Variable Selection in Finite Mixture of Regression Models

Abbas KHALILI and Jiahua CHEN

In the applications of finite mixture of regression (FMR) models, often many covariates are used, and their contributions to the response variable vary from one component to another of the mixture model. This creates a complex variable selection problem. Existing methods, such as the Akaike information criterion and the Bayes information criterion, are computationally expensive as the number of covariates and components in the mixture model increases. In this article we introduce a penalized likelihood approach for variable selection in FMR models. The new method introduces penalties that depend on the size of the regression coefficients and the mixture structure. The new method is shown to be consistent for variable selection. A data-adaptive method for selecting tuning parameters and an EM algorithm for efficient numerical computations are developed. Simulations show that the method performs very well and requires much less computing power than existing methods. The new method is illustrated by analyzing two real data sets.

KEY WORDS: EM algorithm; LASSO; Mixture model; Penalty method; SCAD.

1. INTRODUCTION

Finite mixture models provide a flexible tool for modeling data that arise from a heterogeneous population. They are used in many fields, including biology, genetics, engineering, and marketing. The book by McLachlan and Peel (2000) contains a comprehensive review of finite mixture models. When a random variable with a finite mixture distribution depends on certain covariates, we obtain a *finite mixture of regression* (FMR) model. Jacobs, Jordan, Nowlan, and Hinton (1991) and Jiang and Tanner (1999) have discussed the use of FMR models in machine learning applications under the term mixture of experts models. The books by Wedel and Kamukura (2000) and Skrondal and Rabe-Hesketh (2004), among others, contain comprehensive reviews on the applications of FMR models in market segmentation and the social sciences.

Often, in the initial stage of a study many covariates are of interest, and their contributions to the response variable vary from one component to another of the FMR model. To enhance predictability and to give a parsimonious model, it is common practice to include only the important covariates in the model.

The problem of variable selection in FMR models has received much attention recently. All-subset selection methods, such as the Akaike information criterion (AIC; Akaike 1973), the Bayes information criterion (BIC; Schwarz 1978), and their modifications, have been studied in the context of FMR models; for instance, Wang, Puterman, Cockburn, and Le (1996) used AIC and BIC in finite mixture of Poisson regression models. However, even for FMR models with moderate numbers of components and covariates, all-subset selection methods are computationally intensive.

The least absolute shrinkage and selection operator (LASSO) of Tibshirani (1996) and the smoothly clipped absolute deviation (SCAD) method of Fan and Li (2001, 2002) are new methods for variable selection that have many interesting properties. For example, LASSO has the soft-thresholding property, and SCAD has a type of oracle property as discussed by Fan and Li (2001), and unlike all-subset selection methods, they can be used in reasonably high-dimensional problems. In this article we design a new variable selection procedure for FMR models based on these methods. We propose new class of penalty functions to be used for variable selection in FMR models. We investigate methods for selecting tuning parameters adaptively and develop an EM algorithm for numerical computations. The new method for variable selection is shown to be consistent. The performance of the method is studied theoretically and by simulations. Our simulations indicate that the new method is as good as or better than BIC at selecting correct models, with much less computational effort.

The article is organized as follows. In Section 2 FMR models as well as their identifiability are formally defined. In Section 3 the penalized likelihood-based approach is introduced for variable selection in the FMR models. Section 4 studies large-sample properties of the penalized likelihood-based estimators. A numerical algorithm and a data-adaptive method for choosing tuning parameters are discussed in Section 5. In Section 6 the performance of the new method is studied through simulations, and in Section 7 analysis of two real data sets illustrates the use of the new method. Conclusions are given in Section 8. Some technical details are given in an Appendix. For brevity, more extensive simulation results and more detailed proofs are not included in this article, but these are available in earlier work (Khalili and Chen 2005).

2. FINITE MIXTURE OF REGRESSION MODELS

Let Y be a response variable of interest and let $\mathbf{x} = (x_1, x_2, \ldots, x_P)^\tau$ be the vector of covariates believed to have an effect on Y. The FMR model is defined as follows.

Definition 1. Let $\mathcal{G} = \{f(y; \theta, \phi); (\theta, \phi) \in \Theta \times (0, \infty)\}$ be a family of parametric density functions of Y with respect to a σ-finite measure ν, where $\Theta \subset \mathrm{R}$ and ϕ is a dispersion parameter. We say that $(\mathbf{x}, Y)$ follows a FMR model of order K if the conditional density function of Y given $\mathbf{x}$ has the form

$$f(y; \mathbf{x}, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k f(y; \theta_k(\mathbf{x}), \phi_k) \tag{1}$$

Abbas Khalili is Postdoctoral Researcher, Department of Statistics, The Ohio State University, Columbus, OH 43210 (E-mail: *khalili@stat.ohio-state.edu*). Jiahua Chen is Professor, Department of Statistics, University of British Columbia, Vancouver, BC, Canada, V6T 1Z2 (E-mail: *jhchen@stat.ubc.ca*). The authors thank the joint editors, associate editor, and the referees for their insightful comments and suggestions. The research is supported in part by the Natural Science and Engineering Research Council of Canada. The first author would like to thank the Department of Statistics and Actuarial Sciences, University of Waterloo, Canada, for their support during his doctoral studies.

© 2007 American Statistical Association
Journal of the American Statistical Association
September 2007, Vol. 102, No. 479, Theory and Methods
DOI 10.1198/016214507000000590

with $\theta_k(\mathbf{x}) = h(\mathbf{x}^\tau \boldsymbol{\beta}_k)$, $k = 1, 2, \dots, K$, for a given link function $h(\cdot)$, and for some $\boldsymbol{\Psi} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_K, \boldsymbol{\phi}, \boldsymbol{\pi})$ with $\boldsymbol{\beta}_k = (\beta_{k1}, \beta_{k2}, \dots, \beta_{kP})^\tau$, $\boldsymbol{\phi} = (\phi_1, \phi_2, \dots, \phi_K)^\tau$, $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_{K-1})^\tau$ such that $\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$.

Model (1) can be generalized to allow the π_k values to be functions of $\mathbf{x}$. Here we restrict ourselves to the current model. The density function $f(y; \theta, \phi)$ can take many parametric forms, including binomial, normal, and Poisson. In some FMR models, the dispersion parameters, ϕ_k, are assumed to be equal.

FMR models combine the characteristics of regression models with those of finite mixture models. Like any regression model, the FMR model is used to study the relationship between response variables and a set of covariates. At the same time, the conditional distribution of the response variable Y given the covariates is a finite mixture.

A potential problem associated with finite mixture models is their identifiability, which is the basis for any meaningful statistical analysis. In some classes of finite mixture models, a single density function can have representations corresponding to different sets of parameter values. When no two sets of parameter values specify the same distribution, the model is identifiable. Many finite mixture models, including mixtures of binomial, multinomial, normal, and Poisson distributions, are identifiable under some conditions (see Titterington, Smith, and Markov 1985).

Definition 2. Consider a FMR model with the conditional density function given in (1). For a given design matrix $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^\tau$, the FMR model is said to be identifiable if for any two parameters $\boldsymbol{\Psi}$ and $\boldsymbol{\Psi}^*$,

$$\sum_{k=1}^{K} \pi_k f(y; \theta_k(\mathbf{x}_i), \phi_k) = \sum_{k=1}^{K^*} \pi_k^* f(y; \theta_k^*(\mathbf{x}_i), \phi_k^*)$$

for each $i = 1, \dots, n$ and all possible values of y, implies that $K = K^*$ and $\boldsymbol{\Psi} = \boldsymbol{\Psi}^*$.

When we exchange the order of two regression components, the parameter $\boldsymbol{\Psi}$ changes. In the foregoing definition, we interpret $\boldsymbol{\Psi} = \boldsymbol{\Psi}^*$ up to a permutation. The identifiability of an FMR model depends on several factors, such as component densities $f(y; \theta, \phi)$, the maximum possible order K, and the design matrix $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)^\tau$. Hennig (2000) observed that for fixed designs, a sufficient condition for identifiability is that the design points do not fall in the union of any K linear subspaces of $(P-1)$-dimension, in addition to some usual conditions on the component density. This condition is applicable to Poisson and normal FMR models. If the $\mathbf{x}_i$ values are also a random sample from a marginal density $f(\mathbf{x})$ that does not depend on $\boldsymbol{\Psi}$, then $f(\mathbf{x})$ must not have all of its mass in up to K of $(P-1)$-dimensional linear subspaces. Some discussion has been provided by Wang et al. (1996). In this article we assume that the FMR model under consideration is identifiable with the given or random design.

3. THE METHOD FOR VARIABLE SELECTION

In the case where $\mathbf{x}$ is random, we assume that its density $f(\mathbf{x})$ is functionally independent of the parameters in the FMR model. Thus the statistical inference can be done based purely on the conditional density function specified in Definition 1.

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ be a sample of observations from the FMR model (1). The (conditional) log-likelihood function of $\boldsymbol{\Psi}$ is given by

$$l_n(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k f(y_i; \theta_k(\mathbf{x}_i), \phi_k) \right\}.$$

When the effect of a component of $\mathbf{x}$ is not significant, the corresponding ordinary maximum likelihood estimate is often close to, but not equal to 0. Thus this covariate is not excluded from the model. To avoid this problem, we may study submodels with various components of $\mathbf{x}$ excluded, as is done by AIC and BIC. However, the computational burden of these approaches is heavy and should be avoided. The approach that we consider here is as follows.

We define a penalized log-likelihood function as

$$\tilde{l}_n(\boldsymbol{\Psi}) = l_n(\boldsymbol{\Psi}) - \mathbf{p}_n(\boldsymbol{\Psi}) \tag{2}$$

with the penalty function

$$\mathbf{p}_n(\boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \left\{ \sum_{j=1}^{P} p_{nk}(\beta_{kj}) \right\}, \tag{3}$$

where the $p_{nk}(\beta_{kj})$ values are nonnegative and nondecreasing functions in $|\beta_{kj}|$. By maximizing $\tilde{l}_n(\boldsymbol{\Psi})$ that contains a penalty, there is a positive chance of having some estimated values of β equaling 0 and thus of automatically selecting a submodel. Thus the procedure combines the variable selection and parameter estimation into one step and reduces the computational burden substantially. In (3) we choose the penalty imposed on the regression coefficients within the kth component of the FMR model to be proportional to π_k. This is in line with the common practice of relating the penalty to the sample size. The virtual sample size from the kth subpopulation is proportional to π_k, and this choice enhances the power of the method in our simulations.

When some prior information is available on the importance of a covariate's effects within the components of the FMR model, covariate-specific penalty functions may be used. In general, we should choose appropriate penalty functions to suit the need of the application, under the guidance of statistical theory. The following three penalty functions have been investigated in the literature in a number of contexts, and we use them to illustrate the theory that we develop for the FMR models:

- L_1-norm penalty: $p_{nk}(\beta) = \gamma_{nk}\sqrt{n}|\beta|$
- HARD penalty: $p_{nk}(\beta) = \gamma_{nk}^2 - (\sqrt{n}|\beta| - \gamma_{nk})^2 I(\sqrt{n} \times |\beta| < \gamma_{nk})$
- SCAD penalty: Let $(\cdot)_+$ be the positive part of a quantity,

$$p'_{nk}(\beta) = \gamma_{nk}\sqrt{n} I\{\sqrt{n}|\beta| \le \gamma_{nk}\} + \frac{\sqrt{n}(a\gamma_{nk} - \sqrt{n}|\beta|)_+}{a-1} I\{\sqrt{n}|\beta| > \gamma_{nk}\}.$$

The L_1-norm penalty was used in LASSO by Tibshirani (1996); the other two have been discussed by Fan and Li (2001, 2002). The constants $\gamma_{nk} > 0$ and $a > 2$ are chosen based on how strenuously the procedure tries to eliminate the covariates from the model. In applications, these choices may be based on some prior information; that is, the constants may be chosen subjectively by the data analysts or by a data-driven method. We call the penalty functions $\mathbf{p}_n(\cdot)$ in (3) constructed from LASSO, HARD, and SCAD the MIXLASSO, MIXHARD, and MIXSCAD penalties.

The three penalty functions have similar properties with some subtle differences. Maximizing the penalized likelihood is equivalent to constrained maximization. The penalty function of LASSO is convex and thus advantageous for numerical computation. It tends to reduce all effects by similar amounts until the estimated effect is reduced to 0. When the penalty is increased, SCAD reduces smaller effects faster than larger effects. Intuitively, HARD should work more like SCAD, although less smoothly.

4. ASYMPTOTIC PROPERTIES

We decompose the regression coefficient vector $\boldsymbol{\beta}_k$ in the kth component into $\boldsymbol{\beta}_k^\tau = \{\boldsymbol{\beta}_{1k}^\tau, \boldsymbol{\beta}_{2k}^\tau\}$ such that $\boldsymbol{\beta}_{2k}$ contains the zero effects. In general, the set of nonzero effects $\boldsymbol{\beta}_{1k}$ may depend on k. We choose to not use more complex notation to reflect this fact without loss of generality. Naturally, we split the parameter $\boldsymbol{\Psi}^\tau = (\boldsymbol{\Psi}_1^\tau, \boldsymbol{\Psi}_2^\tau)$ such that $\boldsymbol{\Psi}_2^\tau$ contains all zero effects, namely $\boldsymbol{\beta}_{2k}$, $k = 1, \dots, K$. The vector of true parameters is denoted as $\boldsymbol{\Psi}_0$. The components of $\boldsymbol{\Psi}_0$ are denoted with a superscript, such as β_{kj}^0.

Our asymptotic results are presented with the help of the quantities

$$a_n = \max_{k,j}\{p_{nk}(\beta_{kj}^0)/\sqrt{n} : \beta_{kj}^0 \neq 0\},$$

$$b_n = \max_{k,j}\big\{|p'_{nk}(\beta_{kj}^0)|/\sqrt{n} : \beta_{kj}^0 \neq 0\big\},$$

and

$$c_n = \max_{k,j}\big\{|p''_{nk}(\beta_{kj}^0)|/n : \beta_{kj}^0 \neq 0\big\},$$

where $p'_{nk}(\beta)$ and $p''_{nk}(\beta)$ are the first and second derivatives of the function $p_{nk}(\beta)$ with respect to β. The asymptotic results are based on the following conditions on the penalty functions $p_{nk}(\cdot)$:

P_0. For all n and k, $p_{nk}(0) = 0$, and $p_{nk}(\beta)$ is symmetric and nonnegative. In addition, it is nondecreasing and twice differentiable for all β in $(0, \infty)$ with at most a few exceptions.

P_1. As $n \to \infty$, $a_n = o(1 + b_n)$, and $c_n = o(1)$.

P_2. For $N_n = \{\beta; 0 < \beta \leq n^{-1/2} \log n\}$,

$$\lim_{n\to\infty} \inf_{\beta \in N_n} p'_{nk}(\beta)/\sqrt{n} = \infty.$$

Conditions P_0 and P_2 are needed for sparsity—namely, consistent variable selection. Condition P_1 is used to preserve the asymptotic properties of the estimators of nonzero effects in the model. To develop the asymptotic theory, some commonly used regularity conditions are needed on the joint density function $f(\mathbf{z}; \boldsymbol{\Psi})$ of $\mathbf{Z} = (\mathbf{x}, Y)$; we give these in the Appendix.

Theorem 1. Let $\mathbf{Z}_i = (\mathbf{x}_i, Y_i), i = 1, 2, \dots, n$, be a random sample from the density function $f(\mathbf{z}; \boldsymbol{\Psi})$ that satisfies the regularity conditions A_1–A_4 in the Appendix. Suppose that the penalty functions $p_{nk}(\cdot)$ satisfy conditions P_0 and P_1. Then there exists a local maximizer $\hat{\boldsymbol{\Psi}}_n$ of the penalized log-likelihood function $\tilde{l}_n(\boldsymbol{\Psi})$ for which

$$\|\hat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}_0\| = O_p\{n^{-1/2}(1 + b_n)\},$$

where $\|\cdot\|$ represents the Euclidean norm.

When $b_n = O(1)$, $\hat{\boldsymbol{\Psi}}_n$ has usual convergence rate $n^{-1/2}$. This is the case for MIXHARD, MIXSCAD, and MIXLASSO with the proper choice of tuning parameters. Another important property is sparsity, which enables consistent variable selection. The next theorem proves the sparsity property under some mild conditions.

Theorem 2. Assume that the conditions given in Theorem 1, that the penalty functions $p_{nk}(\cdot)$ satisfy P_0–P_2, and that K is known in parts (a) and (b). We then have the following:

a. For any $\boldsymbol{\Psi}$ such that $\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_0\| = O(n^{-1/2})$, with probability tending to 1,

$$\tilde{l}_n\{(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)\} - \tilde{l}_n\{(\boldsymbol{\Psi}_1, \mathbf{0})\} < 0.$$

b. For any $\sqrt{n}$-consistent maximum penalized likelihood estimator $\hat{\boldsymbol{\Psi}}_n$ of $\boldsymbol{\Psi}$,

1. Sparsity: $P\{\hat{\boldsymbol{\beta}}_{2k} = \mathbf{0}\} \to 1,\ k = 1, 2, \dots, K$ as $n \to \infty$
2. Asymptotic normality:

$$\sqrt{n}\left\{\left[\mathbf{I}_1(\boldsymbol{\Psi}_{01}) - \frac{\mathbf{p}''_n(\boldsymbol{\Psi}_{01})}{n}\right](\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) + \frac{\mathbf{p}'_n(\boldsymbol{\Psi}_{01})}{n}\right\} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{I}_1(\boldsymbol{\Psi}_{01})),$$

where $\mathbf{I}_1(\boldsymbol{\Psi}_{01})$ is the Fisher information computed under the reduced model when all zero effects are removed.

c. If K is estimated consistently by $\hat{K}_n$ separately, then the results in parts a and b still hold when $\hat{K}_n$ is subsequently used in the variable selection procedure.

The derivatives of $p_n(\cdot)$ part b.2 become negligible for some choices of the penalty function, other than providing some finite-sample adjustment. The result suggests the following variance estimator of $\hat{\boldsymbol{\Psi}}_1$:

$$\widehat{\mathrm{var}}(\hat{\boldsymbol{\Psi}}_1) = \{l''_n(\hat{\boldsymbol{\Psi}}_1) - \mathbf{p}''_n(\hat{\boldsymbol{\Psi}}_1)\}^{-1} \times \widehat{\mathrm{var}}\{l'_n(\hat{\boldsymbol{\Psi}}_1)\}\{l''_n(\hat{\boldsymbol{\Psi}}_1) - \mathbf{p}''_n(\hat{\boldsymbol{\Psi}}_1)\}^{-1}. \quad (4)$$

Keribin (2000) showed that under certain regularity conditions, the order of a finite mixture model can be estimated consistently using penalized-likelihood-based approaches such as the BIC. In applications, the BIC or the scientific background may be used to identify the order of the FMR model. However, blind use of a consistent estimator of K, regardless of the sample size n, should be discouraged.

In light of Theorem 2, the asymptotic properties differ slightly when different penalty functions are used. Yet the difference largely ends up in parameter estimation rather than in model selection. Due to the super-efficiency phenomenon associated with model selection, the conclusions on asymptotic bias or variance should be used cautiously (Leeb and Pötscher 2003). Technically, the L_1-norm penalty cannot simultaneously achieve sparsity and maintain root-n consistency, because the bias term $\mathbf{p}'_n(\boldsymbol{\Psi}_{01})/n$ is proportional to $\gamma_{nk}/\sqrt{n}$, and the γ_{nk} must be large enough to achieve sparsity. In the MIXSCAD or

MIXHARD penalties, $\mathbf{p}_n'(\boldsymbol{\Psi}_{01})/n = o(n^{-1/2})$ for a wide range of γ_{nk}. Thus sparsity and root-n consistency can be achieved simultaneously.

Theorem 2 applies to the local maximum that falls into a small neighborhood of the true parameter value. One need not be concerned about its general applicability. Under a very general (albeit long) list of conditions, the global maximum of the penalized likelihood is consistent. Further, by restricting the range of β to a compact region, commonly used models, such as those in this article, satisfy these conditions. As mentioned in Section 1, we have given technical details in earlier work (Khalili and Chen 2005).

5. NUMERICAL SOLUTIONS

We discuss a numerical method that uses the traditional EM algorithm applied to finite mixture models with revised maximization in the M step.

5.1 Maximization of the Penalized Log-Likelihood Function

Let $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ be a random sample of observations from the FMR model (1). In the context of finite mixture models, the EM algorithm of Dempster, Laird, and Rubin (1977) provides a convenient approach to the optimization problem. However, due to condition P_0, which is essential to achieve sparsity, the $p_{nk}(\beta)$'s are not differentiable at $\beta = 0$. The Newton–Raphson algorithm cannot be used directly in the M step of the EM algorithm unless it is properly adapted to deal with the single nonsmooth point at $\beta = 0$. We follow the approach of Fan and Li (2001) and replace $p_{nk}(\beta)$ by a local quadratic approximation,

$$p_{nk}(\beta) \simeq p_{nk}(\beta_0) + \frac{p_n'(\beta_0)}{2\beta_0}(\beta^2 - \beta_0^2),$$

in a neighborhood of β_0. This function increases to infinity whenever $|\beta| \to \infty$, which is more suitable for our application than the simple Taylor's expansion. Let $\boldsymbol{\Psi}^{(m)}$ be the parameter value after the mth iteration. We replace $\mathbf{p}_n(\boldsymbol{\Psi})$ in the penalized log-likelihood function in (2) by the following function:

$$\tilde{\mathbf{p}}_n(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) = \sum_{k=1}^{K} \pi_k \sum_{j=1}^{P} \left\{ p_{nk}(\beta_{jk}^{(m)}) + \frac{p_n'(\beta_{jk}^{(m)})}{2\beta_{jk}^{(m)}}(\beta_{jk}^2 - \beta_{jk}^{(m)2}) \right\}.$$

The revised EM algorithm is as follows: Let the complete log-likelihood function be

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

where the z_{ik}'s are indicator variables showing the component membership of the ith observation in the FMR model and are unobserved imaginary variables. The penalized complete log-likelihood function is then given by $\tilde{l}_n^c(\boldsymbol{\Psi}) = l_n^c(\boldsymbol{\Psi}) - \mathbf{p}_n(\boldsymbol{\Psi})$. The EM algorithm maximizes $\tilde{l}_n^c(\boldsymbol{\Psi})$ iteratively in the following two steps:

- E step. The E step computes the conditional expectation of the function $\tilde{l}_n^c(\boldsymbol{\Psi})$ with respect to z_{ik}, given the data $(\mathbf{x}_i, y_i)$ and assuming that the current estimate $\boldsymbol{\Psi}^{(m)}$ gives the true parameters of the model. The conditional expectation is

$$\begin{aligned} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) = &\sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(m)} \log \pi_k \\ &+ \sum_{i=1}^{n}\sum_{k=1}^{K} w_{ik}^{(m)} \log\{f(y_i; \theta_k(\mathbf{x}_i), \phi_k)\} \\ &- \mathbf{p}_n(\boldsymbol{\Psi}), \end{aligned}$$

where the weights

$$w_{ik}^{(m)} = \frac{\pi_k^{(m)} f(y_i; \theta_k^{(m)}(\mathbf{x}_i), \phi_k^{(m)})}{\sum_{l=1}^{K} \pi_l^{(m)} f(y_i; \theta_l^{(m)}(\mathbf{x}_i), \phi_l^{(m)})} \tag{5}$$

are the conditional expectation of the unobserved z_{ik}.

- M step. The M step on the $(m+1)$th iteration maximizes the function $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$ with respect to $\boldsymbol{\Psi}$. In the usual EM algorithm, the mixing proportions are updated by

$$\pi_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ik}^{(m)}, \qquad k = 1, 2, \ldots, K, \tag{6}$$

which maximize the leading term of $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$. Maximizing $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$ itself with respect to the π_k will be more complex. For simplicity, we use updating scheme (6) nevertheless; this worked well in our simulations.

We now consider that the π_k are constant in $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$, and maximize $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$ with respect to the other parameters in $\boldsymbol{\Psi}$. By replacing $\mathbf{p}_n(\boldsymbol{\Psi})$ by $\tilde{\mathbf{p}}_n(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$ in $Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$, the regression coefficients are updated by solving

$$\sum_{i=1}^{n} w_{ik}^{(m)} \frac{\partial}{\partial \beta_{kj}}\{\log f(y_i; \theta_k(\mathbf{x}_i), \phi_k^{(m)})\} - \pi_k\left\{\frac{\partial}{\partial \beta_{kj}} \tilde{p}_{nk}(\beta_{kj})\right\} = \mathbf{0},$$

where $\tilde{p}_{nk}(\beta_{kj})$ is the corresponding term in $\tilde{\mathbf{p}}_n(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)})$, for $k = 1, 2, \ldots, K$; $j = 1, 2, \ldots, P$. The updated estimates $\phi_k^{(m+1)}$ of the dispersion parameters are obtained by solving

$$\sum_{i=1}^{n} w_{ik}^{(m)} \frac{\partial}{\partial \phi_k}\{\log f(y_i; \theta_k(\mathbf{x}_i), \phi_k)\} = 0, \qquad k = 1, 2, \ldots, K.$$

Starting from an initial value $\boldsymbol{\Psi}^{(0)}$, we iterate between the E and M steps until the Euclidean norm, $\|\boldsymbol{\Psi}^{(m)} - \boldsymbol{\Psi}^{(m+1)}\|$, is smaller than some threshold value, taken as 10^{-8} in our simulation. When the algorithm converges, the equation

$$\left.\frac{\partial l_n(\boldsymbol{\Psi})}{\partial \beta_{kj}}\right|_{\boldsymbol{\Psi} = \hat{\boldsymbol{\Psi}}_n} - \pi_k p_{nk}'(\hat{\beta}_{kj}) = 0 \tag{7}$$

is satisfied (approximately) for the nonzero estimate $\hat{\beta}_{kj}$. At the same time, (7) is not satisfied when the estimated value of β_{kj} is 0. This fact allows us to identify zero estimates. For other issues of numerical implementation, the work of Hunter and Li (2005) will be helpful.

5.2 Choice of the Tuning Parameters

When using MIXLASSO, MIXHARD, MIXSCAD, and other penalty functions, we need to choose the size of the tuning parameters γ_{nk}. The current theory provides only some guidance on the order of γ_{nk}, to ensure the sparsity property. In applications, cross-validation (CV; Stone 1974) or generalized cross validation (GCV; Craven and Wahba 1979) are often used for choosing tuning parameters. Following the example of Tibshirani (1996) and Fan and Li (2001), we develop a componentwise deviance-based GCV criterion for the FMR models.

Let $\tilde{\boldsymbol{\Psi}}$ be the ordinary maximum likelihood estimate of $\boldsymbol{\Psi}$ under the full FMR model. For a given value of γ_{nk}, let $(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k)$ be the maximum penalized likelihood estimates of the parameters in the kth component of the FMR model obtained by fixing the remaining components of $\boldsymbol{\Psi}$ at $\tilde{\boldsymbol{\Psi}}$. Denote the deviance function evaluated at $\hat{\boldsymbol{\theta}}_k$, corresponding to the kth component of the FMR model, by

$$D_k(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k) = \sum_{i=1}^{n} w_{ik}\big[\log\{f(y_i; y_i, \hat{\phi}_k)\} - \log\big\{f(y_i; \hat{\theta}_k(\mathbf{x}_i), \hat{\phi}_k)\big\}\big],$$

where the weights w_{ik} are given in (5) evaluated at $\tilde{\boldsymbol{\Psi}}$. Further, let $l_k''(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k)$ be the second derivative of the log-likelihood function with respect to $\boldsymbol{\beta}_k$ evaluated at $(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k)$. We define a GCV criterion for the kth component of the FMR model as

$$GCV_k(\gamma_{nk}) = \frac{D_k(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k)}{n(1 - e(\gamma_{nk})/n)^2}, \qquad k = 1, 2, \dots, K, \quad (8)$$

where $e(\gamma_{nk})$ is the effective number of regression coefficients, given by

$$e(\gamma_{nk}) = \mathrm{tr}\big\{[l_k''(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k) - \Sigma_k(\hat{\boldsymbol{\beta}}_k)]^{-1} l_k''(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k)\big\},$$

where $\Sigma_k(\hat{\boldsymbol{\beta}}_k) = \hat{\pi}_k \,\mathrm{diag}\{p_{nk}'(\hat{\beta}_{k1})/\hat{\beta}_{k1}, \dots, p_{nk}'(\hat{\beta}_{kP})/\hat{\beta}_{kP}\}$, with tr denoting trace and diag denoting diagonal matrix. The tuning parameters γ_{nk} are chosen one at a time by minimizing $GCV_k(\gamma_{nk})$.

Using the GCV criterion to choose the tuning parameter results in a random tuning parameter. To ensure the validity of the asymptotic results, a common practice is to place a restriction on the range of the tuning parameter (see James, Priebe, and Marchette 2001). For example, by letting $\alpha_n = C_1 n^{-1/2} \log n$, $\beta_n = C_2 n^{-1/2} \log n$ for some constants $0 < C_1 < C_2$, and requiring $\lambda_{nk} = \frac{\gamma_{nk}}{\sqrt{n}} \in (\alpha_n, \beta_n)$, we retain the consistency property.

6. SIMULATION STUDY

The first simulations are based on the normal FMR model $\pi \mathrm{N}(\mathbf{x}^\tau \boldsymbol{\beta}_1, \sigma^2) + (1 - \pi)(\mathbf{x}^\tau \boldsymbol{\beta}_2, \sigma^2)$ with $\sigma^2 = 1$ and $P = 5$. We assume that $K = 2$ is known. When $\pi = .5$, using BIC, we found that $\hat{K} = 2$ in 996 simulations out of 1,000. When $\pi = .1$, the data do not contain sufficient information to choose K consistently.

The covariate $\mathbf{x}$ in the simulation is generated from a multivariate normal with mean 0, variance 1, and two correlation structures: $\rho_{ij} = \mathrm{cor}(x_i, x_j) = (.5)^{|i-j|}$ and $\rho_{ij} = 0$. Table 1 specifies the regression coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ and three choices of mixing proportion π. The M_1 and M_2 represent the FMR models with parameter values given in the table. A total of 1,000 data sets with sample sizes $n = 100$ and $n = 200$ were generated from each FMR model. We also simulated binomial and Poisson FMR models; the outcomes were similar, and thus we do not report the results here.

Table 1. Regression coefficients in the normal FMR models

Parameters	M_1	M_2
$\boldsymbol{\beta}_1$	(1, 0, 0, 3, 0)	(1, .6, 0, 3, 0)
$\boldsymbol{\beta}_2$	(−1, 2, 0, 0, 3)	(−1, 0, 0, 4, .7)
π	.5, .3, .1	.5, .3, .1

We compare the performance of different variable selection methods from a number of angles. First, we used the average correct and incorrect estimated zero effects in each component of the FMR model. Second, we used standard errors of the estimated nonzero regression coefficients. Finally, we generated a set of 10,000 test observations aside from each of the models M_1 and M_2 and, using the test data, computed the log-likelihood value of each submodel selected. We call this the *predictive log-likelihood*. A good variable selection method should consistently produce large predictive log-likelihood values. For either M_1 or M_2, there are a total of 1,024 potential submodels, all of which must be examined by the BIC method. To reduce the computational burden, we considered only a subset of the 266 and 380 most probable submodels of M_1 and M_2. These submodels are those containing at least two variables in each component, at least one of these which has a nonzero coefficient. Because the excluded models are known to be poor, the simulation results are not biased to favor the new method.

Tables 2 (for $\rho_{ij} = .5^{|i-j|}$) and 3 (for $\rho_{ij} = 0$) present the average numbers of correctly and incorrectly estimated zero coefficients. The results are presented in terms of mixture components 1 and 2. The new method is superior computationally, while maintaining similar performance to BIC in this respect. Note that when the sample size increases, all methods improve. When π reduces, all methods for the first component of the FMR model become less satisfactory, due to the lower number of observations from this component. In applications where the fitted mixing proportion is low and the sample size is small, the result of the corresponding regression component should be interpreted with caution.

We also computed the standard error (SD) of the nonzero regression coefficient estimates based on the same 1,000 samples, along with its estimate (SD_m) based on formula (4). Because we examined 5 nonzero coefficients, 2 models and 2 mixing proportions, we have a total 20 cases. The simulated SD of the new method (MS, MH, and ML combined) is smaller than that of BIC in 36 of 20 × 3 comparisons when $\rho = 0$ and is smaller in 47 of 20 × 3 comparisons when $\rho_{ij} = .5^{|i-j|}$. Thus, the new methods do well in this respect compared with BIC, more so when the noise level increases. In addition, the estimates (SD_m) are all reasonable. To save space, we do not present all of the details here.

The ultimate goal of variable selection is to identify a simple submodel with a very good predictive value. The boxplots of the predictive log-likelihood values, after logarithm of their

Table 2. Average numbers of correct and incorrect estimated zero coefficients for normal FMR models with $n = 100$ (200), $\rho_{ij} = (.5)^{|i-j|}$

	Model M_1				Model M_2			
	Component 1		Component 2		Component 1		Component 2	
Method	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
$\pi = .5$								
BIC	$2.85_{(2.92)}$	$.004_{(.000)}$	$1.89_{(1.95)}$	$.012_{(.010)}$	$1.89_{(1.94)}$	$.233_{(.063)}$	$1.90_{(1.93)}$	$.195_{(.021)}$
MS	$2.94_{(2.99)}$	$.024_{(.002)}$	$1.98_{(2.00)}$	$.058_{(.004)}$	$1.85_{(1.96)}$	$.191_{(.122)}$	$1.85_{(1.93)}$	$.168_{(.059)}$
MH	$2.87_{(2.90)}$	$.035_{(.002)}$	$1.92_{(1.92)}$	$.046_{(.000)}$	$1.87_{(1.89)}$	$.262_{(.096)}$	$1.90_{(1.87)}$	$.169_{(.037)}$
ML	$2.52_{(2.75)}$	$.027_{(.054)}$	$1.77_{(1.84)}$	$.078_{(.080)}$	$1.63_{(1.71)}$	$.125_{(.063)}$	$1.82_{(1.89)}$	$.119_{(.054)}$
$\pi = .3$								
BIC	$2.85_{(2.91)}$	$.035_{(.002)}$	$1.92_{(1.95)}$	$.001_{(.013)}$	$1.81_{(1.91)}$	$.607_{(.361)}$	$1.85_{(1.92)}$	$.147_{(.011)}$
MS	$2.84_{(2.96)}$	$.089_{(.025)}$	$1.96_{(2.00)}$	$.024_{(.024)}$	$1.80_{(1.89)}$	$.618_{(.378)}$	$1.94_{(1.94)}$	$.118_{(.045)}$
MH	$2.77_{(2.86)}$	$.088_{(.010)}$	$1.95_{(1.94)}$	$.007_{(.000)}$	$1.77_{(1.86)}$	$.685_{(.364)}$	$1.92_{(1.90)}$	$.193_{(.087)}$
ML	$2.54_{(2.73)}$	$.133_{(.050)}$	$1.78_{(1.84)}$	$.042_{(.053)}$	$1.65_{(1.73)}$	$.524_{(.245)}$	$1.80_{(1.92)}$	$.056_{(.079)}$
$\pi = .1$								
BIC	$2.46_{(2.75)}$	$.415_{(.163)}$	$1.91_{(1.94)}$	$.056_{(.027)}$	$1.47_{(1.69)}$	$1.08_{(.956)}$	$1.70_{(1.82)}$	$.553_{(.322)}$
MS	$2.40_{(2.79)}$	$.577_{(.380)}$	$1.99_{(2.00)}$	$.026_{(.023)}$	$1.22_{(1.53)}$	$.934_{(.811)}$	$1.91_{(1.98)}$	$.059_{(.020)}$
MH	$2.31_{(2.63)}$	$.575_{(.359)}$	$1.93_{(1.95)}$	$.074_{(.054)}$	$1.26_{(1.52)}$	$.983_{(.805)}$	$1.92_{(1.91)}$	$.084_{(.029)}$
ML	$2.58_{(2.78)}$	$.919_{(.625)}$	$1.71_{(1.78)}$	$.044_{(.085)}$	$1.61_{(1.77)}$	$1.59_{(1.37)}$	$1.76_{(1.85)}$	$.052_{(.051)}$

NOTE: MS, MIXSCAD; MH, MIXHARD; ML, MIXLASSO.

absolute values for clarity, shown in Figures 1 and 2 compare these methods from this angle. Due to the transformation, a method is better if it has lower median in the plot. The oracle model (OR; assuming that zero coefficients are known) is the clear winner. The medians of MH and MS are often as small as the median of OR. ML works very well for M_2 and beats BIC in most cases. The difference between MS and BIC are greater than that between OR and MS. Thus, MS is markedly better in this sense. BIC is the most unstable procedure. These findings show that the new method not only has less computational burden, but also has some advantages in selecting a most appropriate model.

In the second stage of the simulation, we investigated the situation where the number of covariates is relatively large by setting it equal to 40. For this case we considered three different parameter settings. Model I has 15 and 30 nonzero coefficients for two components of the normal FMR model. In this model the coefficients are either 0 or substantially different from 0. In models II and III, the regression coefficients were randomly generated from a standard normal distribution and a standard normal distribution plus 1 in mean. To create a meaningful variable selection problem, we set those with absolute value <.3 to 0. In the end, we have six and five zero coefficients in model II and five and four zero coefficients in model III.

Table 3. Average numbers of correct and incorrect zero coefficients for normal FMR models with $n = 100$ (200), $\rho_{ij} = 0$

	Model M_1				Model M_2			
	Component 1		Component 2		Component 1		Component 2	
Method	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
$\pi = .5$								
BIC	$2.88_{(2.92)}$	$.001_{(.000)}$	$1.92_{(1.95)}$	$.014_{(.004)}$	$1.88_{(1.93)}$	$.268_{(.031)}$	$1.89_{(1.92)}$	$.108_{(.017)}$
MS	$2.98_{(3.00)}$	$.021_{(.002)}$	$1.99_{(2.00)}$	$.024_{(.009)}$	$1.85_{(1.93)}$	$.280_{(.060)}$	$1.88_{(1.92)}$	$.131_{(.037)}$
MH	$2.96_{(3.00)}$	$.003_{(.000)}$	$1.97_{(2.0)}$	$.000_{(.000)}$	$1.92_{(1.98)}$	$.276_{(.025)}$	$1.92_{(1.97)}$	$.131_{(.003)}$
ML	$2.93_{(2.98)}$	$.026_{(.011)}$	$1.98_{(1.98)}$	$.025_{(.018)}$	$1.83_{(1.92)}$	$.278_{(.117)}$	$1.83_{(1.92)}$	$.184_{(.169)}$
$\pi = .3$								
BIC	$2.81_{(2.90)}$	$.016_{(.000)}$	$1.93_{(1.95)}$	$.032_{(.019)}$	$1.80_{(1.86)}$	$.624_{(.280)}$	$1.88_{(1.89)}$	$.098_{(.082)}$
MS	$2.81_{(2.88)}$	$.057_{(.034)}$	$1.92_{(1.95)}$	$.044_{(.042)}$	$1.85_{(1.95)}$	$.644_{(.348)}$	$1.94_{(1.97)}$	$.063_{(.041)}$
MH	$2.72_{(2.97)}$	$.011_{(.001)}$	$1.99_{(2.00)}$	$.000_{(.000)}$	$1.82_{(1.93)}$	$.573_{(.247)}$	$1.92_{(2.00)}$	$.144_{(.003)}$
ML	$2.74_{(2.84)}$	$.060_{(.014)}$	$1.94_{(1.97)}$	$.028_{(.014)}$	$1.79_{(1.83)}$	$.699_{(.415)}$	$1.89_{(1.96)}$	$.068_{(.066)}$
$\pi = .1$								
BIC	$2.53_{(2.78)}$	$.363_{(.145)}$	$1.91_{(1.94)}$	$.039_{(.019)}$	$1.44_{(1.52)}$	$1.14_{(.983)}$	$1.75_{(1.78)}$	$.474_{(.487)}$
MS	$2.10_{(2.40)}$	$.246_{(.140)}$	$1.88_{(1.85)}$	$.023_{(.029)}$	$1.29_{(1.58)}$	$.917_{(.850)}$	$1.90_{(1.95)}$	$.046_{(.045)}$
MH	$2.17_{(2.63)}$	$.333_{(.136)}$	$2.00_{(2.00)}$	$.004_{(.000)}$	$1.30_{(1.58)}$	$.917_{(.822)}$	$1.99_{(2.00)}$	$.123_{(.028)}$
ML	$2.64_{(2.76)}$	$.733_{(.455)}$	$1.83_{(1.73)}$	$.040_{(.042)}$	$1.69_{(1.72)}$	$1.64_{(1.34)}$	$1.83_{(1.87)}$	$.078_{(.059)}$

NOTE: MS, MIXSCAD; MH, MIXHARD; ML, MIXLASSO.

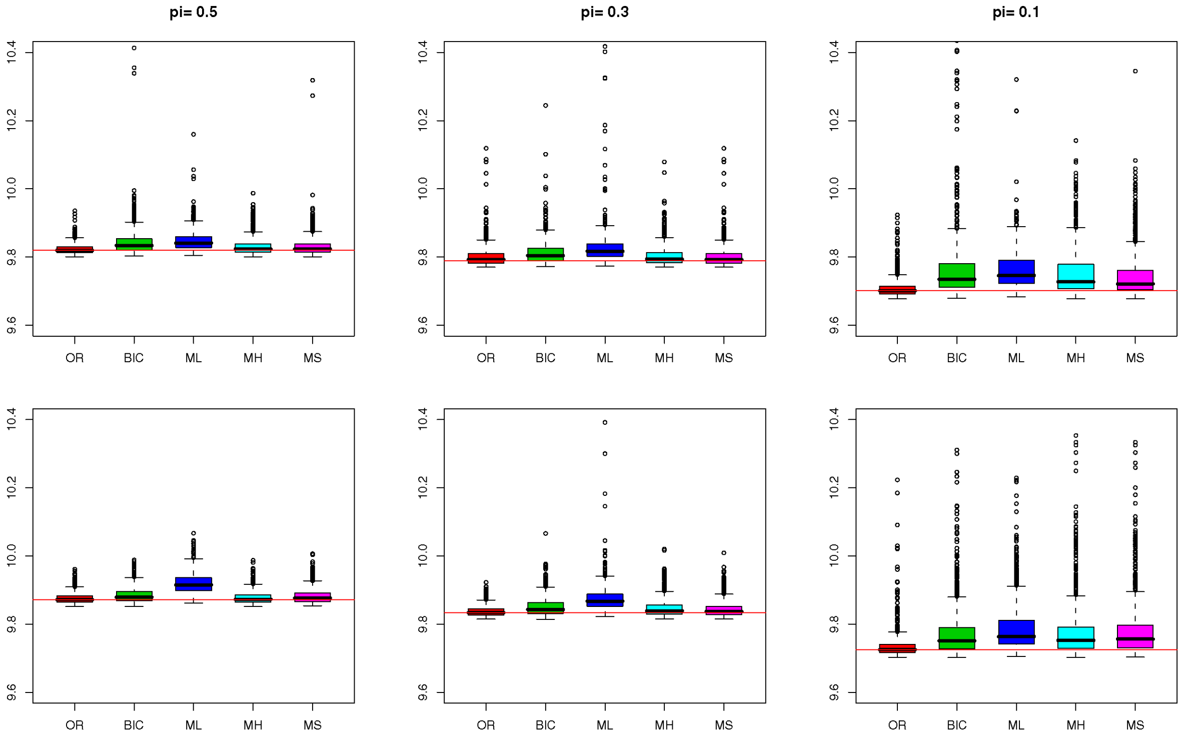

Figure 1. Predictive log-likelihood, model M_1. Top, $\rho_{ij} = (.5)^{|i-j|}$; bottom, $\rho_{ij} = 0$.

The covariates are from multivariate normal distribution with mean 0, variance 1, and two correlation structures: $\rho_{ij} = .5^{|i-j|}$ and $\rho_{ij} = 0$. We chose $n = 300$ and generated, 1,000 samples. In this case, BIC becomes impractical due to the amount of computation.

The most important performance measure is the predictive log-likelihood. We report the boxplots of the predictive log-likelihood values, computed based on a test sets of size 10,000, in Figure 3, and judge the success of the new method against the ideal OR. We notice that both MS and MH do well in this

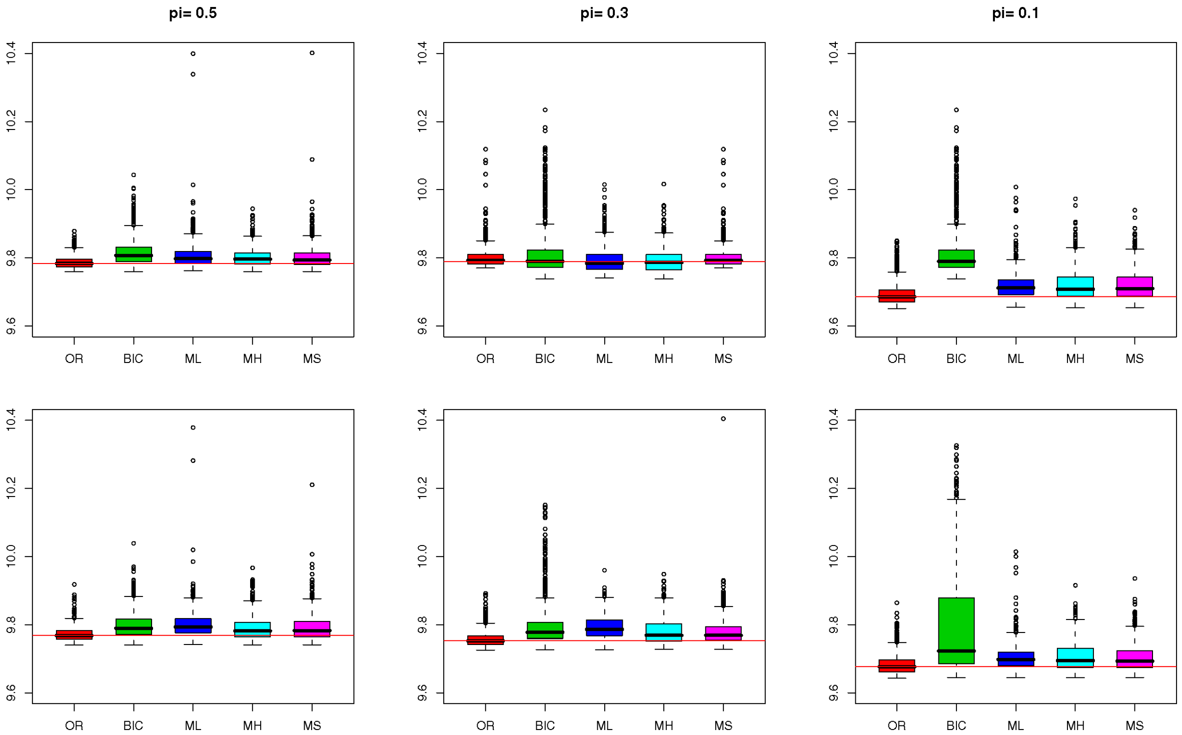

Figure 2. Predictive log-likelihood, model M_2. Top, $\rho_{ij} = (.5)^{|i-j|}$; bottom, $\rho_{ij} = 0$.

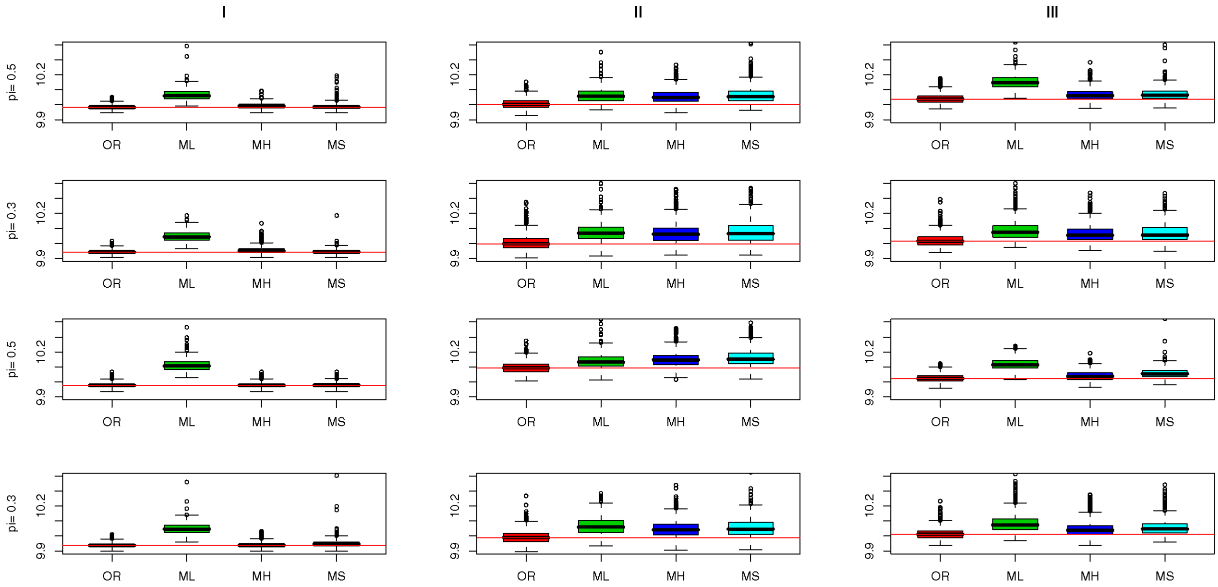

Figure 3. Predictive log-likelihood, models I–III. Top, $\rho_{ij} = (.5)^{|i-j|}$; bottom, $\rho_{ij} = 0$.

respect, for example, in most cases the median of MS is well above the 25th percentile of OR. When the effects have a wide range of sizes, it becomes harder to match up with the OR, namely the true model.

Other simulation results for models I–III have been given previously (Khalili and Chen 2005).

7. REAL DATA ANALYSIS

We now analyze two real data sets to further demonstrate the use of the new method.

7.1 Market Segmentation Analysis

FMR models often have been used in market segmentation analysis. The concept of market segmentation is an essential element in both marketing theory and practice. According to this concept, a heterogeneous market can be divided into a number of smaller homogeneous markets, in response to differing preferences of consumers. FMR models provide a model-based approach for market segmentation (see Wedel and Kamakura 2000).

In market segmentation studies, selected consumers were repeatedly asked to choose one product from a collection of hypothetical products with various features. Such experiments are often called conjoint choice experiments. The data collected from such experiments are analyzed to provide estimates of the market shares of new products. This method gives the researchers an idea of which products are likely to be successful before being introduced to the market.

A conjoint choice experiment was conducted at a large shopping mall in the Netherlands regarding consumer preferences for coffee makers. The data set is available at *www.gllamm.org/books*, provided by Skrondal and Rabe-Hesketh (2004). The main goal of the study was to estimate the market share for coffee makers with different features. The hypothetical coffee makers had five attributes: brand name (three levels), capacity (three levels), price (three levels), thermos (two levels), and filter (two levels). The levels of these attributes are given in Table 4. Each level combination forms a profile of the hypothetical coffee maker. Not all possible profiles are realistic; for example, a coffee maker with all features cannot have the low-

Table 4. Parameter estimates in market segmentation data

		SRH	MS	BIC
Factors	Levels	estimates (SE)	estimates	estimates
Brand	Philips	$-.37^{*}_{(.17)}$	0	0
	Braun	$-.40^{*}_{(.16)}$	0	0
	Moulinex	0	0	0
Capacity	6	$-2.48^{*}_{(.21)}$	$-2.59_{(.08)}$	$-2.59_{(.09)}$
	10	$.06_{(.14)}$	0	0
	15	0	0	0
Price	39	$1.97^{*}_{(.34)}$	$1.91_{(.22)}$	$1.91_{(.23)}$
	69	$1.48^{*}_{(.17)}$	$1.43_{(.12)}$	$1.43_{(.13)}$
	99	0	0	0
Thermos	Yes	$1.14^{*}_{(.18)}$	$1.08_{(.14)}$	$1.08_{(.14)}$
	No	0	0	0
Filter	Yes	$.92^{*}_{(.12)}$	$1.02_{(.11)}$	$1.02_{(.11)}$
	No	0	0	0
Brand	Philips	$.12_{(.21)}$	0	0
	Braun	$-1.43^{*}_{(.31)}$	$-1.61_{(.11)}$	$-1.52_{(.11)}$
	Moulinex	0	0	0
Capacity	6	$-.25_{(.26)}$	0	0
	10	$.07_{(.25)}$	0	0
	15	0	0	0
Price	39	$-.49_{(.32)}$	0	0
	69	$-.04_{(.22)}$	0	0
	99	0	0	0
Thermos	Yes	$.35_{(.20)}$	0	0
	No	0	0	0
Filter	Yes	$1.00^{*}_{(.20)}$	$.54_{(.10)}$	$.76_{(.10)}$
	No	0	0	0

est price. Thus only a total of 16 profiles were constructed by combining the levels of the foregoing attributes.

Two groups of eight choice sets were constructed with each set containing three profiles (alternatives). A set of eight profiles was assigned as the first alternatives in each group. A second alternative in each choice set was chosen differently to provide a realistic situation that resembles the purchasing situation. A base profile was then added as the common third alternative in all choice sets. Statistically, the design matrix also has good property under this setup.

A total of 185 respondents participated in the experiment. They were randomly divided into 2 groups of 94 and 91 subjects. Each respondent was repeatedly asked to make one choice out of each set of three profiles from one of the two groups. The data resulting from the experiment were binary responses from the participants, indicating their profile choices. For subject i, on replication j, we get a three-dimensional response vector; $\mathbf{y}_{ij}^T = (y_{ij1}, y_{ij2}, y_{ij3})$. The five attributes are the covariates in the model.

Skrondal and Rabe-Hesketh (2004) fitted a multinomial logit FMR model with $K = 2$, corresponding to two market segments, to the data arising from the coffee maker conjoint analysis. Mathematically, the FMR model is given by

$$P(\mathbf{y}_i) = P(\mathbf{Y}_i = \mathbf{y}_i) = (1-\pi)P_1(\mathbf{y}_i) + \pi P_2(\mathbf{y}_i),$$

where

$$P_k(\mathbf{y}_i) = \prod_{j=1}^{8}\prod_{a=1}^{3}\left[\frac{\exp\{\mathbf{x}_a^\tau \boldsymbol{\beta}_k\}}{\sum_{l=1}^{3}\exp\{\mathbf{x}_l^\tau \boldsymbol{\beta}_k\}}\right]^{y_{ija}}, \qquad k = 1, 2.$$

The covariate $\mathbf{x}_a^\tau$ is an 8×1 vector of dummy variables corresponding to the five attributes. Because the value of the covariates $\mathbf{x}_a^\tau$ did not change with subjects often enough, to make the parameters identifiable, an intercept term in the linear predictor $\mathbf{x}_a^\tau \boldsymbol{\beta}_k$ was not included.

Skrondal and Rabe-Hesketh (2004) obtained the maximum likelihood estimates (MLEs) of the parameters with $\hat{\pi} = .28$. Thus the estimated size of the first market segment was 72%, and that of the second segment was 28%. The MLEs of the $\boldsymbol{\beta}_k$ are given in Table 4, column "SRH." The coefficient estimate of the first market segment, $\hat{\boldsymbol{\beta}}_1$, is given in the upper half of the table; that of $\hat{\boldsymbol{\beta}}_2$ in the lower half.

Apparently, some of the regression coefficients are not significant, and, thus a variable selection procedure is needed. We applied the MIXLASSO, MIXHARD, and MIXSCAD methods to this data and used the GCV criterion outlined in Section 5.2. The new method with the MIXHARD and MIXSCAD penalties chose the same model with more zero coefficients than the model chosen by the MIXLASSO penalty. We report only the results based on the MIXSCAD penalty in Table 4. The data-adaptive choice of tuning parameters were .1 for the first segment and .27 for the second segment of the FMR model. The mixing proportion π was estimated as 26%. We also applied BIC to the data. In light of the model chosen by the new method with the MIXSCAD penalty, we considered a collection of 12 models to be examined by BIC. Note that total number of possible models is at least 961, which is much larger. The outcome was the same as for the new method with the MIXSCAD penalty. The parameter estimates and their corresponding standard errors are presented in Table 4. We computed the predictive log-likelihood of the models selected based on a small test data set from the same source. The predictive log-likelihood values based on the full model and the two selected models (from MIXSCAD and BIC) are −8.95, −9.65, and −9.65. Unfortunately, we cannot conclude which value is superior based on a single outcome, but they are clearly comparable.

Unlike the full model, the model after variable selection makes it apparent that the brand name has no significant effect in one component, and that its effect in the other component reflects a protest vote against Braun, a German company. Some consumer relationship work is needed. The indifference to capacity and price in one market segment could be an artifact of protest votes. For example, coffee makers with 6-cup capacity probably will find no market share at all, even though capacity is found to be insignificant in one component of the model.

7.2 Baseball Salaries

We analyze another real data set that is available on the website of the *Journal of Statistics Education* (*www.amstat.org/publications/jse*). The data contains the year 1992 salaries (measured in thousands of dollars) along with 16 performance measures from the year 1991 for 337 major league baseball players who played at least one game in both the 1991 and 1992 seasons, excluding pitchers. Of interest is how the performance measures affect salaries.

The performance measures are batting average (x_1), on-base percentage (x_2), runs (x_3), hits (x_4), doubles (x_5), triples (x_6), home runs (x_7), runs batted in (x_8), walks (x_9), strikeouts (x_{10}), stolen bases (x_{11}), and errors (x_{12}); and indicators of free agency eligibility (x_{13}), free agent in 1991/2 (x_{14}), arbitration eligibility (x_{15}), and arbitration in 1991/2 (x_{16}). The four (dummy) variables x_{13}–x_{16} indicate how free each player was to move to an other team. The online article by Watnik (1998) suggests the potential importance of the interaction effects between x_{13}–x_{16} and the quantitative variables x_1, x_3, x_7, and x_8. Note that x_1 and x_7 measure individual performance, whereas x_3 are x_8 are measures of individuals' contributions to the team. This leads to a set of 32 potential covariates affecting each player's salary.

The histogram of the salary is highly right-skewed [see Fig. 4(a)]. Thus we use log(salary) as the response variable. Watnik (1998) discussed the use of a linear regression model; however, the histogram of log(salary) and the corresponding nonparametric density estimator shown overlaid in Figure 4(b) suggests a mixture of two normal linear models,

$$Y = \log(\text{salary}) \sim \pi \mathrm{N}(\mathbf{x}^\tau \boldsymbol{\beta}_1, \sigma^2) + (1-\pi)\mathrm{N}(\mathbf{x}^\tau \boldsymbol{\beta}_2, \sigma^2),$$

with $\mathbf{x}$ a 33×1 vector containing all 32 potential covariates plus an intercept. Table 5 presents the parameter estimates under the simple linear model chosen by BIC and the mixture model chosen by MIXSCAD and MIXLASSO. For the linear model, $\hat{\sigma} = .48$, and the coefficient of determination is $R^2 = .83$. MIXSCAD and MIXHARD choose the same model with $\hat{\pi} = .76$, $\hat{\sigma} = .32$, and componentwise coefficient of determinations $R_1^2 = .94$ and $R_2^2 = .90$. MIXLASSO chooses a model with $\hat{\pi} = .72$, $\hat{\sigma} = .25$, $R_1^2 = .96$, and $R_2^2 = .95$. The R^2 values for the mixture models are computed using weighted

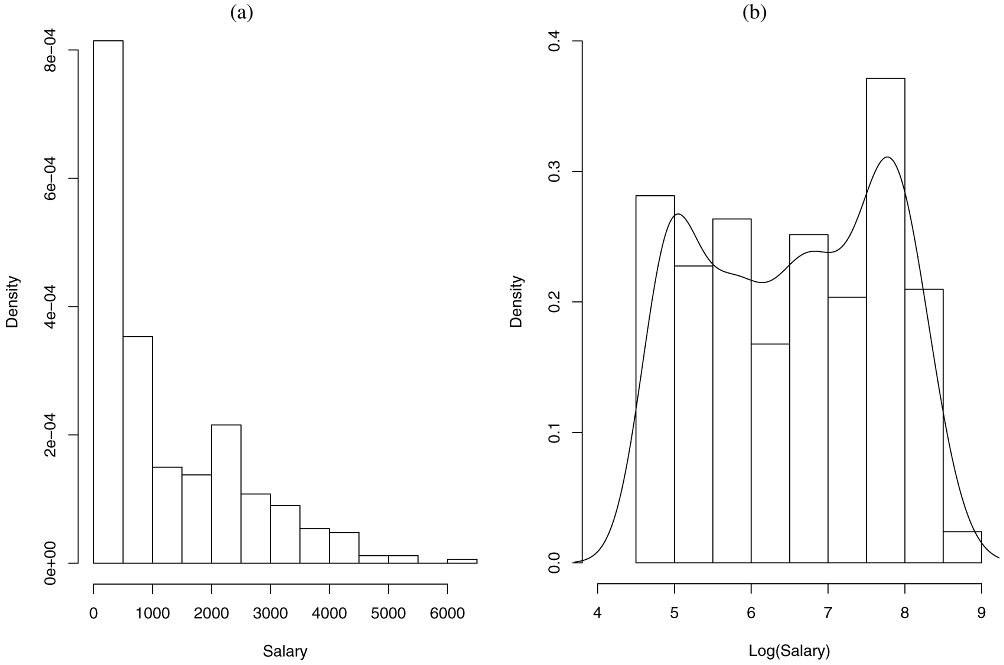

Figure 4. Histograms and density estimate: Baseball salary data. (a) Salary. (b) Log(Salary)

Table 5. Parameter estimates for baseball salary data

Covariates	Linear model (BIC) estimates (SE)	MIXSCAD		MIXLASSO	
		Component 1	Component 2	Component 1	Component 2
x_0	$6.55_{(.03)}$	$6.45_{(.03)}$	$6.87_{(.05)}$	$6.41_{(.01)}$	$7.00_{(.02)}$
x_1					$-.32_{(.02)}$
x_2					$.29_{(.02)}$
x_3					$-.70_{(.02)}$
x_4		$.40_{(.03)}$	$.58_{(.05)}$	$.20_{(.02)}$	$.96_{(.02)}$
x_5	$.27_{(.05)}$				
x_6					
x_7				$-.19_{(.02)}$	
x_8				$.26_{(.03)}$	
x_9	$.17_{(.05)}$				
x_{10}					
x_{11}					
x_{12}					
x_{13}		$1.49_{(.19)}$	$.95_{(.09)}$	$.79_{(.12)}$	$.70_{(.03)}$
x_{14}	$.86_{(.04)}$	$-1.24_{(.23)}$		$.72_{(.15)}$	
x_{15}	$-.18_{(.04)}$	$.68_{(.03)}$		$.15_{(.09)}$	$.50_{(.02)}$
x_{16}	$.54_{(.03)}$				$-.36_{(.02)}$
$x_1 * x_{13}$		$-.93_{(.21)}$		$-.21_{(.13)}$	
$x_1 * x_{14}$		$1.17_{(.23)}$		$.63_{(.18)}$	
$x_1 * x_{15}$				$.34_{(.10)}$	
$x_1 * x_{16}$					
$x_3 * x_{13}$	$-.19_{(.07)}$				
$x_3 * x_{14}$				$.14_{(.05)}$	$-.38_{(.01)}$
$x_3 * x_{15}$					
$x_3 * x_{16}$				$-.18_{(.01)}$	$.74_{(.04)}$
$x_7 * x_{13}$			$.71_{(.15)}$		
$x_7 * x_{14}$					
$x_7 * x_{15}$	$.13_{(.04)}$				$.34_{(.01)}$
$x_7 * x_{16}$					
$x_8 * x_{13}$		$.39_{(.05)}$	$-1.12_{(.19)}$	$.29_{(.03)}$	$-.46_{(.02)}$
$x_8 * x_{14}$				$-.14_{(.04)}$	
$x_8 * x_{15}$					
$x_8 * x_{16}$	$.20_{(.07)}$				

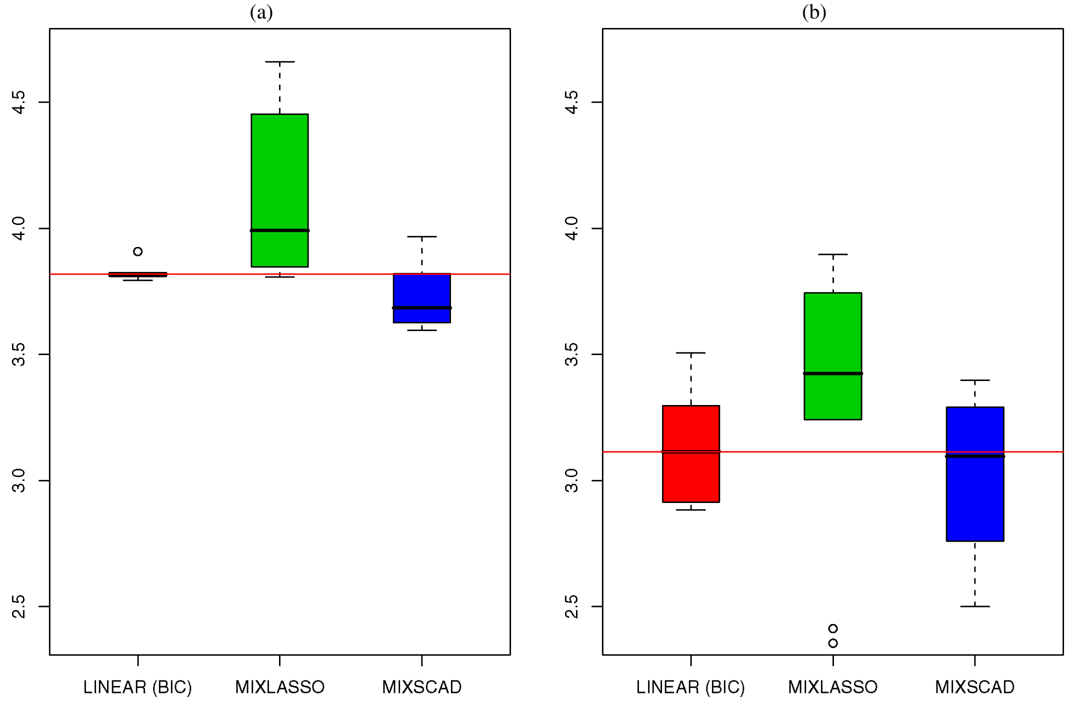

Figure 5. Predictive log-likelihood by CV: Baseball salary data. (a) $d = 5$; (b) $d = 10$.

sum of squares corresponding to each component of the mixture model (as in Wedel and Kamukura 2000). A regression mixture with three components was also fitted and found to be less satisfactory according to a plain BIC comparison with the model with $K = 2$, which in turn is better than nonmixture $K = 1$.

To evaluate the prediction performance of the selected models, we used d-fold cross-validation with $d = 5, 10$, and also Monte Carlo cross-validation (MCCV) (see Shao 1993 for details). In MCCV, the data were partitioned 500 times into disjoint training subsets (with size $n - d$) and test subsets (with size d). The log-likelihoods evaluated for the test data sets over the 500 replications are reported as boxplots (after log-absolute value transformation) in Figures 5 and 6. Apparently, MIXSCAD has superior predicting power than the straightforward linear regression model. MIXLASSO is geared mainly for variable selection rather than prediction, which may explain its lower prediction power.

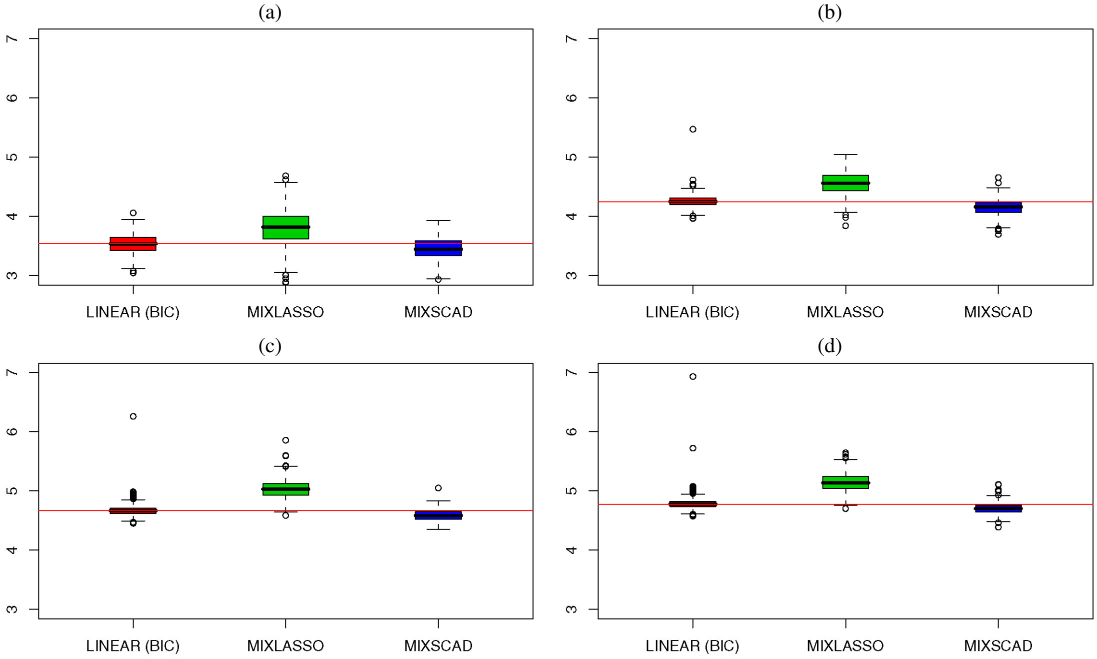

Figure 6. Predictive log-likelihood by MCCV: Baseball salary data. (a) $d = 50$; (b) $d = 100$; (c) $d = 150$; (d) $d = 167$.

Interpreting the outcome of the fit can be a source of controversy. In general, there should be some kind of positive correlation between a baseball player's performance and his salary; the better the player, the better the salary. Yet, the covariates are heavily correlated, which blurs such a correlation. The best that we can claim (according to MIXSCAD) is that a finite mixture of two regressions fits the data better than an ordinary linear model does. The biggest difference between the mixture and nonmixture models is the effect of being a free agent. According to the linear model, being a free agent helps achieve a higher salary, whereas the first component of the mixture model implies that being a free agent does not mean a higher salary for most players. This may explain the Player's Union argument that owners colluded to keep the salary of free agents in lower 1991/2 (Watnik 1998). In addition, according to the mixture model, the arbitration eligibility helps achieve a higher salary, but this is not the case according to the linear model.

The main differences between two components of the selected FMR model are interaction effects. Recall that x_1 and x_7 are individual performances, whereas x_3 and x_8 represent a player's contribution to the team. The effects of $x_1 * x_{13}$ and $x_8 * x_{13}$ in the FMR model imply that for most players, being eligible for free agency enhances the value of their team contribution (x_8) toward a higher salary, but not the value of their individual performance (x_1). However, for other players, we find that the effects are in the opposite direction, as indicated by the coefficients of $x_7 * x_{13}$ and $x_8 * x_{13}$ in the second component of the model.

A general conclusion is that a player's worth is not always reflected in his salary!

8. CONCLUSION

We have introduced the penalized likelihood approach for variable selection in the context of FMR models. The penalty function is designed to be dependent on the size of the regression coefficients and the mixture structure. The new procedure is shown to be consistent in selecting the most parsimonious FMR model. We also have proposed a data-adaptive method for selecting the tuning parameters and demonstrated its use through extensive simulations. The new method performs as well as (or better than in some examples) the BIC method while being computationally much more efficient. In addition, as in the market segmentation application, the new method can be used to suggest a set of plausible models to be examined by BIC if so desired. This helps substantially reduce the computational burden of using BIC, allowing use of the new method in reasonably high-dimensional problems.

APPENDIX: REGULARITY CONDITIONS AND PROOFS

To study the asymptotic properties of the proposed method, some regularity conditions on the joint distribution of $\mathbf{z} = (\mathbf{x}, Y)$ are required. In stating the regularity conditions, we write $\boldsymbol{\Psi} = (\psi_1, \psi_2, \ldots, \psi_v)$, so that v is the total number of parameters in the model. Let $f(\mathbf{z}; \boldsymbol{\Psi})$ be the joint density function of $\mathbf{z}$ and let $\boldsymbol{\Omega}$ be an open parameter space. The regularity conditions are as follows:

A_1. The density $f(\mathbf{z}; \boldsymbol{\Psi})$ has common support in $\mathbf{z}$ for all $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$, and $f(\mathbf{z}; \boldsymbol{\Psi})$ is identifiable in $\boldsymbol{\Psi}$ up to a permutation of the components of the mixture.

A_2. For each $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$, the density $f(\mathbf{z}; \boldsymbol{\Psi})$ admits third partial derivatives with respect to $\boldsymbol{\Psi}$ for almost all $\mathbf{z}$.

A_3. For each $\boldsymbol{\Psi}_0 \in \boldsymbol{\Omega}$, there exist functions $M_1(\mathbf{z})$ and $M_2(\mathbf{z})$ (possibly depending on $\boldsymbol{\Psi}_0$) such that for $\boldsymbol{\Psi}$ in a neighborhood of $N(\boldsymbol{\Psi}_0)$,

$$\left|\frac{\partial f(\mathbf{z}; \boldsymbol{\Psi})}{\partial \psi_j}\right| \le M_1(\mathbf{z}), \qquad \left|\frac{\partial^2 f(\mathbf{z}; \boldsymbol{\Psi})}{\partial \psi_j \, \partial \psi_l}\right| \le M_1(\mathbf{z}),$$

and

$$\left|\frac{\partial^3 \log f(\mathbf{z}; \boldsymbol{\Psi})}{\partial \psi_j \, \partial \psi_l \, \partial \psi_m}\right| \le M_2(\mathbf{z}),$$

such that $\int M_1(\mathbf{z}) \, d\mathbf{z} < \infty$ and $\int M_2(\mathbf{z}) f(\mathbf{z}; \boldsymbol{\Psi}) \, d\mathbf{z} < \infty$.

A_4. The Fisher information matrix,

$$I(\boldsymbol{\Psi}) = E\left\{\left[\frac{\partial}{\partial \boldsymbol{\Psi}} \log f(\mathbf{Z}; \boldsymbol{\Psi})\right]\left[\frac{\partial}{\partial \boldsymbol{\Psi}} \log f(\mathbf{Z}; \boldsymbol{\Psi})\right]^{\tau}\right\},$$

is finite and positive definite for each $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$.

Proof of Theorem 1

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

$$\lim_{n \to \infty} P\left\{\sup_{\|\mathbf{u}\| = M_\epsilon} \tilde{l}_n(\boldsymbol{\Psi}_0 + r_n \mathbf{u}) < \tilde{l}_n(\boldsymbol{\Psi}_0)\right\} \ge 1 - \varepsilon. \qquad \text{(A.1)}$$

Thus, with large probability, there is a local maximum in $\{\boldsymbol{\Psi}_0 + r_n \mathbf{u}; \|\mathbf{u}\| \le M_\epsilon\}$. This local maximizer, say $\hat{\boldsymbol{\Psi}}_n$, satisfies $\|\hat{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}_0\| = O_p(r_n)$.

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

$$\Delta_n(\mathbf{u}) = [l_n(\boldsymbol{\Psi}_0 + r_n \mathbf{u}) - l_n(\boldsymbol{\Psi}_0)] - [\mathbf{p}_n(\boldsymbol{\Psi}_0 + r_n \mathbf{u}) - \mathbf{p}_n(\boldsymbol{\Psi}_0)].$$

From $p_{nk}(0) = 0$, we have $\mathbf{p}_n(\boldsymbol{\Psi}_0) = \mathbf{p}_n(\boldsymbol{\Psi}_{01})$. Because $\mathbf{p}_n(\boldsymbol{\Psi}_0 + r_n \mathbf{u})$ is a sum of positive terms, removing terms corresponding to zero components makes it smaller; thus we have

$$\begin{aligned}\Delta_n(\mathbf{u}) \le{}& [l_n(\boldsymbol{\Psi}_0 + r_n \mathbf{u}) - l_n(\boldsymbol{\Psi}_0)] \\ &- [\mathbf{p}_n(\boldsymbol{\Psi}_{01} + r_n \mathbf{u}_{\mathbf{I}}) - \mathbf{p}_n(\boldsymbol{\Psi}_{01})], \qquad \text{(A.2)}\end{aligned}$$

where $\boldsymbol{\Psi}_{01}$ is the parameter vector with zero regression coefficients removed and $\mathbf{u}_{\mathbf{I}}$ is a subvector of $\mathbf{u}$ with corresponding components. By Taylor's expansion and the triangular inequality,

$$\begin{aligned} l_n(\boldsymbol{\Psi}_0 + r_n \mathbf{u}) - l_n(\boldsymbol{\Psi}_0) ={}& n^{-1/2}(1 + b_n) l_n'(\boldsymbol{\Psi}_0)^T \mathbf{u} \\ &- \frac{(1 + b_n)^2}{2}(\mathbf{u}^\tau I(\boldsymbol{\Psi}_0)\mathbf{u})(1 + o_p(1)) \end{aligned}$$

and

$$\begin{aligned} |\mathbf{p}_n(\boldsymbol{\Psi}_{01} + r_n \mathbf{u}_{\mathbf{I}}) - \mathbf{p}_n(\boldsymbol{\Psi}_{01})| \le{}& d b_n (1 + b_n)\|\mathbf{u}\| + \frac{c_n}{2}(1 + b_n)^2 \|\mathbf{u}\|^2 \\ &+ \sqrt{K} a_n (1 + b_n)\|\mathbf{u}\|, \end{aligned}$$

where $d = \max_k \sqrt{d_k}$ and d_k is the number of true nonzero regression coefficients in the kth component of the FMR model. Regularity conditions imply that $l_n'(\boldsymbol{\Psi}_0) = O_p(\sqrt{n})$ and $I(\boldsymbol{\Psi}_0)$ is positive definite. In addition, by condition P_1 for the penalty function, $c_n = o(1)$, and $a_n = o(1 + b_n)$. The order comparison of the terms in the foregoing two expansions implies that

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

is the sole leading term in the right side of (A.2). Therefore, for any given $\epsilon > 0$, there exists a sufficiently large M_ϵ such that

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

which implies (A.1). This completes the proof.

Proof of Theorem 2

To prove part a, partition $\mathbf{\Psi} = (\mathbf{\Psi}_1, \mathbf{\Psi}_2)$ for any $\mathbf{\Psi}$ in the neighborhood $\|\mathbf{\Psi} - \mathbf{\Psi}_0\| = O(n^{-1/2})$. By the definition of $\tilde{l}_n(\cdot)$, we have

$$\tilde{l}_n\{(\mathbf{\Psi}_1, \mathbf{\Psi}_2)\} - \tilde{l}_n\{(\mathbf{\Psi}_1, \mathbf{0})\} = \big[l_n\{(\mathbf{\Psi}_1, \mathbf{\Psi}_2)\} - l_n\{(\mathbf{\Psi}_1, \mathbf{0})\}\big] - \big[\mathbf{p}_n\{(\mathbf{\Psi}_1, \mathbf{\Psi}_2)\} - \mathbf{p}_n\{(\mathbf{\Psi}_1, \mathbf{0})\}\big].$$

We now find the order of two differences. By the mean value theorem,

$$l_n(\{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}) - l_n\{(\mathbf{\Psi}_1, \mathbf{0})\} = \left[\frac{\partial l_n\{(\mathbf{\Psi}_1, \boldsymbol{\xi})\}}{\partial \mathbf{\Psi}_2}\right]^\tau \mathbf{\Psi}_2 \qquad \text{(A.3)}$$

for some $\|\boldsymbol{\xi}\| \le \|\mathbf{\Psi}_2\| = O(n^{-1/2})$. Furthermore, by A_4 and the mean value theorem,

$$\begin{aligned}
&\left\|\frac{\partial l_n\{(\mathbf{\Psi}_1, \boldsymbol{\xi})\}}{\partial \mathbf{\Psi}_2} - \frac{\partial l_n\{(\mathbf{\Psi}_{01}, \mathbf{0})\}}{\partial \mathbf{\Psi}_2}\right\| \\
&\quad \le \left\|\frac{\partial l_n\{(\mathbf{\Psi}_1, \boldsymbol{\xi})\}}{\partial \mathbf{\Psi}_2} - \frac{\partial l_n\{(\mathbf{\Psi}_1, \mathbf{0})\}}{\partial \mathbf{\Psi}_2}\right\| + \left\|\frac{\partial l_n\{(\mathbf{\Psi}_1, \mathbf{0})\}}{\partial \mathbf{\Psi}_2} - \frac{\partial l_n\{(\mathbf{\Psi}_{01}, \mathbf{0})\}}{\partial \mathbf{\Psi}_2}\right\| \\
&\quad \le \left[\sum_{i=1}^n M_1(\mathbf{z}_i)\right]\|\boldsymbol{\xi}\| + \left[\sum_{i=1}^n M_1(\mathbf{z}_i)\right]\|\mathbf{\Psi}_1 - \mathbf{\Psi}_{01}\| \\
&\quad = \{\|\boldsymbol{\xi}\| + \|\mathbf{\Psi}_1 - \mathbf{\Psi}_{01}\|\}O_p(n) = O_p\big(n^{1/2}\big).
\end{aligned}$$

By the regularity conditions, $\partial l_n\{(\mathbf{\Psi}_{01}, \mathbf{0})\}/\partial \mathbf{\Psi}_2 = O_p(n^{1/2})$, and thus $\partial l_n(\{\mathbf{\Psi}_1, \boldsymbol{\xi}\})/\partial \mathbf{\Psi}_2 = O_p(n^{1/2})$. Applying these order assessments to (A.3), we get

$$l_n(\{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}) - l_n(\{\mathbf{\Psi}_1, \mathbf{0}\}) = O_p(\sqrt{n})\sum_{k=1}^K \sum_{j=d_k+1}^P |\beta_{jk}|$$

for large n. On the other hand,

$$\mathbf{p}_n(\{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}) - \mathbf{p}_n(\{\mathbf{\Psi}_1, \mathbf{0}\}) = \sum_{k=1}^K \sum_{j=d_k+1}^P \pi_k p_{nk}(\beta_{kj}).$$

Therefore,

$$\tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}) - \tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{0}\}) = \sum_{k=1}^K \sum_{j=d_k+1}^P \{|\beta_{kj}|O_p(\sqrt{n}) - \pi_k p_{nk}(\beta_{kj})\}.$$

In a shrinking neighborhood of 0, $|\beta_{kj}|O_p(\sqrt{n}) < \pi_k p_{nk}(\beta_{kj})$ in probability by condition P_2. This completes the proof of part a.

To prove part b.1, consider the partition $\mathbf{\Psi} = (\mathbf{\Psi}_1, \mathbf{\Psi}_2)$. Let $(\hat{\mathbf{\Psi}}_1, \mathbf{0})$ be the maximizer of the penalized log-likelihood function $\tilde{l}_n\{(\mathbf{\Psi}_1, \mathbf{0})\}$, which is considered as a function of $\mathbf{\Psi}_1$. It suffices to show that in the neighborhood $\|\mathbf{\Psi} - \mathbf{\Psi}_0\| = O(n^{-1/2})$, $\tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}) - \tilde{l}_n(\{\hat{\mathbf{\Psi}}_1, \mathbf{0}\}) < 0$ with probability tending to 1 as $n \to \infty$. We have that

$$\begin{aligned}
&\tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}) - \tilde{l}_n(\{\hat{\mathbf{\Psi}}_1, \mathbf{0}\}) \\
&\quad = \big[\tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}) - \tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{0}\})\big] + \big[\tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{0}\}) - \tilde{l}_n(\{\hat{\mathbf{\Psi}}_1, \mathbf{0}\})\big] \\
&\quad \le \big[\tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{\Psi}_2\}) - \tilde{l}_n(\{\mathbf{\Psi}_1, \mathbf{0}\})\big].
\end{aligned}$$

By the result in part a, the last expression is negative with probability tending to 1 as $n \to \infty$.

To prove part b.2, consider $\tilde{l}_n\{(\mathbf{\Psi}_1, \mathbf{0})\}$ a function of $\mathbf{\Psi}_1$. Using the same argument as in Theorem 1, there exists a $\sqrt{n}$-consistent local maximizer of this function, say $\hat{\mathbf{\Psi}}_1$, that satisfies

$$\frac{\partial \tilde{l}_n(\hat{\mathbf{\Psi}}_n)}{\partial \mathbf{\Psi}_1} = \left\{\frac{\partial l_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1} - \frac{\partial \mathbf{p}_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1}\right\}_{\hat{\mathbf{\Psi}}_n = (\hat{\mathbf{\Psi}}_1, \mathbf{0})} = \mathbf{0}. \qquad \text{(A.4)}$$

By the Taylor's series expansion,

$$\left.\frac{\partial l_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1}\right|_{\hat{\mathbf{\Psi}}_n = (\hat{\mathbf{\Psi}}_1, \mathbf{0})} = \frac{\partial l_n(\mathbf{\Psi}_{01})}{\partial \mathbf{\Psi}_1} + \left\{\frac{\partial^2 l_n(\mathbf{\Psi}_{01})}{\partial \mathbf{\Psi}_1 \partial \mathbf{\Psi}_1^\tau} + \mathbf{o}_p(n)\right\}(\hat{\mathbf{\Psi}}_1 - \mathbf{\Psi}_{01})$$

and

$$\left.\frac{\partial \mathbf{p}_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1}\right|_{\hat{\mathbf{\Psi}}_n = (\hat{\mathbf{\Psi}}_1, \mathbf{0})} = \mathbf{p}_n'(\mathbf{\Psi}_{01}) + \{\mathbf{p}_n''(\mathbf{\Psi}_{01}) + \mathbf{o}_p(n)\}(\hat{\mathbf{\Psi}}_1 - \mathbf{\Psi}_{01}),$$

where $\mathbf{p}_n'(\cdot)$ and $\mathbf{p}_n''(\cdot)$ are the first and second derivatives of $\mathbf{p}_n(\cdot)$. Substituting into (A.4), we find that

$$\left\{\frac{\partial^2 l_n(\mathbf{\Psi}_{01})}{\partial \mathbf{\Psi}_1 \partial \mathbf{\Psi}_1^\tau} - \mathbf{p}_n''(\mathbf{\Psi}_{01}) + \mathbf{o}_p(n)\right\}(\hat{\mathbf{\Psi}}_1 - \mathbf{\Psi}_{01}) = \frac{\partial l_n(\mathbf{\Psi}_{01})}{\partial \mathbf{\Psi}_1} - \mathbf{p}_n'(\mathbf{\Psi}_{01}).$$

On the other hand, under the regularity conditions, we have

$$\frac{1}{n}\frac{\partial^2 l_n(\mathbf{\Psi}_{01})}{\partial \mathbf{\Psi}_1 \partial \mathbf{\Psi}_1^\tau} = \mathbf{I}_1(\mathbf{\Psi}_{01}) + \mathbf{o}_p(\mathbf{1}) \qquad \text{and}$$

$$\frac{1}{\sqrt{n}}\frac{\partial l_n(\Psi_{01})}{\partial \mathbf{\Psi}_1} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{I}_1(\mathbf{\Psi}_{01})).$$

Using the foregoing facts and Slutsky's theorem, we have

$$\sqrt{n}\left\{\left[\mathbf{I}_1(\mathbf{\Psi}_{01}) - \frac{\mathbf{p}_n''(\mathbf{\Psi}_{01})}{n}\right](\hat{\mathbf{\Psi}}_1 - \mathbf{\Psi}_{01}) + \frac{\mathbf{p}_n'(\mathbf{\Psi}_{01})}{n}\right\} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \mathbf{I}_1(\mathbf{\Psi}_{01})),$$

which is the result in part b.2.

The proof of part c is obvious under the consistency assumption on $\hat{K}_n$. This completes the proof.

[Received June 2005. Revised April 2007.]

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.

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.

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.

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.

——— (2002), "Variable Selection for Cox's Proportional Hazards Model and Frailty Model," *The Annals of Statistics*, 30, 74–99.

Hennig, C. (2000), "Identifiability of Models for Clusterwise Linear Regression," *Journal of Classification*, 17, 273–296.

Hunter, D. R., and Li, R. (2005), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617–1642.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive Mixture of Local Experts," *Neural Computation*, 3, 79–87.

James, L. F., Priebe, C. E., and Marchette, D. J. (2001), "Consistent Estimation of Mixture Complexity," *The Annals of Statistics*, 29, 1281–1296.

Jiang, W., and Tanner, M. A. (1999), "Hierarchical Mixtures-of-Experts for Exponential Family Regression Models: Approximation and Maximum Likelihood Estimation," *The Annals of Statistics*, 27, 987–1011.

Keribin, C. (2000), "Consistent Estimation of the Order of Mixture Models," *Sankhyā*, Ser. A, 62, 49–66.

Khalili, A., and Chen, J. (2005), "Variable Selection in Finite Mixture of Regression Models," research paper, Department of Statistics and Actuarial Science, University of Waterloo, Canada, available at *www.stats.uwaterloo.ca/stats_navigation/techreports/05workingpapers/2005-03.pdf*.

Leeb, H., and Pötscher, B. M. (2003), "Finite Sample Distribution of Post-Model-Selection Estimates and Uniform versus Non-Uniform Approximations," *Econometric Theory*, 19, 100–142.

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

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*, 88, 486–494.

Skrondal, A., and Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modelling: Multilevel, Longitudinal, and Structural Equation Models*, Boca Raton, FL: Chapman & Hall/CRC.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society*, 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.

Wang, P., Puterman, M. L., Cockburn, I., and Le, N. (1996), "Mixed Poisson Regression Models With Covariate Dependent Rates," *Biometrics*, 52, 381–400.

Watnik, M. R. (1998), "Pay for Play: Are Baseball Salaries Based on Performance?" *Journal of Statistics Education*, 6, available at *www.amstat.org/publications/jse/v6n2/datasets.watnik.html*.

Wedel, M., and Kamakura, W. A. (2000), *Market Segmentation: Conceptual and Methodological Foundations* (2nd ed.), Boston: Kluwer Academic Publishers.