Electronic Journal of Statistics
Vol. 16 (2022) 5149–5181
ISSN: 1935-7524
https://doi.org/10.1214/22-EJS2061

Statistical inference for normal mixtures with unknown number of components*

Mian Huang

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, P. R. China
e-mail: huang.mian@mail.shufe.edu.cn

Shiyi Tang

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, P. R. China
e-mail: tang.shiyi@163.sufe.edu.cn

Weixin Yao†

Department of Statistics, University of California, Riverside, California 92521, U.S.A.
e-mail: weixin.yao@ucr.edu

Abstract: Statistical inference for normal mixture models with unknown number of components has long been challenging due to the issues of nonidentifiability, degenerated Fisher matrix, and boundary parameters. In this paper, a penalized likelihood estimation procedure is proposed for mixtures of normals with unknown number of components to achieve both the order selection consistency and the root-n convergence rate for the component parameters estimators. We show that the proposed new estimator could avoid being trapped in certain degenerated regions of the nonidentifiable subset of the parameter space for over-fitted normal mixture models so that a regular asymptotic quadratic Taylor expansion of the mixture log-likelihood could be derived. With a suitable penalty function on mixing proportions, the new estimator is proved to be consistent on the order selection, and have an asymptotic normal distribution. Our derived sparsity conditions also reveal some surprising but interesting differences among some commonly used penalty functions and explain why the performance of some popularly used penalty functions, such as Lasso and SCAD, provide unsatisfactory results in the order selection. Extensive simulations and a real data analysis are conducted to demonstrate the effectiveness of the newly proposed estimator.

Keywords and phrases: Normal mixture model, penalized estimation, order selection, EM algorithm.

Received May 2021.

1. Introduction

Finite mixture models are important statistical tools to model a heterogeneous population, and can be applied for cluster analysis, latent class analysis, discrim-

*Yao's research is supported in part by NSF grant DMS-2210272.
†Corresponding author.

inant analysis, image analysis, survival analysis, disease mapping, meta analysis, and more. See, e.g. [25, 14, 30]. Compared to the traditional one-component parametric model, the likelihood based estimation and inference for mixture models are more challenging, especially when the number of components is unknown. Difficulties for the mixture model inference arise due to the possible lack of the point identification, the degeneration of the Fisher information matrix, and the true model parameters lying on the boundary of the parameter space. Various approaches are proposed to tackle these issues. [31] proved that the maximum likelihood estimator (MLE) is strongly consistent in the quotient topological space where the likelihood can not be distinguished. [13] obtained parallel results in the Euclidean space, and further provided justification for the bootstrap inference method. How to choose the number of components (also noted as an order selection) for mixture models has long been a challenging problem. See, for example, [9], [26], [2], [5], [7], and [35].

Among many researches, the convergence rate of the mixture model estimation with unknown number of components is notoriously complicated and challenging. Unlike one component parametric models, the minimax convergence rate and the pointwise convergence rate could be quite different. The minimax rate describes the best performance in the worst situation. One illustration example is to consider the difference between $f \sim \phi(x; 0, 1)$ and $f_n = \frac{1}{2}\{\phi(x; n^{-1/3}, 1) + \phi(x; -n^{-1/3}, 1)\}$, where $\phi(x; \mu, \sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2. The fact

$$f_n = f + n^{-2/3}\phi''(x; 0, 1) + o(n^{-2/3}) \tag{1.1}$$

clearly indicates that the minimax convergence rate in this special case is slower than $n^{-1/3}$. [3] proved that the optimal minimax convergence rate of estimating a mixing distribution is $n^{-1/4}$ when the number of components is unknown. This result is challenged by [17] and [16] that the optimal minimax convergence rate could be slower and may depend on the number of overfitted components.

On the other hand, the pointwise convergence rate can be as fast as $n^{-1/2}$ when the number of component is unknown. The pointwise rate is the speed of estimating a fixed but unknown density when the sample size goes to infinity. It is not hard to show that a two-step procedure could yield an $n^{-1/2}$ convergence rate estimator with the first step providing a consistent order selection and a second step estimating the mixture model given the selected number of components in the first step. See for example, [20] and [16]. The number of components can be consistently estimated via the BIC [19] or more complicated penalized likelihood methods such as [4]. [4] applied a penalty to the differences of location parameters to merge components when the differences are shrunk to zero. Both the order of the mixture model and the mixing distribution can thus be consistently estimated. The two step procedure described above suffers from the issues of irregular likelihood function and heavy computation. [18] proposed a new type of penalized likelihood approach with penalty on mixing proportions, which could simultaneously conduct order selection and parameter estimation.

[1] generalized the results of [18] in finite mixture model not limited to mixture of normal, and provided a conventional proof without using complicated local conic representation as in [18].

We consider the estimation of over-fitted mixture models that contain more components than true ones, as the information about an upper bound on the number of components is relatively easy to obtain. Unlike the representative example (1.1), the nonidentifiable subset of the parameter space for the over-fitted normal mixture contains regions where the Fisher matrix is not fully degenerated. The core idea of this paper is to introduce a penalty to prevent the estimator for over-fitted normal mixture models from being in degenerated areas of the nonidentifiable subset so that the mixture log-likelihood has a regular asymptotic quadratic Taylor expansion. As a result, the worst case is avoided and a regular root-n convergence rate estimator can be established for normal mixture models with unknown number of components.

We first propose a penalized likelihood method with a penalty on the difference of component parameters. We prove that the resulting estimator is in the non-degenerated area of the nonidentifiable subset of the parameter space, and establish a root-n convergence rate for the proposed estimator. We further impose an additional penalty on mixing proportions to provide a sparse estimation of the proportion parameters and thus enable an order selection for mixture models. We prove that the order of the mixture model can be selected consistently, and the resulting parameter estimators are root-n consistent and asymptotically normal. An EM algorithm is proposed to find the penalized likelihood estimator, and proved to have an ascent property. Simulation studies and a real-data analysis are conducted to assess the finite sample performance of the penalized estimation.

Our derived sparsity conditions for the penalty on mixing proportions also reveal some surprising but interesting differences among commonly used penalties in constrained penalized MLE optimization. For example, our results reveal why the performance of some popularly used penalty functions, such as the L_1 penalty [34, LASSO] and the smoothly clipped absolute deviation penalty [10, SCAD] provide unsatisfactory results in the order selection. However, some other penalties such as the minimax concave penalty [37, MCP], the truncated L_1-penalty [32, TLP], and the log-type penalties proposed by [18] can satisfy the sparsity condition with properly chosen tuning parameter. Our simulation studies also confirm above findings.

The rest of this paper is organized as follows. Section 2 introduces a root-n consistent penalized likelihood estimation method for overfitted mixtures of normals. In Section 3, we derive the condition for the penalty function to achieve a sparse estimation of mixing proportions. The consistency of the order selection and the asymptotic distribution for the identifiable parameters are established. An EM algorithm is provided and its ascent property is established. In Section 4, simulations and a real-data analysis are conducted to assess the finite sample performance of the proposed penalized likelihood method. Conclusions and discussions are given in Section 5. Detailed proofs are given in the Appendix.

2. Penalized model estimation for over-fitted mixture models

2.1. Motivating example

Suppose that samples $X_1, X_2, \cdots, X_n$ are generated from the normal distribution $N(\mu_1^0, 1)$ with mean μ_1^0 and variance 1. Assuming that μ_1^0 is known, we fit the above samples using the following two-component normal mixture model:

$$f(x;\psi) = (1-\pi)\phi(x;\mu_1^0,1) + \pi\phi(x;\mu_2,1), \tag{2.1}$$

where $\psi = (\pi, \mu_2)^t$. [27] also studied the above model and showed the nonregular asymptotic properties (such as diverging to ∞) of the likelihood ratio test. The nonidentifiable subset of the parameter space corresponding to the true model $N(\mu_1^0, 1)$ is $\Omega_0 = \{\psi : f(x;\psi) = \phi(x;\mu_1^0,1)\} = \{(\pi,\mu_2) : \pi(\mu_2 - \mu_1^0) = 0\}$. Any point lies in the nonidentifiable subset yields the same density function as $\phi(x;\mu_1^0,1)$. If $\pi = 0$, parameter μ_2 is nonidentifiable, and if $\mu_2 = \mu_1^0$, parameter π is nonidentifiable. Given samples $X_1, X_2, \cdots, X_n$, the log-likelihood function is

$$\ell(\psi) = \sum_{i=1}^{n} \log f(X_i;\psi), \tag{2.2}$$

and the Fisher information is given by

$$I(\psi) = \begin{pmatrix} E_\psi \frac{(\phi(x;\mu_2,1)-\phi(x;\mu_1^0,1))^2}{f^2(x;\psi)} & E_\psi \frac{\pi\phi'(x;\mu_2,1)(\phi(x;\mu_2,1)-\phi(x;\mu_1^0,1))}{f^2(x;\psi)} \\ E_\psi \frac{\pi\phi'(x;\mu_2,1)(\phi(x;\mu_2,1)-\phi(x;\mu_1^0,1))}{f^2(x;\psi)} & E_\psi \frac{(\pi\phi'(x;\mu_2,1))^2}{f^2(x;\psi)} \end{pmatrix}.$$

If $\bar{\psi} = (0, \mu_1^0)^t$, the Fisher information is $I(\bar{\psi}) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$. If $\bar{\psi} = (\pi, \mu_1^0)^t$ with $\pi \neq 0$, the Fisher information is

$$I(\bar{\psi}) = \begin{pmatrix} 0 & 0 \\ 0 & E_{\bar{\psi}} \frac{(\pi\phi'(x;\mu_1^0,1))^2}{f^2(x;\bar{\psi})} \end{pmatrix}.$$

If $\bar{\psi} = (0, \mu_2)^t$ with $\mu_2 \neq \mu_1^0$, the Fisher information is

$$I(\bar{\psi}) = \begin{pmatrix} E_{\bar{\psi}} \frac{(\phi(x;\mu_2,1)-\phi(x;\mu_1^0,1))^2}{f^2(x;\bar{\psi})} & 0 \\ 0 & 0 \end{pmatrix}.$$

A Taylor's expansion of (2.2) on $\bar{\psi} \in \Omega_0$ is

$$\ell(\psi) - \ell(\bar{\psi}) = \frac{\partial \ell(\bar{\psi})}{\partial \psi^t}(\psi - \bar{\psi}) + \frac{1}{2} n \cdot (\psi - \bar{\psi})^t(-I(\bar{\psi}) + o_p(1))(\psi - \bar{\psi}) + R(\psi^*), \tag{2.3}$$

where ψ^* lies between ψ and $\bar{\psi}$, and $R(\psi^*)$ is a Lagrangian remainder of the Taylor's expansion. For the traditional likelihood approach, the key to prove the existence of a root-n consistent estimator [24] is that the quadratic term of a Taylor expansion dominates all other terms in the likelihood expansion. This result holds when the information matrix is positive definite under some regularity conditions. However, for model (2.1) and the point $(0, \mu_1^0) \in \Omega_0$ which represents the true density, the Fisher information is fully degenerated and thus the quadratic term does not dominate other terms uniformly over the root-n boundary of the neighborhood of $(0, \mu_1^0)$. This is one of the reasons why statistical inference is not easy for mixture models when the model parameters are not identifiable.

Interestingly, if the Taylor's expansion (2.3) is taken at $\bar{\psi} = (0, \mu_2)$ with μ_2 not converge to μ_1^0, it can be shown that the quadratic term is asymptotically bounded away from 0. More specifically, for $\psi = \bar{\psi} + (\frac{C}{\sqrt{n}}, 0)$,

$$\begin{aligned} n \cdot (\psi - \bar{\psi})^t I(\bar{\psi})(\psi - \bar{\psi}) &= n \cdot (\frac{C}{\sqrt{n}}, 0) I((0, \mu_2))(\frac{C}{\sqrt{n}}, 0))^t \\ &= n(\frac{C}{\sqrt{n}})^2 E_{\bar{\psi}} \frac{\{\phi(x; \mu_2, 1) - \phi(x; \mu_1^0, 1)\}^2}{f^2(x; \bar{\psi})} \\ &\to C^2 E_{\bar{\psi}} \frac{\{\phi(x; \mu_2, 1) - \phi(x; \mu_1^0, 1)\}^2}{f^2(x; \bar{\psi})} > 0. \end{aligned}$$

Therefore, with a sufficient large $|C|$, the quadratic term dominates all other terms in the likelihood expansion. Noted that the quadratic term is also nondegenerate at (π, μ_1^0) when π is bounded away from 0. This would support the methodology of [4] and [29]. See section 5 for further discussion.

Now we define $\Omega^* = \{(0, \mu_2) : \mu_2 \neq \mu_1^0\}$, which is a subset of the nonidentifiable subset Ω_0, and called *target subset* in this paper. From the above analysis we know that the target subset Ω^* is a desired region of estimation which can also identify the true model with the redundant component having a zero component proportion. More importantly, at any point in the newly defined target subset $\Omega^* \subset \Omega_0$, the likelihood can be expanded with a positive quadratic term that dominates all other terms and the traditional technique of the MLE can be employed. Hence, we propose to impose a penalty to prevent the estimate of μ_2 from being close to μ_1^0 so that our target region corresponding to the true model is Ω^* instead of the whole nonidentifiable subset Ω_0. The penalized likelihood function is given by

$$\ell(\pi, \mu_2) = \sum_{i=1}^{n} \log[(1-\pi)\phi(X_i; \mu_1^0, 1) + \pi\phi(X_i; \mu_2, 1)] + \alpha P(|\mu_2 - \mu_1^0|), \quad (2.4)$$

for some $\alpha > 0$, where $P(\cdot)$ is a penalty function imposed to prevent the estimate of μ_2 from being close to μ_1^0 (see comments after (2.10) for more discussion about the penalty function $P(\cdot)$). For example, $P(\cdot) \equiv P_\gamma(\cdot)$ could be a truncated log function which is flat after a threshold $\gamma > 0$, i.e., for $x > 0$, the derivative of the penalty is

$$P'_\gamma(x) = \frac{1}{x} I(x \leq \gamma). \quad (2.5)$$

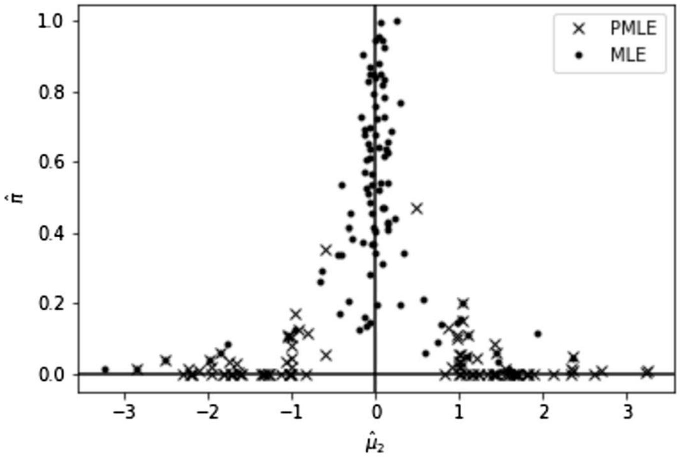

Fig 1. *An illustration of maximum likelihood estimation and penalized maximum likelihood estimation.*

It can be shown that under some mild conditions (see Proposition 1 below), the maximum penalized likelihood estimator (either global or local) of μ_2 does not converge to μ_1^0. Therefore, the quadratic term is positive and dominates other terms in the likelihood expansion. The root-n consistency of the penalized likelihood estimator for π can then be established. See Corollary 1 below for more detail.

We next give a simple illustration of the difference between the standard maximum likelihood estimator and the penalized likelihood estimator (2.4). We set $\mu_1^0 = 0$ and generate 100 observations from one component model $N(0, 1)$. We fit the generated data using a mixture of two normals. The two estimation results are depicted in Figure 1. The maximum likelihood estimate (MLE) converges to the nonidentifiable subset $\Omega_0 = \{(\pi, \mu) : 0 \times (-\infty, +\infty) \text{ or } [0, 1] \times 0\}$, while the penalized maximum likelihood estimate (PMLE) converges to the target subset $\Omega^* = \{(0, \mu_2) : \mu_2 \neq 0\}$. Note, however, many PMLE roots might represent small clusters (with very small proportions) in the tail of the sample. Therefore, the PMLE proposed in this section can not be directly applied to perform model selection, i.e., choose the number of components. In next section, we propose to add a second penalty on mixing proportions to force those small proportions (corresponding to nonexisting components) to exactly 0, and hence provide a sparse estimation of component proportions resulting in automatic choice of number of components.

2.2. *Penalized model estimation and its root-n consistency*

We assume that the density of a finite mixture of normals has the following form

$$f_0(x, \psi_0) = \sum_{m=1}^{M_0} \pi_{m0} \phi(x; \mu_{m0}, \sigma_{m0}^2), \tag{2.6}$$

where $\psi_0 = (\pi_{10}, \cdots, \pi_{M_0,0}, \mu_{10}, \sigma^2_{10}, \cdots, \mu_{M_0,0}, \sigma^2_{M_0,0})^t$, M_0 is the number of components, and all mixing proportions π_{m0} are strictly positive and sum up to 1, i.e., $\pi_{m0} > 0$, and $\sum_{m=1}^{M_0} \pi_{m0} = 1$. In addition, the component parameters $\{\mu_{m0}, \sigma^2_{m0}\}$ are assumed to be different, i.e., no two pairs of $\{\mu_{m0}, \sigma^2_{m0}\}$ are the same. Since the finite mixture of normals is identifiable up to the label switching [33, 36], M_0 in (2.6) is the smallest number of components for a mixture model to have the same density as f_0.

Note that the number of components (the order) is typically unknown in practice. We next propose a simultaneous consistent order selection and root n consistent parameter estimator for the model (2.6). Given an upper bound $M > M_0$, an over-fitted/enlarged normal mixture model has the following form

$$f(x;\psi) = \sum_{m=1}^{M} \pi_m \phi(x; \mu_m, \sigma^2_m), \tag{2.7}$$

where $\psi = (\pi_1, \pi_2, \cdots, \pi_M, \mu_1, \sigma^2_1, \cdots, \mu_M, \sigma^2_M)^t$ takes value in the parameter space $\Omega \subset \mathbb{R}^d$ with $d = 3M-1$, the mixing proportions $\pi_m \geq 0$ and sum up to 1. For simplicity of explanation and the technical proof, we assume the parameter space Ω for ψ is compact. Compactness assumption is necessary for the proofs of theoretical properties in this paper. Combining with other methods to relax the assumption, such as [8] and [6], could be a future research.

As there are multiple ways to represent the true M_0-component mixture by an over-fitted M-component mixture, the larger model is no longer identifiable. Without losing clarity, we reuse the notations Ω_0 and Ω^* in Section 2.1. Define Ω_0 as the nonidentifiable subset that yields the same density as the true density $f_0(x)$ in (2.6), i.e.,

$$\Omega_0 = \left\{ \psi : \sum_{m=1}^{M} \pi_m \phi(x; \mu_m, \sigma^2_m) = \sum_{m=1}^{M_0} \pi_{m0} \phi(x; \mu_{m0}, \sigma^2_{m0}) \right\}.$$

In order to provide consistent order selections and root-n consistent parameter estimates, we restrain our attention on the following target subset of Ω_0:

$$\Omega^* = \{\psi \in \Omega_0 : (\mu_s, \sigma^2_s) \neq (\mu_k, \sigma^2_k) \text{ for } 1 \leq s < k \leq M\}, \tag{2.8}$$

in which no two pairs of component parameters are the same and hence all redundant components have zero component proportions. In Lemma 1, we will prove that at any point in the newly defined target subset $\Omega^* \subset \Omega_0$, the likelihood can be expanded with a positive quadratic term that dominates all other terms, and hence the traditional technique of the MLE can be employed.

The defined target subset in this paper is closely related to the concept of parsimonious subset defined in [1], but they are different in many ways. The parsimonious subset only restricts that no two pairs of component parameters are the same for components with positive proportion parameters, while the component parameters can be the same when their proportions are zero. The new definition of target subset is more concise, and ensures that the proposed penalized estimator in (2.10) below can eliminate unnecessary mixture components with an automatic choice of the number of components.

Remark 1. Because the mixture of normal is identifiable, the mixing distribution is unique. Hence if we force any two pairs of component parameters to be different, the number of nonzero mixing proportions of ψ in Ω^* will be exactly M_0, and the corresponding component parameters will be equal to those true parameters of ψ_0 up to a label switching. There are $M - M_0$ zero mixing proportions, and the corresponding component parameters are arbitrary but no two pairs of component parameters are the same. Without loss of generality, we reorder the parameter in Ω^* such that if $\psi_0^* \in \Omega^*$, then $\psi_0^* = (\pi_{10}, \mathbf{0}, \theta_{10}, \theta_2)$, where $\pi_{10} = (\pi_{10}, \cdots, \pi_{M_0,0})^t$, $\theta_{10} = (\mu_{10}, \sigma_{10}^2, \cdots, \mu_{M_0,0}, \sigma_{M_0,0}^2)^t$, and θ_2 is the arbitrary component parameters corresponding to the zero mixing proportions. Therefore, the M_0-component mixture model (2.6) can be identified by the first M_0 components of Ω^*.

We first propose a penalized method to ensure estimated component parameters to be different. Let $X_1, X_2, \cdots, X_n$ be i.i.d. random samples from an M_0-component mixture (2.6). The log-likelihood function of an over-fitted M-component mixture ($M > M_0$) is

$$\ell(\psi) = \sum_{i=1}^{n} \log \left\{ \sum_{m=1}^{M} \pi_m \phi(X_i; \mu_m, \sigma_m^2) \right\}, \tag{2.9}$$

and the proposed penalized log-likelihood function is

$$\widetilde{\ell}(\psi) = \ell(\psi) + \alpha \sum_{1 \le s < k \le M} P(d(\eta_s, \eta_k)), \tag{2.10}$$

where $\eta_j = (\mu_j, \sigma_j^2)$ and $d(\eta_s, \eta_k)$ is a distance measure between component parameter η_s and η_k, such as the Euclidean distance $|\eta_{sk}| = \{(\mu_k - \mu_s)^2 + (\sigma_k^2 - \sigma_s^2)^2\}^{1/2}$ adopted in this article, $\alpha > 0$ is a tuning parameter, and $P(\cdot)$ is a nondecreasing function defined on $(0, \infty)$ such that $\lim_{x \to 0} P(x) \to -\infty$ to prevent any two component parameters from getting too close. To reduce the estimation bias for separate mixture components, we also assume that $P(d)$ is a flat function when d is larger than a threshold $\gamma > 0$, i.e., $P'(d) = 0$ when $d > \gamma$. Therefore, the penalty will not affect the estimation when $d > \gamma$. Similar ideas have also been adopted by fold concave penalties such as SCAD [10] or MCP [37]. In this article, we will use the truncated log function described in (2.5) for $P(\cdot)$. If $\sigma_m^2 = \sigma^2$ for $m = 1, \cdots, M$, and assumed known, then there is only one parameter μ_m in each component. Note that (2.4) is a special case of (2.10) with $M = 2$.

Let Ω_1 and Ω_2 be two closed sets in $\mathbb{R}^d$, and $D(\Omega_1, \Omega_2)$ be a metric between Ω_1 and Ω_2 such that

$$D(\Omega_1, \Omega_2) = \inf_{\psi_1 \in \Omega_1} \inf_{\psi_2 \in \Omega_2} ||\psi_1 - \psi_2||, \tag{2.11}$$

where $||\cdot||$ is the Euclidean distance, and Ω_1 and Ω_2 are allowed to be single points. When both Ω_1 and Ω_2 are single points, this metric is the same as the classic Euclidean distance.

Proposition 1. *Suppose that $\hat{\psi} = (\hat{\pi}_1, \hat{\pi}_2, \cdots, \hat{\pi}_M, \hat{\mu}_1, \hat{\sigma}_1^2, \cdots, \hat{\mu}_M, \hat{\sigma}_M^2)^t$ is a maximizer of $\widetilde{\ell}(\psi)$, then $D(\hat{\psi}, \Omega_0) = o_p(1)$, and $|\hat{\eta}_{sk}| = \{(\hat{\mu}_k - \hat{\mu}_s)^2 + (\hat{\sigma}_k^2 - \hat{\sigma}_s^2)^2\}^{1/2}$ does not converge to zero in probability for any sub-sequence and any pair of (s,k), $1 \leq s < k \leq M$.*

Proposition 1 states that the penalized maximum likelihood estimate of (2.10) can consistently estimate the mixing distribution and asymptotically has different component parameter estimates. Therefore, if the penalized maximum likelihood estimate gives a consistent distribution estimator of the true distribution, such as in [31] and [13], then the corresponding parameter estimate must be located in the neighborhood of the target subset Ω^*. At the points in Ω^*, we can prove the dominance of the quadratic term of the log-likelihood's Taylor expansion over other terms as we did in the motivating example in Section 2.1, and further prove the root-n consistency of the resulting estimator.

We use and highlight the following notations:

(1) An open subset ω_ϵ that contains Ω^*,

$$\omega_\epsilon = \{\psi \in \Omega : D(\psi, \Omega^*) \leq \epsilon\}, \epsilon > 0.$$

(2) For any point $\psi \in \Omega$, $\psi_0^*(\psi) \in \Omega^*$ is the closest point in Ω^* to ψ based on the Euclidean distance. Therefore, $D(\psi, \Omega^*) = D(\psi, \psi_0^*(\psi))$.

Regularity Conditions for likelihood function

(R1) For all x, the density $f(x;\psi)$ admits all third derivatives for all $\psi \in \omega_\epsilon$,

$$\left| \frac{\partial^3 \log f(x;\psi)}{\partial \psi_j \partial \psi_k \partial \psi_l} \right| \leq M_{jkl}(x),$$

where $E_{\psi_0^*} M_{jkl}(x) \leq +\infty$ for all j,k,l and for all $\psi_0^* \in \Omega^*$.

(R2) $f(x;\psi)$ has a support that does not depend on ψ, and the first and second logarithmic derivatives of $f(x;\psi)$ satisfy the equations

$$E_\psi \left[\frac{\partial \log f(x;\psi)}{\partial \psi_j} \right] = 0 \text{ for } j = 1, \cdots, d,$$

$$I_{jk}(\psi) = E_\psi \left[\frac{\partial \log f(x;\psi)}{\partial \psi_j} \frac{\partial \log f(x;\psi)}{\partial \psi_k} \right] = E_\psi \left[-\frac{\partial^2 \log f(x;\psi)}{\partial \psi_j \partial \psi_k} \right],$$

for $j,k = 1, \cdots, d$.

(R3) For any $\psi \in \omega_\epsilon / \Omega_0$ and its closet point $\psi_0^*(\psi)$ in Ω^*, the information matrix has the property that $(\psi - \psi_0^*(\psi))^t I(\psi_0^*(\psi))(\psi - \psi_0^*(\psi)) > 0$, where

$$I(\psi) = E\left\{ \left[\frac{\partial \log f(x;\psi)}{\partial \psi} \right] \left[\frac{\partial \log f(x;\psi)}{\partial \psi} \right]^t \right\}.$$

Remark 2. The target subset Ω^* is located on the boundary of parameter space, which does not satisfy the standard assumption that the true point is an interior point. In addition, the positive-definiteness of the Fisher information matrix at the truth point does not hold in the whole nonidentifiable subset Ω_0. Condition (R3) states that the quadratic form of the information matrix is positive around Ω^*.

For the mixture of normal distribution, it is not difficult to verify that conditions (R1) and (R2) are satisfied. The following Lemma 1 states that the mixture of normal, including the motivating example in Section 2.1, also satisfies condition (R3). The proof is given in Appendix.

Lemma 1. *Let $f(x, \psi)$ be the mixture of normal defined in (2.7). For any $\psi \notin \Omega_0$ and its closest point $\psi_0^*(\psi)$ in Ω^*,*

$$(\psi - \psi_0^*(\psi))^t I(\psi_0^*(\psi))(\psi - \psi_0^*(\psi)) > 0.$$

Note that any point in Ω_0 can be considered as a *truth point* since they all represent the same true model in (2.6). Given Proposition 1 and Lemma 1, we are able to prove that the penalized maximum likelihood estimate of (2.10) have a root-n convergence rate to the subset $\Omega^* \subset \Omega_0$ that yields the same density as the true density $f_0(x)$ in (2.6). The result is presented in the following theorem.

Theorem 1. *Under conditions (R1)–(R3), there exists a local maximizer $\hat{\psi}$ of $\widetilde{\ell}(\psi)$ in (2.10) such that $D(\hat{\psi}, \Omega^*) = O_p(n^{-1/2})$, i.e., $||\hat{\psi} - \psi_0^*(\hat{\psi})|| = O_p(n^{-1/2})$.*

Note that according to Remark 1, $\psi_0^*(\hat{\psi}) = (\pi_{10}, \mathbf{0}, \theta_{10}, \theta_2(\hat{\psi}))$ for some $\theta_2(\hat{\psi})$ which depends on $\hat{\psi}$, but $\psi_0^*(\hat{\psi})$ produces the true density $f_0(x)$ in (2.6) for any $\hat{\psi}$ since $\psi_0^*(\hat{\psi}) \in \Omega_0$. The similar idea of defining the truth point $\psi_0^*(\hat{\psi})$ has also been used by [31] and [13].

Theorem 1 indicates that a part of $\hat{\psi}$ which forms an M_0-component mixture of normal will converge to the true parameter ψ_0 with a root-n convergence rate. Theorem 1 also indicates that the estimate of mixing proportions on the extra $M - M_0$ components will converge to 0 in probability with a root-n convergence rate, but the estimate of other component parameters, corresponding to the extra non-existing $M - M_0$ components, could be arbitrary. We summarize the results as follows.

Corollary 1. *There exists a partition of $\hat{\psi} = (\hat{\pi}_1, \hat{\pi}_2, \hat{\theta}_1, \hat{\theta}_2)$, such that*

$$||(\hat{\pi}_1, \hat{\pi}_2, \hat{\theta}_1) - (\pi_{10}, \mathbf{0}, \theta_{10})|| = O_p(n^{-1/2}).$$

As explained in Figure 1, the result in Corollary 1 only tells us the estimated component proportions corresponding to those non-existing components will be small (but not exactly 0) and hence the PMLE $\hat{\psi}$ is difficult to be applied directly to choose the number of components in practice. In Section 3, we propose adding a second penalty on mixing proportions to provide a sparse estimation of component proportions resulting in an automatic choice of number of components.

Note that our result does not contradict to the slower minimax rates in [3], [17] and [16]. Taking model (1.1) as an illustration example, the slower minimax rates indicate that with sample size n, we cannot distinguish one component model f from two-component model f_n with very close component means. But if the data is actually generated by f, our estimation procedure ensures a root-n convergence rate. On the other hand, if data is actually generated from a two-component model f_{n_0} with fixed n_0 in (1.1), our estimation could yield a "wrong" one-component model when the sample size n is not significantly larger than n_0. Actually, no estimation procedure works well in this situation, which is the consequence of lower minimax rate. However, when the sample size n is large enough, our estimation will enter into the two-component regime, and the convergence rate is still root-n.

3. Order selection and asymptotic properties

The penalized estimate via $\widetilde{\ell}(\psi)$ does not automatically provide an estimate of the number of components, although the estimated mixing proportions of the redundant $M - M_0$ components are expected to be small. Leveraging the idea of variable selection through penalized regression, we further propose adding a suitable penalty on the mixing proportions to automatically shrink some estimated proportions to 0, thereby resulting an automatic order estimation for the mixture models. The proposed penalized log-likelihood is

$$\ell_p(\psi) = \ell(\psi) + \alpha \sum_{1\le s<k\le M} P(|\eta_{sk}|) - n \sum_{m=1}^{M} p_\lambda(\pi_m), \tag{3.1}$$

where $p_\lambda(\pi_m)$ is a penalty function on mixing proportion π_m. Since $\pi \ge 0$, we will ignore the sign of parameter in $p_\lambda(\cdot)$ and $p'_\lambda(\cdot)$ for simplicity. [18] investigated a special case of (3.1), where $\alpha = 0$ since they did not impose any penalty on the difference of component parameters. The penalties on mixing proportions proposed by [18] have two forms

$$p_{\lambda,\log}(\pi) = \lambda \log\left(\frac{\delta+\pi}{\delta}\right), \tag{3.2}$$

and

$$p_{\lambda,\text{logscad}}(\pi) = \lambda \log\left(\frac{\delta + p_{\lambda,scad}(\pi)}{\delta}\right), \tag{3.3}$$

where $p'_{\lambda,scad}(\pi) = I(\pi \le \lambda) + \frac{(a\lambda-\pi)_+}{(a-1)\lambda} I(\pi > \lambda)$ is proportional to the SCAD penalty function, and δ is a small positive value. Substituting $p_{\lambda,scad}(\pi)$ for π in the second penalty of [18] can avoid excessive penalty for large π. These penalties are quite different from commonly used fold concave penalties such as SCAD or MCP. In this paper, we consider a general form of the penalty on mixing proportions, and investigate conditions for the penalty function that could provide a consistent order selection. The results reveal some new surprising differences among commonly used penalties such as LASSO, SCAD, and MCP, and could be used to guide the choice of penalty for mixing proportions.

We first establish the consistency result of the penalized estimate of (3.1).

Regularity conditions on the penalty on mixing proportions

(R4) $p_\lambda(\pi)$ is second order continuous differentiable for $\pi > 0$ except at a set with 0 measure. In addition, $p'_\lambda(0) := p'_\lambda(0^+)$ and $p_\lambda(0) = 0$.

(R5) $a_n = \max_{1\leq m\leq M_0}\{p'_\lambda(\pi_{m0})\} = O(n^{-1/2})$.

(R6) $b_n = \max_{1\leq m\leq M_0}\{p''_\lambda(\pi_{m0})\} \to 0$ as $n \to 0$.

Remark 3. Conditions (R5)–(R6) are taken from [11]. Condition (R5) guarantees the unbiasedness property for large mixing proportions and the existence of a root-n consistent estimator. Condition (R6) guarantees that the impact of the penalty function on the penalized likelihood estimators is less significant than the likelihood function.

Theorem 2. *Under conditions (R1)–(R6), there exisit a local maximizer $\widetilde{\psi}$ of (3.1) such that $D(\widetilde{\psi}, \Omega^*) = O_p(n^{-1/2})$, i.e., there exists a point $\psi_0^*(\widetilde{\psi}) \in \Omega^*$ such that $||\widetilde{\psi} - \psi_0^*(\widetilde{\psi})|| = O_p(n^{-1/2})$.*

Based on Theorem 2, we can see that the penalized estimate $\widetilde{\psi}$ of (3.1) is also root-n consistent. In Section 3.2, we will establish its oracle properties and order selection consistency for choosing the number of components.

3.1. Computing algorithm and tuning

Next we introduce a computation algorithm and a tuning parameter selection method for the proposed penalized log-likelihood function $\ell_p(\psi)$ in (3.1). We employ an EM gradient algorithm [21] to increase the penalized log-likelihood $\ell_p(\psi)$ after each iteration based on a surrogate function. Note that the EM algorithm can be interpreted as a special case of MM algorithm.

The penalized complete log-likelihood function has the form of

$$\ell_c(\psi) = \sum_{i=1}^{n}\sum_{m=1}^{M}[z_{im}\log\pi_m + z_{im}\log\phi(X_i;\mu_m,\sigma_m^2)] \\ + \alpha\sum_{1\leq s<k\leq M}P_\gamma(|\eta_{sk}|) - n\sum_{m=1}^{M}p_\lambda(\pi_m),$$

where z_{im} are the unobserved indicator variables representing the component-membership of the observation X_i, i.e., $z_{im} = 1$ if X_i belongs to the m-th component and 0, otherwise.

In *E-step*, given the observed data and the current parameter estimation $\psi^{(l)}$, update the conditional expectation of z_{im} given $\psi^{(l)}$ and observations, i.e.,

$$h_{im}^{(l)} = E(z_{im}|\psi^{(l)}) = \frac{\pi_m^{(l)}\phi(X_i;\mu_m^{(l)},\sigma_m^{2(l)})}{\sum_{j=1}^{M}\pi_j^{(l)}\phi(X_i;\mu_j^{(l)},\sigma_j^{2(l)})}. \tag{3.4}$$

In *M-step*, we construct a minorizing function of $\ell_c(\psi)$ in light of the MM-EM relation [22]. Note that $\sum_{i=1}^{n}\sum_{m=1}^{M} h_{im}\log[\pi_m\phi(X_i;\mu_m,\sigma_m^2)]$ is a minorizing function of $\ell(\psi)$ up to a constant. Since $p_\lambda(\cdot)$ is concave,

$$-p_\lambda(\pi_m) \geq -p_\lambda(\pi_m^{(l)}) - p'_\lambda(\pi_m^{(l)})(\pi_m - \pi_m^{(l)})$$

holds for all π_m and the equation holds at $\pi_m^{(l)}$, hence $-p_\lambda(\pi_m^{(l)}) - p'_\lambda(\pi_m^{(l)})(\pi_m - \pi_m^{(l)})$ provides a minorizing function of $-p_\lambda(\pi_m)$.

Based on the convexity of the Euclidean norm $||\cdot||$, we have

$$||x|| \geq ||x_0|| + \frac{(x-x_0)^t x_0}{||x_0||} = \frac{x^t x_0}{||x_0||}.$$

Then, together with the nondecreasing function $P_\gamma(\cdot)$, we have

$$P_\gamma(|\eta_{mj}|) \geq \Big\{\log[(\mu_m^{(l)} - \mu_j^{(l)})(\mu_m - \mu_j) + (\sigma_m^{2(l)} - \sigma_j^{2(l)})(\sigma_m^2 - \sigma_j^2)] \\ - \log(|\eta_{mj}^{(l)}|)\Big\} \times I(|\eta_{mj}^{(l)}| \leq \gamma),$$

which is a minorizing function of the second term of $\ell_p(\psi)$ and has the linear approximation

$$\left\{\frac{(\mu_m^{(l)} - \mu_j^{(l)})}{|\eta_{mj}^{(l)}|^2}(\mu_m - \mu_m^{(l)}) - \frac{(\sigma_m^{2(l)} - \sigma_j^{2(l)})(\sigma_m^{2(l)})^2}{|\eta_{mj}^{(l)}|^2}(\frac{1}{\sigma_m^2} - \frac{1}{\sigma_m^{2(l)}}) - \log(|\eta_{mj}^{(l)}|)\right\} \\ \times I\{|\eta_{mj}^{(l)}| \leq \gamma\}.$$

Hence, we can transfer the maximization of $\ell_p(\psi)$ to the maximization of the surrogate function $Q(\psi|\psi^{(l)})$, in which the first term of $\ell_p(\psi)$ is replaced by $\sum_{i=1}^{n}\sum_{m=1}^{M} h_{im}\log[\pi_m\phi(X_i;\mu_m,\sigma_m^2)]$, the second term of $\ell_p(\psi)$ is replaced by

$$\alpha\sum_{m=1}^{M}\sum_{j\neq m}\left\{\frac{(\mu_m^{(l)} - \mu_j^{(l)})}{|\eta_{mj}^{(l)}|^2}(\mu_m - \mu_m^{(l)}) - \frac{(\sigma_m^{2(l)} - \sigma_j^{2(l)})(\sigma_m^{2(l)})^2}{|\eta_{mj}^{(l)}|^2}(\frac{1}{\sigma_m^2} - \frac{1}{\sigma_m^{2(l)}})\right\} \times \\ I\{|\eta_{mj}^{(l)}| \leq \gamma\},$$

and the third term term of $\ell_p(\psi)$ is replaced by $n\sum_{m=1}^{M} p'_\lambda(\pi_m^{(l)})(\pi_m - \pi_m^{(l)})$. The update of π_m is then

$$\pi_m^{(l+1)} = \frac{\sum\limits_{i=1}^{n} h_{im}^{(l)}}{n - n\sum\limits_{j=1}^{M}\pi_j^{(l)} p'_\lambda(\pi_j^{(l)}) + np'_\lambda(\pi_m^{(l)})}. \tag{3.5}$$

The updated estimation of μ_m and σ_m^2 are

$$\mu_m^{(l+1)} = \frac{\sum\limits_{i=1}^{n} h_{im}^{(l)} X_i + \alpha\sigma_m^{2(l)} d_1^{(l)}}{\sum\limits_{i=1}^{n} h_{im}^{(l)}}, \tag{3.6}$$

and

$$\sigma_m^{2(l+1)} = \frac{\sum\limits_{i=1}^{n} h_{im}^{(l)}(X_i - \mu_m^{(l)})^2 + 2\alpha d_2^{(l)}}{\sum\limits_{i=1}^{n} h_{im}^{(l)}}, \tag{3.7}$$

respectively, where $d_1^{(l)} = \sum_{j \neq m}(\mu_m^{(l)} - \mu_j^{(l)}) I\{|\eta_{mj}^{(l)}| \leq \gamma\}/|\eta_{mj}^{(l)}|^2$, and $d_2^{(l)} = \sum_{j \neq m}(\sigma_m^{2(l)} - \sigma_j^{2(l)})(\sigma_m^{2(l)})^2 I\{|\eta_{mj}^{(l)}| \leq \gamma\}/|\eta_{mj}^{(l)}|^2$.

In M step, $\pi_m^{(l)}$ will be set to zero when it is small enough, e.g., less than 10^{-6}, and then the corresponding component can be eliminated. This algorithm enables us to simultaneously estimate the unknown parameters and the number of components. The above derivation also indicates that the proposed algorithm poses the desired ascent property via a standard argument of MM algorithm.

Proposition 2. *The updating parameter sequence $\psi^{(l)}(l = 1, 2, \cdots)$, based on the EM iteration of (3.4)–(3.7), nondecrease the penalized log-likelihood function $\ell_p(\psi)$ in (3.1), i.e.,*

$$\ell_p(\psi^{(l+1)}) \geq \ell_p(\psi^{(l)}), \quad l = 1, 2, \ldots.$$

Selection of Tuning Parameters

For model (3.1), the selection of tuning parameters α,γ and λ in $\ell_p(\psi)$ can be conducted by the BIC approach:

$$BIC = \sum_{i=1}^{n} \log[\sum_{m=1}^{\widetilde{M}} \widetilde{\pi}_m \phi(X_i; \widetilde{\mu}_m, \widetilde{\sigma}_m^2)] - \frac{1}{2}\widetilde{D}_f \log(n), \tag{3.8}$$

and $\widetilde{D}_f$ is the number of free parameters corresponding to components with positive mixing proportions in the estimated model. For example, if the estimated model has k positive mixing proportions, then the free parameters are $(\mu_1, \ldots, \mu_k, \sigma_1, \ldots, \sigma_k, \pi_1, \ldots, \pi_{k-1})$ with $\widetilde{D}_f = 3k - 1$. But this approach could be computationally expensive. Based on our empirical experience, the estimation is not very sensitive to the choice of γ and a relative large value of γ between 5 and 10 usually works well. Hence, we suggest using a fixed large value of γ, say 5, and then selecting the other two tuning parameters by the above BIC criterion. For model (2.10), we can simply apply the BIC criterion with a 2-dimensional grid search.

3.2. Oracle properties

In this section, we establish the order selection consistency for the number of components and asymptotic normality for the penalized estimates. For each $\psi = (\pi_1, \pi_2, \theta_1, \theta_2)$, we assume it is permuted such that ψ is closest to the truth point $\psi_0^*(\psi) \in \Omega^*$ among all possible component permutations of ψ, where $\psi_0^*(\psi) = (\pi_{10}, \mathbf{0}, \theta_{10}, \theta_2)$. Hence, θ_1 contains the parameters of the first M_0 components, and θ_2 contains the parameters of the last $M - M_0$ components.

Theorem 3. *Under conditions (R1)–(R6), there exists a maximizer $\widetilde{\psi}$ of (3.1) with $\widetilde{\pi} = (\widetilde{\pi}_1, \widetilde{\pi}_2, \cdots, \widetilde{\pi}_M)^t$ satisfying the following results.*

a. Let $\widetilde{M}$ be the estimated number of components corresponding to positive estimated mixing proportions. Given the condition that for some small $\epsilon > 0$

$$p'_\lambda(\pi_l) - \sum_{m=1}^{M_0} \pi_{m0} p'_\lambda(\pi_{m0}) > 0, \tag{3.9}$$

when $\pi_l < \epsilon, l = M_0+1, \cdots, M$, we have $\widetilde{M} \to M_0$ with probability tending to one, and $P(\widetilde{\pi}_l = 0) \to 1, \quad l = M_0+1, \cdots, M$.

b. Asymptotic normality:

$$\sqrt{n}(I(\psi_0) + \Sigma)\{(\widetilde{\pi}_1, \widetilde{\theta}_1) - \psi_0 + (I(\psi_0) + \Sigma)^{-1}\mathbf{b}\} \to N(\mathbf{0}, I(\psi_0)),$$

where $I(\psi_0)$ is the Fisher information matrix of model (2.6),

$$\Sigma = diag\{p''_\lambda(\pi_{10}), \cdots, p''_\lambda(\pi_{M_0,0}), \underbrace{0, \cdots, 0}_{2M_0}\},$$

and

$$\mathbf{b} = (p'_\lambda(\pi_{10}), \cdots, p'_\lambda(\pi_{M_0,0}), \underbrace{0, \cdots, 0}_{2M_0})^t.$$

Theorem 3 states that we can consistently estimate the number of components via the proposed penalized likelihood method, and the estimator is as efficient as if the true number of component were known. It can be verified that folded concave penalties [12], including SCAD and MCP, satisfy condition (3.9), and hence can be employed as an order selection penalty when the sample size is sufficiently large.

However, the Lasso penalty does not satisfy the condition (3.9). Lasso proposed by [34] has the form of $p'_{\lambda,lasso}(\pi) = \lambda$. It can be shown that $p'_{\lambda,lasso}(\pi_l) - \sum_{m=1}^{M_0} \pi_m p'_{\lambda,lasso}(\pi_m) = 0$, which violates the condition (3.9). Our numerical studies also confirm that Lasso dose not work well for all sample sizes considered, even for large sample size such as 2,000 and 20,000.

Based on our empirical experience, for finite sample performance, the order selection/sparsity recovery will not work well if $p'_\lambda(\pi_l) - \sum_{m=1}^{M_0} \pi_{m0} p'_\lambda(\pi_{m0})$ is close to 0 for π_l in a neighborhood region of 0. Therefore, a practical guide is to choose the penalty $p_\lambda(\pi)$ such that $p'_\lambda(\pi)$ decreases fast when π moves away from 0.

For SCAD, the continuous differential function of SCAD penalty is

$$p'_{\lambda,scad}(\pi) = \lambda\{I(\pi \le \lambda) + \frac{(a\lambda - \pi)_+}{(a-1)\lambda} I(\pi > \lambda)\}, \tag{3.10}$$

where $a > 2$ controls the concavity of the penalty function, and $\lambda > 0$ is the maximum value of derivative. Notice that similar to Lasso, the SCAD derivative is λ when $\pi \leq \lambda$, and linearly decreases to zero as π increases. When λ is small, $p'_\lambda(\pi_l) - \sum_{m=1}^{M_0} \pi_{m0} p'_\lambda(\pi_{m0}) < \lambda$ will be also small and thus close to 0. When λ is large such that $\pi_{m0} \leq \lambda$ or $a\lambda$, $p'_\lambda(\pi_l) - \sum_{m=1}^{M_0} \pi_{m0} p'_\lambda(\pi_{m0})$ will be also close to 0. Therefore either when λ is small or large $p'_\lambda(\pi_l) - \sum_{m=1}^{M_0} \pi_{m0} p'_\lambda(\pi_{m0})$ will be close to 0, which makes it difficult for SCAD to choose an appropriate λ if it exists, and explains the unsatisfactory finite sample performance of SCAD in our numerical studies.

[18] proposed two log type penalties given in (3.2) and (3.3), respectively, which have steeper first derivatives than SCAD. The first derivatives are given by

$$p'_{\lambda,\log}(\pi) = \lambda \cdot \frac{1}{\delta + \pi}, \tag{3.11}$$

and

$$p'_{\lambda,logscad}(\pi) = \lambda \cdot \frac{p'_{\lambda,scad}(\pi)}{\delta + p_{\lambda,scad}(\pi)}, \tag{3.12}$$

where δ is a very small positive number, say 10^{-6} or $o(n^{-\frac{1}{2}} \log^{-1} n)$. Compared with SCAD, $p'_{\lambda,\log}(\pi)$ and $p'_{\lambda,logscad}(\pi)$ are strictly decreasing functions of π, and decrease very fast when π is small so that $p'_\lambda(\pi_l) - \sum_{m=1}^{M_0} \pi_{m0} p'_\lambda(\pi_{m0})$ is not close to 0.

It is worth pointing out that the penalty functions, such as MCP [37], with two flexible tuning parameters to control the strength and concavity, respectively, also have good performance on sparse estimation of mixing proportions in finite sample. The MCP has the form

$$p'_{\lambda,mcp}(\pi) = (\lambda - \frac{\pi}{a})_+, \tag{3.13}$$

where $a > 0$. Unlike SCAD, when a is a small value, $p'_{\lambda,mcp}(\pi)$ decreases fast as π increases.

The truncated L_1 penalty [32, TLP], which approximates the L_0 penalty, also performs well on sparse estimation with

$$p'_{\lambda,tlp}(\pi) = \frac{\lambda}{\tau} I(\pi < \tau), \tag{3.14}$$

where $\tau > 0$ is a tuning parameter controlling the degree of approximation. Note that $p'_{\lambda,tlp}(\pi) = 0$ when $\pi \geq \tau$. Therefore, a small τ can be selected such that $\tau < \pi_{m0}$ for some or all m' s so that $p'_{\lambda,tlp}(\pi_l)$ is large when π_l is close to 0 and $\sum_{m=1}^{M_0} \pi_{m0} p'_{\lambda,tlp}(\pi_{m0})$ is much smaller than $p'_{\lambda,tlp}(\pi_l)$.

4. Simulation and application

4.1. Simulation

In this section, we conduct simulations to demonstrate the finite sample performance of the penalized likelihood methods (2.10) and (3.1).

Example 1. We generate observations from a three-component normal mixture model with parameters

$$(\mu_1, \sigma_1, \pi_1) = (-4, 1, 1/3),\ (\mu_2, \sigma_2, \pi_2) = (0, 1.2, 1/3), (\mu_3, \sigma_3, \pi_3) = (4, 1, 1/3).$$

To explore the performance of the penalized estimation (2.10), our studies are based on scenarios with different sample sizes n, with 500 replicates for each scenario. The maximum number of components is set to be $M = 10$. We first obtain initial values by a K-means algorithm. Given the results of K-means clustering, the initial mean, variance and proportion of each component are estimated by the sample mean, sample variance, and the sample mixing proportion, respectively. Table 1 summarizes the mean of parameter estimates for the components of the *three* largest estimated proportions. The values in brackets are the corresponding standard errors of the estimate. We perform 2-dimensional grid search for parameters α and γ, where the grid values are $(1, 3, 5, 7, 14, 18)$ for α, and $(1, 5, 10)$ for γ. Noted that a good choice of γ depends on the scale of the data, and the searching range of tuning parameters should be adjusted according the data scale. Under the BIC criterion, the $\gamma = 5$ is selected, while the selected value of α varies. It can be seen that the sum of three largest component proportions is greater than 0.9 with suitable choices of α and γ. In addition, a larger value of γ may yield better estimation of proportions, while causing larger bias for the estimated location parameters.

Table 1

Parameter estimates with the estimation (2.10) for Example 1.

Sample size	Threshold	μ_1 σ_1 π_1	μ_2 σ_2 π_2	μ_3 σ_3 π_3
n=200	$\gamma = 1$	-3.771 0.604 0.195	0.055 0.680 0.175	3.826 0.593 0.192
		(0.968 0.211 0.049)	(1.376 0.281 0.039)	(0.854 0.191 0.047)
	$\gamma = 5$	-3.946 0.996 0.298	0.020 1.245 0.288	3.928 1.014 0.292
		(0.469 0.300 0.063)	(0.666 0.407 0.073)	(0.395 0.312 0.061)
	$\gamma = 10$	-3.966 1.006 0.295	0.043 1.541 0.328	3.921 1.064 0.294
		(0.511 0.369 0.067)	(0.799 0.520 0.097)	(0.518 0.423 0.075)
n=400	$\gamma = 1$	-3.882 0.642 0.192	0.033 0.687 0.170	3.880 0.636 0.192
		(0.818 0.178 0.041)	(1.393 0.210 0.034)	(0.836 0.184 0.044)
	$\gamma = 5$	-4.025 0.984 0.297	-0.014 1.291 0.299	4.010 0.963 0.298
		(0.308 0.217 0.046)	(0.501 0.292 0.061)	(0.301 0.197 0.045)
	$\gamma = 10$	-4.069 0.967 0.293	-0.034 1.585 0.348	4.078 0.961 0.298
		(0.330 0.219 0.045)	(0.566 0.354 0.070)	(0.273 0.204 0.047)

Given the threshold $\gamma = 1, 5$, and 10, we further implement the penalized estimation (3.1) using the penalty function log-scad, TLP, MCP, SCAD, and Lasso with 500 replicates. Redundant component will be removed if the corresponding estimated mixing proportion is below the pre-determined threshold 10^{-6}. Table 2 shows the accuracy of the order selection, with $*$ representing the results of $\gamma = 5$ that is chosen by BIC. Note that TLP and log-scad penalties satisfy the condition in Theorem 3 and perform well for both $n = 200$ and 400, as expected. However, SCAD and Lasso penalty do not work well which is consistent with the discussions after Theorem 3. Based on Theorem 3 and the discussions followed, Lasso does not satisfy the sparsity condition for the consistent order selection no matter how large the sample size is, and SCAD does not have good finite sample performance. The result of MCP is somewhat unsatisfactory and sensitive to the choice of γ. When the number of components is correctly selected by the TLP penalty, we present the mean and standard deviation of the simulation results in Table 3. It can be seen in Table 3 that the performance is satisfactory and the penalized estimation (3.1) is insensitive to the choice of γ, unlike the estimation (2.10) reported in Table 1.

To better compare the performances of order selection, we further conduct the method (3.1) in the case of $n = 2000, 20000$ and $\gamma = 5$. Figure 2 presents the histograms of the estimated numbers of components for these five penalties. The log-type penalty and TLP give good results as expected. When the sample size increases, SCAD shows some improvement in the order selection, although still not very satisfactory. However, Lasso shows no improvement when the sample size increases.

Table 2

Accuracy of the order selection for Example 1.

Penalty	Threshold	n=200			n=400		
		Underfitted	Correct	Overfitted	Underfitted	Correct	Overfitted
log-scad	$\gamma = 1$	0.020	0.980	0.000	0.000	0.996	0.004
	$\gamma = 5$	0.014*	0.984*	0.002*	0.000*	0.992*	0.008*
	$\gamma = 10$	0.014	0.984	0.002	0.000	0.992	0.008
TLP	$\gamma = 1$	0.032	0.964	0.004	0.004	0.982	0.014
	$\gamma = 5$	0.020*	0.976*	0.004*	0.004*	0.992*	0.004*
	$\gamma = 10$	0.020	0.976	0.004	0.004	0.984	0.012
MCP	$\gamma = 1$	0.118	0.556	0.326	0.008	0.702	0.290
	$\gamma = 5$	0.044*	0.788*	0.168*	0.006*	0.852*	0.142*
	$\gamma = 10$	0.028	0.876	0.096	0.006	0.920	0.074
SCAD	$\gamma = 1$	0.000	0.000	1.000	0.000	0.000	1.000
	$\gamma = 5$	0.000*	0.000*	1.000*	0.000*	0.000*	1.000*
	$\gamma = 10$	0.000	0.060	0.940	0.000	0.012	0.988
Lasso	$\gamma = 1$	0.000	0.000	1.000	0.000	0.000	1.000
	$\gamma = 5$	0.000*	0.000*	1.000*	0.000*	0.000*	1.000*
	$\gamma = 10$	0.000	0.000	1.000	0.000	0.000	1.000

Example 2. We consider a more complicated case where some mixture components overlap, with two of the components having the same mean but different variances. We generate observations from a four-component normal mixture model with

$$(\mu_1, \sigma_1, \pi_1) = (-4, 0.1, 0.2),\ (\mu_2, \sigma_2, \pi_2) = (-4, 1.2, 0.3),\ (\mu_3, \sigma_3, \pi_3) = (0, 1, 0.2),$$

and $(\mu_4, \sigma_4, \pi_4) = (4, 1, 0.3)$. Methods for obtaining initial values and tuning parameters are the same as in Example 1.

Table 4 shows the mean and standard deviation of the parameter estimates for the components corresponding to the four largest estimated proportions, based on estimation method (2.10) and 500 replicates. According to the BIC, the threshold $\gamma = 5$ is selected. Similar to Example 1, when γ and α increase, the mixing proportions of the four largest component proportions increase and the rest six component proportions are shrunk to 0.

TABLE 3
Parameter estimates based on the method (3.1) for Example 1.

Sample size	Threshold	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3
n=200	$\gamma = 1$	-4.017	0.983	0.331	0.014	1.301	0.349	4.008	0.981	0.320
		(0.192	0.146	0.047)	(0.260	0.384	0.071)	(0.207	0.140	0.046)
	$\gamma = 5$	-4.015	0.985	0.331	0.012	1.291	0.348	4.007	0.983	0.321
		(0.191	0.148	0.046)	(0.249	0.323	0.070)	(0.206	0.139	0.046)
	$\gamma = 10$	-4.016	0.984	0.330	0.011	1.306	0.350	4.007	0.981	0.320
		(0.191	0.145	0.048)	(0.251	0.325	0.072)	(0.206	0.140	0.046)
n=400	$\gamma = 1$	-3.996	0.990	0.332	-0.000	1.240	0.337	3.997	1.002	0.331
		(0.135	0.078	0.035)	(0.177	0.206	0.047)	(0.138	0.102	0.031)
	$\gamma = 5$	-3.996	0.990	0.332	-0.000	1.237	0.337	3.997	1.002	0.331
		(0.135	0.077	0.035)	(0.177	0.203	0.047)	(0.137	0.103	0.031)
	$\gamma = 10$	-3.997	0.989	0.331	-0.002	1.246	0.338	3.998	1.002	0.331
		(0.136	0.078	0.036)	(0.181	0.205	0.048)	(0.138	0.101	0.031)

TABLE 4
Parameter estimates based on the method (2.10) for Example 2.

n	Threshold	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3	μ_4	σ_4	π_4
n=200	$\gamma = 1$	-3.983	0.113	0.221	-3.895	0.932	0.210	0.476	0.623	0.131	4.042	0.588	0.169
		(0.263	0.065	0.042)	(0.919	0.220	0.066)	(1.877	0.266	0.030)	(0.709	0.208	0.046)
	$\gamma = 5$	-4.016	0.110	0.220	-3.716	1.472	0.291	0.288	0.961	0.163	4.062	0.971	0.257
		(0.194	0.034	0.046)	(0.654	0.363	0.076)	(0.935	0.504	0.066)	(0.408	0.307	0.059)
	$\gamma = 10$	-3.965	0.117	0.222	-3.408	1.747	0.319	0.508	1.045	0.157	4.082	0.971	0.256
		(0.451	0.074	0.052)	(0.770	0.517	0.087)	(1.194	0.737	0.094)	(0.418	0.366	0.067)
n=400	$\gamma = 1$	-4.000	0.105	0.205	-4.061	0.985	0.218	0.041	0.717	0.135	4.044	0.622	0.171
		(0.018	0.063	0.026)	(0.501	0.150	0.056)	(1.821	0.265	0.031)	(0.667	0.180	0.043)
	$\gamma = 5$	-4.001	0.105	0.211	-3.796	1.472	0.300	0.270	1.023	0.172	4.072	0.931	0.266
		(0.014	0.016	0.023)	(0.346	0.173	0.045)	(0.637	0.371	0.040)	(0.293	0.209	0.050)
	$\gamma = 10$	-4.002	0.112	0.220	-3.597	1.706	0.305	0.430	1.171	0.171	4.101	0.945	0.268
		(0.015	0.027	0.026)	(0.448	0.229	0.054)	(0.689	0.571	0.055)	(0.299	0.227	0.052)

TABLE 5
Accuracy of order selection for Example 2.

Penalty	Threshold	n=200			n=400		
		Underfitted	Correct	Overfitted	Underfitted	Correct	Overfitted
log-scad	$\gamma = 1$	0.204	0.780	0.016	0.020	0.972	0.008
	$\gamma = 5$	0.212*	0.764*	0.024*	0.012*	0.984*	0.004*
	$\gamma = 10$	0.228	0.756	0.016	0.012	0.984	0.004
TLP	$\gamma = 1$	0.064	0.752	0.184	0.020	0.892	0.088
	$\gamma = 5$	0.052*	0.832*	0.116*	0.020*	0.940*	0.040*
	$\gamma = 10$	0.112	0.804	0.084	0.024	0.944	0.032
MCP	$\gamma = 1$	0.004	0.024	0.972	0.004	0.024	0.972
	$\gamma = 5$	0.028*	0.244*	0.728*	0.048*	0.324*	0.628*
	$\gamma = 10$	0.096	0.340	0.564	0.092	0.424	0.484
SCAD	$\gamma = 1$	0.000	0.000	1.000	0.000	0.000	1.000
	$\gamma = 5$	0.000*	0.004*	0.996*	0.000*	0.000*	1.000*
	$\gamma = 10$	0.000	0.016	0.984	0.000	0.004	0.996
Lasso	$\gamma = 1$	0.000	0.000	1.000	0.000	0.000	1.000
	$\gamma = 5$	0.000*	0.000*	1.000*	0.000*	0.000*	1.000*
	$\gamma = 10$	0.000	0.000	1.000	0.000	0.000	1.000

Table 5 presents the order selection results of the method (3.1) for different sample sizes and penalties. The performances of log-scad, TLP and MCP penalties improve as the sample size increases. The results of $\gamma = 5$ (selected by BIC)

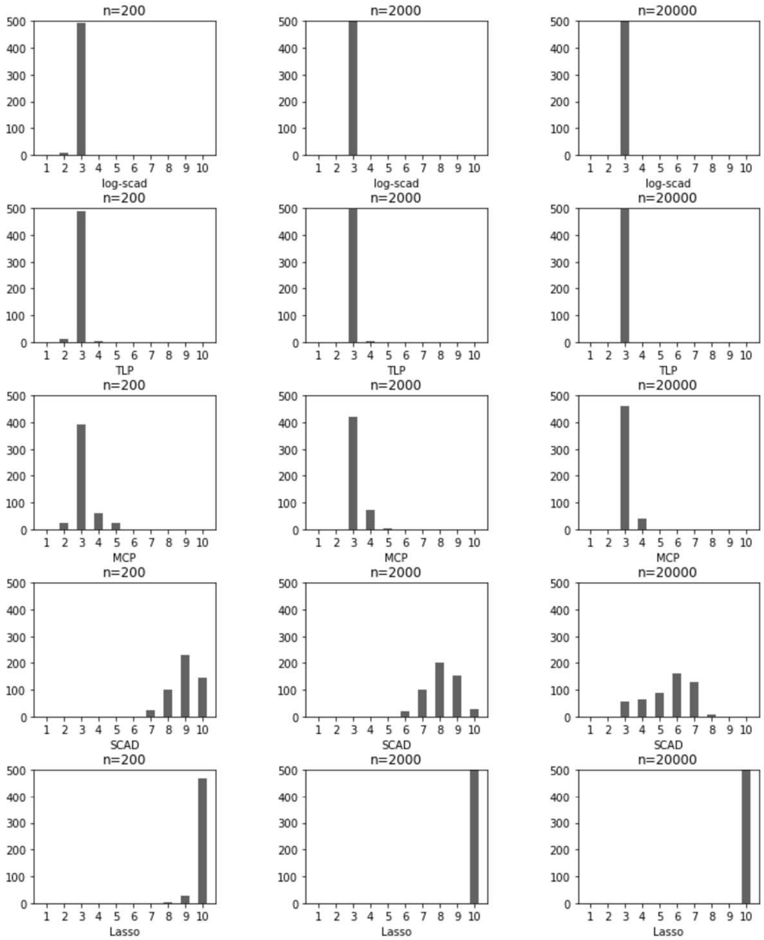

FIG 2. *The histograms of the estimated number of components for Example 1 with $\gamma = 5$.*

TABLE 6
Parameter estimates based on the method (3.1) for Example 2.

n	Threshold	μ_1	σ_1	π_1	μ_2	σ_2	π_2	μ_3	σ_3	π_3	μ_4	σ_4	π_4
n=200	$\gamma = 1$	-4.001	0.101	0.204	-4.002	1.159	0.291	-0.059	1.018	0.210	4.000	0.987	0.293
		(0.020	0.054	0.037)	(0.362	0.229	0.049)	(0.317	0.389	0.057)	(0.187	0.155	0.041)
	$\gamma = 5$	-4.001	0.099	0.205	-3.999	1.186	0.292	-0.027	1.014	0.205	3.998	0.989	0.296
		(0.019	0.018	0.034)	(0.252	0.233	0.049)	(0.321	0.405	0.056)	(0.179	0.144	0.039)
	$\gamma = 10$	-4.001	0.099	0.204	-3.994	1.173	0.293	-0.001	1.008	0.205	4.005	0.980	0.294
		(0.020	0.018	0.034)	(0.278	0.283	0.051)	(0.318	0.397	0.055)	(0.179	0.143	0.040)
n=400	$\gamma = 1$	-4.000	0.100	0.203	-4.014	1.174	0.295	-0.018	1.011	0.202	4.002	0.998	0.300
		(0.014	0.012	0.022)	(0.160	0.125	0.032)	(0.192	0.297	0.036)	(0.124	0.098	0.027)
	$\gamma = 5$	-4.000	0.100	0.203	-3.998	1.195	0.297	0.000	1.001	0.200	4.002	0.998	0.300
		(0.014	0.013	0.022)	(0.172	0.137	0.031)	(0.182	0.287	0.035)	(0.123	0.100	0.026)
	$\gamma = 10$	-4.000	0.100	0.204	-3.997	1.188	0.295	0.004	1.016	0.201	4.006	0.996	0.299
		(0.014	0.013	0.022)	(0.182	0.184	0.033)	(0.204	0.331	0.039)	(0.123	0.101	0.028)

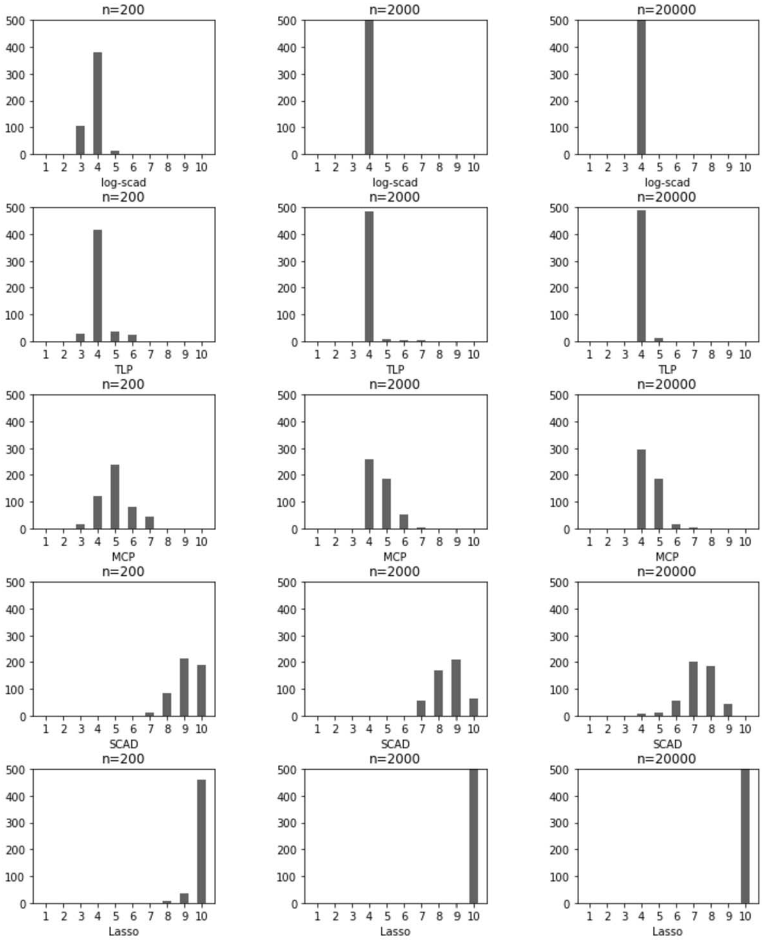

Fig 3. *The histograms of the estimated number of components for Example 2 with $\gamma = 5$.*

are marked by $*$. In addition, MCP is undesirable in detecting the true order while log-scad and TLP provide better results.

Table 6 presents the parameter estimates with the penalty TLP, which demonstrates that the performance of the penalized estimate is satisfactory and insensitive to the choice of γ. To further investigate the performance of SCAD and Lasso, Figure 3 presents the histograms of the estimated number of components for $\gamma = 5$. As the sample size increases, SCAD shows some improvement of order selection, while the Lasso does not show any improvement similar to what we have observed in Example 1.

In terms of order selection, we further compare our method with the traditional BIC method, and the penalized method in [18], which corresponds to (3.1) with $\alpha = 0$. The results of the traditional BIC and [18]'s method (under log-scad and TLP) are presented in Table 7 for both settings of Example 1 and Example 2. Together with the results in Table 2 and Table 5, we can see that BIC only works well in Example 1 with sample size $n = 400$. For Example 2 with more-overlapped component setting, both our method and [18]'s method outperform BIC significantly. Compared with [18]'s method, including a further penalty on component difference enables our method to provide better accuracy in order selection, especially in the case of $n = 200$.

TABLE 7
Order selection accuracy based on BIC and [18] for Examples 1 and 2

	Method	n=200			n=400		
		Underfitted	Correct	Overfitted	Underfitted	Correct	Overfitted
Example 1	BIC	0.054	0.906	0.040	0.000	0.992	0.008
	log-scad	0.034	0.960	0.006	0.000	0.996	0.004
	TLP	0.020	0.886	0.094	0.000	0.948	0.052
Example 2	BIC	0.152	0.234	0.614	0.018	0.274	0.708
	log-scad	0.250	0.728	0.022	0.022	0.966	0.012
	TLP	0.070	0.682	0.248	0.002	0.736	0.262

4.2. Application

We apply our method to the Bean plants data investigated in [28] and [5]. Bean is an autogamous species whose crop (grains or pods) needs fertilization. For the sake of full production in fields, an F1 hybrid, whose female parent is cytoplasmic male sterile, requires nuclear fertility restoration genes-preferably dominant ones-from its male parent. Hence, it is significant to find nuclear genes that induce fertility restoration. [28] conducted an experiment where 150 F2 bean plants were obtained by self-crossing the eight F1 plants, and suggested the analysis made on the square root of the total number of grains per plant. To test whether there exists a major restoration gene in an F2 population, [28] considered a three-component normal mixture model with known mixing proportions and equal variance for the distribution of number of grains per plant,

$$\frac{1}{4}\phi(x;\mu_1,\sigma^2)+\frac{1}{2}\phi(x;\mu_2,\sigma^2)+\frac{1}{4}\phi(x;\mu_3,\sigma^2). \tag{4.1}$$

The requirement of a three-component normal mixture for the data implies the existence of a major restoration gene. They used the LRT to test the null hypothesis $\mu_1 = \mu_2 = \mu_3$ and the resulting p-value is 0.002%. [5] noted that this result may be biased since the simulated null rejection rates are larger than the nominal ones systematically. [5] then carried out the EM-test under a two-component normal mixture model with the equal variance assumption, and found the corresponding p-value is 1.0%. In addition, due to the lack of suitability of the equal variance assumption, they conducted the EM-test with an unequal assumption, resulting in a p-value of around 0.003%, and claimed tha a two-component mixture model with unequal variance can fit the data just as well as the model (4.1) with equal variance suggested by [28].

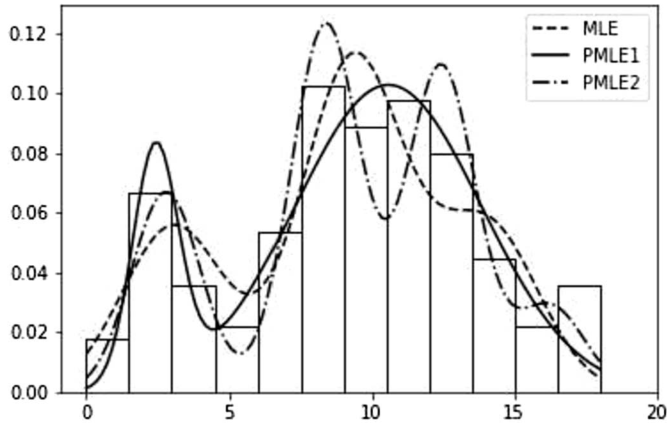

Fig 4. *The histograms of the square root of the total number of grains per plant, the fitted density curve of model (3.1) and (4.1).*

We perform our method under the assumptions of both equal variance and unequal variance settings with $M = 10$. The initial values are obtained by K-means algorithm. The penalized likelihood method selects a four-component normal mixture model with BIC $= -429.7721$ under the equal variance setting, and selects a two-component normal mixture model with BIC $= -428.4997$ under the unequal variance setting. Figure 4 presents the estimated density curve of model (3.1) and (4.1), where PMLE1 represents the fitted curve of unequal variance, PMLE2 represents the fitted curve of equal variance, and MLE represents the fitted curve based on the model (4.1). Our results indicate that a two-component mixture model with unequal variance can fit the data well, which agrees with the results of [5], and provides further justifications for the EM test in the analysis of Bean plants data.

5. Conclusion and discussion

The minimax convergence rate and the pointwise convergence rate are quite different for mixture models with over-fitted components. In this paper, we focus on how to efficiently achieve the best possible root-n pointwise rate, and perform consistent order selection and statistical inference for finite mixture models with unknown number of components. We show that such an optimal convergence rate can be obtained by penalties on the difference between pairs of component parameters for finite normal mixture models. The proposed penalized method is motivated by our new concept of target subset, which ensures the elimination of unnecessary extra components. With unknown number of components but upper bounds for finite normal mixture models, we show that there exists certain parameter space (non-degenerated space) where the over-fitted mixture log-likelihood can still have a regular quadratic Taylor expansion, and thus a root-n consistent estimator is achievable. Our result also indicates that the root-n pointwise convergence rate can be achieved without a first step of order selection or imposing a penalty on mixing proportions. Note that our method could suffer from the issue of unbounded maximal risk [23], e.g., in the case that some proportion parameters being decreased at the $1/\sqrt{n}$ rate as the sample size n increases.

We further propose an order selection method with an additional penalty on the mixing proportions. We show that the proposed order selection method is consistent, and establish the asymptotic normality of the proposed parameter estimators. Noted that [18] provides a $\sqrt{n}$-consistent estimator which is also model selection consistent, without using a first penalty to bound the distance between component parameters. By introducing the first penalty in our work, together with proposition 1, our new method simplifies the proofs so that they are more in line with traditional proofs of the $\sqrt{n}$-consistency of MLE. In addition, based on our numerical studies, including a further penalty on component differences also improves the accuracy in order selection. We provide conditions required for the consistent order selection. These conditions reveal why the performance of some popularly used penalty functions, such as Lasso and SCAD, provide unsatisfactory results in the order selection, while others could be useful, such as TLP and log-type penalties. It will be our future interest to investigate the testing problem under the penalized setting. We expect that the testing distribution follows the traditional chi-square distribution under the proposed penalized method if we can force the estimation in the non-degenerated parameter space. In addition, the extension to multivariate Gaussian mixture deserves further studies, in which the distance computation of component covariance matrices and EM algorithm could be more complicated.

It is of interest to discussion the duality between our methodology and that of [4] and [29]. In our approach, a penalty is introduced to ensure that mixture components are bounded away from each other, and redundant components are eliminated based on the sparse mixing proportions. In the latter approach, a penalty is used to ensure that mixing proportions are bounded away from zero and nearby mixture components are merged. Both approaches lead to consistent parameter estimators, and consistent model selection. In the motivating example, it can be shown that quadratic term of (2.3) is non-degenerate at $(0, \mu_2)$ for μ_2 bounded away from μ_1^0, which is essential for our methedology. It is also non-degenerate at (π, μ_1^0) when π is bounded away from 0, this could be viewed as a support for the latter approaches, that consistent estimation and consistent model selection are reasonable.

Appendix: Proofs

Proof of Proposition 1 We assume that σ_m lies in a compact subset of $(0, \infty)$ for all $m = 1, \ldots, M$, and the whole parameter space of ψ, denoted by Ω, is also compact.

First, we prove that

$$\frac{1}{n}\sum_{i=1}^{n}\log[f(X_i;\psi)/f(X_i;\bar{\psi})] < 0 \tag{A.1}$$

almost surely for any $\psi \notin \Omega_0$ and any $\bar{\boldsymbol{\psi}} \in \Omega^*$. By Jensen's inequality, we have

$$E_{\bar{\psi}}\log[f(x;\psi)/f(x;\bar{\psi})] < \log E_{\bar{\psi}}[f(x;\psi)/f(x;\bar{\psi})] = 0$$

According to the law of large number, there exist some $C > 0$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\log[f(X_i;\psi)/f(X_i;\bar{\psi})] < -C,$$

almost surely for any $\psi \notin \Omega_0$. Because the space of ψ is compact, we have

$$\sup_{\psi\in\Omega_c}\sum_{i=1}^{n}\log[f(X_i;\psi)/f(X_i;\bar{\psi})] < -Cn,$$

for any compact set $\Omega_c \subset \Omega$ that is disjoint of Ω_0.

Because the penalty function $P(|\eta_{sk}|)$ is bounded above (for the truncated log function with a threshold $\gamma > 0$, the upbound is $\log(\gamma)$),

$$\sup_{\psi\in\Omega_c}\widetilde{\ell}(\psi) - \widetilde{\ell}(\bar{\psi}) \le -Cn.$$

Therefore, the penalized maximum log-likelihood estimate $\hat{\psi}$ must satisfy $D(\hat{\psi},\Omega_0) = o_p(1)$, where the distance measure D is defined in (2.11).

Since each component of the true mixture density function is distinct, the penalty term $P(|\eta_{sk}|)$ is constant and hence finite when evaluated at $\bar{\psi}$. Let $\widetilde{\psi}$ be the ordinary MLE of ψ that maximizes $\sum_{i=1}^{n}\log[f(X_i;\psi)]$. Since the parameter space is assumed to be compact, based on the proof of Theorem 2.1 of [15],

$$\sup_{\psi\in\Omega}\left\{\frac{1}{n}\sum_{i=1}^{n}s_\psi(X_i)\right\}^2 = O_p(1) \text{ and } \lim_{n\to\infty}\inf_{\psi\in\Omega}\frac{1}{n}\sum_{i=1}^{n}\{s_\psi^-(X_i)\}^2 > 0.$$

Therefore, based on the Inequality 1.2 of [15],

$$\sum_{i=1}^{n}\log[f(X_i;\widetilde{\psi})/f(X_i;\bar{\psi})] \le \frac{1}{2}\sup_{\psi\in\Omega}\frac{\{\sum_{i=1}^{n}s_\psi(X_i)\}^2}{\sum_{i=1}^{n}\{s_\psi^-(X_i)\}^2} = O_p(1),$$

where

$$s_\psi(x) = \frac{f(x;\psi)/f(x;\bar{\psi}) - 1}{||f(x;\psi)/f(x;\bar{\psi}) - 1||_2},$$

$s_\psi^-(x) = \min\{0, s_\psi(x)\}$, and $\|\cdot\|_2$ is the L^2 normal.

Based on the definition of penalized maximum log-likelihood estimate (PMLE), we have

$$
\begin{aligned}
0 \leq \widetilde{\ell}(\hat{\psi}) - \widetilde{\ell}(\bar{\psi}) = & \sum_{i=1}^{n} \log[f(X_i;\hat{\psi})/f(X_i;\bar{\psi})] \\
& + \alpha \sum_{1\leq s<k\leq M} P(|\hat{\eta}_{sk}|) - \alpha \sum_{1\leq s<k\leq M} P(|\eta_{sk0}|) \\
\leq & \sum_{i=1}^{n} \log[f(X_i;\widetilde{\psi})/f(X_i;\bar{\psi})] + \alpha \sum_{1\leq s<k\leq M} P(|\hat{\eta}_{sk}|) + O_p(1),
\end{aligned}
\tag{A.2}
$$

Therefore,

$$
\alpha \sum_{1\leq s<k\leq M} P(|\hat{\eta}_{sk}|) \geq -\left[\sum_{i=1}^{n} \log\{f(X_i;\widetilde{\psi})/f(X_i;\bar{\psi})\}\right] + O_p(1) = O_p(1),
$$

which implies $\hat{\eta}_{sk}$ does not converge to 0 in probability for any subsequence. This completes the proof.

Proof of Lemma 1 Define $\psi_0^*(\psi) \in \Omega^*$ as the nearest point to ψ in Euclidian distance. According to Remark 1, the form of $\psi_0^*(\psi)$ is $(\pi_{10}, \mathbf{0}, \theta_{10}, \theta_2)$, then $\psi - \psi_0^*(\psi) = (\mathbf{v}^t, \mathbf{0}^t)^t$ with

$$
\mathbf{v} = (\pi_1 - \pi_{10}, \cdots, \pi_{M_0} - \pi_{M_0,0}, \pi_{M_0+1}, \cdots, \pi_M, \mu_1 - \mu_{10}, \cdots, \sigma^2_{M_0} - \sigma^2_{M_0,0})^t.
$$

Let $I(\psi_0^*(\psi))$ be the fisher information and give the corresponding matrix division

$$
I(\psi_0^*(\psi)) = \begin{pmatrix} I_{11}(\psi_0^*(\psi)) & I_{12}(\psi_0^*(\psi)) \\ I_{21}(\psi_0^*(\psi)) & I_{22}(\psi_0^*(\psi)) \end{pmatrix}.
$$

Applying matrix calculation, we have

$$
(\psi - \psi_0^*(\psi))^t I(\psi_0^*(\psi))(\psi - \psi_0^*(\psi)) = \mathbf{v}^t I_{11}(\psi_0^*(\psi))\mathbf{v}.
$$

It suffices to prove that $I_{11}(\psi_0^*(\psi))$ is positive definite, that is, to prove all diagonal elements of $I_{11}(\psi_0^*(\psi))$ are positive. By condition (R2), we have that

$$
I_{\pi_m\pi_m}(\psi_0^*(\psi)) = E_{\psi_0}\left[\frac{\phi(x;\mu_{m0},\sigma^2_{m0})}{f_0}\right]^2, \quad m = 1, \cdots, M_0,
$$

$$
I_{\pi_m\pi_m}(\psi_0^*(\psi)) = E_{\psi_0}\left[\frac{\phi(x;\mu_m,\sigma^2_m)}{f_0}\right]^2, \quad m = M_0+1, \cdots, M.
$$

By exchanging the order of integration and differentiation for the densities

$$
I_{\mu_m\mu_m}(\psi_0^*(\psi)) = E_{\psi_0}\left[\frac{\pi_{m0}\phi'_\mu(x;\mu_{m0},\sigma^2_{m0})}{f_0}\right]^2, \quad m = 1, \cdots, M_0,
$$

$$I_{\sigma_m\sigma_m}(\psi_0^*(\psi)) = E_{\psi_0}\left[\frac{\pi_{m0}\phi'_{\sigma^2}(x;\mu_{m0},\sigma_{m0}^2)}{f_0}\right]^2, \quad m = 1,\cdots,M_0,$$

where $\phi'_\mu(x;\mu,\sigma^2)$ and $\phi'_{\sigma^2}(x;\mu,\sigma^2)$ represent the first derivative of μ and σ^2. This completes the proof.

Proof of Theorem 1 The proof is similar to that of [13] which considers all the points on the boundary of some sufficient small stripe surrounding Ω_0. Following Proposition 1, the maximum likelihood of (2.10) will be located in the neighborhood of Ω^*. Let $\omega_n \subset \omega_\epsilon$ be a $C/\sqrt{n}$-neighborhood of Ω^*, we just need to show that for a large constant C, $\widetilde{\ell}(\psi) < \widetilde{\ell}(\psi_0^*(\psi))$ at all points on the boundary of ω_n.

By the Taylor expansion of the likelihood function, we have

$$\begin{aligned}\widetilde{\ell}(\psi) - \widetilde{\ell}(\psi_0^*(\psi)) =& S_1 + S_2 + S_3 \\ &+ \alpha \sum_{1\le s<k\le M} P(|\eta_{sk}|) - \alpha \sum_{1\le s<k\le M} P(|\eta_{sk0}|),\end{aligned} \tag{A.3}$$

with

$$S_1 = \frac{C}{\sqrt{n}}\frac{\partial \ell(\psi_0^*(\psi))}{\partial \psi^t}\mathbf{u} = C\cdot O_p(1),$$

$$S_2 = \frac{1}{2}\mathbf{u}^t\frac{\partial^2\ell(\psi_0^*(\psi))}{\partial\psi\partial\psi^t}\mathbf{u} = -\frac{C^2}{2}\mathbf{u}^t I(\psi_0^*(\psi))\mathbf{u} + o_p(1),$$

$$S_3 = \frac{C^3}{6n^{3/2}}\sum_{j,k,l=1}^{d} u_j u_k u_l\left(\sum_{i=1}^n \gamma_{jkl}(X_i)M_{jkl}(X_i)\right) = O_p(1/\sqrt{n}),$$

where $\mathbf{u}$ is the unit direction vector of $\psi - \psi_0^*(\psi)$ with $||\mathbf{u}|| = 1$, and $0 \le \gamma_{jkl} \le 1$ by condition (R1). Note that $||\frac{\partial \ell(\psi_0^*(\psi))}{\partial\psi^t}|| = O_p(\sqrt{n})$, since

$$\begin{aligned}P\left(\left\|\frac{\partial\ell(\psi_0^*(\psi))}{\partial\psi^t}\right\| \ge L\sqrt{n}\right) \le \frac{E\left\|\frac{\partial\ell(\psi_0^*(\psi))}{\partial\psi^t}\right\|^2}{L^2 n} &= \frac{\sum_{j=1}^d\sum_{i=1}^n E\left[\left(\frac{\partial f(X_i;\psi_0^*(\psi))}{\partial\psi_j}\right)^2\right]}{L^2 n} \\ &= \frac{n\sum_{j=1}^d I_{jj}(\psi_0^*(\psi))}{L^2 n} \le \frac{dC_1}{L^2},\end{aligned}$$

where the first inequality holds because of Markov's inequality, then the equation holds since $\{X_i\}_{i=1}^n$ are independent, and the last inequality holds due to condition (R3).

Based on Proposition 1, the following term

$$\alpha \sum_{1\le s<k\le M} P(|\eta_{sk}|) - \alpha \sum_{1\le s<k\le M} P(|\eta_{sk0}|)$$

is also bounded by a large positive constant. Hence, by choosing a sufficient large C, the second term of Taylor expansion dominates all other term. This completes the proof.

Proof of Proposition 2 The penalized log-likelihood function is

$$\ell_p(\psi) = \ell(\psi) + \alpha \sum_{1\le s<k\le M} P_\gamma(|\eta_{sk}|) - n\sum_{m=1}^{M} p_\lambda(\pi_m), \tag{A.4}$$

it is sufficient to prove that $Q(\psi|\psi^{(l)})$ can be a minorizing function of the observed log-likelihood $\ell_p(\psi)$.

For the normal mixture model, it is easy to know that

$$\sum_{i=1}^{n}\sum_{m=1}^{M} h_{im}\log[\pi_m\phi(X_i;\mu_m,\sigma_m^2)]$$

is a minorizing function of $\ell(\psi)$. By the property of $p_\lambda(\cdot)$,

$$-p_\lambda(\pi_m) \ge -p_\lambda(\pi_m^{(l)}) - p'_\lambda(\pi_m^{(l)})(\pi_m - \pi_m^{(l)})$$

holds for all π_m and the equation holds at $\pi_m^{(l)}$, so $-p'_\lambda(\pi_m^{(l)})(\pi_m - \pi_m^{(l)})$ provides a minorizing function of $-p_\lambda(\pi_m)$ [22].

In the view of the convexity of the Euclidean norm $||\cdot||$,

$$||x|| \ge ||x_0|| + \frac{(x-x_0)^t x_0}{||x_0||} = \frac{x^t x_0}{||x_0||}.$$

Then, combining with the nondecreasing function $P_\gamma(\cdot)$, we have

$$P_\gamma(|\eta_{sk}|) \ge \Big\{\log[(\mu_s^{(l)} - \mu_k^{(l)})(\mu_s - \mu_k) + (\sigma_s^{2(l)} - \sigma_k^{2(l)})(\sigma_s^2 - \sigma_k^2)] - \log(|\eta_{sk}^{(l)}|)\Big\} \\ \times I(|\eta_{sk}^{(l)}| \le \gamma),$$

which is a minorizing function of the second term of $\ell_p(\psi)$. Hence, we can transfer maximization of $\ell_p(\psi)$ to the surrogate function $Q(\psi|\psi^{(l)})$. The ascending property follows a standard argument of MM algorithm.

Proof of Theorem 2 It is sufficient to show that $\ell_p(\psi) < \ell_p(\psi_0^*(\psi))$ at all points on the boundary of $n^{-1/2}$-neighborhood of Ω^*. Using $p_\lambda(\cdot) \ge 0$, we have

$$\ell_p(\psi) - \ell_p(\psi_0^*(\psi)) \le \widetilde{\ell}(\psi) - \widetilde{\ell}(\psi_0^*(\psi)) - n\sum_{m=1}^{M_0}(p_\lambda(\pi_m) - p_\lambda(\pi_{m0})) \\ \triangleq I_1 + I_2.$$

By Taylor expansion we have

$$\begin{aligned}|I_2| &= |-n\sum_{m=1}^{M_0}[p'_\lambda(\pi_{m0})(\pi_m-\pi_{m0})+\frac{1}{2}p''_\lambda(\pi_{m0})(\pi_m-\pi_{m0})^2(1+o_p(1))]| \\ &\le na_n||\psi-\psi_0^*(\psi)||+\frac{1}{2}nb_n||\psi-\psi_0^*(\psi)||^2(1+o_p(1))\end{aligned}$$

By the proof in Theorem 1 and condition (R5)–(R6), with a sufficient large C, I_2 is dominated by $-\frac{C^2}{2}\mathbf{u}^t I(\psi_0^*(\psi))\mathbf{u}$. This complete the proof.

Proof of Theorem 3 To prove part (a), we first prove that

$$\ell_p(\pi_1,\mathbf{0},\theta_1,\theta_2)=\max_{\pi_2:0\le||\pi_2||\le C/\sqrt{n}}\ell_p(\pi_1,\pi_2,\theta_1,\theta_2)$$

with probability tending to 1, for any $\psi=(\pi_1,\pi_2,\theta_1,\theta_2)\in\omega_n$. It is sufficient to show that

$$\frac{\partial\ell^*(\psi)}{\partial\pi_l}<0 \;\text{ for }\; \pi_l<C/\sqrt{n},$$

where $l=M_0+1,\cdots,M$. The partial derivatives of $\ell^*(\psi)$ about π_l is

$$\frac{\partial\ell^*(\psi)}{\partial\pi_l}=nA_{nl}(\psi)-np'_\lambda(\pi_l)-\beta,$$

where

$$A_{nl}(\psi)=\frac{1}{n}\sum_{i=1}^n\frac{\phi(X_i;\mu_l,\sigma_l^2)}{\sum\limits_{m=1}^M\pi_m\phi(X_i;\mu_m,\sigma_m^2)}.$$

For $l=1,\cdots,M_0$, it is known that $\frac{\partial\ell^*(\widetilde{\psi})}{\partial\pi_l}=0$. Then similar to the derivation of Theorem 3, we have

$$\beta=n\left(1-\sum_{m=1}^{M_0}\pi_m p'_\lambda(\pi_m)+o(1)\right).$$

Therefore, the partial derivative is rewritten as

$$\frac{\partial\ell^*(\psi)}{\partial\pi_l}=n\left\{(A_{nl}(\psi)-1)-\left(p'_\lambda(\pi_l)-\sum_{m=1}^{M_0}\pi_m p'_\lambda(\pi_m)\right)+o(1)\right\}.$$

By the law of large number, we have

$$\begin{aligned}E_0\left(\frac{\phi(X_i;\mu_l,\sigma_l^2)}{\sum_{m=1}^M\pi_m\phi(X_i;\mu_m,\sigma_m^2)}\right)&=1+\int\phi(X_i;\mu_l,\sigma_l^2)\left(\frac{f(X_i;\psi_0^*(\psi))}{f(X_i;\psi)}-1\right)dX_i \\ &\le 1+\max\left|\frac{f(X_i;\psi_0^*(\psi))}{f(X_i;\psi)}-1\right|.\end{aligned}$$

Since $||\psi - \psi_0^*(\psi)|| = O_p(n^{-1/2})$, we have

$$|f(x;\psi_0^*(\psi)) - f(x;\psi)| = O_p(n^{-1/2}).$$

Therefore, as $n \to \infty$ we have $A_{nl}(\psi) \to 1$ with probability tending to one. Hence, the sign of derivative $\frac{\partial \ell^*(\psi)}{\partial \pi_l}$ is completely determined by the sign of $\Big(p'_\lambda(\pi_l) - \sum_{m=1}^{M_0} \pi_m p'_\lambda(\pi_m)\Big)$.

It is obvious that $\ell_p(\pi_1, \mathbf{0}, \theta_1, \theta_2) = \ell_p(\pi_1, \theta_1)$, which does not depend on θ_2. Let $(\widetilde{\pi}_1, \widetilde{\theta}_1)$ be a local maximizer of $\ell_p(\pi_1, \theta_1)$. We now show that $\widetilde{\psi} = (\widetilde{\pi}_1, \mathbf{0}, \widetilde{\theta}_1, \theta_2)$ is a local maximizer of $\ell_p(\psi)$ for any θ_2 such that $\widetilde{\psi}$ lies in the $n^{-1/2}$-neighborhood of Ω^*. For all $\psi \in \omega_n$, we have

$$\begin{aligned}\ell_p(\psi) - \ell_p(\widetilde{\psi}) &= [\ell_p(\psi) - \ell_p(\pi_1, \theta_1)] + [\ell_p(\pi_1, \theta_1) - \ell_p(\widetilde{\pi}_1, \widetilde{\theta}_1)] \\ &\leq \ell_p(\psi) - \ell_p(\pi_1, \theta_1) < 0.\end{aligned}$$

This complete the proof of part (a).

To prove part (b), consider the partial derivative of $\ell_p(\psi)$ about $\varphi = (\pi_1, \theta_1)$

$$\begin{aligned}&\frac{\partial \ell_p(\psi)}{\partial \varphi_j}\bigg|_{\psi=\widetilde{\psi}} \\ &= \frac{\partial \ell(\widetilde{\varphi})}{\partial \varphi_j} + \alpha \left\{ \sum_{1 \leq s < k \leq M_0} \frac{\partial P(|\widetilde{\eta}_{sk}|)}{\partial \varphi_j} \right\} I(M_0 < j \leq 3M_0) - np'_\lambda(\widetilde{\pi}_j) I(1 \leq j \leq M_0) \\ &= 0, \ \text{for } j = 1, \cdots, 3M_0.\end{aligned}$$

Note that $(\widetilde{\pi}_1, \widetilde{\theta}_1)$ is a consistent estimator, by a Taylor's expansion around ψ_0, we have

$$\begin{aligned}&\frac{1}{n}\frac{\partial \ell(\psi_0)}{\partial \varphi_j} + \sum_{i=1}^{3M_0} \left\{ \frac{1}{n} \frac{\partial^2 \ell(\psi_0)}{\partial \varphi_j \partial \varphi_i} + o_p(\frac{1}{n}) \right\} (\widetilde{\varphi}_i - \varphi_{i0}) + O(\frac{\alpha}{n}) I(M_0 < j \leq 3M_0) \\ &\qquad - \{p'_\lambda(\pi_{j0}) + [p''_\lambda(\pi_{j0}) + o_p(1)](\widetilde{\pi}_j - \pi_{j0})\} I(1 \leq j \leq M_0) = 0\end{aligned}$$

By Slutsky's theorem and the central limit theorem, we have

$$\sqrt{n}(I(\psi_0) + \Sigma)\{(\widetilde{\pi}_1, \widetilde{\theta}_1) - \psi_0 + (I(\psi_0) + \Sigma)^{-1}\mathbf{b}\} \to N(\mathbf{0}, I(\psi_0)).$$

This complete the proof of part (b).

Acknowledgment

The authors are grateful to Jiahua Chen for helpful comments and discussions. The authors also thank the editors and the referees for their insightful comments and suggestions, which greatly improved this article.

References

[1] Budanova, S. (2016). Penalized maximum likelihood estimation of finite mixture models, PhD thesis, Working paper, Nothwestern University, Evanston, IL.

[2] Chen, H., Chen, J. and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66** 95–115. MR2035761

[3] Chen, J. (1995). Optimal rate of convergence for finite mixture models. *Ann. Statist.* **23** 221–233. MR1331665

[4] Chen, J. and Khalili, A. (2008). Order selection in finite mixture models with a nonsmooth penalty. *Journal of the American Statistical Association* **103** 1674-1683. MR2722574

[5] Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *Annals of Statistics* **37** 2523–2542. MR2543701

[6] Chen, J., Tan, X. and Zhang, R. (2008). Inference for normal mixtures in mean and variance. *Statistica Sinica* **18** 443–465. MR2432278

[7] Chen, X., Ponomareva, M. and Tamer, E. (2014). Likelihood inference in some finite mixture models. *Journal of Econometrics* **182** 87–99. MR3212763

[8] Ciuperca, G., Ridolfi, A. and Idier, J. (2003). Penalized maximum likelihood estimator for normal mixtures. *Scandinavian Journal of Statistics* **30** 45–59. MR1963892

[9] Dacunha-Castelle, D. and Gassiat, E. (1999). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *The Annals of Statistics* **27** 1178–1209. MR1740115

[10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96** 1348–1360. MR1946581

[11] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32(3)** 928–961. MR2065194

[12] Fan, J., Xue, L. and Zou, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988

[13] Feng, Z. D. and McCulloch, C. E. (1996). Using bootstrap likelihood ratios in finite mixture models. *Journal of the Royal Statistical Society. Series B (Methodological)* **58** 609–617.

[14] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer. MR2265601

[15] Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics* **38** 897–906. Elsevier. MR1955343

[16] Heinrich, P., Kahn, J. et al. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. *The Annals of Statistics* **46** 2844–2870. MR3851757

[17] Ho, N., Nguyen, X. et al. (2016). Convergence rates of parameter estima-

tion for some weakly identifiable finite mixtures. *The Annals of Statistics* **44** 2726–2755. MR3576559

[18] Huang, T., Peng, H. and Zhang, K. (2017). Model selection for Gaussian mixture models. *Statistica Sinica* **27** 147–169. MR3618163

[19] Keribin, C. (2000). Consistent estimate of the order of mixture models. *Sankhyā, Series A* **62** 49–66. MR1769735

[20] Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association* **102** 1025–1038. MR2411662

[21] Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *Journal of the Royal Statistical Society* **57** 425–437. MR1323348

[22] Lange, K., Hunter, D. R. and Yang, I. (2000). Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics* **9** 1–20. MR1819865

[23] Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *Journal of Econometrics* **142** 201–211. MR2394290

[24] Lehmann, E. and Casella, G. (1998). *Theory of point estimation*. Springer. MR1639875

[25] Lindsay, B. G. (1995). Mixture Models: Theory, Geometry, and Applications In *NSF-CBMS Regional Conference Series in Probability and Statistics v 5*. Institure of Mathematical Statistics.

[26] Liu, X. and Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics* **31** 807–832. MR1994731

[27] Liu, X. and Shao, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference* **123** 61–81. MR2058122

[28] Loisel, P., Goffinet, B. and Oca, M. G. M. D. (1994). Detecting a major gene in an F2 population. *Biometrics* **50** 512-516. MR1294684

[29] Manole, T. and Khalili, A. (2021). Estimating the number of components in finite mixture models via the Group-Sort-Fuse procedure. *The Annals of Statistics* **49** 3043–3069. MR4352522

[30] McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York. MR1789474

[31] Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *The Annals of Statistics* **9** 225-228. MR0600553

[32] Shen, X., Pan, W. and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223-232. PMID: 22736876. MR2949354

[33] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of Royal Statistical Association: Series B* **62** 795–809. MR1796293

[34] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* **58** 267–288. MR1379242

[35] Wichitchan, S., Yao, W. and Yang, G. (2019). Hypothesis testing for finite mixture models. *Computational Statistics & Data Analysis* **132**

180–189. MR3913143

[36] Yao, W. and Lindsay, B. G. (2009). Bayesian mixture labeling by highest posterior density. *Journal of American Statistical Association* **104** 758–767. MR2751453

[37] Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701