ORDER SELECTION IN FINITE MIXTURE MODELS [1]

Jiahua Chen, Abbas Khalili
Department of Statistics and Actuarial Science
University of Waterloo

Abstract

A fundamental and challenging problem in the application of finite mixture models is to make inference on the order of the model. In this paper, we develop a new penalized likelihood approach to the order selection problem. The new method deviates from the information-based methods such as AIC and BIC by introducing two penalty functions which depend on the mixing proportions and the component parameters. The new method is shown to be consistent and have other good properties. Simulations show that the method has much better performance compared to a number of existing methods. We further demonstrate the new method by analyzing two well known real data sets.

Short Title: ORDER SELECTION

1. Introduction. Making inference on the number of components of the model is a fundamental and challenging problem in the application of finite mixture models. A mixture model with a large number of components can provide a good fit to the data, but has poor interpretive values. Complex models as such are not favoured in applications in the name of parsimony, and for the sake of preventing over-fitting of the data.

A large number of statistical methods for order selection have been pro-

[1]AMS 2000 subject classifications. Primary 62G05; secondary 62G07.

KEY WORDS: E-M algorithm, finite mixture model, LASSO, penalty method, SCAD.

posed and investigated in the past a few decades. One off-the-shelf method is to use information theoretic approaches such as the Akaike information criterion (AIC, Akaike 1973) and the Bayesian information criterion (BIC, Schwarz 1978). Leroux (1992) discussed the use of AIC and BIC for order selection in finite mixture models. Another class of methods are designed based on some distance measure between the fitted model and the non-parametric estimate of the population distribution; see Chen and Kalbfleisch (1996) and James, Priebe and Marchette (2001). One may also consider testing the hypothesis on the order of finite mixture models. The most influential methods in this class include the $C(\alpha)$ test by Neyman and Scott (1966) and methods based on likelihood ratio techniques, which include Ghosh and Sen (1985), McLachlan (1987), Dacunha-Castelle and Gassiat (1999), Chen and Chen (2001), Chen, Chen and Kalbfleisch (2001, 2004). Charnigo and Sun (2004) propsed an L^2-distance method for testing homogeneity in continuous finite mixture models. The recent paper by Chambaz (2006) studies the asymptotic efficiency of two generalized likelihood ratio tests. Ishwaran, James and Sun (2001) proposed a Bayesian approach.

In this paper, we develop a new order selection method combining the strength of two existing statistical methods. The first was proposed by Chen and Kalbfleisch (1996) which has simple and interesting statistical properties. The second is the variable selection method in the context of regression, such as LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001). We formulate the problem of order selection as a problem of arranging subpopulations (i.e. mixture components) in a parameter space. When the fitted mixture model contains two subpopulations that are close to each other to some degree,

an SCAD-type penalty will merge them. Our procedure starts with a large number of subpopulations and ends up with a mixture model with lower order by merging close subpopulations.

We prove that the new method is consistent in selecting the most parsimonious mixture models. The new method is less computing intensive than many existing methods since the order is determined through a single optimization procedure. Our simulation results are exciting. The new method has a much higher probability of selecting finite mixture models with the proper order when compared to a number of existing methods in the situations that we considered.

The paper is organized as follows. Section 2 introduces the finite mixture model. The new method for order selection is described in Section 3. Asymptotic properties of the new method are studied in Section 4. In Section 5, a computational algorithm is outlined for numerical solution of the optimization problem. The performance of the new method is compared to a number of existing methods through simulations in Section 6. To further demonstrate the use of the new method, a number of well-known real data sets are analyzed in Section 7. A summary and discussion are given in Section 8.

2. The finite mixture model. Let $\mathcal{F} = \{f(y;\theta); \theta \in \Theta\}$ be a known family of parametric (probability) density functions with respect to a σ-finite measure ν. Let Θ be a one-dimensional compact parameter space and $\Theta \subseteq \boldsymbol{R}$. The compactness assumption of Θ is merely a technical requirement used in many papers such as Ghosh and Sen (1985) and Dacunha-Castelle and Gassiat (1999). It is not restrictive in applications since a reasonable

range of the parameter θ can often be specified. The density function of a finite mixture model based on the family $\mathcal{F}$ is given by

$$f(y; G) = \int_{\Theta} f(y; \theta)\ dG(\theta) \tag{1}$$

where $G(\cdot)$ is called *the mixing distribution* and is given by

$$G(\theta) = \sum_{k=1}^{K} \pi_k I(\theta_k \leq \theta). \tag{2}$$

The $I(\cdot)$ is an indicator function, and $\theta_k \in \Theta$, $0 \leq \pi_k \leq 1$ for $k = 1, 2, \ldots, K$. We denote the class of all finite mixing distributions with at most K support points as

$$\mathcal{M}_K = \{G(\theta) = \sum_{k=1}^{K} \pi_k I(\theta_k \leq \theta) : \theta_1 \leq \theta_2 \leq \ldots \leq \theta_K, \sum_{k=1}^{K} \pi_k = 1, \pi_k \geq 0\}.$$

Note that the class $\mathcal{M}_K$ implicitly also contains finite mixing distributions with fewer than K support points. In fact, $\mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \ldots \subseteq \mathcal{M}_{K-1} \subseteq \mathcal{M}_K$. The lower order models are represented in $\mathcal{M}_K$ by allowing the θ_k's to coincide with one another while still maintaining separate π_k's. The class of all finite mixing distributions is given by $\mathcal{M} = \bigcup_{K \geq 1} \mathcal{M}_K$.

Let K_0 be the true number of support points of the finite mixing distribution G in (2). The true value K_0 is the smallest number of support points for G such that all the component densities $f(y; \theta_k)$'s are different and the mixing proportions π_k's are non-zero. We denote the true mixing distribution G_0 as

$$G_0(\theta) = \sum_{k=1}^{K_0} \pi_{0k} I(\theta_{0k} \leq \theta) \tag{3}$$

where $\theta_{01} < \theta_{02} < \ldots < \theta_{0K_0}$ are K_0 distinct interior points of Θ, and $0 < \pi_{0k} < 1$, for $k = 1, 2, \ldots, K_0$, when $K_0 \geq 2$. Note that when $K_0 = 1$, the

population becomes homogeneous. In this case, we denote the true density function of the random variable Y by $f(y;\theta_0)$. We also assume that θ_0 is an interior point of Θ.

3. The new order selection method. Even though the true order of the finite mixture model, i.e. K_0, is not known, we assume that some information is available to provide an upper bound K for K_0. Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from (1) and hence the log-likelihood function of the mixing distribution with order K is given by

$$l_n(G) = \sum_{i=1}^{n} \log f(y_i; G).$$

By maximizing $l_n(G)$ over $\mathcal{M}_K$, the resulting fitted model may over-fit the data with some small values of the mixing proportions (over-fitting type I), and/or with some component densities close to each other (over-fitting type II). These are main causes of difficulties in the order selection problem. Our new approach works by introducing two penalty functions to prevent these two types of overfitting.

Denote $\eta_k = \theta_{k+1} - \theta_k$, for $k = 1, 2, \ldots, K-1$. Also, corresponding to the ordered support points of the true mixing distribution G_0 in (3), denote $\eta_{0k} = \theta_{0,k+1} - \theta_{0k}$, for $k = 1, 2, \ldots, K_0 - 1$, when $K_0 \geq 2$. Define the penalized log-likelihood function as

$$\tilde{l}_n(G) = l_n(G) - \sum_{k=1}^{K-1} p_n(\eta_k) + C_K \sum_{k=1}^{K} \log \pi_k \tag{4}$$

for some $C_K > 0$ and a non-negative function $p_n(\cdot)$. Motivated by LASSO (Tibshirani, 1996) and SCAD (Fan and Li, 2001), the penalty function $p_n(\eta_k)$ is designed so that if any η_k has a small fitted value before penalty, its fitted

value after penalty has a positive chance to be 0. In other words, it prevents the type II over-fitting. The second penalty function in (4) is motivated from Chen and Kalbfleisch (1996). It makes fitted values of π_k's stay away from 0 and hence prevents the type I over-fitting. Its additional utility is to make some fitted values of η_k close to 0 when $K > K_0$ asymptotically, which in turn activates the utility of $p_n(\eta_k)$.

The new order selection method then selects $\hat{G}_n$ that maximizes $\tilde{l}_n(G)$ over the space $\mathcal{M}_K$. When some fitted values of η_k are 0, a mixture model with order lower than K is obtained. We call $\hat{G}_n$ as the maximum penalized likelihood estimator (MPLE), and we show it has desirable asymptotic properties in the next section.

4. Asymptotic properties. Being consistency is often considered as a minimum requirement of a statistical method. In the current context, the consistency expresses itself in two folds. As an estimator of the mixing distribution G_0, the MPLE $\hat{G}_n$ is consistent, but this fact does not imply the order of $\hat{G}_n$ is consistent for K_0. We establish both consistencies in this section. Let us first list the following conditions on the penalty function $p_n(\cdot)$.

P_0. For all n, $p_n(0) = 0$, and $p_n(\eta)$ is a non-decreasing function of η on $(0, \infty)$. It is twice differentiable for η except for a finite number of points.

P_1. For any $\eta \in (0, \infty)$, we have $p_n(\eta) = o(n)$, $p_n(\eta) \to \infty$, and $c_n = \max\{n^{-1}|p_n''(\eta_{0k})| : 1 \le k \le (K_0 - 1)\} = o(1)$.

P_2. Let $N_n = \{\eta; 0 < \eta \le n^{-1/4} \log n\}$, we have $\lim_{n\to\infty} \inf_{\eta \in N_n} \frac{p_n'(\eta)}{\sqrt{n}} = \infty$.

P_3. There exist positive constants $\delta_n = o(1)$, $d_n = o(n)$ such that for all $\eta > \delta_n$, $p_n(\eta) = d_n \to \infty$ as $n \to \infty$.

Since the user has the option of choosing the most appropriate penalty function, the conditions on $p_n(\eta)$ are reasonable as long as the functions satisfying these conditions exist. The following three penalty functions were proposed for variable selection in the regression context.

(a) L_1-norm penalty: $p_n(\eta) = \gamma_n \sqrt{n}|\eta|$.

(b) Hard penalty: $p_n(\eta) = \gamma_n^2 - (\sqrt{n}|\eta| - \gamma_n)^2 \; I\{\sqrt{n}|\eta| < \gamma_n\}$.

(c) SCAD penalty: Let $(\cdot)_+$ be the positive part of a quantity.

$$p_n'(\eta) = \gamma_n \sqrt{n} \; I\{\sqrt{n}|\eta| \le \gamma_n\} + \frac{\sqrt{n}(a\gamma_n - \sqrt{n}|\eta|)_+}{(a-1)} I\{\sqrt{n}|\eta| > \gamma_n\}$$

which is a quadratic spline function, and $a > 2$.

The L_1-norm penalty is used in LASSO by Tibshirani (1996). The other two are discussed in Fan and Li (2001, 2002) and they satisfy conditions P_0-P_3 with proper choice of the tuning parameter γ_n.

We now present the asymptotic properties of the MPLE $\hat{G}_n$ in two general settings: when the true mixing distribution G_0 in (3) is degenerate, i.e. $K_0 = 1$, and when $K_0 \ge 2$. To focus on main results, we leave regularity conditions on the kernel density $f(x;\theta)$ and the proofs in Appendix.

Theorem 1 *(Consistency of $\hat{G}_n$ when $K_0 = 1$). Suppose the kernel density $f(y;\theta)$ satisfies the regularity conditions A_1-A_5, and the penalty function $p_n(\cdot)$ satisfies conditions P_0 and P_1. If the true distribution of Y is homogeneous with density function $f(y;\theta_0)$, then $\hat{\theta}_k \to \theta_0, k = 1, 2, \ldots, K$, in probability, as $n \to \infty$.*

The above theorem shows that introducing penalties to the log-likelihood function does not void the consistency in estimating G_0. The next theorem establishes the consistency for estimating K_0.

Theorem 2 *(Consistency of estimating K_0). Suppose the kernel density $f(y;\theta)$ satisfies regularity conditions A_1-A_5, and the penalty function $p_n(\cdot)$ satisfies conditions P_0-P_2. If the true distribution of Y is homogeneous with density function $f(y;\theta_0)$, then the MPLE $\hat{G}_n$ has the property*

$$P\{\hat{\theta}_{k+1} - \hat{\theta}_k = 0\} \to 1 \ , \ k = 1, 2, \ldots, K-1 \tag{5}$$

as $n \to \infty$.

In what follows we investigate the properties of the MPLE $\hat{G}_n$ when $K_0 \geq 2$. Let $\theta_k^0 = (\theta_{0k} + \theta_{0,k+1})/2$, $k = 1, 2, \ldots, K_0 - 1$, be the middle points between each two consecutive support points of the true mixing distribution G_0. The MPLE $\hat{G}_n$ can then be written as

$$\hat{G}_n(\theta) = \sum_{k=1}^{K_0} \hat{p}_k \hat{G}_k(\theta) \tag{6}$$

where $\hat{G}_1(\theta_1^0) = 1$, $\hat{G}_2(\theta_1^0) = 0$, $\hat{G}_2(\theta_2^0) = 1$, and so on. Note that $\hat{p}_1$ is the probability assigned to the support points smaller than θ_1^0; $\hat{p}_2$ is the probability assigned to the support points between θ_1^0 and θ_2^0; and so on.

Theorem 3 *(Consistency of $\hat{G}_n$ when $K_0 \geq 2$). Suppose the kernel density $f(y;\theta)$ satisfies regularity conditions A_1-A_5, the penalty function $p_n(\cdot)$ satisfies conditions P_0-P_1, and the true distribution of Y is a finite mixture with density function $f(y;G_0)$. Then*

(a) $\hat{G}_n$ is a consistent estimator of G_0, for that for all $k = 1, 2, \ldots, K_0$,

(i) $\hat{p}_k = \pi_{0k} + o_p(1)$,

(ii) $\sup_\theta |\hat{G}_k(\theta) - G_{0k}(\theta)| = o_p(1)$, *where* $G_{0k}(\theta) = I(\theta_{0k} \le \theta)$.

(b) Support points of $\hat{G}_k$ *converge in probability to* θ_{0k} *which is the only support point of* G_{0k}, *for each* $k = 1, 2, \ldots, K_0$.

Let B_k be the event that $\hat{G}_k$ defined in (6) is a degenerate distribution, for $k = 1, 2, \ldots, K_0$. The consistency of estimating K_0 is equivalent to having $P(B_k) \to 1$ for all k which is the result of our next theorem.

Theorem 4 *(Consistency of estimating* K_0*). Suppose the kernel density* $f(y;\theta)$ *satisfies regularity conditions* A_1-A_5, *and the penalty function* $p_n(\cdot)$ *satisfies conditions* P_0-P_3. *Then under the true finite mixture density* $f(y; G_0)$, *if the MPLE* $\hat{G}_n$ *falls into a* $n^{-1/4}$*-neighbourhood of* G_0, *we have*

$$P\Big(\bigcap_{k=1}^{K_0} B_k \Big) \to 1 \quad , \quad n \to \infty.$$

Remark 1 Under some conditions including the strong identifiability in the Appendix, Chen (1995) shows that, when the order of the finite mixture model is unknown, the optimal rate of estimating the finite mixing distribution G is $n^{-1/4}$. Hence our result is applicable to that class of finite mixture models which include many commonly discussed models such as Poisson mixture, Normal mixture in location or scale parameter, and Binomial mixture.

Remark 2 In the light of Theorem 4, our order selection method is consistent with the HARD and SCAD penalty functions with a proper choice of γ_n. For example, letting $\gamma_n = n^{1/4} \log n$ in both penalties will suffice. The LASSO penalty function, however, cannot be made to satisfy all conditions.

Once K_0 is consistently estimated, the asymptotic properties of $\hat{G}_n$ become easier to explore. Denote

$$\mathbf{\Psi} = (\theta_1, \theta_2, \ldots, \theta_{K_0}, \pi_1, \pi_2, \ldots, \pi_{K_0-1})$$

and let $\mathbf{\Psi}_0$ be the vector of true parameters corresponding to G_0. For convenience, in the following we use $\tilde{l}_n(\mathbf{\Psi})$ instead of $\tilde{l}_n(G)$ to denote the penalized log-likelihood function. The following theorem gives the asymptotic properties of the maximizer of $\tilde{l}_n(\mathbf{\Psi})$.

Theorem 5 *Under the standard regularity conditions in the Appendix and conditions P_0-P_1 for the penalty function $p_n(\cdot)$, there exists a local maximizer $\hat{\mathbf{\Psi}}_n$ of the penalized log-likelihood function $\tilde{l}_n(\mathbf{\Psi})$ such that*

$$\|\hat{\mathbf{\Psi}}_n - \mathbf{\Psi}_0\| = O_p\{n^{-1/2}(1 + b_n)\}. \tag{7}$$

where $b_n = \max\{|p_n'(\eta_{0k})|/\sqrt{n} : 1 \leq k \leq (K_0 - 1)\}$.

When $b_n = O(1)$, as in the HARD and SCAD penalties, $\hat{\mathbf{\Psi}}_n$ has usual convergence rate $n^{-1/2}$. This result seems to contradict the conclusion on the optimal rate of $n^{-1/4}$. The seemingly contradiction is a super-efficiency phenomenon. Such properties are sometimes referred as *Oracle* property. In general, estimators with super-efficiency should be used with caution especially for constructing confidence intervals.

5. Numerical solutions. As expected, there are no apparent analytical solutions to the maximization problem posted when applying the new order selection procedure. In this section we discuss a numerical procedure for maximizing the penalized log-likelihood function $\tilde{l}_n(G)$ over the space $\mathcal{M}_K$, for a given K. For convenience, in the following, we use $\tilde{l}_n(\mathbf{\Psi})$ instead of

$\tilde{l}_n(G)$ to denote the penalized log-likelihood function, where $\mathbf{\Psi}$ is the vector of all parameters of the mixture model with order $K \geq K_0$.

5.1. Maximization of the penalized log-likelihood function. A popular numerical method used in finite mixture models is the Expectation-Maximization (EM) algorithm of Dempster, Laird and Rubin (1977). For the current application, the algorithm must be revised in the M-step. The revised EM algorithm is as follows.

Let the complete log-likelihood function be

$$l_n^c(\mathbf{\Psi}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left[\log \ \pi_k + \log\{f(y_i; \theta_k)\}\right]$$

where the z_{ik}'s are indicator variables showing the component-membership of the ith observation in the mixture model. Note that the z_{ik}'s are unobserved. The complete penalized log-likelihood function is then given by

$$\tilde{l}_n^c(\mathbf{\Psi}) = l_n^c(\mathbf{\Psi}) - \sum_{k=1}^{K-1} p_n(\eta_k) + C_K \ \sum_{k=1}^{K} \log \pi_k.$$

The EM algorithm maximizes $\tilde{l}_n^c(\mathbf{\Psi})$ iteratively in two steps as follows.

E-Step: Let $\mathbf{\Psi}^{(m)}$ be the estimate of the parameters after the mth iteration. The E-step of the algorithm computes the conditional expectation of $\tilde{l}_n^c(\mathbf{\Psi})$ with respect to z_{ik}, given the observed data and assuming that the current estimate $\mathbf{\Psi}^{(m)}$ is the true parameter of the model. The conditional expectation is given by

$$\begin{aligned} Q(\mathbf{\Psi}; \mathbf{\Psi}^{(m)}) \ &= \ \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik}^{(m)} \log\{f(y_i; \theta_k)\} - \sum_{k=1}^{K-1} p_n(\eta_k) \\ &\quad + \sum_{i=1}^{n} \sum_{k=1}^{K} [w_{ik}^{(m)} + \frac{C_K}{n}] \log \pi_k \end{aligned}$$

where

$$w_{ik}^{(m)} = \frac{\pi_k^{(m)} f(y_i; \theta_k^{(m)})}{\sum_{l=1}^{K} \pi_l^{(m)} f(y_i; \theta_l^{(m)})} \quad , \quad k = 1, 2, \ldots, K$$

are the conditional expectation of z_{ik} given data and the current estimate $\mathbf{\Psi}^{(m)}$.

M-Step: The M-step on the $(m+1)$th iteration maximizes $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(m)})$ with respect to $\mathbf{\Psi}$. The updated estimate $\pi_k^{(m+1)}$ of the mixing proportion π_k is given by

$$\pi_k^{(m+1)} = \frac{\sum_{i=1}^{n} w_{ik}^{(m)} + C_K}{n + KC_K} \ , \ k = 1, 2, \ldots, K.$$

We need to maximize $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(m)})$ with respect to θ_k next. Due to condition P_0 on the penalty $p_n(\cdot)$, which is essential to achieve consistency in estimating K_0, $p_n(\eta_k)$ is not differentiable at $\eta_k = 0$. Thus, the usual Newton-Raphson method cannot be directly used. However, Fan and Li (2001) suggested of approximating $p_n(\eta)$ by

$$\tilde{p}_n(\eta; \eta_k^{(m)}) = p_n(\eta_k^{(m)}) + \frac{p_n'(\eta_k^{(m)})}{2\eta_k^{(m)}}(\eta^2 - \eta_k^{(m)2}).$$

Unlike a simple Taylor's expansion, this function approximates $p_n(\eta)$ well when η is near $\eta_k^{(m)}$ while it tends to infinity as $|\eta| \to \infty$. With this approximation, the component parameters θ_k are updated by solving

$$\begin{aligned}
\sum_{i=1}^{n} w_{i1}^{(m)} \frac{\partial}{\partial \theta_1} \{\log f(y_i; \theta_1)\} + \frac{\partial \tilde{p}_n(\eta_1; \eta_1^{(m)})}{\partial \theta_1} &= 0, \\
\sum_{i=1}^{n} w_{ik}^{(m)} \frac{\partial}{\partial \theta_k} \{\log f(y_i; \theta_k)\} - \frac{\partial \tilde{p}_n(\eta_{k-1}; \eta_{k-1}^{(m)})}{\partial \theta_k} + \frac{\partial \tilde{p}_n(\eta_k; \eta_k^{(m)})}{\partial \theta_k} &= 0, \\
k = 2, 3, \ldots, K-1, & \\
\sum_{i=1}^{n} w_{iK}^{(m)} \frac{\partial}{\partial \theta_K} \{\log f(y_i; \theta_K)\} - \frac{\partial \tilde{p}_n(\eta_{K-1}; \eta_{K-1}^{(m)})}{\partial \theta_K} &= 0.
\end{aligned}$$

Starting from an initial value $\mathbf{\Psi}^{(0)}$, the iteration between the E-step and M-step continues until some convergence criterion is satisfied. When the algorithm converges, some of the equations

$$\begin{aligned} \frac{\partial l_n(\mathbf{\Psi})}{\partial \theta_1} + \frac{\partial p_n(\eta_1)}{\partial \theta_1} &= 0, \\ \frac{\partial l_n(\mathbf{\Psi})}{\partial \theta_k} - \frac{\partial p_n(\eta_{k-1})}{\partial \theta_k} + \frac{\partial p_n(\eta_k)}{\partial \theta_k} &= 0, \ k = 2, 3, \ldots, K-1, \\ \frac{\partial l_n(\mathbf{\Psi})}{\partial \theta_K} - \frac{\partial p_n(\eta_{K-1})}{\partial \theta_K} &= 0 \end{aligned}$$

are satisfied (approximately) for the corresponding non-zero valued $\hat{\eta}_k$, but not for zero valued $\hat{\eta}_k$'s. This enables us to identify zero estimates of η_k's.

5.2. Choice of the tuning parameters. The next problem in applying our new method is to choose the sizes of the tuning parameters γ_n and C_K. Chen, Chen and Kalbfleisch (2001) reported that the choice of C_K is not crucial which is re-affirmed by our simulations. Nonetheless, in practice, the choice of C_K has some effect on the performance of the method. Chen, Chen and Kalbfleisch (2001) suggested that if the parameters θ_k are restricted to be in $[-M, M]$ or $[M^{-1}, M]$ for large M, then an appropriate choice is $C_K = \log M$.

The current theory provides only some guidance on the order of γ_n to achieve the consistency. In applications, cross validation or CV (Stone, 1974) and generalized cross validation or GCV (Craven and Wahba, 1979) are often used for choosing tuning parameters such as γ_n.

Denote $D = \{y_1, y_2, \ldots, y_n\}$ as the full data set. Let N be the number of partitions of D. For the ith partition, let D_i be the subset of D which is used for evaluation and $D - D_i$ be the rest of the data used for fitting a model. The parts $D - D_i$ and D_i are often called the *training* and *test* data

sets respectively. Let $\hat{\mathbf{\Psi}}_{n,-i}$ be the MPLE of $\mathbf{\Psi}$ based on the training set. Further, let $l_{n,i}(\hat{\mathbf{\Psi}}_{n,-i})$ be the log-likelihood function evaluated on the test set D_i, using the MPLE $\hat{\mathbf{\Psi}}_{n,-i}$, for $i = 1, 2, \ldots, N$. Then, the cross-validation criterion is defined by

$$l_{CV}(\gamma_n) = -\frac{1}{N}\sum_{i=1}^{N} l_{n,i}(\hat{\mathbf{\Psi}}_{n,-i}).$$

The value of γ_n which minimizes $l_{CV}(\gamma_n)$ is chosen as a data-driven choice of γ_n. In particular, the five-fold CV (Zhang, 1993) can be used.

The generalized cross validation (GCV) is computationally cheaper than the CV criterion. The basic idea is to adjust some kind of goodness-of-fit criterion with the effective number of parameters employed in the model corresponding to the current tuning parameter. This method, however, is found not work as well as the simple CV in our simulation.

Using the CV (or GCV) criterion to choose the tuning parameter results in a random γ_n. To ensure the validity of the asymptotic results, a common practice is to place a restriction on the range of the tuning parameter. See for example, James, Priebe and Marchette (2001). The following result is obvious and the proof is omitted.

Theorem 6 *Consider the HARD or SCAD penalty functions given in Section 4. If the tuning parameter* $\lambda_n = \frac{\gamma_n}{\sqrt{n}}$ *is chosen by minimizing the CV or GCV over the interval* $[\alpha_n, \beta_n]$ *such that* $0 \le \alpha_n \le \beta_n$, *and* $\beta_n \to 0$ *and* $\sqrt{n}\alpha_n \to \infty$, *as* $n \to \infty$, *then the results in Theorems 1-5 still hold.*

Let $\alpha_n = C_1 n^{-1/4} \log n, \beta_n = C_2 n^{-1/4} \log n$, for some constants $0 < C1 < C2$. Then (α_n, β_n) meet the conditions in the above theorem.

6. Simulation study. The performance of the new method is compared with the two information-based criteria AIC and BIC and the Bayesian method of Ishwaran, James and Sun (2001) via simulations. We considered the problem of order selection in normal mixture in location parameter and Poisson mixtures. We used the SCAD penalty function in the new method. The simulation results are reported in terms of the estimated number of components of the mixture model, and based on 500 simulated data sets with sample size $n = 100$. The CV criterion were used to choose the tuning parameter γ_n.

Example 1 The density function of the normal mixture in location parameter in our simulation is given by

$$f(y; \mathbf{\Psi}) = \sum_{k=1}^{K} \frac{\pi_k}{\sigma} \, \phi(\frac{y - \theta_k}{\sigma})$$

where $\mathbf{\Psi} = (\sigma, \theta_1, \theta_2, \ldots, \theta_K, \pi_1, \pi_2, \ldots, \pi_{K-1})$, and $\phi(\cdot)$ is the density function for the standard normal $N(0, 1)$. We studied six normal mixtures specified in Ishwaran, James and Sun (2001). The first three mixtures have $K_0 = 2$ and the next three have $K_0 = 4$. The parameter settings are given in Table 1. The plots of mixture densities corresponding to all the experiments are given in Figure 1. A normal mixture model may not have its components appear graphically as separate modes (Figure 1) when their mean difference is smaller than 2σ.

We set $K = 4$ and $K = 8$ in data analysis for the first three and last three models respectively and we considered two cases: σ known ($\sigma = 1$) and unknown. The normal mixture model with unknown σ^2 does not fit into our theoretical development. Generalizing theoretical results is a very interesting but difficult problem which will be discussed further. The new

method can clearly be applied without any obstacles. The simulation results are reported in Tables 2 and 3. Entries in the last four columns are the percentages of times that a model with given candidate order was chosen out of 500 replicates. The values given in brackets correspond to the σ-unknown case. The values in the last column are quoted directly from Ishwaran, James and Sun (2001) based on their Bayesian method called the GWCR method, for the σ-unknown case.

When σ is known, the new method and the AIC and BIC methods have comparable and very good performances for the first three normal mixture models. When σ is unknown, the new method substantially out-performs all other methods. In particular, for the third mixture which has a single mode, the new method detects the correct model with rate as high 53.6% which is 2.3 times the next best. In the rest of mixture models, the new method outperforms all competitors by a big margin when σ is unknown, and is among the best when σ is known.

Example 2 The probability function of the Poisson finite mixture model in our simulation is given by

$$f(y;\mathbf{\Psi}) = \sum_{k=1}^{K} \pi_k \ \frac{\theta_k^y}{y!} \exp(-\theta_k)$$

where $\mathbf{\Psi} = (\theta_1, \theta_2, \ldots, \theta_K, \pi_1, \pi_2, \ldots, \pi_{K-1})$.
We studied two mixtures with $K_0 = 2$, and one with $K_0 = 4$. The parameter settings are given in Table 4.

In our simulation, we set $K = 4$ for the first two models, and $K = 8$ in the last model. The simulation results are reported in Table 5. Similar to Example 1, entries in the last three columns are the percentages of times

that a model with given candidate order is chosen out of 500 samples. It is obvious that the new method have much better performance than all other methods.

7. Application examples. In this section we analyze two well-known real data sets to further demonstrate the use of the new method.

Example 3 (Sodium-Lithium Countertransport (SLC) Data). Suppose that a trait such as blood pressure is determined by a simple mode of inheritance compatible with the action of a single gene with two alleles, A_1 and A_2, which occur with probabilities p and $1-p$. As discussed by Roeder (1994), a finite mixture of normal distributions with common variance is appropriate if each observation is composed of the sum of a genetic component Θ and a normally distributed measurement error. Consider two competing genetic models:

> Model I. (Simple dominance model) Genotypes A_1A_1 and A_1A_2 have phenotype θ_1, whereas A_2A_2 has phenotype θ_2. Hence $P(\Theta = \theta_1) = p^2 + 2p(1-p)$ and $P(\Theta = \theta_2) = (1-p)^2$.
>
> Model II. (Additive model) Each of the three genotypes yields a distinct phenotype with $P(\Theta = \theta_1) = p^2$, $P(\Theta = \theta_2) = 2p(1-p)$ and $P(\Theta = \theta_3) = (1-p)^2$. Furthermore, $\theta_1 < \theta_2 < \theta_3$ and $\theta_3 - \theta_2 = \theta_2 - \theta_1$.

As Roeder (1994) argued, red blood cell SLC is believed to follow one of the above two models. Geneticists are interested in SLC because it is correlated with blood pressure and hence may be an important cause of hypertension.

The data set considered in this example consists of red blood cell SLC activity measured on 190 individuals. Figure 2 gives a histogram of the SLC

measurements. Roeder (1994) fitted a mixture of normal of order three to this data. Her fit in fact corresponds to the additive model (model II above).

Using the new approach we fitted the following model

$$f(y;\hat{\boldsymbol{\Psi}}_n) = \frac{1}{0.57}\left\{0.75\,\boldsymbol{\phi}\left(\frac{y-2.21}{0.57}\right)+0.22\,\boldsymbol{\phi}\left(\frac{y-3.72}{0.57}\right)+0.03\,\boldsymbol{\phi}\left(\frac{y-5.64}{0.57}\right)\right\}$$

A plot of the above density is given in Figure 2. The figure also shows the density function of a mixture model with two components. As Roeder (1994) argued, the model with three components corresponds to the additive model with $\hat{\theta}_2 - \hat{\theta}_1 \approx \hat{\theta}_3 - \hat{\theta}_2$. Ishwaran, James and Sun (2001) also reported a model of order three.

Example 4 (Number of Death Notices Data). This data set has been discussed several times in the literature, see Hasselblad (1969), Titterington, Smith and Markov(1985) and Böhning (2000). The data are shown in Table 6. The table gives the numbers of death notices of women eighty years of age and over, appearing in *The Times* of London, on each day for three consecutive years, namely 1910-1912. Figure 3 shows a histogram of the observed data. Since the data are counts, one may initially think of fitting a homogeneous Poisson model to the data. The third column of Table 6 gives the expected frequency obtained from fitting a homogeneous Poisson model to the data. The Pearson χ^2-value of 26.97 provides strong evidence against the homogeneous model.

However, after a closer look at the data, we can see that the observed frequencies for 0, 1 and 2 death notices, compared with the rest, are inflated. Intuitively, this might be considered evidence for non-homogeneity of the distribution of the variable under study.

Hasselblad (1969) fitted a Poisson mixture model with two components to this data. Titterington, Smith and Markov (1985) commented that a Poisson mixture with two components fits the data quite well. Using the new penalized likelihood approach we also fitted a finite mixture of Poisson distributions to the data. We maximized the function $\tilde{l}_n(\mathbf{\Psi})$ over the space $\mathcal{M}_6$ of finite mixing distributions with at most six support points. We used the SCAD penalty function. The maximum was obtained at a finite mixing distribution with two components. The fitted mixture model is

$$f(y;\hat{\mathbf{\Psi}}) = 0.34\ \frac{e^{-1.23}\ (1.23)^y}{y!} + 0.64\ \frac{e^{-2.64}\ (2.64)^y}{y!}.$$

The fourth column of Table 6 gives the expected frequency obtained from fitting the above mixture model to the data. The Pearson χ^2-value of 1.29 shows that the Poisson mixture model fits the data quite well. Figure 4 shows the empirical density and two fitted densities: the homogenous Poisson and the Poisson mixture model with two components. We can see how well the Poisson mixture model fits the data. Titterington, Smith and Markov(1985) fitted a Poisson mixture model with order 2 which is very similar to ours. Böhning (2000) reported the nonparametric maximum likelihood estimate of the mixing distribution which has an additional third support point at zero with the small mass 0.0068. However, he pointed out that the difference in the log-likelihood function between the fitted models with orders 2 and 3 is negligable. The real-life interpretation of the above fitted mixture model is that there could be different patterns of death in winter and summer.

8. Conclusion and further discussion. We developed a new order selection method for finite mixture models. Under certain regularity conditions on the kernel density function, and with appropriate choice of the

penalty function $p_n(\cdot)$, the method results in consistent estimators for both mixing distribution, and the order of the mixture model.

An EM algorithm was outlined for the maximization problem involved together with a likelihood-based CV method for choosing the tuning parameters. The performance of the new method was investigated via simulations and compared with AIC, BIC and the Bayesian method of Ishwaran, James and Sun (2001). The simulation results indicated that the new method performs very well compared to these methods. We also analyzed two well-known data sets to further demonstrate the application of the new method. Our findings from these data sets are in agreement with the existing analysis in the literature.

We observe that in contrast to AIC and BIC methods where all candidate orders must be fitted, the new method fits a model with maximum possible number of components and achieves the aim of order selection via merging these components. Hence, the new method also has a major advantage in the computational simplicity.

Clearly, the new method is readily applicable to the mixture of multi-parameter models and to the mixture models with the presence of some structural parameters. The statistical methodology can be carried to more general cases easily. However, in the case $K_0 \geq 2$, the consistency result is obtained under an $n^{-1/4}$-convergence rate assumption. By changing the order of the tuning parameters in the penalty function, more general results are not hard to obtain but the results become tedious. We welcome other researchers to join our effort to work on this very interesting and challenging problem.

APPENDIX: Regularity Conditions and Proofs

To establish the asymptotic properties of the MPLE $\hat{G}_n$, some regularity conditions are needed on $f(y;\theta)$. The expectations in the regularity conditions are taken under the true distribution of the data with true mixing distribution G_0.

Regularity Conditions

A_1. (Wald's Integrability Conditions):

(i) $E(|\log f(y;\theta)|) < \infty$, $\forall\ \theta \in \Theta$.

(ii) There exists $\rho > 0$ such that for each $\theta \in \Theta$, $f(y;\theta,\rho)$ is measurable and $E(|\log f(y;\theta,\rho)|) < \infty$, where

$$f(y;\theta,\rho) = 1 + \sup_{|\theta'-\theta|\leq\rho} f(y;\theta').$$

A_2. (Smoothness) The kernel density $f(y;\theta)$ is differentiable with respect to $\theta \in \Theta$ to order 3. Furthermore, the derivatives $f^{(j)}(y;\theta)$ are jointly continuous in y and θ.

A_3. (Strong Identifiability) The finite mixture model is strongly identifiable. That is, for any $m \leq 2K$ distinct values $\theta_1, \theta_2, \ldots, \theta_m$,

$$\sum_{j=1}^{m}\{a_j f(y;\theta_j) + b_j f'(y;\theta_j) + c_j f''(y;\theta_j)\} = 0 \quad , \quad \forall y$$

implies that $a_j = b_j = c_j = 0$, for $j = 1, 2, \ldots, m$.

A_4. For $i = 1, 2, \ldots, n;\ j = 1, 2, 3$, define

$$U_{ij}(\theta_1,\theta_2) = \frac{f^{(j)}(Y_i;\theta_1)}{f(Y_i;\theta_2)}; \quad U_{ij}(\theta, G) = \frac{f^{(j)}(Y_i;\theta)}{f(Y_i;G)}.$$

There exist a small neighborhood for each support point of G_0 and a function $q(Y)$ with $E\{q^2(Y)\} < \infty$ such that for $\theta_1, \theta_2, \theta_1', \theta_2'$ in this neighborhood, we have

$$|U_{ij}(\theta_1, \theta_2) - U_{ij}(\theta_1', \theta_2')| \leq q^2(Y_i)\{|\theta_1 - \theta_1'| + |\theta_2 - \theta_2'|\}.$$

Furthermore, $U_{ij}(\theta, G_0)$ has finite second moment for all θ in the same neighborhood of support points of G_0.

A_5. For any two mixing distribution with support points in small neighborhood of those of G_0, there exists a function $q(Y)$ with $E\{q^2(Y)\} < \infty$ such that

$$\left|\frac{f(Y; G_1)}{f(Y; G_2)} - 1\right| \leq q(Y)\|G_1 - G_2\|.$$

Condition A_4 implies that the processes $n^{-1/2}\sum_{i=1}^{n} U_{ij}(\theta, G_0)$ $(j = 1, 2, 3)$ are tight in small neighbourhoods of the support points θ_{0k} and therefore are all of order $O_p(1)$.

Conditions A_1-A_5 also imply that the finite mixture model with known order K_0 satisfies the standard regularity conditions. Hence the ordinary maximum likelihood estimator of G (with K_0 known) is $\sqrt{n}$-consistency and asymptotically normal; see Lehman (1983) and Render and Walker (1984).

We establish a lemma first before the proof of Theorem 1.

Lemma 1 *Suppose the kernel density $f(y; \theta)$ satisfies regularity conditions A_1-A_4, and the penalty function $p_n(\eta)$ satisfies conditions P_0-P_1. If the true distribution of Y is homogeneous with density function $f(y; \theta_0)$, then the MPLE $\hat{G}_n$ has the properties*

(a) $\sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1)$,

(b) $\hat{\eta}_k = o_p(1)$, *for* $k = 1, 2, \ldots, K-1$.

Proof of Lemma 1: Let $\hat{\theta}_n$ be the usual MLE of θ when $K = 1$, and $\bar{G}_n$ be the usual MLE of G in $\mathcal{M}_K$. Recall that $\hat{G}_n$ denotes the MPLE of G in $\mathcal{M}_K$. Let

$$\tilde{R}_n = 2\{\tilde{l}_n(\hat{G}_n) - \tilde{l}_n(\hat{\theta}_n)\}$$

and

$$R_n = 2\{l_n(\bar{G}_n) - l_n(\hat{\theta}_n)\}.$$

It is clear that

$$0 \leq \tilde{R}_n \leq R_n.$$

By Dacunha-Castelle and Gassiat (1999), the ordinary likelihood ratio statistic $R_n = O_p(1)$ under certain conditions which are satisfied here. Consequently, we also have $\tilde{R}_n = O_p(1)$. From

$$0 \leq \tilde{R}_n + 2\left[\sum_{k=1}^{K-1} p_n(\hat{\eta}_k) - C_K \sum_{k=1}^{K} \log \hat{\pi}_k\right] \leq R_n,$$

we conclude that

$$\sum_{k=1}^{K-1} p_n(\hat{\eta}_k) - C_K \sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1). \tag{8}$$

Since both terms in (8) are non-negative, we must have

$$-C_K \sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1).$$

This proves (a).

Further, (8) implies that

$$p_n(\hat{\eta}_k) = O_p(1)\ ,\ k = 1, 2, \ldots, K-1.$$

Consequently, Conditions P_0 and P_1 on the penalty function $p_n(\cdot)$ imply that

$$\hat{\eta}_k = o_p(1) \ , \ k = 1, 2, \ldots, K-1.$$

This completes the proof. ♠

Result (b) in Lemma 1 shows that all $\hat{\eta}_k$ values converge to zero under homogeneous models. For the purpose of consistent order selection, the $\hat{\theta}_k$'s must be equal and converge to θ_0. These are the conclusions of Theorems 1 and 2 to be proved.

Proof of Theorem 1: Let us denote the Kullback-Leibler information as

$$H(G; \theta_0) = E_0 \left\{ \log \frac{f(Y; G)}{f(Y; \theta_0)} \right\}$$

where the expectation is under the true density $f(y; \theta_0)$. By Condition P_1 on $p_n(\cdot)$ and Condition A_4, we have

$$\frac{1}{n} \left\{ \tilde{l}_n(G) - \tilde{l}_n(\theta_0) \right\} \rightarrow H(G; \theta_0) \tag{9}$$

almost surely and uniformly over the compact parameter region $\pi_k \in [\delta_1, \delta_2]$, and $\theta_k \in \Theta$, for $k = 1, 2, \ldots, K$, and for any two constants $0 < \delta_1 < \delta_2 < 1$. Let

$$A = \{G \in \mathcal{M}_K : \pi_k \in [\delta_1, \delta_2], |\theta_k - \theta_0| > \delta, \eta_k < \delta, k = 1, 2, \ldots, K-1\}$$

for some $0 < \delta_1 < \delta_2 < 1$, $\delta > 0$. Note that G_0, which is a degenerate distribution with a single support point θ_0, does not belong to A.

Suppose that the claim of the theorem is not true. Due to the compactness of the parameter space Θ, and results (a)-(b) in Lemma 1, there must exist a corresponding subsequence n' of n such that

$$P(\hat{G}_{n'} \in A) > \varepsilon$$

for some constants $0 < \delta_1 < \delta_2 < 1$, $\delta > 0$ and $\varepsilon > 0$, and for all large n'. Hence

$$P\left\{ \frac{1}{n'}\left[\tilde{l}_{n'}(\hat{G}_{n'}) - \tilde{l}_{n'}(\theta_0) \right] = \sup_{G \in A} \frac{1}{n'}\left[\tilde{l}_{n'}(G) - \tilde{l}_{n'}(\theta_0) \right] \right\} > \varepsilon$$

for all large n'. On the other hand, for any $G \in A$, due to the strong identifiability, $H(G; \theta_0) < 0$. This implies that, from (9) and the above inequality,

$$P\left\{ \frac{1}{n'}\left[\tilde{l}_{n'}(\hat{G}_{n'}) - \tilde{l}_{n'}(\theta_0) \right] < 0 \right\} > \varepsilon$$

for all large n'. Thus, $\hat{G}_{n'}$ cannot be the maximizer of the function $\tilde{l}_n(G)$, which is a contradiction. ♠

We now get ready to prove Theorem 2. The following useful result is from Serfling (1980, page 253).

Lemma 2 *Let $g(y; \theta)$ be continuous at θ_0, uniformly in y. Let F be a distribution function for which $\int |g(y; \theta_0)| dF(y) < \infty$. Let $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)$ be a random sample from F and suppose that $T_n = T_n(\boldsymbol{Y})$ is a function of the sample such that $T_n \to \theta_0$ in probability. Then, also in probability, we have*

$$\frac{1}{n}\sum_{i=1}^{n} g(Y_i; T_n) \to E_0\{g(Y; \theta_0)\}.$$

Proof of Theorem 2: Note that the MLE $\hat{\theta}_n$ under homogeneous model satisfies $\hat{\theta}_n - \theta_0 = o_p(1)$. Theorem 1 shows that the MPLE $\hat{G}_n$ has all its support points converge to θ_0. Our strategy for the proof is to consider all mixing distributions $G \in \mathcal{M}_K$ with their support points in a small enough neighbourhood of $\hat{\theta}_n$. We show that among them, only those with equal θ_k can possibly be the MPLE.

For G with unequal θ_k's and in a small enough neighbourhood of $\hat{\theta}_n$, let us tentatively claim that

$$l_n(G) - l_n(\hat{\theta}_n) = O_p(n^{1/2}) \sum_{i<j}^{K} \pi_i \pi_j (\theta_i - \theta_j)^2. \tag{10}$$

If so, Condition P_2 on the penalty function $p_n(\cdot)$ implies that

$$\tilde{l}_n(G) - l_n(\hat{\theta}_n) \leq n^{1/2} \sum_{i<j} (\theta_i - \theta_j)^2 \left\{ O_p(1) - \frac{1}{|\theta_i - \theta_j|} \right\} < 0$$

in probability as $|\theta_i - \theta_j|$ is small. That is, none of the $G \in \mathcal{M}_K$, $K \geq 2$, can be the MPLE by definition and hence the conclusion of the theorem must be true.

Thus it suffices to prove (10). Define

$$\delta_i = \sum_{k=1}^{K} \pi_k \left[\frac{f(Y_i; \theta_k)}{f(Y_i; \hat{\theta}_n)} - 1 \right], \ i = 1, 2, \ldots, n.$$

We may then write

$$l_n(G) - l_n(\hat{\theta}_n) = \sum_{i=1}^{n} \log(1 + \delta_i).$$

By inequality $\log(1+x) \leq x - \frac{x^2}{2} + \frac{x^3}{3}$, we have

$$l_n(G) - l_n(\hat{\theta}_n) \leq \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} \frac{\delta_i^2}{2} + \sum_{i=1}^{n} \frac{\delta_i^3}{3}. \tag{11}$$

We study each term on the right-hand side of the above inequality separately.

Denote

$$\begin{aligned} m_1(\hat{\theta}_n) = m_1 &= \sum_{k=1}^{K} \pi_k \ (\theta_k - \hat{\theta}_n), \\ m_2(\hat{\theta}_n) = m_2 &= \sum_{k=1}^{K} \pi_k \ (\theta_k - \hat{\theta}_n)^2. \end{aligned}$$

Note that

$$m_2 - m_1^2 = \sum_{i<j}^{K} \pi_i \pi_j (\theta_i - \theta_j)^2$$

which is in fact the variance of the mixing distribution G.

By the standard Taylor's expansion, we have

$$\begin{aligned} \sum_{i=1}^{n} \delta_i &= m_1 \sum_{i=1}^{n} U_{i1}(\hat{\theta}_n, \hat{\theta}_n) + \frac{1}{2}\, m_2 \sum_{i=1}^{n} U_{i2}(\hat{\theta}_n, \hat{\theta}_n) \\ &\quad + \frac{1}{6} \left[\sum_{k=1}^{K} \pi_k (\theta_k - \hat{\theta}_n)^3 \sum_{i=1}^{n} U_{i3}(\xi_k, \hat{\theta}_n) \right] \end{aligned}$$

where ξ_k is between θ_k and $\hat{\theta}_n$, for $k = 1, 2, \ldots, K$.

Since $\hat{\theta}_n$ is the MLE under $K_0 = 1$, $\hat{\theta}_n - \theta_0 = O_p(n^{-1/2})$. Together with Condition A_4, it is simple to see that

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{i2}(\hat{\theta}_n, \hat{\theta}_n) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{i2}(\theta_0, \theta_0) \\ &\quad + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \{ U_{i2}(\hat{\theta}_n, \hat{\theta}_n) - U_{i2}(\theta_0, \theta_0) \} = O_p(1) \end{aligned}$$

and similarly

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} U_{i3}(\xi_k; \hat{\theta}_n) = O_p(1) \quad , \quad k = 1, 2, \ldots, K.$$

Thus there exists some constant C_0 such that for the first term in (11),

$$\sum_{i=1}^{n} \delta_i \leq C_0 \sqrt{n}\; m_2 \tag{12}$$

in probability.

Using Taylor's expansion again, we have

$$\begin{aligned}\sum_{i=1}^{n}\delta_i^2 &= \sum_{i=1}^{n}\Big\{m_1U_{i1}(\hat\theta_n,\hat\theta_n)+\frac{1}{2}\,m_2U_{i2}(\hat\theta_n,\hat\theta_n)\\ &\quad +\frac{1}{6}\Big[\sum_{k=1}^{K}\pi_k(\theta_k-\hat\theta_n)^3\;U_{i3}(\xi_{i,k},\hat\theta_n)\Big]\Big\}^2\\ &= (I)+(II)+(III)\end{aligned}$$

where $\xi_{i,k}$ is between θ_k and $\hat\theta_n$ for $k=1,2,\ldots,K$, and

$$\begin{aligned}(I) &= \sum_{i=1}^{n}\Big\{m_1U_{i1}(\hat\theta_n,\hat\theta_n)+\frac{1}{2}\,m_2U_{i2}(\hat\theta_n,\hat\theta_n)\Big\}^2,\\ (II) &= \frac{1}{36}\sum_{i=1}^{n}\Big\{\sum_{k=1}^{K}\pi_k(\theta_k-\hat\theta_n)^3\;U_{i3}(\xi_{i,k},\hat\theta_n)\Big\}^2,\\ (III) &= \frac{1}{3}\sum_{i=1}^{n}\Big\{m_1U_{i1}(\hat\theta_n,\hat\theta_n)+\frac{1}{2}\,m_2U_{i2}(\hat\theta_n,\hat\theta_n)\Big\}\Big\{\sum_{k=1}^{K}\pi_k(\theta_k-\hat\theta_n)^3\;U_{i3}(\xi_{i,k},\hat\theta_n)\Big\}.\end{aligned}$$

By Lemma 2, for $j=1,2$,

$$n^{-1}\sum U_{ij}^2(\hat\theta_n,\hat\theta_n)\to E_0\{U_{ij}^2(\theta_0,\theta_0)\}.$$

That is, $n^{-1}(I)$ with fixed m_1 and m_2 converges to quadratic form in (m_1,m_2) which is positive definite due to the strong identifiability condition. That is, for some positive constant $C_1<C_2$, we have

$$C_1n(m_1^2+m_2^2)\le(I)\le C_2n(m_1^2+m_2^2)$$

in probability. On the other hand, Condition A_4 implies that for some $\epsilon>0$,

$$(II)\le\epsilon\;nm_2^2\le\epsilon\;n\;(m_1^2+m_2^2)$$

in probability. From the above two inequalities and the Cauchy-Schwarz inequality, we further obtain

$$|(III)|\le\sqrt{\epsilon\;C_2}\;n\;(m_1^2+m_2^2)$$

in probability. Combining the above three inequalities, in probability, we conclude that for any small constant $\epsilon > 0$,

$$\sum_{i=1}^{n} \delta_i^2 \geq (C_1 - \sqrt{\epsilon\ C_2})\ n\ (m_1^2 + m_2^2). \tag{13}$$

We now work on the third term in (11). By Taylor's expansion,

$$\begin{aligned}
\sum_{i=1}^{n} \delta_i^3 &= \sum_{i=1}^{n} \left\{ m_1 U_{i1}(\hat{\theta}_n, \hat{\theta}_n) + \frac{1}{2} \sum_{k=1}^{K} \pi_k (\theta_k - \hat{\theta}_n)^2\ U_{i2}(\xi_{i,k}, \hat{\theta}_n) \right\}^3 \\
&\leq 8\ |m_1|^3\ \sum_{i=1}^{n} \left| U_{i1}(\hat{\theta}_n, \hat{\theta}_n) \right|^3 + \sum_{i=1}^{n} \left| \sum_{k=1}^{K} \pi_k (\theta_k - \hat{\theta}_n)^2\ U_{i2}(\xi_{i,k}, \hat{\theta}_n) \right|^3 \\
&\leq 8^{K-1}\ n\ \left\{ |m_1|^3 + \sum_{k=1}^{K} \pi_k^3 |\theta_k - \hat{\theta}_n|^6 \right\}
\end{aligned}$$

where $\xi_{i,k}$ is between θ_k and $\hat{\theta}_n$ for $k = 1, 2, \ldots, K$. Thus

$$\sum_{i=1}^{n} \delta_i^3 \leq \epsilon\ n\ (m_1^2 + m_2^2) \tag{14}$$

in probability. The inequalities in (13) and (14) imply that $\sum_{i=1}^{n} \delta_i^2$ dominates $\sum_{i=1}^{n} \delta_i^3$. Thus, from (11) we have

$$l_n(G) - l_n(\hat{\theta}_n) = \sum_{i=1}^{n} \log(1 + \delta_i) \leq \sum_{i=1}^{n} \delta_i - (\frac{1}{2} \sum_{i=1}^{n} \delta_i^2)\ (1 + o_p(1)).$$

Thus by using (12), (13) and the above inequality, and using some generic constants, we have

$$\begin{aligned}
l_n(G) - l_n(\hat{\theta}_n) &\leq C_0 \sqrt{n}\ m_2 - C_3\ n\ (m_1^2 + m_2^2) \\
&= C_0 \sqrt{n} (m_2 - m_1^2) - C_3\ n\ \{(m_1^2 + m_2^2) - \frac{C_0}{C_3 \sqrt{n}} m_1^2\} \\
&\leq C_0 \sqrt{n} (m_2 - m_1^2) \\
&= C_0 \sqrt{n} \sum_{i<j} \pi_i \pi_j (\theta_i - \theta_j)^2
\end{aligned}$$

in probability. Hence,

$$l_n(G) - l_n(\hat{\theta}_n) = O_p(\sqrt{n}) \sum_{i<j} \pi_i \pi_j (\theta_i - \theta_j)^2$$

which is (10) and this completes the proof of the theorem. ♠

In the proof of Lemma 4 bellow we need the following Lemma which is from the discussion part of the paper by Wald (1949, pages 601-602). In a simplistic way, the result states that the likelihood ratio decreases at exponential rate when a neighborhood of the true value is excluded in its definition.

Lemma 3 *Let η and ϵ be given, arbitrarily small, positive numbers. Let $S(\theta_0, \eta)$ be the open sphere with center θ_0 and radius η, and let $\Omega(\eta) = \Omega - S(\theta_0, \eta)$. Let Wald's Assumptions hold. There exists a number $h(\eta)$, $0 < h < 1$, and another positive number $N(\eta, \epsilon)$ such that, for any $n > N(\eta, \epsilon)$,*

$$P_0 \left\{ \frac{\sup_{\theta \in \Omega(\eta)} \prod_{i=1}^n f(Y_i; \theta)}{\prod_{i=1}^n f(Y_i; \theta_0)} > h^n \right\} < \epsilon$$

where P_0 is the probability of the relation in braces according to $f(y; \theta_0)$.

Lemma 4 *Suppose the kernel density $f(y; \theta)$ satisfies regularity conditions A_1-A_4, and the penalty function $p_n(\eta)$ satisfies conditions P_0, P_1 and P_3. If the true distribution of Y is a finite mixture with density function $f(y; G_0)$, then the MPLE $\hat{G}_n$ has the property*

$$\sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1) \quad , \quad as \quad n \to \infty.$$

Proof . By Lemma 3, the difference $l_n(G) - l_n(G_0)$ is negative with order n, uniformly for any G outside a neighbourhood of G_0. On the other hand, due to condition P_1 on the penalty function $p_n(\cdot)$, $\sum_{k=1}^{K-1} p_n(\eta_k) - \sum_{k=1}^{K_0-1} p_n(\eta_{0k}) =$

$o(n)$, where $\eta_k = \theta_{k+1} - \theta_k, k = 1, 2, \ldots, K-1$, correspond to the support points of the G. Thus, $\tilde{l}_n(G) - \tilde{l}_n(G_0)$ is negative also with order n, uniformly for any G outside a given neighbourhood of G_0. Hence, the MPLE $\hat{G}_n$ must be in a small neighborhood of G_0. This implies that $\hat{G}_n$ has at least K_0 distinct support points. Thus, by condition P_3 on the penalty function $p_n(\cdot)$, for large n,

$$\sum_{k=1}^{K-1} p_n(\hat{\eta}_k) - \sum_{k=1}^{K_0-1} p_n(\eta_{0k}) \geq 0 \tag{15}$$

in probability. Let $\bar{G}_n$ be the ordinary MLE of G which has at most K support points. By the definition of $\tilde{l}_n(G)$ and (15) we have that

$$\begin{aligned}
0 &\leq \tilde{l}_n(\hat{G}_n) - \tilde{l}_n(G_0) \\
&= \left\{ l_n(\hat{G}_n) - l_n(G_0) \right\} - \left\{ \sum_{k=1}^{K-1} p_n(\hat{\eta}_k) - \sum_{k=1}^{K_0-1} p_n(\eta_{0k}) \right\} \\
&\quad + \left\{ C_K \sum_{k=1}^{K} \log \hat{\pi}_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k} \right\} \\
&\leq \left\{ l_n(\hat{G}_n) - l_n(G_0) \right\} + \left\{ C_K \sum_{k=1}^{K} \log \hat{\pi}_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k} \right\} \\
&\leq \left\{ l_n(\bar{G}_n) - l_n(G_0) \right\} + \left\{ C_K \sum_{k=1}^{K} \log \hat{\pi}_k - C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k} \right\}.
\end{aligned}$$

From Dacunha-Castelle and Gassiat (1999), $l_n(\bar{G}_n) - l_n(G_0) = O_p(1)$. Also $C_K \sum_{k=1}^{K} \log \hat{\pi}_k$ is a negative quantity and $C_{K_0} \sum_{k=1}^{K_0} \log \pi_{0k}$ is constant with respect to n. Knowing that $0 \leq \tilde{l}_n(\hat{G}_n) - \tilde{l}_n(G_0)$ implies

$$C_K \sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1).$$

This completes the proof. ♠

Proof of Theorem 3. Part(a). Denote

$$H(G;G_0) = E_0\left\{\log\frac{f(Y;G)}{f(Y;G_0)}\right\}$$

where the expectation is under the true density $f(y;G_0)$. By condition P_1 of $p_n(\cdot)$ and Condition A_4, we have

$$\frac{1}{n}\left\{\tilde{l}_n(G) - \tilde{l}_n(G_0)\right\} \to H(G;G_0) \tag{16}$$

almost surely and uniformly over the compact space of the finite mixing distribution G. Denote the set

$$A = \left\{G \in \mathcal{M}_K;\ \pi_l \in [\delta_{1l}, \delta_{2l}], 1 \le l \le K, \|G_k - G_{0k}\| > \delta,\ \ |p_k - \pi_{0k}| > \delta, 1 \le k \le K_0\right\}$$

for some $0 < \delta_{1l} < \delta_{2l} < 1$ and $\delta > 0$. Note that $G_0 \notin A$. Suppose that the claim in part (a) of the theorem is not true. Then, in the light of Lemma 4 and compactness of the parameter space Θ, there must exist a subsequence $\hat{G}_{n'}$ of $\hat{G}_n$ such that

$$P(\hat{G}_{n'} \in A) > \epsilon$$

for some positive $\epsilon > 0$, and for large n'. Hence we have that

$$P\left\{\frac{1}{n'}\left[\tilde{l}_{n'}(\hat{G}_{n'}) - \tilde{l}_{n'}(G_0)\right] = \sup_{G\in A}\frac{1}{n'}\left[\tilde{l}_{n'}(G) - \tilde{l}_{n'}(G_0)\right]\right\} > \varepsilon$$

for all large n'. On the other hand for any $G \in A$, due to the identifiability condition A_4, $H(G, G_0) < 0$. This implies that, from (16) and the above inequality,

$$P\left\{\frac{1}{n'}\left[\tilde{l}_{n'}(\hat{G}_{n'}) - \tilde{l}_{n'}(G_0)\right] < 0\right\} > \varepsilon$$

for all large n'. Thus, $\hat{G}_{n'}$ cannot be the maximizer of the function $\tilde{l}_n(G)$, which is a contradiction. Hence, the result in part (a) holds.

Part(b). From Part(a)-(ii), we have that

$$|\hat{G}_k(\theta) - G_{0k}(\theta)| = |\hat{G}_k(\theta) - I(\theta_{0k} \leq \theta)| = o_p(1), \quad \forall\ \theta \in \Theta.$$

By Lemma 4, the mixing proportion on each support point of the MPLE $\hat{G}_n$ is positive in probability. These facts imply that the support points of $\hat{G}_k$ must converge to θ_{0k} in probability. ♠

Proof of Theorem 4. Let $\hat{G}_0$ be the maximizer of the penalized log-likelihood function $\tilde{l}_n(G)$ among those with exactly K_0 support points. We need only to show that in probability, for any mixing distribution $G \in \mathcal{M}_K$ in a $n^{-1/4}$-neighbourhood of G_0 and with true order larger than K_0, we must have

$$\Delta_n(K, K_0) = \tilde{l}_n(G) - \tilde{l}_n(\hat{G}_0) < 0 \tag{17}$$

as $n \to \infty$ and therefore they cannot be the MPLE. We proceed as follows.

For any G in the $n^{-1/4}$-neighbourhood of G_0, with at most K but more than K_0 support points, and with properties specified by Theorem 3, we write

$$G(\theta) = \sum_{k=1}^{K_0} p_k G_k(\theta). \tag{18}$$

Let $\tilde{G}_0$ be the maximizer of $\tilde{l}_n(\cdot)$ over the space of finite mixing distributions with exactly K_0 support points while the mixing proportions are fixed at $p_1, p_2, \ldots, p_{K_0}$ given in the above G. Since G is in a shrinking neighbourhood of G_0, so must be its corresponding parameters. In that sense, the support points of $\tilde{G}_0$ are also consistent estimators of the support points of the true mixing distribution G_0. By definition, $\tilde{l}_n(\tilde{G}_0) \leq \tilde{l}_n(\hat{G}_0)$ which implies

$$\Delta_n(K, K_0) = \tilde{l}_n(G) - \tilde{l}_n(\hat{G}_0) \leq \tilde{l}_n(G) - \tilde{l}_n(\tilde{G}_0) = \tilde{\Delta}_n(K, K_0). \tag{19}$$

Thus, our task can be replaced by showing $\tilde{\Delta}_n(K, K_0) < 0$.

It is seen that

$$\begin{aligned}\tilde{\Delta}_n(K, K_0) &= \left[l_n(G) - l_n(\tilde{G}_0)\right] - \left[\sum_{k=1}^{K-1} p_n(\eta_k) - \sum_{k=1}^{K_0-1} p_n(\tilde{\eta}_{0k})\right] \\ &+ \left[C_K \sum_{k=1}^{K} \log \pi_k - C_{K_0} \sum_{k=1}^{K_0} \log p_k\right].\end{aligned}$$

Since $K > K_0$ and by (18), each p_k is the sum of some mixing proportions π_j corresponding to G_k. Thus, the third term on the right-hand side of the above expression is negative. Therefore,

$$\tilde{\Delta}_n(K, K_0) \leq \left[l_n(G) - l_n(\tilde{G}_0)\right] - \left[\sum_{k=1}^{K-1} p_n(\eta_k) - \sum_{k=1}^{K_0-1} p_n(\tilde{\eta}_{0k})\right]. \tag{20}$$

We first investigate the second term in the above inequality. The quantities η_k can be divided into two groups; group one consists of differences of supports of G_k, and group two consists of differences between the largest support of G_k and the smallest support of G_{k+1}. By consistency, η_k's in the second group converge to their corresponding $\eta_{0k} \neq 0$. Thus, by Condition P_3 for $p_n(\cdot)$,

$$\sum_{k=1}^{K-1} p_n(\eta_k) - \sum_{k=1}^{K_0-1} p_n(\tilde{\eta}_{0k}) = \sum_{k=1}^{K_0} \sum_{j \in I_k} p_n(\eta_{j_k})$$

with probability approaching one, where I_k are pairs of neighboring support points of G_k.

For the first term in (20), similar to what we did earlier,

$$l_n(G) - l_n(\tilde{G}_0) \leq \sum_{i=1}^{n} \delta_i - \frac{1}{2} \sum_{i=1}^{n} \delta_i^2 + \frac{1}{3} \sum_{i=1}^{n} \delta_i^3,$$

with

$$\delta_i = \frac{f(Y_i; G) - f(Y_i; \tilde{G}_0)}{f(Y_i; \tilde{G}_0)} = \sum_{k=1}^{K_0} p_k \, \frac{f(Y_i; G_k) - f(Y_i; \tilde{\theta}_{0k})}{f(Y_i; \tilde{G}_0)}.$$

For θ such that $\theta - \tilde{\theta}_{0k} = o_p(n^{-1/4})$, we have

$$\begin{aligned}\sum_{i=1}^{n} \frac{f(Y_i;\theta) - f(Y_i;\tilde{\theta}_{0k})}{f(Y_i;\tilde{G}_0)} &= (\theta - \tilde{\theta}_{0k}) \sum_{i=1}^{n} U_{i1}(\tilde{\theta}_{0k}, \tilde{G}_0) + \frac{1}{2}(\theta - \tilde{\theta}_{0k})^2 \sum_{i=1}^{n} U_{i2}(\tilde{\theta}_{0k}, \tilde{G}_0) \\ &+ \frac{1}{6}(\theta - \tilde{\theta}_{0k})^3 \sum_{i=1}^{n} U_{i3}(\xi_k, \tilde{G}_0)\end{aligned}$$

for some ξ_k between θ and $\tilde{\theta}_{0k}$. Letting $m_j(\tilde{\theta}_k) = m_{jk} = \int (\theta - \tilde{\theta}_{0k})^j dG_k(\theta)$ for $j = 1, 2, 3$, we get the expansion

$$\begin{aligned}\sum_{i=1}^{n} \frac{f(Y_i;G_k) - f(Y_i;\tilde{\theta}_{0k})}{f(Y_i;\tilde{G}_0)} &= m_{1k} \sum_{i=1}^{n} U_{i1}(\tilde{\theta}_{0k};\tilde{G}_0) + \frac{1}{2}\, m_{2k} \sum_{i=1}^{n} U_{i2}(\tilde{\theta}_{0k};\tilde{G}_0) \\ &+ \frac{1}{6} \int (\theta - \tilde{\theta}_{0k})^3 \sum_{i=1}^{n} U_{i3}(\xi_k;\tilde{G}_0)\; dG_k(\theta)\end{aligned}$$

for $k = 1, 2, \ldots, K_0$. Therefore,

$$\begin{aligned}\sum_{i=1}^{n} \delta_i &= \sum_{k=1}^{K_0} p_k \left\{ m_{1k} \sum_{i=1}^{n} U_{i1}(\tilde{\theta}_{0k}, \tilde{G}_0) + \frac{1}{2}\, m_{2k} \sum_{i=1}^{n} U_{i2}(\tilde{\theta}_{0k}, \tilde{G}_0) \right. \\ &+ \left. \frac{1}{6} \int (\theta - \tilde{\theta}_{0k})^3 \sum_{i=1}^{n} U_{i3}(\xi_k;\tilde{G}_0)\; dG_k(\theta) \right\}. \qquad (21)\end{aligned}$$

Since $\tilde{G}_0$ is the MPLE with K_0 support points, it must satisfy the following (score-type) equations:

$$\begin{aligned}\sum_{i=1}^{n} p_1\; U_{i1}(\tilde{\theta}_{01};\tilde{G}_0) + p_n'(\tilde{\eta}_{01}) &= 0, \\ \sum_{i=1}^{n} p_{K_0}\; U_{i1}(\tilde{\theta}_{0K_0};\tilde{G}_0) - p_n'(\tilde{\eta}_{0,K_0-1}) &= 0,\end{aligned}$$

and for $k = 2, 3, \ldots, K_0 - 1$,

$$\sum_{i=1}^{n} p_k\; U_{i1}(\tilde{\theta}_{0k};\tilde{G}_0) - p_n'(\tilde{\eta}_{0,k-1}) + p_n'(\tilde{\eta}_{0k}) = 0.$$

By the consistency of $\tilde{G}_0$, we have

$$\tilde{\eta}_{0k} = \tilde{\theta}_{0k} - \tilde{\theta}_{0,k-1} \to \eta_{0k} \neq 0$$

in probability, which implies, with probability tending to one, $p_n'(\tilde{\eta}_{0k}) = 0$ by condition P_3 on $p_n(\cdot)$. The score-type equations hence reduce to

$$\sum_{i=1}^{n} U_{i1}(\tilde{\theta}_{0k}; \tilde{G}_0) = 0$$

for all $k = 1, 2, \ldots, K_0$. This fact then simplifies (21) into

$$\sum_{i=1}^{n} \delta_i = \sum_{k=1}^{K_0} p_k \left\{ \frac{1}{2}\, m_{2k} \sum_{i=1}^{n} U_{i2}(\tilde{\theta}_{0k}; \tilde{G}_0) + \frac{1}{6} \int (\theta - \tilde{\theta}_{0k})^3 \sum_{i=1}^{n} U_{i3}(\xi_k; \tilde{G}_0)\, dG_k(\theta) \right\}$$

in probability. Note that

$$\sum_{i=1}^{n} U_{i2}(\tilde{\theta}_{0k}; \tilde{G}_0) = \sum_{i=1}^{n} U_{i2}(\tilde{\theta}_{0k}; G_0) + \sum_{i=1}^{n} \{U_{i2}(\tilde{\theta}_{0k}; \tilde{G}_0) - U_{i2}(\tilde{\theta}_{0k}; G_0)\}.$$

It is seen that the first term is $\sum_{i=1}^{n} U_{i2}(\tilde{\theta}_{0k}; G_0) = O_p(n^{1/2})$. For the second term, we have

$$\begin{aligned} \sum_{i=1}^{n} |U_{i2}(\tilde{\theta}_{0k}; \tilde{G}_0) - U_{i2}(\tilde{\theta}_{0k}; G_0)| &= \sum_{i=1}^{n} |U_{i2}(\tilde{\theta}_{0k}; G_0)| \left\{ \left| \frac{f(Y_j; G_0)}{f(Y_j; \tilde{G}_0)} - 1 \right| \right\} \\ &\leq \sum_{i=1}^{n} q(Y_i) |U_{i2}(\tilde{\theta}_{0k}; G_0)| \|G_0 - \tilde{G}_0\| \\ &= O_p(n^{3/4}) \end{aligned}$$

by Conditions A_4 and A_5. Hence,

$$\sum_{i=1}^{n} U_{i2}(\tilde{\theta}_{0k}; \tilde{G}_0) = O_p(n^{3/4}).$$

Similarly,

$$\sum_{i=1}^{n} \tilde{U}_{i3}(\xi_k; \tilde{G}_0) = O_p(n^{3/4}).$$

Thus for large n, there exist some constant C_0 such that

$$\sum_{i=1}^{n} \delta_i \leq C_0 \; n^{3/4} \sum_{k=1}^{K_0} p_k m_{2k} \tag{22}$$

in probability.

Now we focus on the quadratic term $\sum_{i=1}^{n} \delta_i^2$. By Taylor's expansion,

$$\begin{aligned}
\sum_{i=1}^{n} \delta_i^2 &= \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K_0} p_k \left[m_{1k} U_{i1}(\tilde{\theta}_{0k}; \tilde{G}_0) + \frac{1}{2} m_{2k} U_{i2}(\tilde{\theta}_{0k}; \tilde{G}_0) + \right. \right. \\
&\qquad \left. \left. \frac{1}{6} \int (\theta - \tilde{\theta}_{0k})^3 U_{i3}(\xi_{ik}; \tilde{G}_0) dG_k(\theta) \right] \right\}^2 \\
&= (I) + (II) + (III)
\end{aligned}$$

where

$$\begin{aligned}
(I) &= \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K_0} p_k \left[\tilde{m}_{1k} U_{i1}(\tilde{\theta}_{0k}; \tilde{G}_0) + \frac{1}{2} \tilde{m}_{2k} U_{i2}(\tilde{\theta}_{0k}; \tilde{G}_0) \right] \right\}^2, \\
(II) &= \frac{1}{36} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K_0} p_k \int (\theta - \tilde{\theta}_{0k})^3 U_{i3}(\xi_{ik}; \tilde{G}_0) \; dG_k(\theta) \right\}^2, \\
(III) &= \frac{1}{3} \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K_0} p_k \left[\tilde{m}_{1k} U_{i1}(\tilde{\theta}_{0k}; \tilde{G}_0) + \frac{1}{2} \tilde{m}_{2k} \tilde{U}_{i2}(\tilde{\theta}_{0k}; \tilde{G}_0) \right] \right\} \\
&\quad \times \left\{ \sum_{k=1}^{K_0} p_k \int (\theta - \tilde{\theta}_{0k})^3 \tilde{U}_{i3}(\xi_{ik}; \tilde{G}_0) \; dG_k(\theta) \right\}.
\end{aligned}$$

Using completely the same arguments as in the proof of Theorem 2, it is seen that there exist some positive constants C_1 and C_2 such that

$$C_1 \; n \sum_{k=1}^{K_0} (m_{1k}^2 + m_{2k}^2) \leq (I) \leq C_2 \; n \sum_{k=1}^{K_0} (m_{1k}^2 + m_{2k}^2)$$

$$(II) \leq \epsilon \; n \sum_{k=1}^{K_0} (m_{1k}^2 + m_{2k}^2)$$

and

$$|(III)| \leq \sqrt{C_2\epsilon}\ n\ \sum_{k=1}^{K_0}(m_{1k}^2 + m_{2k}^2).$$

Thus combining the above inequalities, we have

$$\sum_{i=1}^{n}\delta_i^2 \geq (C_1 - \sqrt{\epsilon\ C_2})\ n\ \sum_{k=1}^{K_0}(\tilde{m}_{1k}^2 + \tilde{m}_{2k}^2) \tag{23}$$

in probability. It further implies

$$l_n(G) - l_n(\tilde{G}_0) = \sum_{i=1}^{n}\log(1+\delta_i) \leq \sum_{i=1}^{n}\delta_i - (\frac{1}{2}\sum_{i=1}^{n}\delta_i^2)\ (1 + o_p(1)).$$

Substituting order assessments we have obtained, for some generic constant C,

$$l_n(G) - l_n(\tilde{G}_0) \leq C\ n^{3/4}\sum_{k=1}^{K_0}\sum_{i<j}(\theta_{ik} - \theta_{jk})^2 \leq C\ n^{1/2}\sum_{k=1}^{K_0}\sum_{i<j}|\theta_{ik} - \theta_{jk}|$$

in probability. Thus, we get

$$\tilde{\Delta}_n(K, K_0) = C_0\ \sqrt{n}\sum_{k=1}^{K_0}\sum_{i<j}|\theta_{ik} - \theta_{jk}| - \sum_{k=1}^{K_0}\sum_{j\in I_k}p_n(\eta_{j_k})$$

in probability. Condition P_2 on $p_n(\cdot)$ is designed to make the right-hand side of the above inequality negative for large n. Thus by (19),

$$\Delta_n(K, K_0) \leq \tilde{\Delta}_n(K, K_0) < 0$$

for large n. This completes the proof. ♠

Proof of Theorem 5. Let $r_n = n^{-1/2}(1 + b_n)$. It suffices to show that for any given $\varepsilon > 0$, there exists a constant M_ε such that

$$P\left\{\sup_{\|\boldsymbol{u}\|=M_\varepsilon}\tilde{l}_n(\boldsymbol{\Psi}_0 + r_n\boldsymbol{u}) < \tilde{l}_n(\boldsymbol{\Psi}_0)\right\} > 1 - \varepsilon. \tag{24}$$

This implies that with probability at least $1-\varepsilon$, a local maximum of the function is in the ball $\{\boldsymbol{\Psi}_0 + r_n\boldsymbol{u}; \|\boldsymbol{u}\| \leq M_\varepsilon\}$. Thus this local maximizer satisfies (7).

Let $\Delta_n(\boldsymbol{u}) = \tilde{l}_n(\boldsymbol{\Psi}_0 + r_n\boldsymbol{u}) - \tilde{l}_n(\boldsymbol{\Psi}_0)$. By definition of the penalized log-likelihood function $\tilde{l}_n(\cdot)$,

$$\Delta_n(\boldsymbol{u}) \leq \{l_n(\boldsymbol{\Psi}_0+r_n\boldsymbol{u})-l_n(\boldsymbol{\Psi}_0)\}-\sum_{k=1}^{K_0-1}\{p_n(\eta_{0k}+r_nu_k)-p_n(\eta_{0k})\}-C_{K_0}\sum_{k=1}^{K_0}\log\pi_{0k}. \tag{25}$$

By the standard Taylor's expansion, we have

$$l_n(\boldsymbol{\Psi}_0+r_n\boldsymbol{u})-l_n(\boldsymbol{\Psi}_0) = n^{-1/2}(1+b_n)\,[l_n'(\boldsymbol{\Psi}_0)]^\tau\,\boldsymbol{u}-\frac{(1+b_n)^2}{2}[\boldsymbol{u}^\tau I(\boldsymbol{\Psi}_0)\boldsymbol{u}]\,(1+o_p(1)),$$

$$\left|\sum_{k=1}^{K_0-1}\{p_n(\eta_{0k}+r_nu_k)-p_n(\eta_{0k})\}\right| \leq \sqrt{K_0-1}\,b_n(1+b_n)\|\boldsymbol{u}\|+\frac{c_n}{2}(1+b_n)^2\|\boldsymbol{u}\|^2.$$

By the standard regluarity conditions, $l_n'(\boldsymbol{\Psi}_0) = O_p(\sqrt{n})$ and that $I(\boldsymbol{\Psi}_0)$ is positive definite. In addition, $c_n = o(1)$. An order comparison of the terms in the above two expressions implies that

$$-\frac{1}{2}(1+b_n)^2[\boldsymbol{u}^\tau I(\boldsymbol{\Psi}_0)\boldsymbol{u}](1+o_p(1))$$

which is the sole leading term on the right-hand side of (25). Therefore, for any given $\varepsilon > 0$, there exists a sufficiently large M_ε such that

$$\lim_{n\to\infty} P\left\{\sup_{\|\boldsymbol{u}\|=M_\varepsilon}\Delta_n(\boldsymbol{u}) < 0\right\} > 1-\epsilon$$

which implies (24), and this completes the proof. ♠

REFERENCES

Akaike, H. (1973), Information Theory and an Extension of the Maximum Likelihood Principle, in *Second International Symposium on Information Theory*, eds. B.N. Petrox and F. Caski. Budapest: Akademiai Kiado, page 267.

Böhning, D. (2000). *Computer-Assisted Analysis of Mixtures and Applications: Meta Analysis, Disease Mapping and Others.* New York: Chapman & Hall/CRC.

Chambaz, A. (2006), Testing the order of a model, *Ann. Statist.*, **34**, 2350-2383.

Charnigo, R. and Sun, J. (2004). Testing homogeneity in a mixture distribution via the L^2-distance between competing models. *J. Amer. Statist. Assoc.*, **99**, 488-498.

Chen, H. and Chen, J. (2001). The likelihood ratio test for homogeneity in the finite mixture models. *Canad. J. Statist.*, **29**, 201-216.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2001). A modified likelihood ratio test for homogeneity in finite mixture models. *J. Roy. Statist. Soc. Ser. B*, **63**, 19-29.

Chen, H., Chen, J. and Kalbfleisch, J. D. (2004). Testing for a finite mixture model with two components. *J. Roy. Statist. Soc. Ser. B*, **66**, 95-115.

Chen, J. (1995). Optimal rate of convergence in finite mixture models. *Ann. Statist.*, **23**, 221-234.

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

Craven, P., Wahba, G. (1979), Smoothing noisy data with Spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation, *Numerische Mathematika*, **31**, 377-403.

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.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977), Maximum likelihood from incomplete data via the EM algorithm, (with discussion), *J. Roy. Statist. Soc. Ser. B*, **39**, 1-38.

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

Fan, J. and Li, R. (2002), Variable selection for Cox's proportional hazards model and frailty model, *Ann. Statist.*, **30**, 74-99.

Ghosh, J. K. and Sen, P. K. (1985). On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results, in *Proc. Berkeley Conf. in Honor of J. Neyman and Kiefer, Volume 2*, eds L. LeCam and R. A. Olshen, 789-806.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *J. Amer. Statist. Assoc.*, **64**, 1459-1471.

Ishwaran, H., James, L. F., Sun, J. (2001), Bayesian model selection in finite mixtures by marginal density decompositions, *J. Amer. Statist. Assoc.*, **96**, 1316-1332.

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

Lehmman, E. L. (1983). *Theory of Point Estimation.* New York: Wiley.

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

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistics for the number of components in a normal mixture. *Appl. Statist.*, **36**, 318-324.

Neyman, J. and Scott, E. L. (1966). On the use of $C(\alpha)$ optimal tests of composite hypothesis. *Bull. Inst. Int. Statist.*, **41(I)**, 477-497.

Render, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195-239.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.*, **89**, 487-500.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.*, **6**, 461-464.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Stone, M. (1974), Cross-validatory choice and assessment of statistical predictions, (With discussion), *J. Roy. Statist. Soc. Ser. B*, **36**, 111-147.

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

Titterington, D. M., Smith, A. F. M., and Markov, U. E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: Wiley.

Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Ann. Math. Statist.*, **20**, 595-602.

Zhang, P. (1993), Model selection via multifold cross-validation, *Ann. Statist.*, **21**, 229-231.

Table 1: Parameter values in Example 1.

	Parameter Values			
Model	(π_1, θ_1)	(π_2, θ_2)	(π_3, θ_3)	(π_4, θ_4)
1	(1/3, 0)	(2/3, 3)		
2	(0.5, 0)	(0.5, 3)		
3	(0.5, 0)	(0.5, 1.8)		
4	(0.25, 0)	(0.25, 3)	(0.25, 6)	(0.25, 9)
5	(0.25, 0)	(0.25, 1.5)	(0.25, 3)	(0.25, 4.5)
6	(0.25, 0)	(0.25, 1.5)	(0.25, 3)	(0.25, 6)

Table 2: Simulation results of Example 1 (Models 1-3).

Model	K_0	# Modes	K	AIC	BIC	NEW	GWCR
1	2	2	1	0.000 (0.024)	0.000 (0.150)	0.006 (0.010)	(0.018)
			2	0.952 (0.862)	0.994 (0.838)	0.988 (0.966)	(0.920)
			3	0.048 (0.072)	0.006 (0.012)	0.006 (0.024)	(0.058)
			4	0.000 (0.042)	0.000 (0.000)	0.000 (0.000)	(0.004)
2	2	2	1	0.000 (0.028)	0.000 (0.224)	0.006 (0.026)	(0.030)
			2	0.962 (0.874)	0.996 (0.772)	0.988 (0.918)	(0.916)
			3	0.036 (0.054)	0.004 (0.004)	0.006 (0.054)	(0.054)
			4	0.002 (0.044)	0.000 (0.000)	0.000 (0.002)	(0.000)
3	2	1	1	0.006 (0.668)	0.062 (0.950)	0.038 (0.392)	(0.868)
			2	0.978 (0.234)	0.938 (0.048)	0.924 (0.536)	(0.130)
			3	0.016 (0.052)	0.000 (0.002)	0.038 (0.072)	(0.002)
			4	0.000 (0.046)	0.000 (0.000)	0.000 (0.000)	(0.000)

The values in brackets are results for σ-unknown case.

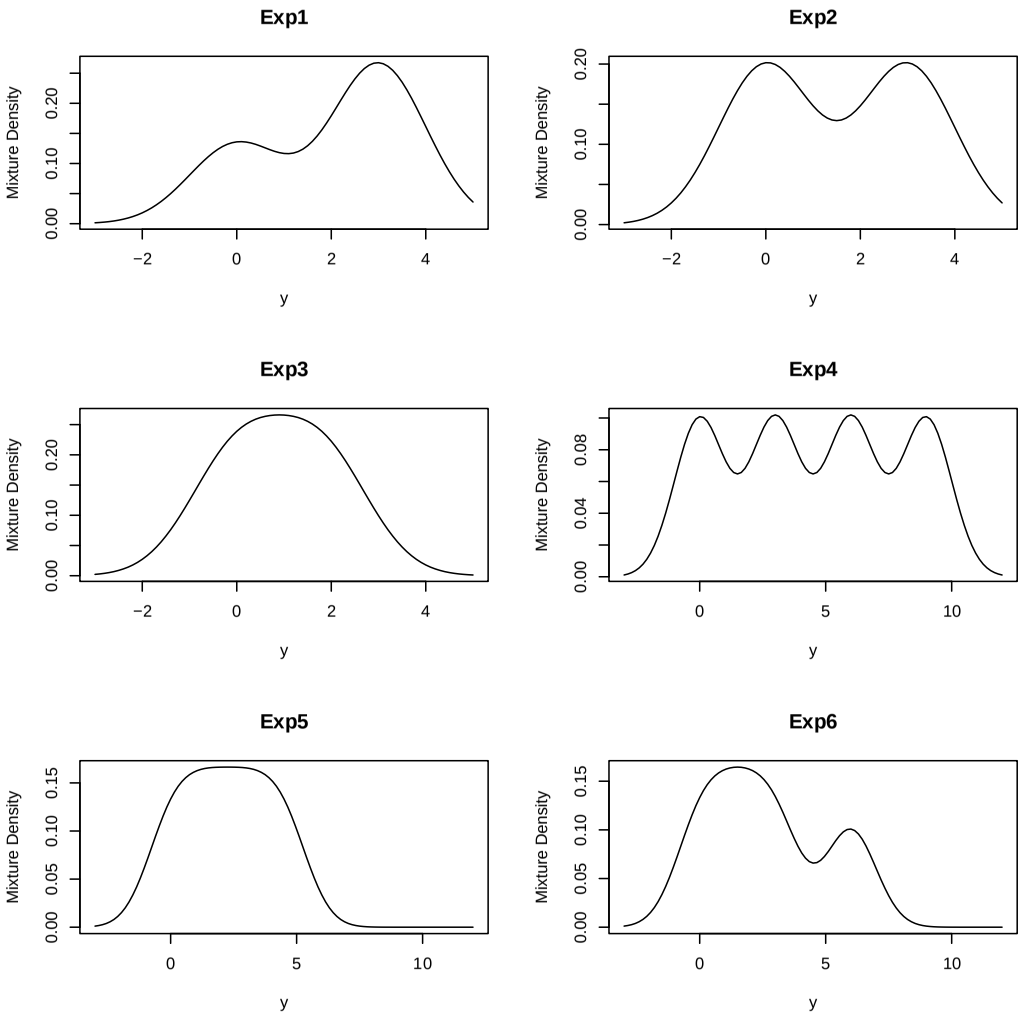

Figure 1: The mixture densities.

Table 3: Simulation results of Example 1 (Models 4-6).

Model	K_0	# Modes	K	AIC	BIC	NEW	GWCR
4	4	4	1	0.000 (0.000)	0.000 (0.110)	0.000 (0.000)	(0.000)
			2	0.000 (0.178)	0.000 (0.596)	0.000 (0.044)	(0.102)
			3	0.008 (0.110)	0.076 (0.110)	0.044 (0.154)	(0.554)
			4	0.976 (0.674)	0.924 (0.182)	0.908 (0.772)	(0.306)
			5	0.016 (0.038)	0.000 (0.002)	0.048 (0.030)	(0.038)
			6	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	(0.000)
			7	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	(0.000)
			8	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	(0.000)
5	4	1	1	0.000 (0.244)	0.000 (0.748)	0.000 (0.066)	(0.144)
			2	0.284 (0.556)	0.670 (0.246)	0.046 (0.450)	(0.818)
			3	0.704 (0.142)	0.330 (0.004)	0.744 (0.374)	(0.032)
			4	0.012 (0.044)	0.000 (0.002)	0.210 (0.092)	(0.006)
			5	0.000 (0.014)	0.000 (0.000)	0.000 (0.018)	(0.000)
			6	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	(0.000)
			7	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	(0.000)
			8	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	(0.000)
6	4	2	1	0.000 (0.016)	0.000 (0.188)	0.000 (0.006)	(0.000)
			2	0.006 (0.474)	0.036 (0.698)	0.020 (0.288)	(0.612)
			3	0.944 (0.392)	0.960 (0.106)	0.818 (0.572)	(0.368)
			4	0.050 (0.102)	0.004 (0.008)	0.158 (0.114)	(0.020)
			5	0.000 (0.014)	0.000 (0.000)	0.004 (0.018)	(0.000)
			6	0.000 (0.000)	0.000 (0.000)	0.000 (0.002)	(0.000)
			7	0.000 (0.002)	0.000 (0.000)	0.000 (0.000)	(0.000)
			8	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	(0.000)

Table 4: Parameter values in Poisson mixture model.

	Parameter Values			
Experiment	(π_1, θ_1)	(π_2, θ_2)	(π_3, θ_3)	(π_4, θ_4)
1	(1/3, 4)	(2/3, 6)		
2	(0.5, 4)	(0.5, 6)		
3	(0.25, 4)	(0.25, 6)	(0.25, 10)	(0.25, 15)

Table 5: Simulation Results for Poisson Mixture Models

Model	K_0	K	AIC	BIC	NEW
1	2	1	0.724	0.958	0.462
		2	0.274	0.042	0.532
		3	0.002	0.000	0.006
		4	0.000	0.000	0.000
2	2	1	0.684	0.938	0.450
		2	0.316	0.062	0.544
		3	0.000	0.000	0.006
		4	0.000	0.000	0.000
3	4	1	0.000	0.000	0.000
		2	0.706	0.940	0.112
		3	0.290	0.060	0.608
		4	0.004	0.000	0.238
		5	0.000	0.000	0.040
		6	0.000	0.000	0.002
		7	0.000	0.000	0.000
		8	0.000	0.000	0.000

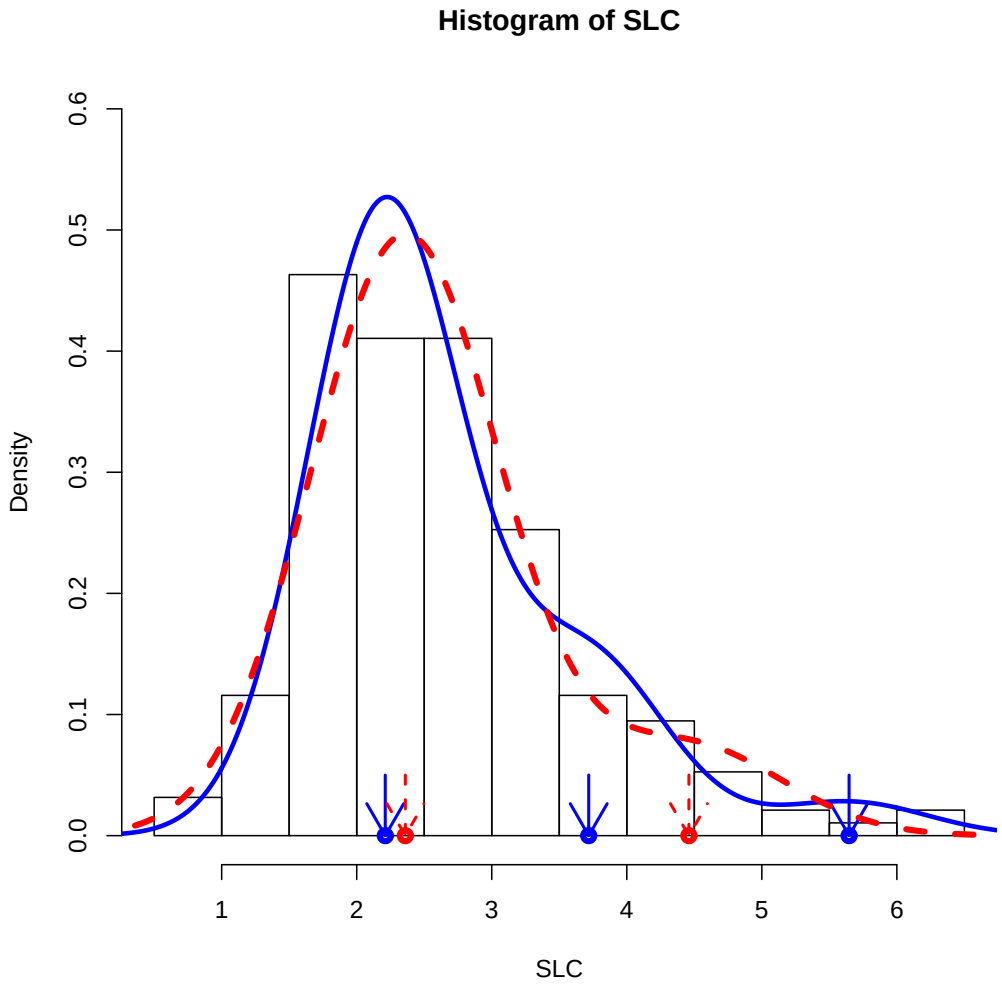

Figure 2: Dashed line: Fitted normal mixture model of order two; Solid line: Fitted normal mixture model of order three.

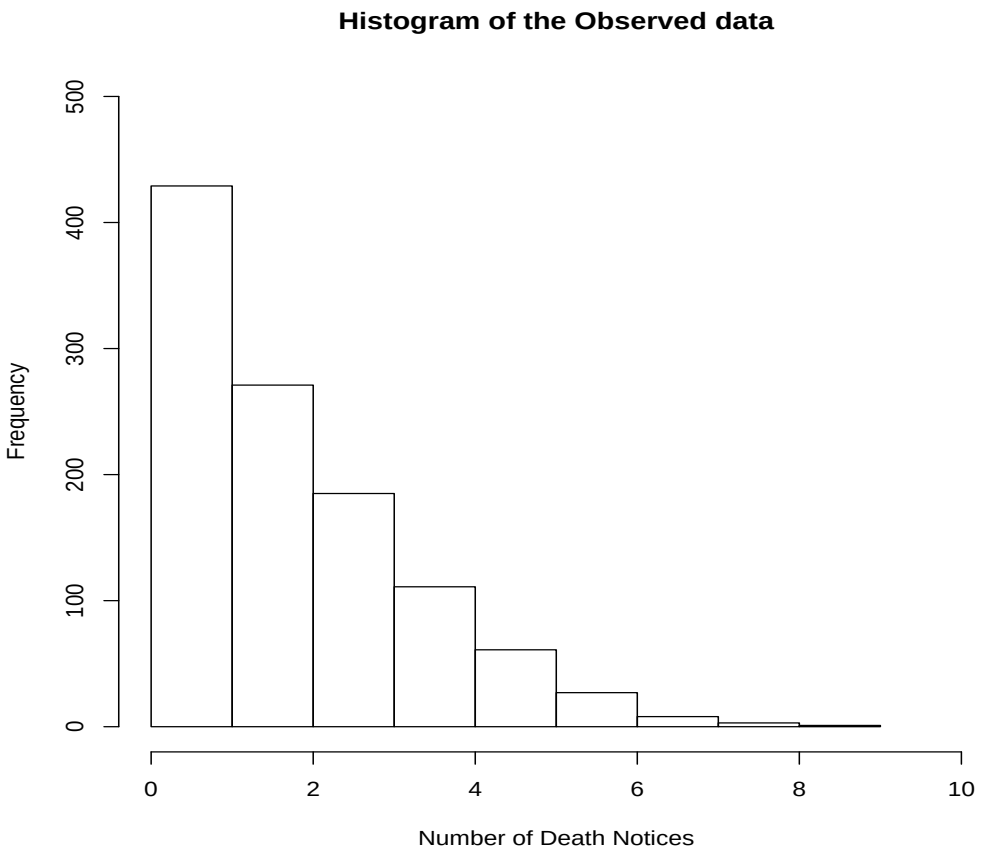

Figure 3: Histogram of the observed frequency of number of death notices.

Table 6: Number of death notices and the results of fitting two models to the data: a homogeneous Poisson and the Poisson mixture fitted by the new method.

Number of Death Notices	Observed Frequency	Expected Frequency Homogeneous Poisson	Expected Frequency Poisson Mixture
0	162	126.78	160.77
1	267	273.46	270.09
2	271	294.92	261.97
3	185	212.04	191.97
4	111	114.34	114.94
5	61	49.32	57.83
6	27	17.73	24.88
7	8	5.46	9.29
8	3	1.47	3.05
9	1	0.35	0.89
		$\chi^2_6 = 26.97$	$\chi^2_4 = 1.29$

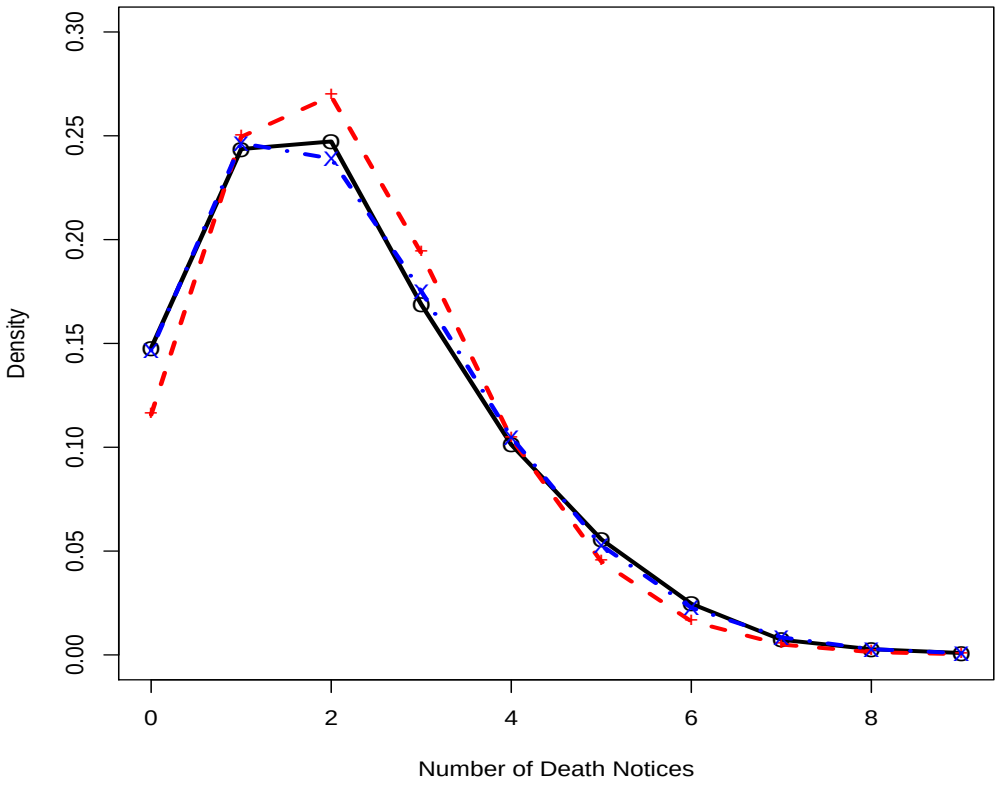

Figure 4: Empirical density: Solid line (O); Estimated Poisson density: dashed line (+); Estimated Poisson mixture density: dashed-dot line (X).