JIRSS (2011)
Vol. 10, No. 2, pp 201-235

An Overview of the New Feature Selection Methods in Finite Mixture of Regression Models

Abbas Khalili

Department of Mathematics and Statistics, McGill University, Canada.

Abstract. Variable (feature) selection has attracted much attention in contemporary statistical learning and recent scientific research. This is mainly due to the rapid advancement in modern technology that allows scientists to collect data of unprecedented size and complexity. One type of statistical problem in such applications is concerned with modeling an output variable as a function of a small subset of a large number of features. In certain applications, the data samples may even be coming from multiple subpopulations. In these cases, selecting the correct predictive features (variables) for each subpopulation is crucial. The classical best subset selection methods are computationally too expensive for many modern statistical applications. New variable selection methods have been successfully developed over the last decade to deal with large numbers of variables. They have been designed for simultaneously selecting important variables and estimating their effects in a statistical model. In this article, we present an overview of the recent developments in theory, methods, and implementations for the variable selection problem in finite mixture of regression models.

Keywords. EM algorithm, mixture models, mixture of regression models, penalized likelihood, regularization, variable selection.

MSC: Primary: 62F12, 62J02, 62J07; Secondary: 62P10, 62P20.

Abbas Khalili (✉)(khalili@math.mcgill.ca)

Received: April, 2011; October, 2011

1 Introduction

Feature selection has become a ubiquitous statistical activity in regression modeling in recent years. Rapid advancement in modern technology has led to many types of high-throughput data. In genome-wide association studies, geneticists nowadays routinely genotype half of a million single nucleotide polymorphisms (SNPs) over the whole genome in hope of identifying a handful of SNPs that play a major role in the genetic variation of a quantitative trait or a disease status (Chanock and Hunter, 2008). In functional genomics, one of the aims is to find a subset of the candidate motifs, out of hundreds or thousands, that highly contributes to the gene expression variations (Conlon et al., 2003 and Zhong et al., 2005). Other examples of high-throughput data include high-resolution images, high-frequency financial data, functional and longitudinal data, among others. In such applications, variable (feature) selection is the key statistical issue. Since the number of candidate variables, say p, is large classical variable selection methods such as the Akaike information criterion (AIC; Akaike 1973), the Bayesian information criterion (BIC; Schwarz 1978) or the Mallows' C_p (Mallows 1973) become computationally intractable and not possible to use in practice. As a result, new methods need to be developed.

The problem is one of the most actively researched topics in recent statistical literature. There have been many recent advances on the variable selection problem for linear and generalized linear regression models. The Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996), and the Smoothly Clipped Absolute Deviation (SCAD) method by Fan and Li (2001, 2002, 2004) are the new regularization or penalty methods proposed for variable selection with many interesting properties. For example, LASSO has the soft-thresholding property and SCAD has an oracle property as discussed in Fan and Li (2001). Unlike classical variable selection methods, LASSO and SCAD can be used in reasonably high dimensional problems. Efron et al. (2004) developed a revolutionary algorithm, called Least Angle Regression (LARS), that allows fast execution of LASSO, and Zou and Li (2008) proposed a fast one-step algorithm for SCAD. What distinguishes the new methods from the classical variable selection methods is the nature of their penalty functions. New methods incorporate penalty functions which are continuous functions of the regression coefficients, while the penalty functions in the classical methods are functions of the number of variables included in a submodel. In what follows, we give a summary of the new variable selection methods in multiple linear regression models.

Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ be a sample of observations governed by the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i,$$

for some additive error ε_i with $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. Suppose that $\beta_j = 0$, for some j; the goal is to identify these coefficients. In LASSO, the regression coefficients are estimated through the minimization problem

$$\min_{\beta_0, \beta_1, \ldots, \beta_p} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \boldsymbol{\beta}^\top x_i)^2 + n \sum_{j=1}^{p} \lambda_n |\beta_j| \right\}, \tag{1}$$

where λ_n is a tuning parameter that controls the amount of shrinkage of each regression coefficient β_j. In (1), the first term is the regular residual sum of squares (RSS), a measure of goodness-of-fit, and the second term is proportional to the L_1-norm of the vector $\boldsymbol{\beta}$, and is to control the complexity of the model. Since the L_1-norm in (1) spikes at $\beta_j = 0$, the solution $\hat{\boldsymbol{\beta}}(\lambda_n)$ has some of its elements equal to zero when the shrinkage parameter λ_n is large enough. Thus, the goal of variable selection is achieved without fitting all possible submodels, which in turn reduces the computational burden of the problem significantly.

Figure 1 shows the contours of the RSS and the L_1-norm for the case $p = 2$. The solution to (1) is the first point where the elliptical contours of RSS hit those of the L_1-norm. Since the contour of L_1 has corners, if the solution occurs at a corner, then it has one parameter estimator $\hat{\beta}_j(\lambda_n)$ equal to zero for some j. The choice of λ_n is thus important. The LARS algorithm of Efron et al. (2004) provides the entire solution path $\{\hat{\boldsymbol{\beta}}(\lambda_n); \lambda_n > 0\}$ to the LASSO minimization problem (1). The LARS algorithm is implemented as an R package named *lars* which is very easy to use. See also Rosset and Zhu (2007) on the piecewise linear solutions path of the new regularization methods.

It is worth noting that the L_1-norm penalty in (1) belongs to the family of so-called bridge functions

$$\sum_{j=1}^{p} \lambda_n |\beta_j|^\kappa, \qquad \kappa > 0. \tag{2}$$

The solution to the minimization problem (1) has the variable selection property using the penalty (2) for any $0 < \kappa \leq 1$. However,

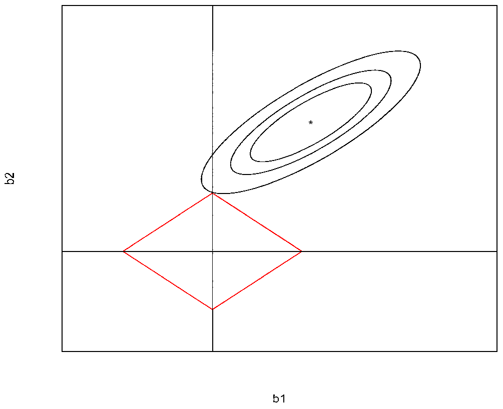

Figure 1: The black ellipses are the contours of the RSS and the red diamond is one level of the contours of $|\beta_1| + |\beta_2|$.

when $0 < \kappa < 1$, the function becomes nonconvex and this makes the optimization problem more difficult. The case $\kappa = 1$ (LASSO) is thus desirable. Note that if $\kappa = 0$, then the penalty in (2) reduces to the number of non-zero elements of the vector $\boldsymbol{\beta}$, and thus by the choice $\lambda_n = 2$ or $\lambda_n = \log n$ we will have the AIC or BIC penalty. The case $\kappa = 2$ corresponds to the well-known ridge penalty proposed by Hoerl and Kennard (1970), and it is a well-known fact that it does not have the variable selection property. Figure 2 shows one level of the contours of the bridge function for $\kappa = 0.5, 1.0, 2.0$ for the case $p = 2$. For a more in-depth discussion of the bridge function see Hastie, Tibshirani and Friedman (2009).

Fan and Li (2001) proposed the SCAD penalty which leads to estimators with desirable statistical properties. Consider the general penalized residual sum of squares

$$\sum_{i=1}^{n}(y_i - \beta_0 - \boldsymbol{\beta}^\top x_i)^2 + \sum_{j=1}^{p} p_n(\beta_j; \lambda_n). \tag{3}$$

As discussed in Antoniadis and Fan (2001), a *good* penalty function in (3) should result in estimators $\hat{\boldsymbol{\beta}}(\lambda_n)$ with three properties:

∗ *Unbiasedness*: the estimator is (approximately) unbiased when the true unknown parameter is large. This property avoids modeling bias when it is unnecessary.

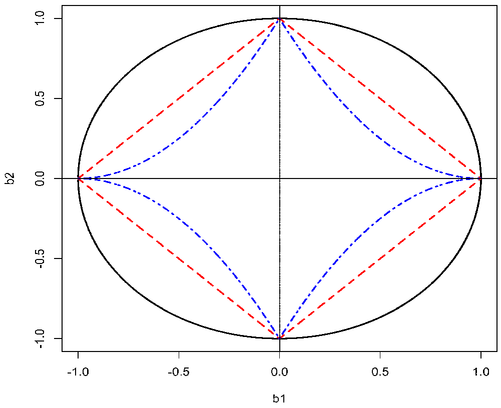

Figure 2: Solid curve: ridge penalty; Dashed-line: L_1-norm; Dashed-dotted line: bridge with $a = 0.5$.

- *Sparsity*: the estimator will automatically set small estimates $\hat{\beta}_j(\lambda_n)$ to zero in order to give a parsimonious model.
- *Continuity*: the estimator is continuous in the data to avoid instability (Breiman, 1996) in the model selection.

Designing a penalty function which results in estimators with the aforementioned properties is a challenging task. The bridge penalties (2) result in estimators that have some of the properties 1-3. The ridge estimators are continuous in data but do not have properties 1-2. For $0 < \kappa \leq 1$, the estimators are continuous and also have the sparsity property, but they introduce unnecessary bias in the estimators.

The SCAD penalty of Fan and Li (2001) results in estimators with all three properties. The penalty function is

$$p_n(\beta_j; \lambda_n)/n = \begin{cases} \lambda_n|\beta_j| & , |\beta_j| \leq \lambda_n \\ -(\beta_j^2 - 2a\lambda_n|\beta_j| + \lambda_n^2)/[2(a-1)] & , \lambda_n < |\beta_j| \leq a\lambda_n \\ \lambda_n^2(a+1)/2 & , |\beta_j| > a\lambda_n \end{cases}$$

for some constant $a > 2$. Through a Bayesian risk analysis, Fan and Li (2001) showed that the value $a = 3.7$ minimizes a Bayes risk criterion for β_j, and they argued that this choice of a gives good practical performance in various variable selection problems. We have used the SCAD penalty with $a = 3.7$, and it worked well in our simulation study.

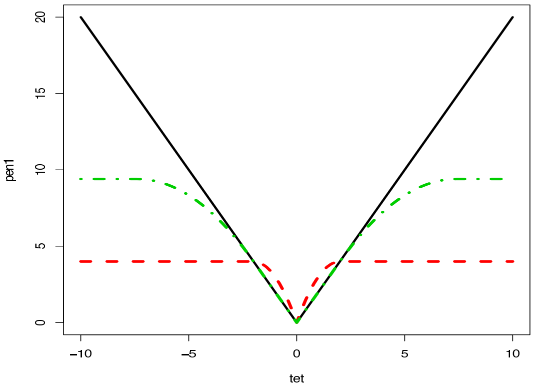

Figure 3: L_1-norm (solid line); SCAD (dash-dotted line); HARD (dashed line).

Figure 3 shows the plot $p_n(\beta_j;\lambda_n)$ versus β_j for three penalty functions LASSO, SCAD, and HARD (Antoniadis, 1997) which is given by

$$p_n(\beta_j;\lambda_n)/n = \lambda_n^2 - (|\beta_j| - \lambda_n)^2 I_{\{|\beta_j|<\lambda_n\}}.$$

LASSO is convex and thus beneficial for numerical computation. It tends to reduce all effects by similar amounts until the estimated effect is set to zero. When the penalty increases, SCAD reduces smaller effects faster than larger effects, which is also the case in HARD.

Fan and Li (2001) showed that, under standard regularity conditions, the SCAD estimator $\hat{\boldsymbol{\beta}}(\lambda_n)$ has what is called the *oracle property*, which means the estimator performs similarly to the estimator when the true submodel is known in advance. Their results have been extended to a broad class of models, including generalized linear models, Cox's proportional hazard models, frailty models, and semi-parametric modeling in longitudinal data analysis.

When the number of observations n is less than the potential number of variables, p, consistent estimation of all regression coefficients is impossible unless the model has what is called the sparsity property. That is, either the number of non-zero regression coefficients, or the sum of the absolute value of the regression coefficients remains finite as the sample size increases. There has been a lot of research done on sparse linear and generalized linear regression models. Meinshausen and Bühlmann (2006), Huang, Ma and Zhang (2009), Meinshausen and Yu (2009), Zhang and Huang (2008), and others studied properties of LASSO and adaptive LASSO in high-dimensional sparse linear regression models. Kim, Choi and Oh (2008), and Xie and Huang (2009) studied scad in high-dimensional linear and partial linear models. Van de Geer (2008) investigated lasso in high dimensional generalized linear models. Fan and Lv (2010) provides a comprehensive review of the recent developments in theory, methods, and implementations for the variable selection problem in linear and generalized linear regression models in high-dimensional feature spaces.

A complicating factor in some of the aforementioned applications is the underlying heterogeneity of the population from which the data are obtained. Such problems may be approached by finite mixture models. In general, finite mixture models are used to model data that arise from a heterogeneous population. When a response variable Y with a finite mixture distribution depends on certain covariates (features), a finite mixture of regression (fmr) model is obtained. In particular, an fmr model segments the population into subpopulations and models each subpopulation by a distinct linear or generalized linear regression model. For instance, in market segmentation studies (Wedel and Kamakura, 2000), fmr models allow researchers to investigate the phenomenon that different features of a certain product may appeal to different consumers. Identifying these subsets together with the number of segmentations (or submarkets) provides information on which products are likely to be successful before being introduced to the market. In medicine, different groups of patients with Parkinson's disease may show different trajectories of their disease progress depending on their economical, social and family situations. In motif gene expression research, the set of regulating motifs varies from one subpopulation to another. Such genetic phenomena can be well captured by an fmr model. See McLachlan and Peel (2000), Skrondal and Rabe-Hesketh (2004), and Frühwirth-Schnatter (2006) for more examples.

Variable selection in fmr models is a challenging task since the contributions of the covariates toward the response variable may vary from one component (or subpopultation) to another of these models. To enhance the predictability and to provide a more parsimonious model, it is then logical to only include the most significant variables in the model in a component-wise manner. This variable selection problem has received much attention recently. All-subset selection methods such as aic, bic and their modifications have been studied in the context of fmr models. However, even for fmr models with moderate numbers of components and covariates, all-subset selection methods are computationally intensive. Despite the many recent advances on the variable selection problem for linear and generalized linear models, the research on this problem in FMR models is still in its early stages of development. In this paper, we provide an overview of the recent advances on the variable selection problem in FMR models.

The rest of the paper is organized as follows. The definition of FMR models and their identifiability are provided in Section 2. In Section 3, the variable selection problem in FMR models is outlined, followed by a short review of the classical all-subset selection methods and their associated drawbacks. New variable selection methods and their numerical implementations are discussed in Sections 4 and 5. Statistical properties of these methods are provided in Section 6. In Section 7, we discuss model-based prediction in FMR models. Some simulation results on the performance of the new variable selection methods in FMR models are provided in Section 8. Variable selection in FMR models in high-dimensional feature spaces is then discussed in Section 9. Section 10 concerns a generalization of FMR models, called mixture-of-experts (MOE). Finally, in Section 11, we conclude with an emphasis on the issues in variable selection problems of FMR and MOE models that remain to be addressed in future research.

2 Finite Mixture of Regression Models

2.1 Definition

Consider a response variable Y with possible values in $\mathcal{Y} \subset \mathbb{R}$, and a p-dimensional vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p) \in \mathcal{X} \subset \mathbb{R}^p$ of covariates (features) that may affect Y. Let $\mathcal{F} = \{h(y; \theta, \phi); (\theta, \phi) \in \Theta \times (0, \infty), \Theta \subset \mathbb{R}\}$ be a parametric family of density functions with respect to a σ-finite measure. In the ordinary regression context, a universal linear or nonlinear regression model $h(y; \theta(\boldsymbol{x}), \phi)$, with a known real-valued link function $\theta(\boldsymbol{x}) = \mathcal{L}(\beta_0 + \boldsymbol{x}^\top \boldsymbol{\beta})$ and a dispersion parameter ϕ, is used to describe the relationship between Y and $\boldsymbol{x}$ across all members of a population. However, in some applications the relationship between Y and $\boldsymbol{x}$ may differ across different parts or members of a population. Finite mixture of regression (FMR) models provide a natural way of modelling such unobserved heterogeneous relationships. More formally, in a population made up of K subpopulations, the conditional density (or probability) function of Y given $\boldsymbol{x}$ is postulated to be

$$f(y; \boldsymbol{x}, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k \; h(y; \theta_k(\boldsymbol{x}), \phi_k), \tag{4}$$

where $\theta_k(\boldsymbol{x}) = \mathcal{L}(\beta_{0k} + \boldsymbol{x}^\top \boldsymbol{\beta}_k)$ and $\pi_k, \beta_{0k}, \boldsymbol{\beta}_k, \phi_k, k = 1, 2, \ldots, K$, such that $\sum_{k=1}^{K} \pi_k = 1$, are unknown parameters. One may use a common dispersion parameter $\phi_1 = \phi_2 = \ldots = \phi_K = \phi$ across all the K components of the model. The master vector of all parameters is given by

$$\boldsymbol{\Psi} = (\beta_{01}, \boldsymbol{\beta}_1, \beta_{02}, \boldsymbol{\beta}_2, \ldots, \beta_{0K}, \boldsymbol{\beta}_K, \phi_1, \phi_2, \ldots, \phi_K, \pi_1, \pi_2, \ldots, \pi_K)^\top$$

The $h(y; \theta_k(\boldsymbol{x}), \phi_k)$ are referred to as component density functions that belong to a parametric family. The most popular FMR models are based on well-known families such as normal, Poisson, and Binomial distributions. The $0 < \pi_k < 1$ are called the mixing probabilities and can be viewed as the proportion or contribution of the k-th subpopulation in a population that is made up of K subpopulations.

In some applications of FMR models, the number of components K (or the order of the model) is known a priori, while in others it needs to be estimated based on the data.

Example. For illustrative purposes, consider the FMR model

$$f(y; x) = 0.5\ \phi(y; x - 1, 1) + 0.5\ \phi(y; 2 + 2x, 1), \tag{5}$$

where $\phi(y; \mu, \sigma^2)$ stands for the normal density function with mean μ and variance σ^2; that is, a mixture of two simple normal linear regression models. Figure 4 shows a scatter plot of 600 random pairs (x_i, Y_i), where given x_i, each Y_i is randomly generated from model (5). The x_i's were randomly generated from the standard normal distribution $N(0, 1)$. From the scatter plot, there seems to be two groups of points. Ignoring this fact, and fitting a simple linear regression model to the (x_i, Y_i) leads to the poor least squares fit that is shown by the solid line in Figure 4. The dashed-line and the dashed-dotted line are the least squares fits to each of the two groups of the points. It can be seen that fitting an FMR model provides a much better fit than the ordinary simple linear regression fit.

2.2 Identifiability of an FMR Model

In using FMR models in data analysis, one needs be careful about a potential inferential problem about identifiability or uniqueness of the parameters of the model. In some families of mixture models it is possible to find two sets of parameter values that give the same density

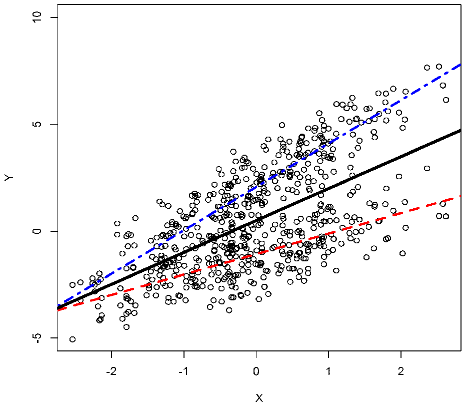

Figure 4: Scatter plot of a random sample from model (5). Solid-line: simple linear least squares fit. Dashed- and dashed-dotted lines: groupwise least squares fits.

function in (4), which is then called a non-identifiable model; McLachlan and Peel (2000). Identifiability is necessary for a valid statistical inference. Formally, an FMR model is called identifiable if, for a given design matrix $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$,

$$f(y; \boldsymbol{x}_i, \boldsymbol{\Psi}_1) = f(y; \boldsymbol{x}_i, \boldsymbol{\Psi}_2) \ , \ i = 1, 2, \ldots, n$$

for all values of $y \in \mathcal{Y}$, implies that $K_1 = K_2$ and $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2$, up to a permutation.

In general, identifiability of an FMR model depends on the family $f(y; \theta, \phi)$, the order K, and the design matrix $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. It is important to note that identifiability of a classical finite mixture model does not necessarily indicate identifiability of the corresponding FMR model, as falsely claimed in DeSarbo and Cron (1988). For example, the mixture of normal distributions $\sum_{k=1}^{K} \pi_k \ \phi(y; \mu_k, \sigma_k^2)$ is identifiable, but this is not necessarily true when the mean parameters μ_k's are functions of $\boldsymbol{x}$. This was first noticed by Hennig (2000) who studied identifiability of mixture of normal linear regression models and pointed out that, unlike ordinary regression models, a full rank design matrix does not guarantee identifiability of an FMR model. He showed 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-dimension. Loosely speaking, an FMR model may not be identifiable if covariates show little variability; for example, if the covariates are dummy variables or categorical with few levels. In the rest of this paper, we assume that the FMR models under study are identifiable.

3 The Variable Selection Problem in FMR Models

Let $\mathcal{S} = \{1, 2, \ldots, p\}$ be the index set representing all the p covariates, and let s be any subset of $\mathcal{S}$. We denote $\boldsymbol{x}[s]$ and $\boldsymbol{\beta}[s]$ as the subvectors of $\boldsymbol{x}$ and $\boldsymbol{\beta}$, respectively, where $\beta_j = 0$ if $j \notin s$.

For any K subsets $s_1, s_2, \ldots, s_K$, an FMR submodel is given by

$$f(y; \boldsymbol{x}, \boldsymbol{\Psi}, s_1, s_2, \ldots, s_K) = \sum_{j=1}^{K} \pi_k \; h(y; \theta_k(\boldsymbol{x}[s_k]), \phi_k) \tag{6}$$

with $\theta_k(\boldsymbol{x}[s_k]) = \mathcal{L}(\beta_{0k} + \boldsymbol{x}[s_k]^\top \boldsymbol{\beta}_k[s_k])$ and $\boldsymbol{\beta}_k[s_k]$ is a subvector of $\boldsymbol{\beta}_k$.

Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ be a sample of observations governed by the FMR model (4) or its submodel (6). The (conditional) log-likelihood function of the parameter $\boldsymbol{\Psi}$ is given by

$$l_n(\boldsymbol{\Psi}; s_1, s_2, \ldots, s_K) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{x}_i, \boldsymbol{\Psi}, s_1, s_2, \ldots, s_K) \tag{7}$$

The variable selection problem aims at selecting the subsets $s_1, s_2, \ldots,$ s_K such that the resulting FMR submodel best balances the model complexity and the goodness-of-fit to the data. In general, increasing the size of each s_k (i.e. adding more and more covariates to the model) will lead to an increase in the value of the log-likelihood function $l_n(\boldsymbol{\Psi}; s_1, s_2, \ldots,$ $s_K)$. As in any model selection problem, the log-likelihood function is thus not a good criterion for model selection purposes since the more complex the model, the larger the value of $l_n(\boldsymbol{\Psi}; s_1, s_2, \ldots, s_K)$. This is obviously a complex combinatorial optimization problem which becomes computationally more extensive as K and p increase. Penalization or regularization (Bickel and Li, 2006) methods are used for such model selection problems. In these methods, as the name indicates, the log-likelihood function, or any other measure of the goodness-of-fit, is combined with a penalty function that controls the complexity of the model, and is used to select a parsimonious model that also provides a good fit to the data. What follows is a review of these methods.

3.1 Classical Variable Selection Methods in FMR Models

The information theoretic approaches such AIC and BIC and their modifications have been used for model selection purposes in FMR models. This has been particularly the case in marketing (Wedel and Kamakura, 2000) and biostatistical applications (Wang, Puterman, Cockburn and Le, 1996) of FMR models. See also Skrondal and Rabe-Hesketh (2004) and Frühwirth-Schnatter (2006) for more applications.

Given a random sample of observations $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)$ from the FMR model (4), an information-based criterion is

$$\begin{aligned}&\text{GIC}(\hat{\boldsymbol{\Psi}}_{\text{MLE}}; s_1, s_2, \ldots, s_K)\\ &\qquad = -2l_n(\hat{\boldsymbol{\Psi}}_{\text{MLE}}; s_1, s_2, \ldots, s_K) + c_n \ \text{DF}(s_1, s_2, \ldots, s_K),\end{aligned}$$

where given the index sets $s_1, s_2, \ldots, s_K$, $\hat{\boldsymbol{\Psi}}_{\text{MLE}}$ is the maximum likelihood estimate of the parameters of the corresponding FMR submodel, c_n is called a penalty parameter, and $\text{DF}(s_1, s_2, \ldots, s_K)$ represents the complexity (degrees of freedom) of the FMR submodel which is equal to the total number of parameters included in the submodel. The $\text{GIC}(s_1, s_2, \ldots, s_K)$ selects the best submodel out of a pool of possible FMR submodels. The most popular choices of the penalty parameter c_n are $c_n = 2$ in AIC and $c_n = \log n$ in BIC. However, despite the popularity of these criteria, there does not exist a rigorous mathematical proof about potential optimal statistical properties of these criteria in an FMR model context. Most of the existing literature on properties of GIC in the context of the FMR model is mainly based on simulation studies. Cross-validation, d-fold cross-validation and generalized cross validation are also used for model selection problems.

For small or moderate values of p and also K, GIC is computationally manageable. Based on our experience in such cases, BIC often selects the best submodel. However, in an application with, say $p = 20$ potential covariates, for an FMR model with $K = 3$, there are about $2^{20} \times 2^{20} \times 2^{20}$ possible submodels that need to be examined by the GIC in order to select the best submodel. This is a computationally complex combinatorial optimization problem. Furthermore, for higher order problems it may not even be possible to perform the computations. There are two main challenges. First, as previously mentioned, it is computationally impractical to examine all possible FMR submodels. Second, as pointed out by Chen and Chen (2008), and Khalili, Chen and Lin (2011), even if the computation is feasible, the classical all-subset selection methods are too liberal. They select a model with more spurious features than warranted. This motivates the use of the new regularization techniques such as LASSO and SCAD for such complex problems.

4 New Variable Selection Approaches in FMR Models

Consider the FMR model (4). Let $(\boldsymbol{x}_1, Y_1), (\boldsymbol{x}_2, Y_2), \ldots, (\boldsymbol{x}_n, Y_n)$ be a random sample of observations from the model with given order K. The (conditional) log-likelihood function of $\boldsymbol{\Psi}$ based on the full FMR model containing all the covariates is

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

Khalili and Chen (2007) proposed to estimate $\boldsymbol{\Psi}$ by maximizing the penalized log-likelihood function

$$pl_n(\boldsymbol{\Psi}) = l_n(\boldsymbol{\Psi}) - \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p} p_n(\beta_{kj}; \lambda_{nk}), \tag{8}$$

where $p_n(\beta_{kj}; \lambda_{nk})$ could be one of the penalty functions discussed in the introduction, with the component-wise tuning parameters λ_{nk}. Let $\hat{\boldsymbol{\Psi}}_n$ be the maximizer of $pl_n(\boldsymbol{\Psi})$. The hope is that by the proper choice of the penalty function $p_n(\beta_{kj}; \lambda_{nk})$, if some of the regression coefficients β_{kj} are zero, then their corresponding estimators $\hat{\beta}_{kj}$ are also zero. This is the sparsity property. On the other hand, we would like the estimators of the true non-zero regression coefficients to perform similarly to their regular maximum likelihood estimators when the true model is known in advance. Thus, unlike the all-subset selection methods, the new method combines the variable selection and estimation into one step and reduces the computational burden substantially. Statistical properties of the estimator $\hat{\boldsymbol{\Psi}}_n$ are further discussed later on in this paper.

It is worth noting that the amount of penalty on each regression coefficient β_{kj} in the k-th component of the mixture model is proportional to the mixing probability π_k, which is a common practice in relating the amount of penalty to the sample size. In the mixture model setting, the virtual sample size from the k-th component of the model is proportional to π_k. This enhances the power of the method, especially in finite sample situations.

5 Numerical Computations and Tuning Parameter Selection

5.1 Maximization Algorithm

It is challenging to maximize the penalized log-likelihood function $pl_n(\mathbf{\Psi})$ in (8). Clearly, the problem does not have a closed-form solution, and we must resort to numerical methods. The expectation-maximization (EM) algorithm of Dempster, Laird and Rubin (1977), combined with the Newton-Raphson algorithm, is very popular and convenient for parameter estimation in mixture models. However, in the current setting, the EM-Newton-Raphson cannot be used directly due to the non-differentiability of the penalty function $p_n(\beta_{kj}; \lambda_{nk})$ at $\beta_{kj} = 0$, a property that is required for the variable selection property of the method. There have been a number of suggestions to deal with this issue. Fan and Li (2001) suggested the following locally quadratic approximation (LQA) to the penalty function:

$$p_n(\beta; \lambda) \approx p_n^*(\beta; \lambda) = p_n(\beta_0; \lambda) + \frac{p_n'(\beta_0; \lambda)}{2\beta_0}(\beta^2 - \beta_0^2), \quad \text{for } \beta \approx \beta_0. \tag{9}$$

Hunter and Li (2005) studied convergence properties of the LQA approximation. Zou and Li (2008) used a locally linear approximation (LLA) of the penalty, which results in a penalty function similar to the adaptive LASSO. In the current review, we focus on the LQA approximation.

Applying the LQA to the penalty function in (8) results in a ridge-type penalty function, which in turn allows the use of the EM-Newton-Raphson algorithm. Note that the coefficient $p_n'(\beta_0; \lambda)/2\beta_0$ in (9) distinguishes the LQA from the regular ridge penalty. The parameter $\mathbf{\Psi}$ is then estimated by maximizing the (approximated) penalized log-likelihood function

$$pl_n^*(\mathbf{\Psi}) = l_n(\mathbf{\Psi}) - \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p} p_n^*(\beta_{kj}; \lambda_{nk})$$

The EM algorithm is used to approximate the maximizer of $pl_n^*(\mathbf{\Psi})$. The algorithm uses the complete data $(\boldsymbol{x}_i, \boldsymbol{z}_i, y_i), i = 1, 2, \ldots, n$, where $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iK})$ is the vector of missing binary labels z_{ik} which shows the component membership of the ith observation in the mixture model and $\sum_{k=1}^{K} z_{ik} = 1$ for each i. The complete log-likelihood function constructed based on the complete data is then

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

In the ordinary EM algorithm, the function $l_n^c(\mathbf{\Psi})$ is maximized, which also leads to the maximization of the main objective function $l_n(\mathbf{\Psi})$. This is the well-known property of the EM; Wu (1983). In the current setting, the EM works with the penalized complete log-likelihood function

$$pl_n^{c*}(\mathbf{\Psi}) = l_n^c(\mathbf{\Psi}) - \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p} p_n^*(\beta_{kj}; \lambda_{nk}).$$

Often, the penalized log-likelihood function also increases after each EM iteration (Green, 1990) and the algorithm converges as quickly as the algorithm applied to the unpenalized log-likelihood. The algorithm maximizes $pl_n^{c*}(\mathbf{\Psi})$ iteratively in two steps:

E-Step: Let $\mathbf{\Psi}^{(m)}$ be the current estimate of the parameters. In this step, the algorithm computes the conditional expectation of $pl_n^{c*}(\mathbf{\Psi})$ with respect to z_{ik}, given the observed data $(\boldsymbol{x}_i, y_i)$ and $\mathbf{\Psi}^{(m)}$. The expectation is

$$\begin{aligned} Q(\mathbf{\Psi}; \mathbf{\Psi}^{(m)}) &= \sum_{k=1}^{K} \sum_{i=1}^{n} \omega_{ik}^{(m)} \Big[\log \pi_k + \log h(y_i; \theta_k(\boldsymbol{x}_i), \phi_k) \Big] \\ &\quad - \sum_{k=1}^{K} \pi_k \sum_{j=1}^{p} p_n^*(\beta_{kj}; \lambda_{nk}) \end{aligned}$$

where

$$\omega_{ik}^{(m)} = E(z_{ik} | \text{data}, \mathbf{\Psi}^m) = \frac{\pi_k^{(m)} h(y_i; \theta_k^{(m)}(\boldsymbol{x}_i), \phi_k^{(m)})}{f(y_i; \boldsymbol{x}_i, \mathbf{\Psi}^{(m)})}. \qquad (10)$$

This step in fact comes down to the computation of the weights $\omega_{ik}^{(m)}$.

M-Step: Given the weights, the function $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(m)})$ is maximized with respect to the parameters $(\boldsymbol{\beta}_k, \phi_k, \pi_k)$ of the model. One may need to use, for example, the Newton-Raphson algorithm to perform the maximization. The maximization of $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(m)})$ with respect to the mixing probabilities π_k is particularly difficult so to avoid computational issues, we suggest the use of the updated estimates

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

which are the maximizers of the leading term in $Q(\mathbf{\Psi}; \mathbf{\Psi}^{(m)})$. Simulation studies in Khalili and Chen (2007) and Khalili, Chen and Lin (2011) have shown that this suggestion works well in applications.

The updated estimates $\beta_{kj}^{(m+1)}$ of the regression coefficients β_{kj} are obtained by solving the equations

$$\sum_{i=1}^{n} \omega_{ik}^{(m)} \, \frac{\partial \log h(y_i; \theta_k(\boldsymbol{x}_i), \phi_k)}{\partial \beta_{kj}} - \pi_k \times \frac{\partial p_n^*(\beta_{kj}; \lambda_{nk})}{\partial \beta_{kj}} = 0,$$

for $j = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, K$.

The dispersion parameters ϕ_k are updated by solving the equations

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

Starting from an initial value $\boldsymbol{\Psi}^{(0)}$, the algorithm iterates between the E- and M-steps until, for example, $\|\boldsymbol{\Psi}^{(m)} - \boldsymbol{\Psi}^{(m+1)}\| \leq \delta$, for a prespecified value δ. When the EM converges, a coefficient β_{kj} is declared zero if its corresponding estimate $|\hat{\beta}_{kj}|$ is smaller than a threshold value, taken as 10^{-5} in our simulations. To avoid numerical instability of the algorithm due to very small values of some of the $\hat{\beta}_{kj}$'s in the denominator of the approximation (9), as suggested by Hunter and Li (2005), we replace β_0 by $\beta_0 + \varepsilon$ for a given small value $\varepsilon > 0$.

It is well known that the success of an EM-type algorithm depends heavily on suitable starting values, especially when the likelihood surface is of multimodality, as is the case in our setting. To increase the chance of finding the global maximum or at least a good local one, it is often recommended that multiple (random) starting points be used (McLachlan and Peel, 2000). More computational details are provided in Khalili and Chen (2007), and Khalili, Chen and Lin (2011).

5.2 Tuning Parameter Selection

The choice of the tuning parameters λ_{nk} is important in the penalized likelihood approach. Large values of the λ_{nk} tend to select a simpler model whose parameter estimates have smaller variances, whereas small values of the tuning parameters lead to more complex models which means smaller modeling biases. The trade-off between the biases and variances yields an optimal choice of λ_{nk}.

In general, the most popular criteria used for tuning parameter selection in penalized likelihood approaches are d-fold cross validation (d-CV) and generalized cross validation (GCV). See Tibshirani (1996) and Fan and Li (2001). Khalili and Chen (2007) proposed deviance-based GCV criteria for component-wise selection of the tuning parameters in FMR models as follows.

Let $\tilde{\boldsymbol{\Psi}}_n$ be the ordinary MLE of the parameter $\boldsymbol{\Psi}$ which maximizes $l_n(\boldsymbol{\Psi})$ under the full model. Under the standard regularity conditions, we have $\|\tilde{\boldsymbol{\Psi}}_n - \boldsymbol{\Psi}^0\| = O_p(n^{-1/2})$. Let $\tilde{\omega}_{ik}$ be the estimated posterior probabilities in (10) evaluated at the MLE $\tilde{\boldsymbol{\Psi}}_n$, and $n_k = \sum_{i=1}^n \tilde{\omega}_{ik}$ be the expected number of observations generated from the kth component of the FMR model, which remain fixed throughout the tuning parameter selection process. For a given value of λ_{nk}, let $(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k)$ be the maximizer of the Q-function in the M-step of the EM algorithm. We definte the component-wise likelihood-based deviance statistics as

$$\mathcal{D}_k(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k) = \sum_{i=1}^{n} \tilde{\omega}_{ik}\{\log h(y_i; y_i, \hat{\phi}_k) - \log h(y_i; \hat{\eta}_k(\boldsymbol{x}_i), \hat{\phi}_k)\}.$$

and the component-wise GCV

$$\text{GCV}_k(\lambda_{nk}) = \frac{\mathcal{D}_k(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k)}{n_k(1 - \text{DF}_k/n_k)^2}\ ,\ k = 1, 2, \ldots, K,$$

where DF_k is the number of nonzero elements of the vector $\hat{\boldsymbol{\beta}}_k$. The tuning parameters $\lambda_{nk}, k = 1, 2, \ldots, K$, are chosen one at a time by minimizing $\text{GCV}_k(\lambda_{nk})$ over a plausible range of λ_{nk} values.

A recent study by Wang, Li and Tsai (2007) for multiple linear regression models has shown that the model corresponding to the tuning parameter selected by GCV for the SCAD penalty may contain some unimportant variables among the set of significant covariates. They suggested the use of BIC for tuning parameter selection, while they also showed that the model selected by using BIC achieves the model selection consistency. In Khalili and Lin (2011), we proposed a component-wise BIC for selection of the tuning parameters λ_{nk}. The component-wise BIC for the kth component of the FMR model is defined as

$$\text{BIC}_k(\lambda_{nk}) = 2\mathcal{D}_k(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k) + \log n_k \times \text{DF}_k. \qquad (11)$$

The tuning parameters $\lambda_{nk}, k = 1, 2, \ldots, K$, are chosen one at a time by minimizing $\text{BIC}_k(\lambda_{nk})$ over a plausible range of λ_{nk} values. If the factor $\log n_k$ in (11) is replaced by 2, then the AIC-type criterion

$$\text{AIC}_k(\lambda_{nk}) = 2\mathcal{D}_k(\hat{\boldsymbol{\beta}}_k, \hat{\phi}_k) + 2 \times \text{DF}_k$$

is obtained.

At the moment, theoretical properties of the λ_{nk} chosen by the above criteria is unknown, but the simulation studies show that the proposed criteria performs reasonably well in selecting appropriate values of the tuning parameters, though BIC$_k$ outperforms the other two criteria in selecting the correct model. A comprehensive study on statistical properties of AIC, GCV and BIC in a linear regression context is provided in Shao (1997). The recent work by Zhang, Li and Tsai (2010) introduces a generalized information criterion GIC for tuning parameter selection in new variable selection methods for generalized linear models.

6 Statistical Properties of the Penalized Likelihood Estimators in FMR Models

There is a considerable amount of research done over the last decade on the asymptotic properties of the penalized least squares estimators $\hat{\boldsymbol{\beta}}_n$ resulting from solving the minimization problem in (3) for different choices of the penalty $p_n(\beta_j; \lambda_n)$, such as LASSO and SCAD. See Fan and Lv (2010) and the references within. The two important properties of $\hat{\boldsymbol{\beta}}_n$ that are studied most often are: (1) $\sqrt{n}$-consistency and asymptotic normality of the estimators $\hat{\beta}_j$ of the true nonzero regression coefficients $\beta_j \neq 0$; (2) sparsity or consistency in the variable selection property of $\hat{\boldsymbol{\beta}}_n$, implying that coefficients whose true values are zero in the model (i.e. $\beta_j = 0$) have corresponding estimators that are also zero (i.e. $\hat{\beta}_j = 0$) with probability tending to one as $n \to \infty$. The estimator $\hat{\boldsymbol{\beta}}_n$ with properties (1)-(2) is referred to as the oracle estimator as discussed in Fan and Li (2001).

Khalili and Chen (2007) studied conditions under which the maximizer $\hat{\boldsymbol{\Psi}}_n$ of the penalized likelihood function $pl_n(\boldsymbol{\Psi})$ in (8) has the oracle property as discussed below.

Consider the partitioning $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, \boldsymbol{\beta}_{k2})$ of $\boldsymbol{\beta}_k$, where $\boldsymbol{\beta}_{k2} = \mathbf{0}$ is the vector of zero regression coefficients in the k-th component of the mixture model. The master vector of parameters $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_1, \boldsymbol{\Psi}_2)$ is also partitioned such that $\boldsymbol{\Psi}_2$ contains all the zero vectors $\boldsymbol{\beta}_{k2} = \mathbf{0}$ across all K components of the mixture model. Let $\boldsymbol{\Psi}_0$ be the true vector of parameters in the true FMR model underlying the data. Under standard regularity conditions as well as certain conditions on the penalty function, Khalili and Chen (2007) showed that as $n \to \infty$, the estimator $\hat{\boldsymbol{\Psi}}_n$ has the following two properties:

(a) *Sparsity*: $P(\hat{\boldsymbol{\beta}}_{k2} = \mathbf{0}) \to 1$, for $k = 1, 2, \ldots, K$.

(b) *Asymptotic normality*:

$$\sqrt{n}\left\{\left[\boldsymbol{I}(\boldsymbol{\Psi}_{01})-\frac{\boldsymbol{p}''(\boldsymbol{\Psi}_{01})}{n}\right](\hat{\boldsymbol{\Psi}}_{n1}-\boldsymbol{\Psi}_{01})+\frac{\boldsymbol{p}'(\boldsymbol{\Psi}_{01})}{n}\right\}\to^{d} N(\mathbf{0},\boldsymbol{I}(\boldsymbol{\Psi}_{01}))$$

where $\boldsymbol{p}'(\cdot)$ and $\boldsymbol{p}''(\cdot)$ are the first and second derivatives of the penalty function $\boldsymbol{p}_n(\boldsymbol{\Psi})=\sum_{k=1}^{K}\pi_k\sum_{j=1}^{p}p_n(\beta_{kj};\lambda_{nk})$ with respect to β_{kj}'s, and $\boldsymbol{I}(\boldsymbol{\Psi}_{01})$ is the Fisher information matrix under the true model with all zero effects removed.

By proper choices of the penalty function, the derivatives of the penalty in Part (b) will be negligible asymptotically, and we have that

$$\sqrt{n}(\hat{\boldsymbol{\Psi}}_{n1}-\boldsymbol{\Psi}_{01})\to^{d} N(\mathbf{0},\boldsymbol{I}^{-1}(\boldsymbol{\Psi}_{01})).$$

This is similar to the ordinary MLE of the parameter $\boldsymbol{\Psi}_{01}$ when we know in advance which β_{kj} is zero in the model.

It is important to mention that in the above discussion it is assumed that the true order K of the model is known. In applications, one may use the BIC or the scientific background to select K. Keribin (2000) showed that under certain regularity conditions, the number of components of a finite mixture model can be estimated consistently by using penalized-likelihood approaches such as BIC. However, blind use of a consistent estimator of K, regardless of the sample size n, should be discouraged. More general discussions and references related to the order estimation of a mixture model can be found in McLachlan and Peel (2000) and Chen and Khalili (2008).

Simulation studies in Khalili and Chen (2007) showed that when the number of potential covariates p is small and the mixture model is balanced, BIC is highly reliable for choosing the correct order K of the FMR model. The order estimation is performed by fitting full FMR models (which include all covariates) of different orders $K=1,2,\ldots,$ to the data and then selecting K based on BIC. However, when the number of covariates p is large, fitting full FMR models is not an easy task. Gupta and Ibrahim (2007) proposed a Bayesian approach for this problem. At the moment, there is no satisfactory non-Bayesian solution available to the problem of simultaneous order and variable selection in FMR models, especially when p is large. In these situations, one may fit FMR models with different orders $K=1,2,\ldots$, through the penalized likelihood approach outlined in Section 4, and using the extended Bayesian information criterion (EBIC) of Chen and Chen (2008) to select the final model. This is a problem requiring future investigation.

7 Model-Based Prediction

The ultimate goal of variable selection in most of the regression problems is to identify a submodel with a good predictive value. In linear regression models, after variable selection, for a given value $\boldsymbol{x}_0$ of the vector of selected covariates, the predicted response is

$$\hat{Y}_0 = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}};$$

i.e. the mean or expected value of the response variable, where $\hat{\boldsymbol{\beta}}$ is the vector of estimated coefficients of the selected covariates based on the current data. However, in FMR models the expected value of the response variable is not appropriate for prediction purposes since it ignores any information provided by the shape of the FMR distribution. For instance, if the density function of the FMR model is multimodal then clearly the expectation will not be a good representation of the distribution. A more reasonable prediction approach in FMR models is the so-called *predictive distribution*, which is also very common in a nonlinear time series context; see Wong and Li (2000). After selecting the final FMR model through the methodology outlined in Section 4, given a value $\boldsymbol{x}_0$ of the vector of selected covariates, the predictive density function of the future observation Y_0 is defined as

$$f(y; \boldsymbol{x}_0, \hat{\boldsymbol{\Psi}}_n) = \sum_{k=1}^{K} \hat{\pi}_k \; h(y; \hat{\boldsymbol{\theta}}_k(\boldsymbol{x}_0), \hat{\phi}_k),$$

where $\hat{\boldsymbol{\Psi}}_n$ is the maximum penalized likelihood estimator of the parameters based on the current data $(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, 2, \ldots, n$. In Khalili and Chen (2007) the predictive power of the selected FMR models by SCAD, LASSO and BIC are compared through the predictive distribution by extensive simulation studies. It turns out that the new variable selection methods, such as LASSO and SCAD, outperform the classical all-subset selection methods, such as BIC, when it comes to prediction.

8 Simulation Study

In this section we show the performance of the new penalized likelihood approach in FMR models with some simulations. The component-wise BIC_k in (11) is used for tuning parameter selection. We also tried the GCV_k and AIC_k for tuning parameter selection and noticed that the BIC_k

performs somewhat better and so we only report the simulation results based on the BIC$_k$.

The vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ is generated from a multivariate normal distribution with mean zero and the pairwise correlation between the covariates is $\text{corr}(x_i, x_j) = \rho^{|i-j|}$, for $i \neq j$ and $\rho = 0.5, 0.75$. For a given sample size n, the covariate vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are generated and form a design matrix which will remain unchanged in 1000 replications.

For a given design matrix, the response variable Y is generated from the two-component normal FMR model

$$\pi\ \phi(y; 1 + \boldsymbol{x}^\top \boldsymbol{\beta}_1, \sigma^2) + (1-\pi)\phi(y; 2 + \boldsymbol{x}^\top \boldsymbol{\beta}_2, \sigma^2)$$

with the parameter values $\pi = .15, .30, .50$, $\sigma = 1$, and

$$\boldsymbol{\beta}_1^\top = (1.5, 2.5, 0, 0, 1.7, 0, 0, 0),\ \ \boldsymbol{\beta}_2^\top = (-1.8, 0, 2.0, -1.5, 0, 0, 0, 0)^\top.$$

In this case, if one wants to use the all-subset selection methods, such as BIC, for variable selection, there are $2^{16} = 65536$ potential FMR submodels to be examined in order to select the best submodel, which obviously involves a lot of computations. Thus, the new methods such as LASSO and SCAD are particularly advantageous in such situations. We call the penalty function in (8) constructed from LASSO and SCAD the MIXLASSO and MIXSCAD penalties.

The simulation results are reported as the proportions of: correctly estimated zero coefficients (*specificity*; S_1), and correctly estimated nonzero coefficients (*sensitivity*; S_2). Ideally, the specificity and sensitivity values should be 1. The results in Table 1 are based on 1000 data sets with sample sizes $n = 100, 150$ generated from the normal FMR model.

From Table 1, we can see that overall the new method performs reasonably well. MIXLASSO selects less-sparse models compared to MIXSCAD. Variable selection becomes more difficult as the correlation between the x_j's increases. As the sample size increases, the method improves, as expected.

Extensive simulation studies can be found in the cited papers by Khalili et al.

9 Variable Selection in High-Dimensional Feature Spaces

Modern scientific research in biology, engineering, medicine, economics, finance and machine learning often require the analysis of high dimen-

Table 1: Specificity (S_1) and Sensitivity (S_2) summaries. ($\sigma = 1$).

	Setting			Mixture	$\pi = .15$		$\pi = .30$		$\pi = .5$	
Method	n	p	ρ	Components	S_1	S_2	S_1	S_2	S_1	S_2
MIXSCAD	100	8	.50	Com$_1$	.921	.922	.975	.985	.993	.998
			.75		.811	.860	.898	.954	.948	.972
			.50	Com$_2$	.989	.977	.999	.996	.993	.981
			.75		.951	.910	.981	.960	.967	.941
MIXLASSO			.50	Com$_1$	.808	.880	.920	.990	.967	1.00
			.75		.705	.835	.799	.984	.880	.999
			.50	Com$_2$	.986	.972	.997	.990	.984	.947
			.75		.969	.970	.981	.972	.954	.903
MIXSCAD	150	8	.50	Com$_1$	.960	.947	.991	.990	.999	.999
			.75		.890	.943	.954	.982	.977	.987
			.50	Com$_2$	.999	.997	1.00	.996	.999	.980
			.75		.994	.964	.996	.971	.985	.964
MIXLASSO			.50	Com$_1$	.823	.902	.924	.994	.979	1.00
			.75		.766	.970	.886	.999	.944	1.00
			.50	Com$_2$	.993	.978	.999	.945	.995	.835
			.75		.993	.996	.986	.994	.961	.958

sional data. For instance, geneticists nowadays routinely genotype half of a million single nucleotide polymorphisms (SNPs) over the whole genome. The goal is to establish which SNPs are influential in the genetic diversity of a quantitative trait or a disease status. From a statistical modeling point of view, simultaneous variable selection and parameter estimation play a central role in such investigations.

A major advantage of the new regularization techniques, such as LASSO and SCAD, is their applicability to high-dimensional problems when there is a large number of features $x_1, x_2, \ldots, x_p$ in the data. The number of features p is sometimes comparable or even larger than the sample size n; we refer to these problems as large-p-small-n problems, which have been extensively studied by many researchers in the recent literature. See for example Meinshausen and Bühlmann (2006), Kim, Choi and Oh (2008), and Wasserman and Roeder (2008), among many other papers cited in Fan and Lv (2010).

However, for the variable selection problem in ultra-high dimensional situations, when $p >> n$, even the new regularization methods are not computationally efficient. Fan and Lv (2008) suggested the use of a so called sure screening procedure to first reduce the number of potential features from a large or huge scale to a relatively large scale by a fast, reliable, and efficient method, so that well-developed variable selection techniques, such as LASSO and SCAD, can be applied to the reduced feature space. This provides a powerful tool for variable selection in ultra-high dimensional feature space. In multiple linear regression models, the screening is based on the magnitude of the marginal correlation of each covariate x_j with the response variable Y, and in generalized linear models the screening is based on the marginal likelihood function as discussed in Fan and Song (2010). See also Fan, Samworth and Wu (2009). Under certain regularity conditions, it is shown that the variable screening procedure has a sure screening property, meaning that it retains all the important variables with probability tending to one. The method is very effective in ultra-high dimensional problems.

9.1 High Dimensionality in FMR Models

Some of the applications of FMR models involve high dimensional data analysis. For example, in functional genomics, hundreds or even thousands of candidate motifs may be examined to find a small subset that highly contributes to gene expression variations. The set of regulating motifs may vary from one group of genes (or subpopulation) to another. Such genetic phenomena can be well captured by an FMR model. Variable selection is the key statistical issue in these kinds of applications. The problem compared to linear or generalized linear models becomes even more complex since out of possibly hundreds or thousands of potential covariates, different (small) subsets of the covariates may be significant between different regression components of an FMR model.

Almost all the existing literature on variable selection problems for high dimensional data is in the context of linear and generalized linear models (Van de Geer, 2008) and mixed effect models (Schelldorfer, Bühlmann and Van de Geer, 2011). The two recent papers by Städler, Bühlmann and Van de Geer (2010) and Khalili and Lin (2011) consider feature selection in FMR models where the dimension p is considered as a function of the sample size n, say p_n, and allowed to increase with n in a polynomial order.

The variable selection problem under FMR models poses serious statistical and computational challenges when p is comparable to the sample size n or when $n < p$ in the more extreme situation. The modified EM algorithm of Khalili and Chen (2007) in high dimensions is too computationally intensive and should be avoided. In fact, in an application when the number of potential features is comparable to or larger than the sample size, it is not even clear how to fit an FMR model by direct use of their method. The key issues are numerical instability and the uncertainty of being able to find the global maximizer (or even just a reasonable local one) as the penalized log-likelihood surface is likely to be extremely flat given the huge number of predictors. Even if hundreds or thousands of initial starting values are tried, the high dimensionality makes it difficult to find a proper maximizer. It is a well-known fact that even in a relatively low dimensional parameter space, maximization of a likelihood function of a mixture model is challenging (McLachlan and Peel, 2000).

To overcome computational difficulties and large false discovery rates caused by the large dimensionality, Khalili, Chen and Lin (2011) proposed a 2-stage procedure for variable selection in finite mixture of sparse normal linear (FMSL) models. First, to deal with the curse of dimensionality, a likelihood-based boosting (Bühlmann and Yu, 2003) is designed to effectively reduce the number of candidate features. This is the key thrust of the new method. Such a screening method, which selects variables without simultaneously fitting the final model, is known as a filtering method in machine learning applications. The greatly reduced set of features is then subjected to a sparsity-inducing procedure via the penalized likelihood approach outlined in Section 4. This second stage procedure is a so-called wrapper method in machine learning terminology. The screening algorithm is briefly outlined in the following section.

9.2 Feature Screening in FMSL Models in Large p-small-n Situations

The screening is based on the idea of boosting in ordinary regression, where it starts with a weak learner (or fit) and improves the fit in a sequential manner by considering the addition of one variable at a time, albeit with a mixture of regressions rather than a single regression. This "screening stage" will identify a set of variables that is potentially important in the final model. The focus of the following algorithm is the FMR model (4) with the normal component density functions $\phi(y; \beta_{0k} + \boldsymbol{x}^\top \boldsymbol{\beta}_k, \sigma_k^2)$, componentwise mean $\beta_{0k} + \boldsymbol{x}^\top \boldsymbol{\beta}_k$ and variance σ_k^2.

It is a well-known fact that the likelihood function of a mixture of normal distributions with unequal component variances σ_k^2 is unbounded. To avoid the unboundedness of the likelihood, instead of working with the log-likelihood function in (7), we work with the adjusted log-likelihood function

$$\tilde{l}_n(\mathbf{\Psi}; s_1, s_2, \ldots, s_K) = l_n(\mathbf{\Psi}; s_1, s_2, \ldots, s_K) - \sum_{k=1}^{K} a_n(\sigma_k), \qquad (12)$$

where $a_n(\cdot)$ is a non-negative penalty function. Chen et al. (2008) suggested

$$a_n(\sigma_k) = c\left\{\frac{S_n^2}{\sigma_k^2} - \log(\frac{S_n^2}{\sigma_k^2})\right\},$$

for some positive value $c > 0$, and S_n^2 is the sample variance of y_i's. It is seen that since $a_n(\sigma_k) \to \infty$ as $\sigma_k^2 \to 0$, the maximizer of $\tilde{l}_n(\mathbf{\Psi}; s_1, s_2, \ldots, s_K)$ does not have estimates of σ_k^2 close to zero.

A summary of the screening procedure developed in Khalili, Chen and Lin (2011) is as follows:

1. Initialization. Fit the submodel (6) with the active sets $s = s_1 = \cdots = s_K$ being empty. That is, we the find the adjusted maximum likelihood estimate (AMLE) of $\mathbf{\Psi}$ under the classical normal mixture model $\sum_{k=1}^{K} \pi_k\ \phi(y; \beta_{0k}, \sigma_k^2)$, by maximizing the adjusted log-likelihood $\tilde{l}_n(\cdot)$ in (12). We denote the empty active set by $s^{(0)}$. Let $\hat{\mathbf{\Psi}}^{(0)}$ be the vector of AMLEs of $\pi_k, \beta_{0k}, \sigma_k^2$, for $k = 1, \ldots, K$.

Let $\hat{\mu}_k^{(0)} = (\hat{\beta}_{0k}, \hat{\beta}_{0k}, \ldots, \hat{\beta}_{0k})^\top$, for $k = 1, \ldots, K$, be an $n \times 1$ vector. In what follows $\hat{\mu}_{k,i}^{(m)}$ refers to the ith element of the vector $\hat{\mu}_k^{(m)}$ in the mth iteration.

2. Boosting. Given the active set $s^{(m)}$ obtained from the last iteration or from the initialization, together with the corresponding $\hat{\mathbf{\Psi}}^{(m)}$ and $\hat{\mu}_k^{(m)}$, the fit is updated as follows. For each $j = 1, 2, \ldots, p$, we fit single-variable normal mixture regression models through the adjusted log-likelihood function

$$\tilde{l}_n(\pi, \beta_0, \beta, \sigma^2; j) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \pi_k\ \phi(y_i; \mu_{k,i}[j], \sigma_k^2)\right\} - \sum_{k=1}^{K} a_n(\sigma_k),$$

with

$$\mu_{k,i}[j] = \hat{\mu}_{k,i}^{(m)} + \beta_{0k} + \beta_k \boldsymbol{x}_i[j],$$

and $\pi, \beta_0, \beta, \sigma^2$ are vectors of length K.

Let $l_n(j) = \sup\{\tilde{l}_n(\pi, \beta_0, \beta, \sigma^2; j) : \alpha, \beta_0, \beta, \sigma^2\}$. Suppose j_0 satisfies $l_n(j_0) = \max_j l_n(j)$. The active set is then updated by $s^{(m+1)} = s^{(m)} \cup$

$\{j_0\}$, and we boost the fit by

$$\hat{\mu}_{k,i}^{(m+1)} = \hat{\mu}_{k,i}^{(m)} + \nu\ (\hat{\beta}_{0k} + \hat{\beta}_k\ \boldsymbol{x}_i[j_0]), k = 1, \ldots, K, \tag{13}$$

where $\hat{\beta}_{0k}$ and $\hat{\beta}_k$ are the adjusted maximum likelihood estimates corresponding to $l_n(j_0)$, and $0 < \nu \leq 1$ is a pre-specified stepsize (or shrinkage) parameter, as discussed further below.

3. Iteration: The boosting procedure is repeated with a pre-determined number of iterations M, or until, for example, the condition: $|s^{(M)}| < n/K$ is violated, where $|s^{(M)}|$ denotes the size of the active set and K is the order of the FMSL model. The iteration stops when the aforementioned condition is violated or when the maximum number of predetermined iterations is reached, whichever happens first.

The step size parameter ν in (13) controls the contribution of a selected variable in every update of the fit so that other potentially important variables are also given a chance to be selected in subsequent iterations. As mentioned in the literature (Bühlmann, 2006), the choice of ν is less crucial as long as its value is small, for example $\nu = 0.1$.

The outcome of the above algorithm is the final active set $s^{(M)}$. It contains at most M or n/K variables and it is the same for all components of the FMSL model. This facilitates the optimization problem involved in the second stage, which fits a model by maximizing the penalized likelihood (8) and using the variables in $s^{(M)}$ in order to select the final sparse FMR model.

Extensive simulation studies have shown that the above algorithm is very effective in greatly screening out a large number of potential features in the data while also retaining the important features in the active set $s^{(M)}$. This is a property that is referred to as sure screening in Fan and Lv (2008). Our hope is to be able to investigate conditions under which the algorithm has the sure screening property.

10 Future Directions and Extensions

A generalization of the FMR models are called mixture-of-experts (MOE) models in which both the mixing probabilities and the mixture components are functions of the covariates $x_1, x_2, \ldots, x_p$. The MOE models were first introduced by Jacob et al. (1991) in machine learning applications. They have often been applied in a problem decomposition context, where a complex problem is divided into a set of simpler subproblems, based on a divide-and-conquer principle, and then one or more specialized problem-solving tools or experts are assigned to each of the subproblems. Statistically, MOE models provide a rich class of regression models. They have have been extensively used in statistics, health sciences, bioinformatics, and many other disciplines; see Khalili (2010) and references within.

In a MOE model with K components, the conditional density function of Y given $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ is given by

$$f(y; \boldsymbol{x}, \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k(\boldsymbol{x}; \boldsymbol{\alpha})\ h(y; \theta_k(\boldsymbol{x}), \phi_k), \tag{14}$$

where $\theta_k(\boldsymbol{x})$ are the same as in the FMR model (4), and the mixing probabilities $\pi_k(\boldsymbol{x}; \boldsymbol{\alpha})$ are modeled in the multinomial logistic regression fashion

$$\log\left\{\frac{\pi_k(\boldsymbol{x}; \boldsymbol{\alpha})}{\pi_K(\boldsymbol{x}; \boldsymbol{\alpha})}\right\} = \alpha_{0k} + \boldsymbol{x}^\top \boldsymbol{\alpha}_k, \quad k = 1, 2, \ldots, K-1. \tag{15}$$

Let $\boldsymbol{\alpha}_k = (\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kp})$, then the vector of regression coefficients in the mixing probabilities is $\boldsymbol{\alpha} = (\alpha_{01}, \boldsymbol{\alpha}_1, \alpha_{02}, \boldsymbol{\alpha}_2, \ldots, \alpha_{0,K-1}, \boldsymbol{\alpha}_{K-1})$.

In MOE literature, the component density functions $h(y; \theta_k(\boldsymbol{x}), \phi_k)$ are called *experts* and the mixing probabilities $\pi_k(\boldsymbol{x}; \boldsymbol{\alpha})$ are called the *gating network*.

Figure 5 shows the architecture of a MOE model with $K = 4$ experts. In this model, based on the input variables $\boldsymbol{x}$, the gating network assigns an incoming task with probability $\pi_k(\boldsymbol{x})$ to experts k, for $k = 1, 2, 3, 4$. The output of each expert is represented by $\mu_k(\boldsymbol{x})$, which, in this example, is the conditional expectation $E(Y|\boldsymbol{x}, k)$. The overall output of the MOE model is then given by $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x}) = \sum_k \pi_k(\boldsymbol{x})\mu_k(\boldsymbol{x})$.

Due to (15), the MOE models provide a great deal of flexibility from a statistical modeling point of view. Jordan and Jacobs (1994), Peng, Jacobs and Tanner (1996), and Jiang and Tanner (1999) studied statistical inference and numerical computations in (hierarchical) MOE models. Carvalho and Tanner (2005), and Ge and Jiang (2006) studied these models in time series and classification problems.

Like any regression model, it is also natural to consider the feature selection problem in MOE models. The problem in MOE models becomes even more complex when subsets of significant features vary between the experts and also the gating network. Evidently, including all the features produces an undesirably large and complex MOE model. In applications such as market segmentation (Wedel and Kamakura, 2000), usually subsets of features are included in the mixture components. Siegmund et al.

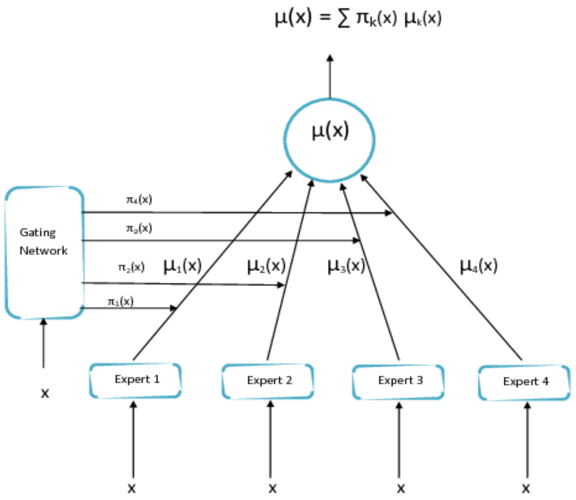

Figure 5: A MOE model with $K = 4$ experts.

(2006) used a MOE model to estimate the association between exposure and latent disease subtype measured by DNA methylation profiles; only the mixing probabilities were modeled as a function of the p potential exposures. Similarly, in a climatological application, Jeffries and Pfeiffer (2001) used such a model in estimating the distribution of rain rate. In machine learning applications (Jacobs, Peng and Tanner, 1997), different subsets of the features are included in both the mixing probabilities and the mixture components. However, a unified data-dependent approach that can automatically identify significant features in different parts of a MOE architecture is still lacking. Due to the intrinsic computational complexity, few statistical papers have addressed this problem. Jacobs, Peng and Tanner (1997) and the recent paper by Chung and Dunson (2009) proposed Bayesian approaches for model selection in MOE models. The Akaike information criterion (Akaike, 1973) and the Bayes information criterion (Schwarz, 1978) become computationally intensive for feature selection in MOE models even for moderate numbers of mixture components and features. It is also difficult to study the statistical properties of the final selected MOE model.

The recent work of Khalili (2010) studied the feature selection problem in MOE models using an extension of the penalized likelihood approach outlined in Section 4. The performance of the method is studied both theoretically and by simulations, and it is shown that the method is very promising in a MOE context. It is worth mentioning that the proposed methodology is also applicable for feature selection problems in MOE models for high dimensional data. However, the study of the statistical properties of the method in high-dimensional situations requires more advanced theoretical developments which is the subject of future research.

11 Concluding Remarks

The current state of the research on variable selection problems in mixture of regression (FMR) and mixture-of-experts (MOE) models indicates that the story is far from complete, particularly for the high-dimensional problems. Statistical techniques for high-dimensional data analysis are developed to address the challenges emerging in many scientific disciplines. Due to the complex nature of the new data, a single linear or generalized linear model may not be appropriate for modelling the data under study. The usage of FMR and MOE models thus continues to grow due to their flexibility in modeling, and new innovative techniques are needed for fitting these models to the data. Issues which remain to be addressed include the characterization of optimality properties, the selection of data-driven penalty functions and parameters, the confidence in selected models and estimated parameters, inference after model selection, the incorporation of information on covariates, and development of robust and user-friendly algorithms and software.

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.

Antoniadis, A. (1997), Wavelets in statistics: a review (with discussion). J. Italian Statist. Assoc., **96**, 97-144.

Antoniadis, A. and Fan, J. (2001), Regularization of wavelets approximations (with discussion). J. Amer. Statist. Assoc., **96**, 939-967.

Bickel, P. J. and Li, B. (2006), Regularization in statistics (with discussion). Test, **15**, 271-344.

Breiman, L. (1996), Heuristics of instability and stabilization in model selection. Ann. Statist., **24**, 2350-2383.

Bühlmann, P. (2006), Boosting for high-dimensional linear models. Ann. Statist., **34**, 559-583.

Bühlmann, P. and Yu, B. (2003), Boosting with the L2 loss: regression and classification. J. Amer. Statist. Assoc., **98**, 324-339.

Carvalho, A. X. and Tanner, M. A. (2005), Mixtures-of-experts of autoregressive time series: asymptotic normality and model specification. IEEE Transactions on Neural Networks, **16**, 39-56.

Chanock, S. J. and Hunter, D. J. (2008), Genomics: when the smoke clears. Nature, **452**, 537-538.

Chen, J. and Chen, Z. (2008), Extended Bayesian information criteria for model selection with large model spaces. Biometrika, **95**, 759-771.

Chen, J. and Khalili, A. (2008), Order selection in finite mixture models with a non-smooth penalty. J. Amer. Statist. Assoc., **103**, 1674-1683.

Chen, J., Tan, X., and Zhang, R. (2008), Consistency of penalized MLE for normal mixtures in mean and variance. Statist. Sinica, **18**, 443-465.

Chung, Y. and Dunson, D. B. (2009), Nonparametric Bayes conditional distribution modeling with variable selection. J. Amer. Statist. Assoc., **104**, 1646-1660.

Conlon, E., Liu, X., Lieb, J., and Liu, J. (2003), Integrating regulatory motif discovery and genome-wide expression analysis. Proceedings of the National Academy of Sciences, USA, **100**, 3339-3344.

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

DeSarbo, W. S. and Cron, W. L. (1988), A maximum likelihood methodology for clusterwise linear regression. Journal of Classification, **5**, 249-282.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), Least angle regression (with discussion). Ann. Statist., **32**, 407-499.

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

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

Fan, J. and Li, R. (2004), New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. J. Amer. Statist. Assoc., **99**, 710-723.

Fan, J. and Lv, J. (2010), A selective overview of variable selection in high dimensional feature space. Statist. Sinica, **20**, 101-148

Fan, J. and Lv, J. (2008), Sure independence screening for ultra-high dimensional feature space (with discussion). J. R. Statist. Soc. B, **70**, 849-911.

Fan, J. and Peng, H. (2004), Nonconcave penalized likelihood with a diverging number of parameters. Ann. Statist., **32**, 928-961.

Fan, J., Samworth, R., and Wu, Y. (2009), Ultrahigh dimensional variable selection: beyond the linear model. J. Mach. Learn. Res., **10**, 1829-1853.

Fan, J. and Song, R. (2010), Sure independence screening in generalized linear models with NP-dimensionality. Ann. Statist., **38**, 3567-3604.

Frühwirth-Schnatter, S. (2006), Finite Mixture and Markov Switching Models, New York: Springer Series in Statistics.

Ge, Y. and Jiang, W. (2006), On consistency of Bayesian inference with mixtures of logistic regression. Neural Comp., **18**, 224-243.

Green, P. J. (1990), On use of the EM algorithm for penalized likelihood estimation. J. R. Statist. Soc. B., **52**, 443-452.

Gupta, M. and Ibrahim, J. G. (2007), Variable selection in regression mixture modeling for the discovery of gene regulatory networks. J. Amer. Statist. Assoc., **102**, 867-880.

Hastie, T., Tibshirani, R., and Friedman, J. (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., New York: Springer Series in Statistics.

Hennig, C. (2000), Identifiability of models for clusterwise linear regression. J. Classific., **17**, 273-296.

Hoerl, A. E. and Kennard, R. W. (1970), Ridge regression: applications to non-orthogonal problems. Technometrics, **1**, 69-82.

Huang, J., Ma, S. G., Xie, H. L., and Zhang, C.-H. (2009), A group bridge approach for variable selection. Biometrika, **96**, 339-355.

Hunter, D. R. and Li, R. (2005), Variable selection using MM algorithms. Ann. Statist., **33**, 1617-1642.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), Adaptive mixture of local experts. Neural Computat., **3**, 79-87.

Jacobs, R. A., Peng, F., and Tanner, M. A. (1997), A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. Neural Networks, **10**, 231-241.

Jeffries, N. and Pfeiffer, R. (2001), A mixture model for the probability distribution of rain rate. Environmetrics, **12**, 1-10.

Jiang, W. and Tanner, M. A. (1999), Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. Ann. Statist., **27**, 987-1011.

Jordan, M. and Jacobs, R. (1994), Hierarchical mixtures-of-experts and the EM algorithm. Neural Computat., **6**, 181-214.

Keribin, C. (2000), Consistent estimation of the order of mixture models. Sankhya Series A, **62**, 49-66.

Khalili, A. and Chen, J. (2007), Variable selection in finite mixture of regression models. J. Amer. Statist. Assoc., **102**, 1025-1038.

Khalili, A. (2010), New estimation and feature selection methods in mixture-of-experts models. The Can. J. Statist., **38**, 519-539.

Khalili, A., Chen, J., and Lin, S. (2011), Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. Biostatistics, **12**, 156-172.

Khalili, A. and Lin, S. (2011), Regularization in finite mixture of regression models with diverging number of parameters. Submitted Manuscript.

Kim, Y., Choi, H., and Oh, H. S. (2008), Smoothly clipped absolute deviation on high dimensions. J. Amer. Statist. Assoc., **103**, 1665-1673.

Mallows, C. L. (1973), Some Comments on Cp. Technometrics, **15**, 661-675.

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

Meinshausen, N. and Buhlmann, P. (2006), High dimensional graphs and variable selection with the Lasso. Ann. Statist., **34**, 1436-1462.

Meinshausen, N. and Yu, B. (2009), Lasso-type recovery of sparse representations for high-dimensional data. Ann. Statist., **37**, 246-270.

Peng, F., Jacobs, R. A., and Tanner, M. A. (1996), Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. J. Amer. Statist. Assoc., **91**, 953-960.

Rosset, S. and Zhu, J. (2007), Piecewise linear regularized solution paths. Ann. Statist., **35**, 1012-1030.

Schelldorfer, J., Bühlmann, P., and van de Geer, S. (2011), Estimation for high-dimensional linear mixed-effects models using L1-penalization. To appear in Scand. J. Statist..

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

Shao, J. (1997), An asymptotic theory for linear model selection. Statistica Sinica, **7**, 221-264.

Siegmund, K. D., Levine, A. J., Chang, J., and Laird, P. W. (2006), Modeling exposures for DNA methylation profiles. Cancer Epidemiology Biomarkers and Prevention, **15**, 567-572.

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

Städler, N., Bühlmann, P., and van de Geer, S. (2010), l1-penalization for mixture regression models (with discussion). Test, **19**, 209-285.

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

Wang, H., Li, R., and Tsai, C.-L. (2007), Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika, **94**, 553-568.

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.

Wong, C. S. and Li, W. K. (2000), On a mixture autoregressive model. J. R. Statist. Soc. B, **62**, 95-115.

Wu, C. F. J. (1983), On the convergence properties of the EM algorithm. Ann. Statist., **11**, 95-103.

Van de Geer, S. (2008), High-dimensional generalized linear models and the lasso. Ann. Statist., **36**, 614-645.

Xie, H. L. and Huang, J. (2009), SCAD-penalized regression in high-dimensional partially linear models. Ann. Statist., **37**, 673-696.

Zhang, C.-H. and Huang, J. (2008), The sparsity and bias of the LASSO selection in high-dimensional linear regression. Ann. Statist., **36**, 1567-1594.

Zhang, Y., Li, R., and Tsai, C.-L. (2010), Regularization parameter selections via generalized information criterion. J. Amer. Statist. Assoc., **105**, 312-323.

Zhong, W., Zeng, P., Ma, P., Liu, J. S., and Zhu, Y. (2005), RSIR: regularized sliced inverse regression for motif discovery. Bioinformatics, **21**, 4169-4175.

Zou, H. and Hastie, T. (2005), Regularization and variable selection via the elastic net. J. R. Statist. Soc. B, **67**, 301-320.

Zou, H. and Li, R. (2008), One-step sparse estimates in nonconcave penalized likelihood models (with discussion). Ann. Statist., **36**, 1509-1566.