ORDER SELECTION IN FINITE MIXTURE MODELS WITH A NON-SMOOTH PENALTY

Jiahua Chen and Abbas Khalili *

September 2, 2008

Abstract

Order selection is a fundamental and challenging problem in the application of finite mixture models. In this paper, we develop a new penalized likelihood approach. The new method (MSCAD) deviates from information-based methods such as AIC and BIC by introducing two penalty functions which depend on the mixing proportions and the component parameters. It is consistent at estimating both the order of the mixture model and the mixing distribution. Simulations show that MSCAD has much

*Jiahua Chen is Professor, Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z2 (Email: *jhchen@stat.ubc.ca*). Abbas Khalili is Postdoctoral Researcher, Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z2 (Email: *khalili@stat.ubc.ca*).
The authors thank the editor, associate editor and the referees for their insightful comments and suggestions. The authors also thank Bradley Efron in the department of Statistics at Stanford University for providing the data for the HIV study in Example 1. The research is supported by the Natural Science and Engineering Research Council of Canada and by MITACS.

better performance than a number of existing methods. Two real-data examples are examined to illustrate the performance of MSCAD.

KEY WORDS: EM algorithm, Finite mixture model, Penalty method, SCAD.

1. INTRODUCTION

Order selection is a fundamental and challenging problem in the application of finite mixture models. A mixture model with a large number of components provides a good fit but may have poor interpretive value. Complex models are not favored in applications for the sake of parsimony and the prevention of over-fitting.

Many statistical methods have been proposed in the past 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 is based on distance measures between the fitted model and the non-parametric estimate of the population distribution; see Chen and Kalbfleisch (1996), James, Priebe and Marchette (2001), and Woo and Sriram (2006, 2007). One may also consider hypothesis testing 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, such as Ghosh and Sen (1985), McLachlan (1987), Dacunha-Castelle and Gassiat (1999), Chen and Chen (2001), and Chen, Chen, and Kalbfleisch (2001, 2004). Charnigo and Sun (2004) proposed an L^2-distance method for testing homogeneity in continuous finite mixture models. Chambaz (2006) studied the asymptotic efficiency of

two generalized likelihood ratio tests. Ishwaran, James, and Sun (2001) proposed a Bayesian approach. Ray and Lindsay (2008) investigated model selection in multivariate mixtures.

In this paper, we develop a new order selection method combining the strength of two existing statistical methods. The first is the Modified likelihood proposed by Chen and Kalbfleisch (1996). The second is the variable selection method called the smoothly clipped absolute deviation or SCAD, by 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. The penalty introduced by the modified likelihood clusters the fitted subpopulations around the true subpopulations. A SCAD-type penalty will merge each cluster of subpopulations into a single subpopulation. Hence, our procedure starts with a large number of subpopulations and obtains a mixture model with lower order by clustering and then merging subpopulations. For this reason, the new method is called MSCAD. Compared to six existing methods, simulations show that MSCAD consistently has the best or almost the best performance in terms of identifying the true order of the finite mixture model. In some cases, MSCAD is by far the best.

The rest of the paper is organized as follows. We introduce MSCAD in Section 2. Asymptotic properties of MSCAD are presented in Section 3. In Section 4, we present the computational method. The simulation results and two real-data examples are given in Section 5. Some conclusions are given in Section 6. Some brief proofs are in the Appendix and the detailed proofs are in a supplementary document at

www.amstat.org/publications/jasa/supplemental_materials.

2. MIXTURE MODEL AND NEW ORDER SELECTION METHOD

Let $\mathcal{F} = \{f(y;\theta,\sigma); \theta \in \Theta, \sigma \in \Omega\}$ be a family of parametric (probability) density functions with respect to a σ-finite measure ν, Θ be a compact subset of the real line $\boldsymbol{R}$, and $\sigma \in \Omega$ be a structure parameter where $\Omega \subset (0,\infty)$. For parametric families without a structure parameter, the value of σ is regarded as known. We assume that Ω is also compact. The compact assumption on Ω can often be relaxed. For example, Chen and Chen (2003) showed that the structure parameter under a normal mixture model is consistently estimated by the maximum likelihood estimator even when the order of the mixture is unknown. Under the model where σ is also a mixing parameter, placing a small positive lower bound on the ratio of two component parameters $\sigma_{k_1}/\sigma_{k_2}$ restores the consistency of the maximum likelihood estimator without any compactness condition (Hathaway, 1985). To avoid being overwhelmed by technicality, we choose to retain the compactness assumption.

The density function of a finite mixture model based on the family $\mathcal{F}$ is given by

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

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 $\theta_k \in \Theta$ are atoms of $G(\cdot)$, $0 \leq \pi_k \leq 1$ for $k = 1, 2, \ldots, K$.

Let K_0 be the smallest number of atoms θ_k of $G(\cdot)$ such that all the component densities $f(y;\theta_k,\sigma)$ are different and the mixing proportions π_k 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 $0 < \pi_{0k} \leq 1$, for $k = 1, 2, \ldots, K_0$.

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. Often scientists have a candidate order in mind. For example, geneticists may suspect that there are a few major genes behind a quantitative trait. If there is only one gene, then the order of the mixture model will be two or three; if there are two, the order can be at most six. In general, they will not consider models of a higher order. In such applications, an upper bound of $K = 12$ will be sufficient. In addition, due to the nature of the new method, a large upper bound can always be used. Nevertheless, if the method selects an order that is close to the upper bound, we can re-analyze the data with an increased upper bound. Finally, if a finite mixture model with a very high order is needed in an application, knowing the exact order of the model is likely less crucial.

Let $y_1, y_2, \ldots, y_n$ be a random sample from (1). The log-likelihood function of the mixing distribution with order K, and σ is given by

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

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

Assume that $\theta_1 \le \theta_2 \le \cdots \le \theta_K$, and denote $\eta_k = \theta_{k+1} - \theta_k$, for $k = 1, 2, \ldots, K-1$, and $\eta_{0k} = \theta_{0,k+1} - \theta_{0k}$, for $k = 1, 2, \ldots, K_0 - 1$, where $K_0 \ge 2$. Define the penalized log-likelihood function as

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

for some $C_K > 0$ and a non-negative function $p_n(\cdot)$. The first penalty function is from the modified likelihood of Chen and Kalbfleisch (1996) which forces the estimated values of π_k away from 0 to prevent type I over-fitting. Consequently, the atoms of any fitted G of order K will form K_0 clusters tightly around the true atoms θ_{0k}.

We choose the additional penalty function $p_n(\eta)$ such that it has a spike at $\eta = 0$. It is well known that such a penalty shrinks near-zero η values to exactly zero with positive probability. We focus on the SCAD penalty proposed by Fan and Li (2001) which is most conveniently characterized through its derivative:

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

for some $a > 2$, where $(\cdot)_+$ is the positive part of a quantity. The method is not sensitive with respect to a wide range of the choice of a or the C_K. The choice of γ_n is important. More discussion is in Section 4.

Let $(\hat{G}_n, \hat{\sigma})$ be the maximizer of $\tilde{l}_n(G, \sigma)$ (or σ assumes a known value). When some $\hat{\eta}_k = 0$, the actual number of atoms of $\hat{G}_n$ can be smaller than K and this is taken as the new order estimator. This is how the procedure achieves order selection without explicit maneuvers. We call $\hat{G}_n$ the maximum penalized likelihood estimator (MPLE), and we now show that it has desirable asymptotic properties.

3. ASYMPTOTIC PROPERTIES

Write $\hat{G}_n = \sum_{j=1}^{K} \hat{\pi}_j I(\hat{\theta}_j \le \theta)$ and define $I_k = \{j : \theta_{0,k-1} + \theta_{0,k} < 2\hat{\theta}_j \le \theta_{0,k} + \theta_{0,k+1}\}$ for $k = 1, 2, \ldots, K_0$ with $\theta_{0,0} = -\infty$ and $\theta_{0,K_0+1} = \infty$. Further, define

$$\hat{H}_k(\theta) = \sum_{j \in I_k} \hat{\pi}_j I(\hat{\theta}_j \le \theta) / \sum_{j \in I_k} \hat{\pi}_j$$

and hence,

$$\hat{G}_n(\theta) = \sum_{k=1}^{K_0} \hat{\alpha}_k \hat{H}_k(\theta) \tag{5}$$

with $\hat{\alpha}_k = \sum_{j \in I_k} \hat{\pi}_j$. In other words, $\hat{H}_k$ is a part of $\hat{G}_n$ containing atoms near θ_{0k}. The main idea of the MSCAD method is to use the modified likelihood to squeeze the atoms of $\hat{H}_k$ into a small neighborhood of θ_{0k} and to use the SCAD penalty to further shrink them into a single atom.

Theorem 1 *Assume that $f(y; \theta, \sigma)$ satisfies regularity conditions A_1-A_4 in the Appendix, the true distribution of Y is a finite mixture with density function $f(y; G_0, \sigma_0)$, and we apply the SCAD penalty with $\gamma_n = n^{1/4} \log n$. Then*

(a) For any continuous point θ of G_0, $\hat{G}_n(\theta) \to G_0(\theta)$ in probability, as $n \to \infty$, and $\hat{\alpha}_k = \pi_{0k} + o_p(1)$ for each $k = 1, 2, \ldots, K_0$.

(b) All atoms of $\hat{H}_k$ converge in probability to θ_{0k} for each $k = 1, 2, \ldots, K_0$.

If any $\hat{H}_k$ in Theorem 1 has more than one atom, the order K_0 is still over-estimated. We show that $\hat{H}_k$ has a single atom with probability tending to one for each $k = 1, 2, \ldots, K_0$, and therefore $\hat{G}_n$ is consistent in estimating K_0.

Theorem 2 *(Consistency of estimating K_0). Assume the same conditions as in Theorem 1. Under the true finite mixture density $f(y; G_0, \sigma_0)$, if $(\hat{G}_n, \hat{\sigma})$ falls into an $n^{-1/4}$-neighborhood of (G_0, σ_0), then $\hat{G}_n$ has K_0 atoms with probability tending to one.*

Under some conditions, Chen (1995) shows that when the order of the finite mixture model is unknown, the optimal rate of estimating the finite mixing distribution G_0 is $n^{-1/4}$. Hence, our result is applicable to the class of finite mixture models that includes Poisson mixtures, normal mixtures in location or scale parameter, and Binomial mixtures.

4. NUMERICAL SOLUTION

Denote $\mathbf{\Psi} = (\theta_1, \theta_2, \ldots, \theta_K, \pi_1, \pi_2, \ldots, \pi_{K-1}, \sigma)$, and let $\mathbf{\Psi}_0$ be the vector of true parameters corresponding to G_0 and σ_0. For convenience, in the following we use $\tilde{l}_n(\mathbf{\Psi})$ instead of $\tilde{l}_n(G, \sigma)$ to denote the penalized log-likelihood function. We first present a revised EM-algorithm (Dempster, Laird, and Rubin, 1977) for maximizing $\tilde{l}_n(\mathbf{\Psi})$ with given K.

Algorithm: Let the complete log-likelihood function be

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

where the z_{ik} are unobserved indicator variables showing the component-membership of the ith observation in the mixture. Then the penalized complete log-likelihood function is

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

The EM algorithm maximizes $l_n^c(\mathbf{\Psi})$ 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 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

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

where

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

are the conditional expectations of z_{ik} given the 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

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

Due to non-smoothness of $p_n(\cdot)$, the usual Newton-Raphson method cannot be directly used for maximization with respect to θ_k. However, Fan and Li (2001) suggested approximating $p_n(\eta)$ by

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

With this approximation, the component parameters θ_k and the structure parameter σ are

updated by solving

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

Starting from an initial value $\mathbf{\Psi}^{(0)}$, the iteration between the E and M steps continues until some convergence criterion is satisfied. For example, for a pre-specified value $\epsilon > 0$, the algorithm will stop if $\|\mathbf{\Psi}^{(m+1)} - \mathbf{\Psi}^{(m)}\| < \epsilon$. When the algorithm converges, the equations

$$\begin{aligned}
\frac{\partial l_n(\hat{\mathbf{\Psi}}_n)}{\partial \theta_1} - \frac{\partial p_n(\hat{\eta}_1)}{\partial \theta_1} &= 0, \\
\frac{\partial l_n(\hat{\mathbf{\Psi}}_n)}{\partial \theta_k} - \frac{\partial p_n(\hat{\eta}_{k-1})}{\partial \theta_k} - \frac{\partial p_n(\hat{\eta}_k)}{\partial \theta_k} &= 0,\ k = 2,3,\ldots,K-1, \\
\frac{\partial l_n(\hat{\mathbf{\Psi}}_n)}{\partial \theta_K} - \frac{\partial p_n(\hat{\eta}_{K-1})}{\partial \theta_K} &= 0
\end{aligned}$$

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

To see this, recall that when a local maximum is attained, all $\hat{\theta}_j$ are stationary points of $\tilde{l}_n(\Psi)$. Thus, if $\hat{\theta}_j$ is at a smooth point of $\tilde{l}_n(\Psi)$, we get a zero derivative. However, because of the non-smoothness of SCAD at $\eta = 0$, the derivative of $\tilde{l}_n(\Psi)$ does not exist in theory at $\hat{\eta}_k = 0$, or the above equation fails to hold in numerical computation.

The initial values of G are chosen to be a discrete uniform distribution on the $100(k -$

$1/2)/K\%$ sample quantiles. We used σ_0 as the initial σ value in our simulations. In applications, one may use the sample variance or its 0.8, 1.2, etc. multiples as initial values.

Next, we discuss the choice of the tuning parameters γ_n and a in SCAD and C_K. We let $a = 3.7$ as suggested in Fan and Li (2001). Chen et al. (2001) reported that the choice of C_K is not crucial, and this is re-affirmed by our simulations. They 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 $C_K = \log M$. In our simulations, we choose $C_K = \log 20$ for both the normal and Poisson mixture models.

The theory provides merely some guidance on the order of γ_n to achieve consistency. In applications, cross validation or CV (Stone, 1974) and generalized cross validation or GCV (Craven and Wahba, 1979) are often used. 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, for a given γ_n. Further, let $l_{n,i}(\hat{\mathbf{\Psi}}_{n,-i})$ be the log-likelihood function evaluated on the test set D_i, using $\hat{\mathbf{\Psi}}_{n,-i}$, for $i = 1, 2, \ldots, N$. Then the cross-validation criterion is defined by

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

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

In our implementation, we delete one observation at a time and the CV is calculated on a sequence of γ_n values over a specified range. For the normal mixture model, the

range of $\gamma_n/\sqrt{n}$ was chosen as $[.2, 1.5]$, and for the Poisson mixture model the range was $[.4, 1.6]$. These choices meet the conditions specified in the theorems for the sample sizes under consideration. In applications, some trial runs can be used to identify a proper range before a formal analysis.

On a typical Unix machine, it took about 45 seconds to complete the analysis of one simulated data set for the most difficult normal mixture model with 7 components and sample size $n = 400$.

The generalized cross validation (GCV) is computationally cheaper than the CV criterion. Yet its derivation often requires regularity conditions that may not be satisfied by mixture models. We find that the GCV does not work well in our simulations. Thus we do not recommend its use.

5. SIMULATIONS AND EXAMPLES

We compared MSCAD with a number of existing methods in the literature for order selection under normal mixtures in location parameter and Poisson mixtures. For the normal mixtures, MSCAD is compared to six methods: two information-based criteria AIC and BIC; the Bayesian method GWCR by Ishwaran et al. (2001); the Kullback-Leibler (KL) distance method by James et al. (2001); the Hellinger distance (HD) method by Woo and Sriram (2006); and the method of Lassoing by Xing and Rao (2008). For the Poisson mixtures, MSCAD is compared to AIC and BIC, the sequential testing procedure based on the likelihood ratio test (LRT) by Karlis and Xekalaki (1999), and the Hellinger distance (HD) method of Woo and Sriram (2007). We report the percentage of times the estimated order equals a number of values out of 500 replications with sample sizes $n = 100, 400$ for

the normal mixtures, and $n = 100, 500$ for the Poisson mixtures.

For normal mixtures, we have

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

with $\phi(\cdot)$ being the density of the standard normal $N(0, 1)$. We regard σ as an unknown parameter and generate data from the ten normal mixtures discussed in Ishwaran et al. (2001) with the parameter values given in Table 1. Note that the number of modes of the mixture models does not necessarily reflect the order of the model.

The results for AIC, BIC, and GWCR are quoted from Ishwaran et al. (2001), and the results for Lassoing are quoted from Xing and Rao (2008). The results for the two distance methods KL and HD are based on our own implementations. Since both methods involve integrations, numerical approximations are used. For the HD method, we used the EM-type algorithm outlined in Section 4.1 of Cutler and Cordero-Brana (1996). The non-parametric density estimate required in these two methods was computed by a standard function in the R software. Our codes may not be as efficient as the authors' codes but the latter are not publicly available. Similarly to Ishwaran et al. (2001), we set the upper bound $K = 15$. The simulation results are reported in Tables 3, 4, and 5.

Under the first two models, all the methods did well except for KL and Lassoing. MSCAD is not the best but nearly so. Under the third model, MSCAD is the indisputable best followed by Lassoing and then KL. The AIC, BIC, GWCR, and HD were all misled by the number of modes. Under model 4, MSCAD did well and outperformed all the other methods by good margins. Under models 5 and 6, none of the methods were effective at detecting the

true order but MSCAD had the highest success rate. In addition, in as many as 60% of the instances, MSCAD underestimated the true order of four by only one, which was unmatched by other methods. Under model 7, MSCAD is simply the best again. Under models 8-10, none of the methods were very effective at detecting the true order. However, MSCAD provided a closer order estimate most often. Overall, MSCAD is consistently the best or close to the best under all models. Compared to the other methods, MSCAD tends to give a higher order estimate, yet it does not seem to overestimate according to this simulation study.

We next simulate data from Poisson mixture models with density function

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

We generate data from seven selected models from Woo and Sriram (2007). The parameters are given in Table 2. We exclude the models with very close components because all methods are expected to be poor in such situations. We let the upper bound $K = 15$ for all models, simulated with two sample sizes $n = 100, 500$, and with 500 replicates. The simulation results are reported in Tables 6, 7, and 8.

Under models 1-2 and both sample sizes, all the methods did very well. Under model 3 with $n = 100$, the two HD methods did poorly. Woo and Sriram (2007) also noted that HD methods are not effective when a component has a large mean but a small mixing proportion. Under models 4 and 5, the BIC and HD methods trailed behind when $n = 100$. Under models 6 and 7, MSCAD again outperformed by a good margin all the other methods.

The simulation results indicate that even AIC underestimates the order of the mixture

model. MSCAD corrects this problem to a large degree. Yet the MSCAD estimate is not always larger than the AIC estimate, as in the Poisson example to be presented.

Example 1. (Continuous Data) Efron (2004) provided an empirical Bayes approach for estimating an appropriate null hypothesis when many hypotheses are tested simultaneously. The motivating example was an HIV study of 1391 patients, investigating which of 6 Protease Inhibitor drugs cause mutations at which 74 sites on the viral genome. Using logistic regression analysis, 444 z-values were calculated, each for testing a null hypothesis that the specific drug does not cause mutation at the specific site. Figure 1 contains a histogram of these z-values, with negative values indicating greater mutational effects.

The usual goal of a large-scale testing problem is to identify a small percentage of interesting cases that deserve further investigation. Efron (2004) fitted the z-values with an 8-component normal mixture model with common unit variance as in Figure 1. In this fit, about 60.8% of the 444 cases were from the mixture component $N(0, 1)$ and were clearly "uninteresting" or "non-significant". Note that $N(0, 1)$ is the theoretical null distribution. This proportion is relatively small for a typical screening problem. It results in too many interesting cases to be handled in the second-stage analysis. This genetic background motivated Efron (2004) to combine the four middle components of the fitted model to form the "uninteresting" class containing 89.4% of the cases. The new class is well approximated by $N(-0.34, 1.19^2)$ and the original fitted mixture reduced to a 5-component mixture model, revealing the crucial importance of replacing the theoretical null $N(0, 1)$ and the empirical null in a large scale problem.

We applied MSCAD to re-analyze the data with an upper bound of $K = 15$. The analysis

concluded with a 4-component normal mixture model, with the parameter estimates given in Table 9. The estimated common variance is $\hat{\sigma}^2 = 1.25^2$. See also Figure 1 for the fitted model. The third component of our fit is $N(-0.38, 1.25^2)$ which is close to Efron's combined null $N(-0.34, 1.19^2)$. Efron's reduced-order mixture model was motivated from the genetic background as observed earlier. Coupled with the MSCAD analysis, the conclusion can be more revealing and insightful.

We applied the other methods employed in the simulations to re-analyze the z-values. They all resulted in 4-component normal mixture models with slightly different parameter estimates. The results are reported in Table 9.

Example 2. (Count Data) In this example we re-analyze the count data in Table 1 of Karlis and Xekalaki (2001). The data concern the number of defaulted installments in a financial institution in Spain. There is a high degree of over-dispersion and a large number of zero counts (3002 out of 4691). Due to over-dispersion, Karlis and Xekalaki (2001) suggested a Poisson mixture model for this data, and fitted a six-component Poisson mixture. The unusually high percentage of zero counts also suggests fitting a 0-inflated Poisson mixture model (McLachlan and Peel, 2000; Böhning, 2000). Using a robust procedure, Woo and Sriram (2007) selected a 4-component 0-inflated model (W&S). We analyzed the data with $K = 15$ and obtained $\hat{K}_0 = 4$, and under the 0-inflated model we obtained $\hat{K}_0 = 5$. The traditional AIC and BIC methods were also applied. The resulting fits are given in Table 10, and some expected frequencies obtained are given in Table 11.

It is seen that MSCAD fits the data reasonably well under both model assumptions. Interestingly, both estimates contain a component with mixing parameter 0.7% in spite of

the built-in penalty on small mixing probabilities. It is natural to question whether such a fit is supported by the data. After this component is removed, the mean of the remaining Poisson mixture is only about 1.37, and the chance that a single observation is larger than 15 is about 2×10^{-12}. Having 33 or more observations out of 4691 being larger than 15 is practically impossible. Hence, we believe that MSCAD as well as most other methods have rightfully selected a model with this component. The robust W&S procedure gains by denying an unreliable component mean estimation, but at the cost of a poorer Pearson's goodness-of-fit measure, which is 45.6 compared to 34.8 and 34.7 for MSCAD.

6. CONCLUSIONS

We have developed a new penalized likelihood approach, MSCAD, for order selection in univariate finite mixture models. Under some conditions, MSCAD is consistent and has much better performances than six existing methods. In addition, it avoids fitting mixture models of various orders repeatedly.

The MSCAD formulates the problem of order selection as a problem of arranging subpopulations (i.e., mixture components) in a parameter space. The penalty introduced by the modified likelihood clusters the fitted subpopulations around the true subpopulations. A SCAD-type penalty merges each cluster of subpopulations into a single subpopulation. The procedure starts with a large number of subpopulations and obtains a mixture model with a proper order by clustering and merging subpopulations. Thus, MSCAD remains effective even if only a conservative large upper bound is available.

As pointed out by a referee, the influence of the largest observation on the largest $|\hat{\theta}_k|$ with small $\hat{\pi}_k$ can be large for MSCAD. Our additional analysis of two real data sets confirmed

this insightful observation, but showed that the influence is not drastic. It is still advisable to be cautious when interpreting the meaning of the largest $\hat{\theta}$. Since MSCAD reduces the excessive number of components in the initial model by merging near subpopulations, it tends to under-estimate the order when the true model contains near subpopulations. However, this is more a problem of poor identifiability than the ineffectiveness of the method. We are not aware of any methods immune from this problem.

APPENDIX: Regularity Conditions and Proofs

The proofs will be brief and heuristic. For detailed proofs, see the supplementary paper at the *JASA* website. The expectations below are under (G_0, σ_0).

Regularity Conditions

A_1. (i) $E(|\log f(y;\theta,\sigma)|) < \infty$, $\forall\ \theta$ and σ.

(ii) There exists $\rho > 0$ such that for each θ, σ, $f(y;\theta,\sigma,\rho)$ is measurable and $E(|\log f(y;\theta,\sigma,\rho)|) < \infty$, where $f(y;\theta,\sigma,\rho) = 1 + \sup_{|\theta'-\theta|+|\sigma'-\sigma|\le\rho} f(y;\theta',\sigma')$.

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

A_3. For $i = 1, 2, \ldots, n;\ j = 1, 2, 3$, define

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

For each atom of G_0, θ_{0k}, there exists a small neighborhood of (θ_{0k}, σ_0) and a function $q(Y)$ with $E\{q^2(Y)\} < \infty$ such that for $G, G', \theta_1, \theta_1'$, and σ, σ' in this neighborhood,

we have

$$|U_{ij}(\theta_1, G, \sigma) - U_{ij}(\theta_1', G', \sigma')| \leq q(Y_i)\{|\theta_1 - \theta_1'| + \|G - G'\| + |\sigma - \sigma'|\}.$$

A_4. The matrix with the (k_1, k_2)th element

$$E\{U_{11}(\theta_{0k_1}, G_0, \sigma_0)U_{11}(\theta_{0k_2}, G_0, \sigma_0)\}$$

is finite and positive definite.

Conditions A_1-A_4 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 $\sqrt{n}$-consistent and asymptotically normal.

Lemma 1 *Suppose the component density $f(y; \theta, \sigma)$ satisfies A_1-A_4. Then the MPLE $\hat{G}_n$ has the property $\sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1)$.*

Proof. We provide a very intuitive proof. If (G, σ) is the MPLE, the penalized likelihood function must be larger at (G, σ) than at (G_0, σ_0). It can be verified that this is possible only if (G, σ) is in a small neighborhood of (G_0, σ_0). When this is the case, the SCAD penalty at G can be shown to be more severe than the SCAD penalty at G_0. Thus, $\sum_{k=1}^{K} \log \pi_k$ of G cannot exceed $l_n(G, \sigma) - l_n(G_0, \sigma_0) = O_p(1)$. This is why $\sum_{k=1}^{K} \log \hat{\pi}_k = O_p(1)$. ♠

Proof of Theorem 1. Part (a). Note that the consistency of $(\hat{G}_n, \hat{\sigma})$ is loosely justified in Lemma 1. The other conclusion is a consequence.

Part (b). By Lemma 1, the mixing proportion on each atom of the $\hat{G}_n$ is positive in probability. Thus the atom of $\hat{H}_k$ must converge to θ_{0k} in probability. ♠

Proof of Theorem 2. Suppose (G, σ) is a candidate MPLE with more than K_0 atoms within an $n^{-1/4}$ neighborhood of (G_0, σ_0). It can then be decomposed as $\sum_{k=1}^{K_0} \alpha_k H_k$ so that each H_k has its atoms in the $n^{-1/4}$ neighborhood of θ_{0k}. Since G has more than K_0 atoms the variance of H_k, $m_{2k} > 0$ for at least one k.

Let $\tilde{G} = \sum_{k=1}^{K_0} \tilde{\alpha}_k I(\theta_k \leq \theta)$ be a mixing distribution that maximizes $\tilde{l}_n(G, \sigma)$ with respect to θ_k, $k = 1, 2, \ldots, K_0$, with the same σ. It turns out that $l_n(\tilde{G}, \sigma)$ is smaller than $l_n(G, \sigma)$ by at most a quantity of order $n^{3/4} \sum_{k=1}^{K_0} m_{2k}$. At the same time, the SCAD penalty at (G, σ) is larger than the SCAD penalty at $(\tilde{G}, \sigma)$ by a quantity larger than $n^{3/4} \sum_{k=1}^{K_0} m_{2k}$. Thus, (G, σ) cannot possibly be the MPLE. That is, the MPLE must have exactly K_0 atoms as claimed. ♠

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, pp. 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," *The Annals of Statistics*, **34**, 2350-2383.

Charnigo, R. and Sun, J. (2004). "Testing homogeneity in a mixture distribution via the L^2-distance between competing models," *Journal of the American Statistical Association*, **99**, 488-498.

Chen, H. and Chen, J. (2001). "The likelihood ratio test for homogeneity in finite mixture models," *The Canadian Journal of Statistics*, **29**, 201-216.

Chen, H. and Chen, J. (2003). "Test for homogeneity in normal mixtures in the presence of a structural parameter," *Statistica Sinica*, **13**, 351-365.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2001). "A modified likelihood ratio test for homogeneity in finite mixture models," *Journal of the Royal Statistical Society, Ser.* B, **63**, 19-29.

Chen, H., Chen, J., and Kalbfleisch, J. D. (2004). "Testing for a finite mixture model with two components," *Journal of the Royal Statistical Society, Ser.* B, **66**, 95-115.

Chen, J. (1995). "Optimal rate of convergence in finite mixture models," *The Annals of Statistics*, **23**, 221-234.

Chen, J. and Kalbfleisch, J. D. (1996). "Penalized minimum-distance estimates in finite mixture models," *The Canadian Journal of Statistics*, **24**, 167-175.

Craven, P. and 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.

Cutler, A. and Cordero-Brana, O. I. (1996). "Minimum Hellinger distance estimation for finite mixture models," *Journal of the American Statistical Association*, **91**, 1716-1723.

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.

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

Efron, B. (2004). "Large-scale simultaneous hypothesis testing: The choice of a null hypothesis," *Journal of the American Statistical Association*, **99**, 96-104.

Fan, J. and Li, R. (2001). "Variable selection via non-concave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, **96**, 1348-1360.

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 *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Volume 2*, eds L. LeCam and R. A. Olshen, 789-806.

Hathaway, R. J. (1985). "A constrained formulation of maximum-likelihood estimation for normal mixture distributions," *The Annals of Statistics*, **13**, 795-800.

Ishwaran, H., James, L. F., and Sun, J. (2001). "Bayesian model selection in finite mixtures by marginal density decompositions," *Journal of the American Statistical Association*, **96**, 1316-1332.

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

Karlis, D. and Xekalaki, E. (1999). "On testing for the number of components in a mixed Poisson model," *Annals of the Institute of Statistical Mathematics*, **51**, 149-162.

Karlis, D. and Xekalaki, E. (2001). "Robust inference for finite Poisson mixtures," *Journal of Statistical Planning and Inference*, **93**, 93-115.

Leroux, B. G. (1992). "Consistent estimation of a mixing distribution," *The Annals of Statistics*, **20**, 1350-1360.

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

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models.* New York: Wiley.

Neyman, J. and Scott, E. L. (1966). "On the use of $C(\alpha)$ optimal tests of composite hypotheses," *Bulletin de l'Institut International de Statistique*, **41(I)**, 477-497.

Ray, S. and Lindsay, B. G. (2008). "Model selection in High-Dimensions: A Quadratic-risk Based Approach," *Journal of the Royal Statistical Society, Ser.* B, **70**, 95-118.

Schwarz, G. (1978). "Estimating the dimension of a model," *The Annals of Statistics*, **6**, 461-464.

Stone, M. (1974). "Cross-validatory choice and assessment of statistical predictions (with discussion)," *Journal of the Royal Statistical Society, Ser.* B, **36**, 111-147.

Woo, M. and Sriram, T. N. (2006). "Robust estimation of mixture complexity," *Journal of the American Statistical Association*, **101**, 1475-1485.

Woo, M. and Sriram, T. N. (2007). "Robust estimation of mixture complexity for count data," *Computational and Statistical Data Analysis*, **51**, 4379-4392.

Xing, G. and Rao, J. S. (2008). "Lassoing mixtures," *Technical report*, Department of Epidemiology and Biostatistics, Case Western Reserve University.

Zhang, P. (1993). "Model selection via multifold cross-validation," *The Annals of Statistics*, **21**, 229-231.

Table 1: Parameter values in simulation studies for the normal mixtures.

	(π_1, θ_1)	(π_2, θ_2)	(π_3, θ_3)	(π_4, θ_4)	(π_5, θ_5)	(π_6, θ_6)	(π_7, θ_7)
Model	Normal Mixtures						
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)			
7	(1/7, 0)	(1/7, 3)	(1/7, 6)	(1/7, 9)	(1/7, 12)	(1/7, 15)	(1/7, 18)
8	(1/7, 0)	(1/7, 1.5)	(1/7, 3)	(1/7, 4.5)	(1/7, 6)	(1/7, 7.5)	(1/7, 9)
9	(1/7, 0)	(1/7, 1.5)	(1/7, 3)	(1/7, 4.5)	(1/7, 6)	(1/7, 9.5)	(1/7, 12.5)
10	(1/7, 0)	(1/7, 1.5)	(1/7, 3)	(1/7, 4.5)	(1/7, 9)	(1/7, 10.5)	(1/7, 12)

Table 2: Parameter values in simulation studies for the Poisson mixtures.

	(π_1, θ_1)	(π_2, θ_2)	(π_3, θ_3)	(π_4, θ_4)
Model	Poisson Mixtures			
1	(0.5, 1)	(0.5, 9)		
2	(0.8, 1)	(0.2, 9)		
3	(0.95, 1)	(0.05, 10)		
4	(0.45, 1)	(0.45, 5)	(0.1, 10)	
5	(1/3, 1)	(1/3, 5)	(1/3, 10)	
6	(0.3, 1)	(0.4, 5)	(0.25, 9)	(0.05, 15)
7	(0.25, 1)	(0.25, 5)	(0.25, 10)	(0.25, 15)

Table 3: Simulation results for normal mixture models (1-3; $n = 100$).

Model	K_0	# Modes	$\hat{K}_0$	AIC	BIC	GWCR	KL	HD	Lassoing		MSCAD
									CDF	PDF	
			1	.018	.150	.018	.000	.102	.050	.074	.056
1	2	2	2	**.896**	**.838**	**.920**	.220	**.898**	**.584**	**.432**	**.900**
			3	.062	.012	.058	.318	.000	.186	.220	.044
			4	.024	.000	.004	**.368**	.000	.068	.136	.000
			1	.022	.212	.030	.000	.078	.022	.016	.032
2	2	2	2	**.900**	**.780**	**.916**	**.484**	**.922**	**.510**	**.386**	**.898**
			3	.050	.006	0.054	.236	.000	.232	.238	.070
			4	.028	.002	0.000	.234	.000	.132	.172	.000
			1	**.702**	**.968**	**.868**	.000	**.824**	.106	.194	.280
3	2	1	2	.264	.030	.130	**.354**	.176	**.572**	**.416**	**.634**
			3	.024	.002	.002	.258	.000	.186	.184	.086
			4	.000	.000	.000	.242	.000	.044	.096	.000

Table 4: Simulation results for normal mixture models (4-6; $n = 100$).

Model	K_0	# Modes	$\hat{K}_0$	AIC	BIC	GWCR	KL	HD	Lassoing		MSCAD
									CDF	PDF	
			1	.000	.110	.000	.000	.026	.014	.034	.000
			2	.178	**.596**	.102	.062	**.964**	**.348**	.080	.002
			3	.110	.110	**.554**	.174	.010	.198	.090	.104
4	4	4	4	**.674**	.182	.306	**.700**	.000	.194	**.268**	**.824**
			5	.038	.002	.038	.064	.000	.078	.182	.070
			1	.244	**.748**	.144	.000	**.504**	.028	.060	.000
			2	**.556**	.246	**.818**	.292	.496	**.494**	**.312**	.072
			3	.142	.004	.032	.254	.000	.230	.274	**.604**
5	4	1	4	.044	.002	.006	**.310**	.000	.122	.142	.314
			5	.014	.000	.000	.140	.000	.040	.086	.010
			1	.016	.188	.000	.000	.120	.022	.060	.000
			2	**.474**	**.698**	**.612**	.154	**.880**	**.476**	.216	.052
			3	.392	.106	.368	**.504**	.000	.208	**.280**	**.654**
6	4	2	4	.102	.008	.020	.266	.000	.106	.184	.280
			5	.014	.000	.000	.070	.000	.054	.120	.014

Table 5: Simulation results for normal mixture models (7-10; $n = 400$).

Model	K_0	# Modes	$\hat{K}_0$	AIC	BIC	GWCR	KL	HD	Lassoing		MSCAD
									CDF	PDF	
			1	.004	**.816**	.000	.000	.072	.008	.026	.000
			2	.000	.000	.000	.000	.002	**.260**	.218	.000
			3	.000	.000	.010	.000	.126	.228	**.232**	.000
7	7	7	4	.302	.168	.188	.004	**.800**	.234	.206	.000
			5	.212	.016	**.424**	.480	.000	.160	.086	.002
			6	.098	.000	.178	**.516**	.000	.062	.054	.074
			7	**.326**	.000	.114	.000	.000	.032	.116	**.848**
			1	.030	**.538**	.000	.000	.000	.002	.014	.000
			2	**.684**	.462	.078	.034	.164	**.354**	.282	.000
			3	.000	.000	**.590**	.132	**.824**	.252	**.326**	.002
8	7	1	4	.248	.000	.272	.318	.012	.234	.246	.012
			5	.000	.000	.048	**.376**	.000	.096	.094	.184
			6	.012	.000	.008	.134	.000	.036	.026	**.502**
			7	.024	.000	.004	.006	.000	.016	.010	.270
			1	.002	**.458**	.000	.000	.060	.014	.101	.000
			2	.000	.000	.002	.028	.014	**.384**	**.334**	.000
			3	.144	.398	.120	**.550**	**.716**	.220	.268	.000
9	7	3	4	**.460**	.138	**.408**	.026	.210	.196	.170	.010
			5	.308	.006	.312	.382	.000	.106	.138	.398
			6	.048	.000	.128	.014	.000	.046	.058	**.506**
			7	.016	.000	.024	.000	.000	.020	.014	.080
			1	.000	.000	.000	.000	.000	.010	.014	.000
			2	**.496**	**.992**	.020	.006	.224	**.292**	.232	.000
			3	.000	.000	.370	**.846**	**.616**	.220	**.310**	.006
10	7	2	4	.302	.006	**.466**	.112	.160	.256	.242	.248
			5	.118	.002	.128	.034	.000	.112	.150	**.506**
			6	.064	.000	.010	.000	.000	.060	.026	.234
			7	.016	.000	.006	.002	.000	.034	.010	.006

Table 6: Simulation results for the 2-component Poisson mixture models (1-3)

	$n = 100$			Model 1, $K_0 = 2$	$n = 500$		
Method	1	2	3		1	2	3
AIC	.000	**.938**	.062		.000	**.924**	.076
BIC	.000	**.998**	.002		.000	**.998**	.002
$HD_{2/n}$	.000	**.998**	.002		.000	**1.00**	.000
$HD_{\log n/n}$	.000	**1.00**	.000		.000	**1.00**	.000
LRT	.000	**.950**	.050		.000	**.960**	.000
MSCAD	.000	**.988**	.012		.000	**1.00**	.000

	$n = 100$			Model 2, $K_0 = 2$	$n = 500$		
Method	1	2	3		2	3	4
AIC	.000	**.958**	.042		**.950**	.042	.008
BIC	.000	**.994**	.006		**1.00**	.000	.000
$HD_{2/n}$	.000	**.998**	.002		**1.00**	.000	.000
$HD_{\log n/n}$	.002	**.998**	.000		**1.00**	.000	.000
LRT	.000	**.950**	.050		**.960**	.040	.000
MSCAD	.002	**.986**	.012		**.990**	.008	.000

	$n = 100$			Model 3, $K_0 = 2$	$n = 500$		
Method	1	2	3		2	3	4
AIC	.012	**.948**	.036		**.950**	.048	.002
BIC	.026	**.972**	.002		**.998**	.002	.000
$HD_{2/n}$	**.616**	.384	.000		**1.00**	.000	.000
$HD_{\log n/n}$	**.946**	.054	.000		**.994**	.000	.000
LRT	.000	**.930**	.070		**.950**	.050	.000
MSCAD	.052	**.868**	.080		**.994**	.004	.000

Table 7: Simulation results for the 3-component Poisson mixture models (4-5)

	$n = 100$				Model 4, $K_0 = 3$	$n = 500$			
Method	1	2	3	4		1	2	3	4
AIC	.000	.410	**.590**	.000		.000	.006	**.972**	.022
BIC	.000	**.778**	.222	.000		000	.100	**.900**	.000
$HD_{2/n}$	.000	**.966**	.034	.000		.000	.162	**.838**	.000
$HD_{\log n/n}$	.000	**1.00**	.000	.000		.000	**.846**	.154	.000
LRT	.000	.390	**.580**	.020		.000	.000	**.940**	.060
MSCAD	.000	.280	**.692**	.028		.000	.082	**.896**	.022

	$n = 100$				Model 5, $K_0 = 3$	$n = 500$			
Method	1	2	3	4		1	2	3	4
AIC	.000	.274	**.720**	.006		.000	.000	**.974**	.026
BIC	.000	**.684**	.316	.000		.000	.026	**.974**	.000
$HD_{2/n}$	.000	**.840**	.160	.000		.000	.018	**.982**	.000
$HD_{\log n/n}$	.000	**.988**	.012	.000		.000	.462	**.538**	.000
LRT	.000	.300	**.660**	.030		.000	.000	**.940**	.060
MSCAD	.000	.200	**.780**	.020		.000	.016	**.964**	.020

Table 8: Simulation results for the 4-component Poisson mixture models (6-7)

	$n = 100$				Model 6, $K_0 = 4$	$n = 500$			
Method	2	3	4	5		2	3	4	5
AIC	.080	[.878]	.042	.000		.000	[.644]	.356	.000
BIC	.316	[.680]	.004	.000		.000	[.974]	.026	.000
$HD_{2/n}$	[.718]	.282	.000	.000		.000	[.956]	.044	.000
$HD_{\log n/n}$	[.962]	.038	.000	.000		.060	[.940]	.000	.000
LRT	.090	[.780]	.130	.000		.000	[.590]	.380	.030
MSCAD	.010	[.666]	.320	.004		.000	.366	[.624]	.010

	$n = 100$				Model 7, $K_0 = 4$	$n = 500$			
Method	2	3	4	5		2	3	4	5
AIC	.010	[.918]	.072	.000		.000	[.592]	.408	.000
BIC	.134	[.858]	.008	.000		.000	[.970]	.030	.000
$HD_{2/n}$	.182	[.812]	.006	.000		.000	[.924]	.076	.000
$HD_{\log n/n}$	[.718]	.282	.000	.000		.000	[1.00]	.000	.000
LRT	.020	[.860]	.120	.000		.000	[.590]	.400	.010
MSCAD	.000	[.512]	.460	.028		.000	.110	[.812]	.078

Table 9: The 4-component normal mixture model fit in Example 1 by different methods.

Method	k	1	2	3	4
MSCAD	$\hat{\pi}_k$	.017	.075	.884	.024
	$\hat{\theta}_k$	-9.97	-5.06	-0.38	3.53
GWCR	$\hat{\pi}_k$	.017	.075	.892	.016
	$\hat{\theta}_k$	-9.91	-5.03	-0.38	4.03
AIC - BIC	$\hat{\pi}_k$	.013	.075	.892	.016
	$\hat{\theta}_k$	-10.12	-5.11	-0.37	4.01
HD	$\hat{\pi}_k$	.012	.077	.878	.033
	$\hat{\theta}_k$	-9.64	-5.01	-0.42	2.36
KL	$\hat{\pi}_k$	.013	.076	.894	.017
	$\hat{\theta}_k$	-10.08	-5.10	-0.39	3.92

Table 10: Parameter estimates for the Poisson mixture models in Example 2.

Estimates	$(\hat{\pi}_1, \hat{\theta}_1)$	$(\hat{\pi}_2, \hat{\theta}_2)$	$(\hat{\pi}_3, \hat{\theta}_3)$	$(\hat{\pi}_4, \hat{\theta}_4)$	$(\hat{\pi}_5, \hat{\theta}_5)$
AIC	(.736, .147)	(.194, 3.95)	(.057, 8.91)	(.011, 14.83)	(0.001, 28.84)
BIC	(.739, .150)	(.205, 4.15)	(.053, 10.55)	(.003, 24.09)	—
MSCAD	(.733, .147)	(.200, 3.98)	(.060, 9.52)	(.007, 19.72)	—
(0-Inflated, AIC-BIC)	(.314, 0)	(.435, .298)	(.200, 4.37)	(.048, 10.99)	(.002, 26.15)
(0-Inflated, MSCAD)	(.328, 0)	(.417, .302)	(.193, 4.19)	(.055, 9.78)	(.007, 20.01)
(0-Inflated) W&S	(.373, 0)	(.385, .36)	(.199, 4.52)	(.043, 11.26)	—

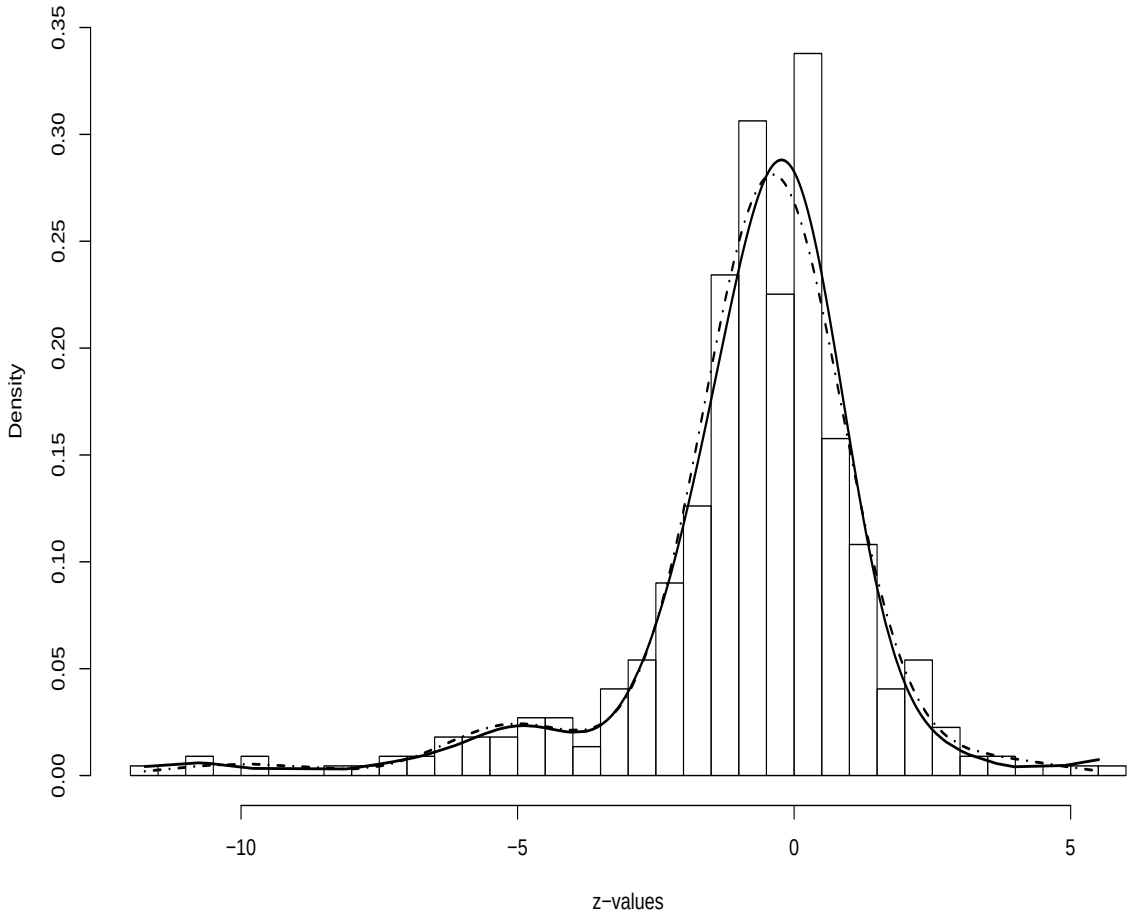

Figure 1: Histogram of the z-values in Example 1; Solid curve: density of the 8-component normal mixture of Efron; Dashed-point curve: Density of the 4-component normal mixture selected by MSCAD.

Table 11: Observed number of defaulted installments and the results for three fitted models.

Num. Defaults	Obs. Freq.	Expec. Freq. $\hat{K_0} = 4$	Expec. Freq. (0-inflated) $\hat{K_0} = 5$	Expec. Freq. (0-inflated) W&S
0	3002	2986.0	2998.6	3019.9
1	502	506.3	494.4	499.6
2	187	171.9	187.0	185.7
3	138	188.7	177.0	166.9
4	233	190.4	182.2	179.4
5	160	159.4	158.5	163.8
6	107	118.1	120.8	127.8
7	80	84.1	86.5	89.6
8	59	62.0	62.6	60.6
9	53	48.8	48.1	42.9
10	41	39.8	38.7	33.4
11	28	32.3	31.4	28.1
12	34	25.1	24.7	24.1
13	10	18.7	18.7	20.0
14	13	13.3	13.6	15.9
15	11	9.4	9.7	11.8
≥ 16	33	36.7	38.3	21.5