{\phi_m(\mu_m, \boldsymbol{\Sigma}_m)}{\sum_{i=1}^{M} \hat{\pi}_i \phi_i(\mu_i, \boldsymbol{\Sigma}_i)} - n\lambda D_f \frac{1}{\epsilon + \hat{\pi}_m} - \beta, \tag{A.6}$$

where $m \leq M-q$ and $\hat{\pi}_m < \varepsilon_n$. The first term and β in (A.6) are of order $O_p(n)$. For the second term, because $\pi_m = O_p(1/\sqrt{n})$, $\sqrt{n}\lambda \to \infty$, and ϵ is sufficient small, we have

$$\frac{\{n\lambda D_f[1/(\epsilon + \pi_m)]\}}{n} = \lambda D_f \frac{1}{\epsilon + \pi_m} = O_p(\sqrt{n}\lambda) \to \infty$$

with probability tending to one. Hence the second term in (A.6) dominates the first and the third terms. Therefore we prove (A.4), or, equivalently, $\hat{\pi}_m = 0$, $m = 1, \ldots, M-q$ with probability tending to one when $n \to \infty$.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* Springer.

Brand, M. E. (1999). Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Comput.* **11**, 1155-1182.

Bunea, F., Tsybakov, A. B., Wegkamp, M. and Barbu, A. (2010). SPADES and mixture models. *Ann. Statist.* **38**, 2525-2558.

Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23**, 221-233.

Chen, J. and Kalbfleisch, J. D. (1996). Penalized minimum-distance estimates in finite mixture models. *Canad. J. Statist.* **24**, 167-175.

Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *J. Amer. Statist. Assoc.* **104**, 187-196.

Corduneanu, A. and Bishop, C. M. (2001). Variational Bayesian model selection for mixture distributions. In *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, 27-34. Morgan Kaufmann.

Dacunha-Castelle, D. and Gassiat, E. (1997). Testing in locally conic models and application to mixture models. *ESAIM: Probability and Statistics*, 285-317.

Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes, *Ann. Statist.* **27**, 1178-1209.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Figueiredo, M. and Jain, A. (2002). Unsupervised learning on finite mixture models, *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 381-396.

Ghahramani, Z. and Hinton, G. E. (1997). The EM algorithm for mixtures of factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, Canada.

Hathaway, R. J. (1985). A constrained maximum likelihood estimation for normal mixture distribution. *Ann. Statist.* **13**, 795-780.

Ho, N. and Nguyen, X. (2015). Identifiability and optimal rates of convergence for parameters of multiple types in finite mixtures. arXiv:1501.02497.

James, L. F., Priebe, C. E. and Marchette, D. J. (2001). Consistent estimation of mixture complexity, *Ann. Statist.* **29**, 1281-1296.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā* A **62**, 49-66.

Leroux, B. (1992). Consistent estimation of a mixing distribution. *Ann. Statist.* **20**, 1350-1360.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications.* Institute for Mathematical Statistics, Hayward, CA.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* John Wiley & Sons, New York.

Nguyen, X. (2013). Convergence of latent mixture measures in finite and infinite mixture models. *Ann. Statist.* **41**, 370-400.

Ormoneit, D. and Tresp, V. (1998). Averaging, maximum penalized likelihood and Bayesian estimation for improving Gaussian mixture probability density estimates. *IEEE Trans. Neural Networks* **9**, 1045-9227.

Ray, S. and Lindsay, B. G. (2008). Model selection in high-dimensions: A quadratic-risk based approach. *J. Roy. Statist. Soc. Ser. B* **70**, 95-118.

Roeder, K. and Wasserman, L. (1997). Practical density estimation using mixtures of normals. *J. Amer. Statist. Assoc.* **92**, 894-902.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Ueda, N., Nakano, R., Ghahramani, Z. and Hinton, G. E. (1999). SMEM algorithm for mixture models. *Adv. Neural Inform. Process. Systems* **11**, 299-605.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Wang, H., Li, R. and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553-568.

Woo, M. and Sriram, T. N. (2006). Robust estimation of mixture complexity. *J. Amer. Statist. Assoc.* **101**, 1475-1485.

Zivkovic, Z. and van der Heijden, F. (2004). Recursive unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 651-656.

School of Statistics and Management, Shanghai University of Finance and Economics, 777 Guoding Road, Shanghai 200433, China.

E-mail: huang.tao@mail.shufe.edu.cn

Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong.

E-mail: hpeng@math.hkbu.edu.hk

Department of Philosophy, Carnegie Mellon University & Max Planck, Institute for Intelligent Systems, Baker Hall 161, 5000 Forbes Avenue Pittsburgh, PA 15213 USA.

E-mail: kunz1@cmu.edu

(Received March 2014; accepted January 2016)