VARIABLE SELECTION
IN FINITE MIXTURE OF REGRESSION MODELS

Abbas Khalili and Jiahua Chen [1]

Department of Statistics and Actuarial Science, University of Waterloo

Abstract: In the applications of finite mixture of regression models, a large number of covariates are often used and their contributions toward the response variable vary from one component to another of the mixture model. This creates a complex variable selection problem. Existing methods, such as AIC and BIC, are computationally expensive as the number of covariates and the components in the mixture model increase. In this paper, we introduce a penalized likelihood approach for variable selection in finite mixture of regression models. The new method introduces a penalty which depends on the sizes of regression coefficients and the mixture structure. The new method is shown to have the desired sparsity property. A data adaptive method for selecting tuning parameters, and an EM-algorithm for efficient numerical computations are developed. Simulations show that the method has very good performance with much lower demand on computing power. The new method is also illustrated by analyzing a real data set in marketing applications.

[1]Address for correspondence: Jiahua Chen, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON. Canada, N2L 3G1. 519-888-4567-ex5506, jhchen@uwaterloo.ca

KEY WORDS: E-M algorithm, LASSO, Mixture model, Penalty method, SCAD.

1. INTRODUCTION

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

Often, in the initial stage of a study, a large number of 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 a common practice to include only important covariates in the model.

The problem of variable selection in FMR models has received much attention recently. All-subset selection methods such as Akaike Information Criterion (AIC); Akaike (1973), Bayes Information Criterion (BIC); Schwartz (1978), and their mod-

ifications 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. In addition, these methods are unstable due to their inherited discreteness; Breiman (1996). It is also more difficult to study the theoretical sampling properties of resulting parameter estimators.

Due to these difficulties, the new generation of variable selection methods, such as the Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996) and Smoothly Clipped Absolute Deviation (SCAD) method by Fan and Li (2001, 2002), are particularly advantageous for variable selection in the context of FMR models. The LASSO and SCAD are different from the traditional variable selection methods in that they delete the non-significant covariates in the model by estimating their effects as 0. In this paper, we design a new variable selection procedure for FMR models based on these methods. A new class of penalty functions to be used for variable selection in FMR models is proposed. We investigate the 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 and computationally efficient. The performance of the method is studied theoretically

and via simulations. Our simulations indicate that the new method has power in selecting correct models similar to or better than BIC, with much less computational effort.

The paper 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 via simulations, and Section 7 presents a real data analysis to illustrate the use of the new method. Section 8 contains conclusions.

2. FINITE MIXTURE OF REGRESSION MODELS

Let Y be a response variable of interest and $\boldsymbol{x} = (x_1, x_2, \ldots, x_P)^T$ be the vector of covariates which are believed to have effect on Y. The finite mixture of regression 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 $\Theta \subset \boldsymbol{R}$, and ϕ is a dispersion parameter. We say that $(\boldsymbol{x}, Y)$ follows a finite mixture of regression model*

of order K if the conditional density function of Y given $\boldsymbol{x}$ has the form

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

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

Model (1) can be generalized to allow π_k to be functions of $\boldsymbol{x}$. We decide to 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's are assumed equal.

The FMR models combine the characteristics of the regression models with those of the finite mixture models. Like any regression model, the FMR models are 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 base 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. Many finite mixture models, including the mixtures

of Binomial, Multi-nomial, Normal, and Poisson distributions, are identifiable, under some conditions. See Titterington, Smith and Markov (1985).

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

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

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

When we exchange the order of two regression components, the parameter $\boldsymbol{\Psi}$ changes. In the above definition, we interpret $\boldsymbol{\Psi} = \boldsymbol{\Psi}^*$ up to a permutation. In general, 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 $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. Hennig (2000) pointed out that for fixed designs, a sufficient condition for identifiability is that the design points spread over a set that cannot be covered by K $(P-1)$-dimensional linear sub-spaces in addition to some usual conditions on the component density. This condition is applicable to Poisson and Normal FMR models. If $\boldsymbol{x}_i$'s are also a random sample from a marginal density $f(\boldsymbol{x})$ which does not depend on $\boldsymbol{\Psi}$, then $f(\boldsymbol{x})$ must not have all its mass in up to K of $(P-1)$-dimensional linear sub-spaces. Some discussions can also be found in Wang et al.

(1996). In this paper, we assume the FMR model under consideration is identifiable with the given or random design.

3. The METHOD FOR VARIABLE SELECTION

In the case when $\boldsymbol{x}$ is random, we assume that its density $f(\boldsymbol{x})$ 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 $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ be a sample of observations from the FMR model (1). The (conditional) log-likelihood function of $\boldsymbol{\Psi}$ is given by

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

When the effect of a component of $\boldsymbol{x}$ is not significant, the corresponding ordinarily maximum likelihood estimate is often close but not equal to 0. Thus, this covariate is not excluded from the model. To avoid this problem, one may study sub-models with various components of $\boldsymbol{x}$ excluded as is done by AIC and BIC. The computational burden of these approaches is however heavy and is to be avoided. The approach we consider in this paper is as follows.

We define a penalized log-likelihood function as

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

with the penalty function

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

where $p_{nk}(\beta_{kj})$'s are non-negative and non-decreasing functions in $|\beta_{kj}|$. By maximizing $\tilde{l}_n(\boldsymbol{\Psi})$ that contains a penalty, there is a positive chance to have some estimated values of β equaling zero and hence automatically select a sub-model. Thus, the procedure combines the variable selection and parameter estimation in one step and reduces the computational burden substantially. In (3), we choose the amount of penalty imposed on the regression coefficients within the kth component of the FMR model to be proportional to π_k. This is inline with the common practice to relate the amount of 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 are available on the importance of 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, and under the guidance of the statistical theory. The following three penalty functions have been investigated in the literature in a number of contents, and will be used to illustrate the theory we develop for the FMR models.

(a) L_1-norm penalty: $p_{nk}(\beta) = \gamma_{nk}\sqrt{n}|\beta|$.

(b) HARD penalty: $p_{nk}(\beta) = \gamma_{nk}^2 - (\sqrt{n}|\beta| - \gamma_{nk})^2 I(\sqrt{n}|\beta| < \gamma_{nk})$.

(c) 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 is used in LASSO by Tibshirani (1996). The other two are discussed in Fan and Li (2001, 2002). The constants $\gamma_{nk} > 0$ and $a > 2$ are chosen based on how hard the procedure tries to eliminate the covariates from the model. In applications, their choices may be determined based on some prior information, i.e. subjectively by the data analysts or by some data-driven methods. We call the penalty function $\boldsymbol{p}_n(\cdot)$ in (3) constructed from the LASSO, HARD and SCAD, as MIXLASSO, MIXHARD and MIXSCAD penalties, respectively.

The three penalty functions have similar properties with some subtle differences. Maximizing the penalized likelihood is equivalent to constrained maximization. When the constraint is tightened, SCAD quickly removes variables with smaller effects and leaves larger effects untouched while LASSO reduces all effects at the same rate. Thus, SCAD estimates non-zero effects with high efficiency while LASSO may not even maintain the best possible convergence rate. Intuitively, HARD should work more like SCAD except 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 0 effects. In general, the set of non-zero effects $\boldsymbol{\beta}_{1k}$ may depend on k. We choose not to 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 $\boldsymbol{\beta}_{2k} : k = 1, \ldots, K$. 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:

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

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

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 will be 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 non-negative. In addition, it is non-decreasing and two times differentiable for β in $(0, \infty)$ with at most a finite number of exceptions.

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

P_2. For $N_n = \{\beta; 0 < \beta \leq n^{-1/2} \log n\}$, $\lim_{n\to\infty} \inf_{\beta \in N_n} \frac{p'_{nk}(\beta)}{\sqrt{n}} = \infty$.

Conditions P_0 and P_2 are needed for sparsity. Condition P_1 is used to preserve the asymptotic properties of the estimators of non-zero effects in the model. To develop

asymptotic theory, some commonly used regularity conditions are needed on the joint density function $f(\boldsymbol{z};\boldsymbol{\Psi})$ of $\boldsymbol{Z}=(\boldsymbol{x},Y)$. To focus on the main results, we left them in Appendix.

Theorem 1 *Let $\boldsymbol{Z}_i=(\boldsymbol{x}_i,Y_i), i=1,2,\ldots,n$, be a random sample from the density function $f(\boldsymbol{z};\boldsymbol{\Psi})$ that satisfies the regularity conditions A1-A5 in the Appendix. Suppose that the penalty functions $p_{nk}(\cdot)$'s 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*

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

When $b_n=O(1)$ such as in the cases of MIXHARD and MIXSCAD, $\hat{\boldsymbol{\Psi}}$ has usual convergence rate $n^{-1/2}$. This property is lost if $b_n\to\infty$ which is likely the case of MIXLASSO as we will see.

The penalized likelihood method estimates some regression parameters exactly 0 with positive probability. This leads to sparsity which is sometimes referred as oracle property although Donoho and Jonhstone (1994) introduced this terminology in a different context. The oracle property is a *super-efficiency phenomenon* first noticed by Hodges; see Ferguson (1996). Being super-efficient does not in general help in terms of accuracy of confidence intervals. See Leeb and Pöscher(2003) for through discussion. Yet it is the key for variable selection. The next Theorem proves the oracle property under some mild conditions.

Theorem 2 *Assume conditions in Theorem 1, the penalty functions $p_{nk}(\cdot)$ satisfy P_0-P_2, and K is known in parts (a) and (b) below. We have*

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

(i) Sparsity: $P\{\hat{\boldsymbol{\beta}}_{2k} = \mathbf{0}\} \to 1$, $k = 1, 2, \ldots, K$ as $n \to \infty$.

(ii) Asymptotic normality:

$$\sqrt{n}\left\{\left[\boldsymbol{I_1(\Psi_{01})} - \frac{\boldsymbol{p}_n''(\boldsymbol{\Psi}_{01})}{n}\right](\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) + \frac{\boldsymbol{p}_n'(\boldsymbol{\Psi}_{01})}{n}\right\} \longrightarrow^d N(\mathbf{0}, \boldsymbol{I_1(\Psi_{01})})$$

where $\boldsymbol{I_1(\Psi_1)}$ 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 $p_n(\cdot)$ in (b)-(ii) become negligible by some choices of the penalty function other than providing some finite sample adjustment. The result suggests a variance estimator of $\hat{\boldsymbol{\Psi}}_1$ as follows.

$$\hat{Var}(\hat{\boldsymbol{\Psi}}_1) = \{l_n''(\hat{\boldsymbol{\Psi}}_1) - \boldsymbol{p}_n''(\boldsymbol{\Psi})\}^{-1}\hat{Var}\{l_n'(\hat{\boldsymbol{\Psi}}_1)\}\{l_n''(\hat{\boldsymbol{\Psi}}_1) - \boldsymbol{p}_n''(\boldsymbol{\Psi})\}^{-1}. \tag{4}$$

Keribin (2000) showed that under certain regularity conditions, the order of a finite mixture model can be estimated consistently by using penalized-likelihood-based approaches such as the BIC criterion. In applications, one can first use the BIC or the scientific background to first identify the order of the full FMR model. Most statistical methods have some limitations. We should not overly rely on a computer to produce a perfect model in applications (Burnham and Anderson, 2002, page 15). When K cannot be reliably determined, one must be very cautious in using variable selection procedures. Only after the order is reliably estimated, a variable selection procedure is recommended.

In the light of this theorem, the method has different asymptotic properties when we use different penalty functions. It is impossible to choose a γ_{nk} in the L_1-norm penalty function to achieve both sparsity and to maintain root-n consistency of the parameter estimators. By choosing proper γ_{nk} in the other two penalty functions, however, the sparsity and root-n consistency can be achieved simultaneously. For example, choosing $\gamma_{nk} = \log n$ in MIXSCAD or MIXHARD penalties will do.

5. NUMERICAL SOLUTIONS

There are no apparent analytical solutions to the maximization problem posted when applying the new variable selection procedure. We discuss a numerical method that combines the traditional EM algorithm applied to finite mixture models, and

the revised maximization in the M-step.

5.1 Maximization of the Penalized Log-likelihood Function

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ 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, $p_{nk}(\beta)$'s are not differentiable at $\beta = 0$. The Newton-Raphson algorithm can not be directly used in the M-step of the EM algorithm unless it is properly adopted to deal with the single non-smooth point at $\beta = 0$. We follow 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 infinite whenever $|\beta| \to \infty$ which is more suitable to our application than the simple Taylor's expansion. Let $\boldsymbol{\Psi}^{(m)}$ be the parameter value after the mth iteration. We replace $\boldsymbol{p}_n(\boldsymbol{\Psi})$ in the penalized log-likelihood function in (2) by the following function:

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

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

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

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

E-Step: Let $\boldsymbol{\Psi}^{(m)}$ be the estimate of the parameters after the mth iteration. The E-step computes the conditional expectation of the function $\tilde{l}_n^c(\boldsymbol{\Psi})$ with respect to z_{ik}, given the data $(\boldsymbol{x}_i, y_i)$, and assume the current estimate $\boldsymbol{\Psi}^{(m)}$ are the true parameters of the model. The conditional expectation is found to be

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

where the weights

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

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 a usual EM-algorithm, the mixing proportions are updated by

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

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

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

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

where $\tilde{p}_{nk}(\beta_{kj})$ is the corresponding term in $\tilde{\boldsymbol{p}}_n(\mathbf{\Psi}; \mathbf{\Psi}^{(m)})$, 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 the equations

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

Starting from an initial value $\mathbf{\Psi}^{(0)}$, we iterate between the E and M-steps until some convergence criterion is satisfied. When the algorithm converges, the equation

$$\frac{\partial l_n(\mathbf{\Psi}_n)}{\partial \beta_{kj}} - p'_{nk}(\beta_{kj}) = 0 \tag{7}$$

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

5.2 Choice of the Tuning Parameters

In using MIXLASSO, MIXHARD, MIXSCAD and other penalty functions, we need to choose the sizes of some tuning parameters γ_{nk}. The current theory only

provides some guidance on the order of γ_{nk} to ensure the sparsity property. In applications, the cross-validation (CV); Stone (1974), or generalized cross validation (GCV); Craven and Wahba (1979), are often used for choosing tuning parameters. Following the examples 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 MLE 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 by fixing the rest of 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 as

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

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

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

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

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

where $\Sigma_k(\hat{\boldsymbol{\beta}}_k) = \hat{\pi}_k \; diag\{p'_{nk}(\hat{\beta}_{k1})/\hat{\beta}_{k1}, \ldots, p'_{nk}(\hat{\beta}_{kP})/\hat{\beta}_{kP}\}$, and tr stands for trace and

diag for diagonal matrix. The tuning parameters, γ_{nk}'s, 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 for example, James, Priebe and Marchette (2001). The following result is obvious and the proof is omitted.

Theorem 3 *Consider the MIXSCAD or MIXHARD penalty functions given in Section 3. If the tuning parameter* $\lambda_{nk} = \frac{\gamma_{nk}}{\sqrt{n}}$ *is chosen by minimizing the CV or GCV over the interval* $[\alpha_n, \beta_n]$ *such that* $0 \leq \alpha_n \leq \beta_n$, *and* $\beta_n \to 0$ *and* $\sqrt{n}\alpha_n \to \infty$, *as* $n \to \infty$, *then the results in Theorems 1 and 2 still hold.*

Let $\alpha_n = C_1 n^{-1/2} \log n$, $\beta_n = C_2 n^{-1/2} \log n$ for some constants $0 < C_1 < C_2$. Then (α_n, β_n) will met the conditions in the above theorem.

6. SIMULATION STUDY

Our simulations are based on the Normal FMR model $\pi N(\boldsymbol{x}^\tau \boldsymbol{\beta}_1, \sigma^2) + (1 - \pi)N(\boldsymbol{x}^\tau \boldsymbol{\beta}_2, \sigma^2)$ with $\sigma^2 = 1$ and $P = 5$. We assume $K = 2$ is known. When K is unknown, one may use BIC to select K under the full regression model. When $\pi = 0.5$ we found that $\hat{K} = 2$ in 996 simulations out of 1000. When $\pi = 0.1$, the data do not contain enough information to choose K consistently.

The covariate $\boldsymbol{x}$ in the simulation is generated from multivariate normal with mean 0, variance 1, and correlation $\text{Cor}(x_i, x_j) = (0.5)^{|i-j|}$. Table 1 specifies the regression coefficients $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and three choices of mixing proportion π. The M_1 and M_2 represent the FMR models with parameter values given in the table. One thousand data sets with sample sizes $n = 100, 200$ from each FMR model were generated. We also simulated Binomial FMR models and the outcomes are similar and the results are not reported.

We compare the performance of different variable selection methods from a number of angles. The first is the average correct and incorrect estimated zero effects in each component of the FMR model. The second is standard errors of the estimated non-zero regression coefficients. At last, we generated a set of 10,000 test observations aside from each model, and computed the log-likelihood values of each submodel selected. A good variable selection method should consistently produce large log-likelihood values based on the test data set. For the current model, there are a total of 1024 potential sub-models all of which have to be examined by the BIC method. To reduce the computational burden, we only considered a set of 182 most probable models.

Table 2 contains the average numbers of correctly and incorrectly estimated zero coefficients with MIXSCAD shortened as MS and so on. Based on these results, BIC,

MH and MS have similar performances, and they all outperform the ML. When the sample size increases, all methods improve, and the performance of the ML becomes reasonable. 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, when a fitted mixing proportion is low combined with small sample size, one should be cautious in interpreting the result of the corresponding regression component. Table 3 reports the standard error (SD) of the non-zero regression coefficients based on the same 1000 samples as in Table 2, and its estimate (SD_m) based on formula (4). For robustness, SD is computed as median absolute deviation scaled by a factor of 0.6745 as in Fan and Li (2001). We observe that the methods under consideration do not differ substantially in this respect, and the variance estimators are all reasonably accurate. Other than MIXLASSO, the biases for estimating the non-zero coefficients are very low and the details are omitted here. The ultimate goal of variable selection is to identify a submodel with the best predictive value. The predicted log-likelihood values reported in Table 7 compare these methods from this angle. Based on the 25%, 50% and 75% quantiles, the oracle model is the clear winner followed by MIXSCAD. The BIC method turns out to be the worst.

Tables 3, 7 omit some results for $n = 200$, and $\pi = 0.1$. Let us mention only that increasing the sample size improves the performance of all methods but does not

change their comparison. In the same vein, reducing π makes all methods poor but does not change their comparison either.

Lastly, we investigate the situation when the number of covariates is relatively large by setting the number of covariates equal 40 with 15 and 30 non-zero coefficients for two components of the normal FMR model. The parameter values are

$$\begin{aligned}\beta_1 &= (1.5, 2, 1.5, 3, 2, 0, 0, \ldots, 0, 2, 3, 1.5, 1.5, 2, 2, 3, 2, 2, 2, 0, 0, \ldots, 0)\\ \beta_2 &= (0, 0, 0, 0, 0, -2, 2, 2, 1.5, 1.5, 2, 3, -2, 2, 3, 0, 0, 0, 0, 0, 2, 2, \ldots, 2).\end{aligned}$$

The covariates are from auto-regression model with mean 0, variance 1 and with correlation coefficient $\text{cor}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0.5^{|i-j|}$. We choose $n = 300$ and generated 1000 samples. In this case, BIC becomes impractical due to the amount of computation. The simulation results of other methods are reported in Tables 5 and 6. The MIXSCAD and MIXHARD still have good performances.

7. REAL DATA ANALYSIS

We analyzed a real data set from marketing applications to further demonstrate the use of the new method. The FMR models have often 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. The FMR models provide a model-based approach for

market segmentation; See Wedel and Kamakura (2000).

In marketing research, selected consumers are repeatedly asked to choose one product from a collection of hypothetical products with various features. The data collected from such experiments are analyzed to provide estimates of the market shares of new products. This method gives the researchers some idea on which products are likely to be successful before they are introduced to the market. The data set is from a conjoint choice experiment conducted in the Netherlands and is available on the website *www.gllamm.org/books* provided by Skrondal and Rabe-Hesketh (2004) . The authors analyzed the data by fitting a multi-nomial logit FMR model. The variable selection problem presents itself naturally but was not discussed in their book.

7.1 Data: Consumer Preferences for Coffee Makers

A conjoint choice experiment was conducted at a large shopping mall in the Netherlands regarding *the consumer preferences for coffee makers*. The main goal of the study was to estimate the market share for coffee makers with different features. The hypothetical coffee-makers have five attributes: brand name (3 levels), capacity (3 levels), price (3 levels), thermos (2 levels) and filter (2 levels). The levels of the attributes are given in Table 8.

A total of 16 profiles were constructed by combining the levels of the above attributes. Two groups of 8 choice sets were constructed with each set containing three

profiles (alternatives). One of the three profiles is common to provide a base choice. There were 185 respondents participating in the experiment. They were randomly divided into two 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 resulted from the above experiment were binary responses from the participants, indicating their profile choices. For subject i, on replication j, we get a three dimensional response vector: $\boldsymbol{y}_{ij}^T = (y_{ij1}, y_{ij2}, y_{ij3})$. The five attributes are the covariates in the model.

7.2 Model and Data Analysis

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

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

where

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

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

Skrondal and Rabe-Hesketh (2004) obtained MLEs of the parameters with $\hat{\pi} = 0.28$. Thus, the estimated size of the first market segment as 72% and that of the second segment as 28%. The MLEs of $\boldsymbol{\beta}_k$'s are given in Table 8, column SRH. The coefficient estimate of the first market segment, $\hat{\boldsymbol{\beta}}_1$, is given in the top half, and $\hat{\boldsymbol{\beta}}_2$ is in the lower half of the table.

Apparently, some of the regression coefficients are not significant and a variable selection procedure is needed. We applied MIXLASSO, MIXHARD and MIXSCAD methods to this data and used the GCV criterion outlined in Section 5.2. The new method with MIXHARD and MIXSCAD penalties chose the same model with more zero coefficients than the model chosen by MIXLASSO penalty. We only reported the results based on the MIXSCAD penalty in Table 8. The data adaptive choice of tuning parameters were 0.1 and 0.27 for the first and the second segments of the FMR model. The mixing proportion π was estimated as 26%. We also applied the BIC criterion to the data. In the light of the model chosen by the new method with MIXSCAD penalty, we considered a collection of 12 models to be examined by the BIC. Note that total number of possible models is at least 961, which is much larger. The outcome was the same as the new method with the MIXSCAD penalty. The parameter estimates and their corresponding standard errors are presented in Table 8. 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, respectively. They are clearly comparable in this respect.

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

8. CONCLUSION

We introduced the penalized likelihood approach for variable selection in the context of finite mixture of regression. 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 proposed a data adaptive method for selecting the tuning parameters and demonstrate the usage by extensive simulations. The new method with the MIXHARD and MIXSCAD penalty functions performed as well as the BIC method while it is computationally much more efficient. In addition, as in the example of market segmentation

application, the new method can also be used to suggest a set of plausible models to be examined by the BIC method if desired. This helps to reduce the computational burden of using the BIC, substantially.

Acknowledgment: The authors wish to thank the associate editor and the referees for constructive comments which lead to clarifications of important concepts and a much improved paper. The research is partially supported by the Natural Science and Engineering Research Council of Canada.

APPENDIX: Regularity Conditions and Proofs

To study the asymptotic properties of the proposed method, some regularity conditions on the joint distribution of $\boldsymbol{z} = (\boldsymbol{x}, Y)$ 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 $f(\boldsymbol{z}; \boldsymbol{\Psi})$ be the joint density function of $\boldsymbol{z}$ and $\boldsymbol{\Omega}$ be an open parameter space.

Regularity Conditions:

A_1 The density $f(\boldsymbol{z}; \boldsymbol{\Psi})$ has common support in $\boldsymbol{z}$ for all $\boldsymbol{\Psi} \in \boldsymbol{\Omega}$, and $f(\boldsymbol{z}; \boldsymbol{\Psi})$ 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 $f(\boldsymbol{z}; \boldsymbol{\Psi})$ admits third partial derivatives with respect to $\boldsymbol{\Psi}$ for almost all $\boldsymbol{z}$.

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

$$\left|\frac{\partial f(\boldsymbol{z};\boldsymbol{\Psi})}{\partial \psi_j}\right| \leq M_1(\boldsymbol{z}) \ , \quad \left|\frac{\partial^2 f(\boldsymbol{z};\boldsymbol{\Psi})}{\partial \psi_j \partial \psi_l}\right| \leq M_1(\boldsymbol{z}) \ , \quad \left|\frac{\partial^3 \log f(\boldsymbol{z};\boldsymbol{\Psi})}{\partial \psi_j \partial \psi_l \partial \psi_m}\right| \leq M_2(\boldsymbol{z})$$

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

A_4 The Fisher information matrix

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

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

$$\lim_{n\to\infty} P\left\{\sup_{\|\boldsymbol{u}\|=M_\epsilon} \tilde{l}_n(\boldsymbol{\Psi}_0 + r_n\boldsymbol{u}) < \tilde{l}_n(\boldsymbol{\Psi}_0)\right\} \geq 1 - \varepsilon \tag{9}$$

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

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

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

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

$$\Delta_n(\boldsymbol{u}) \leq [l_n(\boldsymbol{\Psi}_0 + r_n\boldsymbol{u}) - l_n(\boldsymbol{\Psi}_0)] - [\boldsymbol{p}_n(\boldsymbol{\Psi}_{01} + r_n\boldsymbol{u}_{\boldsymbol{I}}) - \boldsymbol{p}_n(\boldsymbol{\Psi}_{01})] \tag{10}$$

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

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

$$|\boldsymbol{p}_n(\boldsymbol{\Psi}_{01}+r_n\boldsymbol{u_I}) - \boldsymbol{p}_n(\boldsymbol{\Psi}_{01})| \le d\ b_n(1+b_n)\ \|\boldsymbol{u}\| + \frac{c_n}{2}(1+b_n)^2\|\boldsymbol{u}\|^2 + \sqrt{K}a_n(1+b_n)\|\boldsymbol{u}\|$$

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

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

is the sole leading in the right hand side of (10). Therefore, for any given $\epsilon > 0$, there exists a sufficiently large M_ϵ such that

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

which implies (9), and this completes the proof. ♠

Proof of Theorem 2: (a). Partition $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)$ for any $\boldsymbol{\Psi}$ in the neighborhood

$\|\boldsymbol{\Psi} - \boldsymbol{\Psi}_0\| = O(n^{-1/2})$. By the definition of $\tilde{l}_n(\cdot)$, we have

$$\begin{aligned}
&\tilde{l}_n\{(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)\} - \tilde{l}_n\{(\boldsymbol{\Psi}_1, \mathbf{0})\} \\
&\quad = \quad [l_n\{(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)\} - l_n\{(\boldsymbol{\Psi}_1, \mathbf{0})\}] - [\boldsymbol{p}_n\{(\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)\} - \boldsymbol{p}_n\{(\boldsymbol{\Psi}_1, \mathbf{0})\}]
\end{aligned}$$

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

$$l_n(\{\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2\}) - l_n\{(\boldsymbol{\Psi}_1, \mathbf{0})\} = \left[\frac{\partial l_n\{(\boldsymbol{\Psi}_1, \boldsymbol{\xi})\}}{\partial \boldsymbol{\Psi}_2}\right]^{\tau} \boldsymbol{\Psi}_2 \tag{11}$$

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

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

By the regularity conditions, $\partial l_n\{(\boldsymbol{\Psi}_{01}, \mathbf{0})\}/\partial \boldsymbol{\Psi}_2 = O_p(n^{1/2})$, thus $\partial l_n(\{\boldsymbol{\Psi}_1, \boldsymbol{\xi}\})/\partial \boldsymbol{\Psi}_2 = O_p(n^{1/2})$. Applying these order assessment to (11), we get

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

for large n. On the other hand,

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

Therefore,

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

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 (a).

(b). (i). 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 regarded 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 one as $n \to \infty$. We have that

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

By the result (a), the last expression is negative with probability tending to one as $n \to \infty$. This completes the proof of (i).

(ii). Regard $\tilde{l}_n\{(\mathbf{\Psi}_1, \mathbf{0})\}$ as 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 $\hat{\mathbf{\Psi}}_1$, which satisfies

$$\frac{\partial \tilde{l}_n(\hat{\mathbf{\Psi}}_n)}{\partial \mathbf{\Psi}_1} = \left\{ \frac{\partial l_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1} - \frac{\partial \boldsymbol{p}_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1} \right\}_{\hat{\mathbf{\Psi}}_n = (\hat{\mathbf{\Psi}}_1, \mathbf{0})} = \mathbf{0} \tag{12}$$

By the Taylor's series expansion,

$$\begin{aligned}
\left.\frac{\partial l_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1}\right|_{\hat{\mathbf{\Psi}}_n = (\hat{\mathbf{\Psi}}_1, \mathbf{0})} &= \frac{\partial l_n(\mathbf{\Psi}_{01})}{\partial \mathbf{\Psi}_1} + \left\{ \frac{\partial^2 l_n(\mathbf{\Psi}_{01})}{\partial \mathbf{\Psi}_1 \partial \mathbf{\Psi}_1^\tau} + \mathbf{o_p(n)} \right\} (\hat{\mathbf{\Psi}}_\mathbf{1} - \mathbf{\Psi}_\mathbf{01}) \ , \\
\left.\frac{\partial \boldsymbol{p}_n(\mathbf{\Psi})}{\partial \mathbf{\Psi}_1}\right|_{\hat{\mathbf{\Psi}}_n = (\hat{\mathbf{\Psi}}_1, \mathbf{0})} &= \boldsymbol{p}_n'(\mathbf{\Psi}_{01}) + \left\{ \boldsymbol{p}_n''(\mathbf{\Psi}_{01}) + \mathbf{o_p(n)} \right\} (\hat{\mathbf{\Psi}}_\mathbf{1} - \mathbf{\Psi}_\mathbf{01})
\end{aligned}$$

where $\boldsymbol{p}'_n(\cdot)$ and $\boldsymbol{p}''_n(\cdot)$ are the first and second derivatives of $\boldsymbol{p}_n(\cdot)$. Substituting to (12), we find

$$\left\{\frac{\partial^2 l_n(\boldsymbol{\Psi}_{01})}{\partial\boldsymbol{\Psi}_1\partial\boldsymbol{\Psi}_1^\tau} - \boldsymbol{p}''_n(\boldsymbol{\Psi}_{01}) + \mathbf{o_p(n)}\right\}(\hat{\boldsymbol{\Psi}}_\mathbf{1} - \boldsymbol{\Psi}_\mathbf{01}) = \frac{\partial\mathbf{l_n}(\boldsymbol{\Psi}_\mathbf{01})}{\partial\boldsymbol{\Psi}_\mathbf{1}} - \boldsymbol{p}'_\mathbf{n}(\boldsymbol{\Psi}_\mathbf{01}).$$

On the other hand, under the regularity conditions,

$$\frac{1}{n}\frac{\partial^2 l_n(\boldsymbol{\Psi}_{01})}{\partial\boldsymbol{\Psi}_1\partial\boldsymbol{\Psi}_1^\tau} = \boldsymbol{I_1(\Psi_{01})} + \mathbf{o_p(1)} \quad , \quad \frac{\mathbf{1}}{\sqrt{\mathbf{n}}}\frac{\partial\mathbf{l_n}(\boldsymbol{\Psi}_\mathbf{01})}{\partial\boldsymbol{\Psi}_\mathbf{1}} \longrightarrow^{\mathbf{d}} \mathbf{N}(\mathbf{0}, \boldsymbol{I_1(\Psi_{01})}).$$

Using the above facts and the Slutsky's Theorem, we have

$$\sqrt{n}\left\{\left[\boldsymbol{I_1(\Psi_{01})} - \frac{\boldsymbol{p}''_n(\boldsymbol{\Psi}_{01})}{n}\right](\hat{\boldsymbol{\Psi}}_1 - \boldsymbol{\Psi}_{01}) + \frac{\boldsymbol{p}'_n(\boldsymbol{\Psi}_{01})}{n}\right\} \longrightarrow^d N(\mathbf{0}, \boldsymbol{I_1(\Psi_{01})})$$

which is the result in (ii).

(c). The proof is obvious under the consistency assumption on $\hat{K}$. This completes the proof. ♠

REFERENCES

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

Burnham, K. P. and Anderson, D. R. (2002). *Model Selection and Multimodel Inference, A practical Information-Theoretic Approach.* 2nd ed, Springer.

Breiman, L. (1996), "Heuristics of instability and stabilization in model selection," *The Annals of Statistics*, 24, 2350-2383.

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

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.

Donoho, D. L. and Johnstone, I. M. (1994), "Ideal spatial adaptation by wavelet shrinkage," *Biometrika*, 81, 425-455.

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.

Fan, J. and Li, R. (2002), "Variable selection for Cox's proportional hazards model and frailty model," *The Annals of Statistics*, 30, 74-99.

Ferguson, T. S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, New York.

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," *Sankhya, ser. A*, 62, 49-66.

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.

Skrondal, A. and Rabe-Hesketh, S. (2004), *Generalized Latent Variable Modelling: Multilevel, Longitudinal, and Structural Equation Models*, 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.

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

Table 1: Regression coefficients in the Normal FMR models

Parameters	M_1	M_2
$\boldsymbol{\beta}_1$	(1, 0, 0, 3, 0)	(1, 0.6, 0, 3, 0)
$\boldsymbol{\beta}_2$	(-1, 2, 0, 0, 3)	(-1, 0, 0, 4, 0.7)
π	0.5, 0.3, 0.1	0.5, 0.3, 0.1

Table 2: Average numbers of correct and incorrect estimated zero coefficients Normal FMR models with n = 100 (200)

Method	Model M_1				Model M_2			
	Cor.	Inc.	Cor.	Inc.	Cor.	Inc.	Cor.	Inc.
	Com.1		Com.2		Com.1		Com.2	
	$\pi = 0.5$							
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)
	$\pi = 0.3$							
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)
	$\pi = 0.1$							
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)

Table 3: The standard errors of $\hat{\beta}$'s and their estimates
(Normal FMR models with $n = 100$)

Model M_1 , $\boldsymbol{\pi} = 0.5$										
	$\hat{\beta}_{11}$		$\hat{\beta}_{14}$		$\hat{\beta}_{21}$		$\hat{\beta}_{22}$		$\hat{\beta}_{25}$	
	SD	SD_m	SD	SD_m	SD	SD_m	SD	SD_m	SD	SD_m
BIC	.210	.157	.245	.160	.189	.176	.220	.183	.214	.167
MS	.187	.166	.166	.173	.172	.185	.218	.195	.192	.181
MH	.172	.172	.171	.175	.169	.193	.214	.198	.185	.183
ML	.179	.147	.194	.147	.197	.148	.229	.195	.197	.146
OR	.178	.171	.154	.173	.177	.194	.197	.198	.181	.181
Model M_1 , $\boldsymbol{\pi} = 0.3$										
BIC	.282	.220	.339	.232	.159	.142	.175	.141	.170	.130
MS	.231	.241	.260	.257	.156	.152	.147	.148	.139	.138
MH	.224	.249	.266	.268	.153	.154	.150	.150	.144	.139
ML	.254	.193	.293	.198	.169	.123	.166	.119	.144	.119
OR	.233	.249	.235	.259	.154	.155	.145	.150	.139	.139
Model M_2 , $\boldsymbol{\pi} = 0.5$										
	$\hat{\beta}_{11}$		$\hat{\beta}_{14}$		$\hat{\beta}_{21}$		$\hat{\beta}_{24}$		$\hat{\beta}_{25}$	
	SD	SD_m	SD	SD_m	SD	SD_m	SD	SD_m	SD	SD_m
BIC	.212	.176	.250	.171	.195	.166	.253	.190	.238	.200
MS	.214	.193	.209	.188	.188	.178	.247	.207	.205	.215
MH	.226	.195	.213	.191	.188	.181	.260	.211	.195	.211
ML	.214	.169	.211	.173	.211	.149	.235	.150	.234	.155
OR	.194	.199	.183	.190	.184	.179	.216	.215	.220	.221
Model M_2 , $\boldsymbol{\pi} = 0.3$										
BIC	.332	.228	.384	.235	.176	.136	.189	.154	.189	.161
MS	.296	.265	.341	.271	.157	.145	.231	.160	.198	.165
MH	.323	.271	.356	.287	.158	.148	.230	.167	.174	.171
ML	.308	.222	.367	.233	.179	.124	.180	.125	.190	.130
OR	.290	.293	.289	.285	.152	.142	.167	.166	.178	.171

Table 4: The predicted log-likelihoods for Normal FMR models ($n = 100$).

Method	$\pi = 0.5$			$\pi = 0.3$			$\pi = 0.1$		
	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$
	Model M_1								
BIC	-19020	-18650	-18400	-18500	-18110	-17850	-17680	-16890	-16510
MS	-18740	-18460	-18290	-18210	-17900	-17720	-17330	-16660	-16380
MH	-18730	-18460	-18280	-18270	-17920	-17730	-17650	-16770	-16430
ML	-19140	-18780	-18530	-18750	-18340	-18070	-17860	-17080	-16690
OR	-18590	-18390	-18270	-18020	-17830	-17690	-16550	-16330	-16180
	Model M_2								
BIC	-18600	-18160	-17830	-18440	-17850	-17530	-21000	-16970	-16290
MS	-18280	-17910	-17680	-18180	-17720	-17380	-17030	-16480	-16120
MH	-18300	-17960	-17720	-18200	-17790	-17410	-17060	-16440	-16120
ML	-18380	-17990	-17730	-18210	-17750	-17440	-16900	-16510	-16190
OR	-17970	-17740	-17560	-17600	-17360	-17180	-16420	-16080	-15850

Table 5: Average numbers of correct and incorrect estimated zero coefficients Normal FMR model with large number of covariates, $n = 300$

Method	$\pi = 0.5$				$\pi = 0.3$			
	Cor.	Inc.	Cor.	Inc.	Cor.	Inc.	Cor.	Inc.
	Com.1		Com.2		Com.1		Com.2	
MS	24.95	.068	10.00	.134	24.66	.277	9.99	.447
MH	24.62	.000	9.72	.000	24.27	.000	9.89	.000
ML	24.40	.600	9.58	1.16	23.43	1.02	9.81	1.52

Table 6: Selected standard deviations of $\hat{\beta}$ and their estimates
Normal FMR model with large number of covariates, $n = 300$

	$\hat{\beta}_{11}$		$\hat{\beta}_{15}$		$\hat{\beta}_{26}$		$\hat{\beta}_{2,10}$		$\hat{\beta}_{2,40}$	
	SD	SD_m	SD	SD_m	SD	SD_m	SD	SD_m	SD	SD_m
$\boldsymbol{\pi} = 0.5$										
MS	.121	.115	.103	.099	.106	.123	.125	.133	.109	.123
MH	.119	.115	.104	.098	.109	.123	.127	.132	.108	.121
ML	.137	.079	.122	.082	.194	.073	.193	.076	.171	.076
OR	.121	.115	.102	.099	.106	.123	.124	.133	.109	.122
$\boldsymbol{\pi} = 0.3$										
MS	.157	.170	.133	.148	.091	.091	.088	.099	.093	.090
MH	.163	.173	.140	.151	.091	.090	.089	.098	.093	.089
ML	.199	.105	.164	.108	.118	.064	.114	.068	.112	.068
OR	.158	.171	.132	.148	.091	.091	.089	.099	.093	.091

Table 7: The predicted log-likelihoods for Normal FMR model with large number of covariates, $n = 300$.

Method	$\pi = 0.5$			$\pi = .3$		
	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$
	Model M_1					
MS	-21940	-21670	-21460	-21050	-20800	-20590
MH	-22110	-21810	-21550	-21250	-20940	-20680
ML	-24050	-23380	-22940	-23620	-22990	-22530
OR	-21890	-21640	-21450	-21020	-20780	-20580

Table 8: Parameter estimates in the real data example

Factors	Levels	SRH Estimates (SE)	MS Estimates	BIC Estimates
Brand	Philips	-0.37^* (0.17)	0	0
	Braun	-0.40^* (0.16)	0	0
	Moulinex	0	0	0
Capacity	6	-2.48^* (0.21)	-2.59 (0.08)	-2.59 (0.09)
	10	0.06 (0.14)	0	0
	15	0	0	0
Price	39	1.97^* (0.34)	1.91 (0.22)	1.91 (0.23)
	69	1.48^* (0.17)	1.43 (0.12)	1.43 (0.13)
	99	0	0	0
Thermos	yes	1.14^* (0.18)	1.08 (0.14)	1.08 (0.14)
	no	0	0	0
Filter	yes	0.92^* (0.12)	1.02 (0.11)	1.02 (0.11)
	no	0	0	0
Brand	Philips	0.12 (0.21)	0	0
	Braun	-1.43^* (0.31)	-1.61 (0.11)	-1.52 (0.11)
	Moulinex	0	0	0
Capacity	6	-0.25 (0.26)	0	0
	10	0.07 (0.25)	0	0
	15	0	0	0
Price	39	-0.49 (0.32)	0	0
	69	-0.04 (0.22)	0	0
	99	0	0	
Thermos	yes	0.35 (0.20)	0	0
	no	0	0	0
Filter	yes	1.00^* (0.20)	0.54 (0.10)	0.76 (0.10)
	no	0	0	0