Statistics and Its Interface Volume 8 (2015) 239–254

Componentwise variable selection in finite mixture regression

Bin Chen* and Keying Ye

The finite mixture regression is a method to account for heterogeneity in relationship between the response variable and the predictor variables. The goal of this research is to investigate the variable selection issue within each component in the finite mixture regression. This has not been studied much in the literature from a Bayesian perspective. We propose an approach by embedding variable selection into the data augmentation method that iteratively updates estimation in two steps: estimate parameters for each component and determine the latent membership of each observation. Componentwise variable selection is realized by imposing special priors or procedures designed for parsimony in the first step. Due to separation of the two steps, our approach provides a freedom to choose from a wide variety of variable selection techniques. In particular, we illustrate how two popular variable selection techniques can be embedded in the proposed approach: g-prior and Stochastic Search Variable Selection. A simulation study is conducted to assess performance of the proposed approach under a variety of scenarios through investigating accuracy of variable selection and clustering. Results show that the proposed approach successfully identifies important variables even in noisy scenarios. The proposed approach is also applied to a real data set from bioinformatics and the results provide evidence to an existing hypothesis.

Keywords and phrases: Bayesian, Mixture regression, Componentwise, Variable selection.

1. INTRODUCTION

One of the important problems with a regression model is variable selection. Existing methodologies often perform variable selection based on whole observed sample. However, under many circumstances this practice is inadequate since the sample may come from a heterogeneous population that is composed of several subpopulations. For instance, one of the important econometric topics is to assess which factors have significant impact on GDP growth. The list of candidate factors includes GDP level, life expectancy, primary school enrollment rate, and so on, e.g., see Fernández et al. (2001). The standard treatment is to collect data during a certain time period and/or from multiple regions and then fit one single regression-type model for the whole collected sample. But we might want to raise the question: is regression of GDP heterogeneous in nature, that is, can selection of explanatory factors vary by region or by time?

Dealing with heterogeneous variable selection in regression involves two inference problems: first, separating the whole sample into multiple components (Clustering), and second, each component has its own variable selection (Componentwise Variable Selection) as well as parameter estimation. Statistical literature has seen numerous studies on clustering and variable selection on their own. Clustering methods can be commonly categorized into three major groups: partitioning algorithms, hierarchical algorithms (e.g., *k-means* in MacQueen, 1967), and model-based methods (e.g., finite mixture model in McLachlan and Peel, 2000). Variable selection techniques could be roughly classified into four categories: classical methods (e.g., Mallows' C_p, R^2, and subset selection algorithm), information criteria (e.g., AIC and BIC), shrinkage methods (e.g., LASSO in Tibshirani, 1996) and Least Angle Regression in Efron et al., 2004), and Bayesian methods (Stochastic Search Variable Selection (George and McCulloch, 1993), g-prior (Zellner, 1986), and Bayesian LASSO (Park and Casella, 2008).

Although literature is rich in clustering and variable selection, there is little existing research in tackling the two subjects simultaneously to disclose heterogeneous variable selection in regression models. Relevant studies such as Gupta and Ibrahim (2007), who, however, select variables that are shared by all the components. Khalili and Chen (2007) propose a frequentist method that truly solves this problem. In this paper we propose an approach to incorporate variable selection techniques with the finite mixture regression under a Bayesian framework.

The paper is organized as follows. In Section 2 we propose a Bayesian approach for componentwise variable selection in finite mixture regression. Two commonly used variable selection techniques are then illustrated to fit into the approach. A series of simulation studies are conducted to evaluate performance of the proposed approach under various scenarios in Section 3. The proposed method is applied to a high-dimension real dataset from bioinformatics in Section 4. Lastly Section 5 concludes with discussions.

*Corresponding author.

2. COMPONENTWISE VARIABLE SELECTION IN FINITE MIXTURE REGRESSION

There exist papers regarding variable selection in heterogeneous regression in the last decade. We classify these studies into two categories: global variable selection and componentwise variable selection. Global variable selection methods (e.g., Gupta and Ibrahim, 2007) identify significant variables that are shared by all the components, whereas the goal of componentwise variable selection is to select variables within each component. The latter is the interest of this paper.

Compared to global variable selection, componentwise variable selection is a better solution because in many situations each component may have its own choice of explanatory variables. An early study by Wang et al. (1996) takes a two-stage approach: first to determine the number of components with all the variables included, and then to perform variable selection within each component using AIC or BIC. A new-generation study of simultaneously implementing clustering and variable selection with more realistic computation is done by Khalili and Chen (2007). They propose a frequentist method by replacing the regular likelihood function in the mixture likelihood with the penalized log-likelihood function. The EM algorithm is then used to maximize the mixture likelihood function. They claim that their method is consistent in selecting the most parsimonious mixture regression model. With the same research goal as Khalili and Chen (2007), we propose a Bayesian approach in this section.

2.1 Data augmentation approach for componentwise variable selection

We begin with discussing the Bayesian approach to linear regression, which will be utilized for componentwise estimation later on. Suppose the data have the response variable $\boldsymbol{y}$ and a set of candidate covariates $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_p$, which are all $n \times 1$ vectors. The likelihood density of a linear regression model is

$$\boldsymbol{y}|X, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}\left(X\boldsymbol{\beta}, \sigma^2 \mathrm{I}\right), \tag{1}$$

where $X = (\mathbf{1}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_q)$ and $\mathcal{N}(\cdot)$ is the normal distribution density with the parameters $\boldsymbol{\beta}$ and σ^2. The central issue of the Bayesian approach to linear regression is the choice of priors for the parameters $\boldsymbol{\beta}$ and σ^2. We will briefly describe three types of priors: conjugate prior, noninformative prior, and g-prior. Detailed expositions can be found in excellent references such as Box and Tiao (1973) and Marin and Robert (2007).

A conjugate prior is assigned for computational convenience. The normal-inverse-gamma structure is the common conjugate prior for Bayesian analysis in linear regression, that is, imposing a multivariate normal prior distribution on $\boldsymbol{\beta}$,

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}\left(\boldsymbol{b}_0, \sigma^2 \mathrm{B}_0\right), \text{ and } \sigma^2 \sim \mathcal{G}^{-1}\left(n_0/2, S_0/2\right), \tag{2}$$

where $\mathcal{G}^{-1}(\cdot)$ is the inverse-gamma density function with the parameters n_0 and S_0.

Choices of the hyperparameters $\boldsymbol{b}_0$, B_0, n_0 and S_0 in (2) are important, but they are not always easy to assign in practice. This is one of the reasons that people often use noninformative (or weakly informative, e.g., see Gelman et al. 2009) priors so that hyperparameters are assigned automatically or semi-automatically. One choice of non-informative priors is Jeffreys' prior. For the linear regression model (1), the independent Jeffreys' prior is

$$P(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}. \tag{3}$$

One problem with Jeffreys' prior is its impropriety, that is, it can not be normalized to become a distribution. To overcome this drawback, Zellner (1986) proposed the g-prior:

$$\boldsymbol{\beta}|\sigma^2, g \sim \mathcal{N}\left(\boldsymbol{\beta}_0, g\sigma^2\left(X^{'}X\right)^{-1}\right), \tag{4}$$

where g is a scalar.

Usually we assume that the regression parameters in (1) are homogenous over the population. The inadequacy of this assumption arises in many areas such as economy, marketing, and biology. A reasonable alternative is to assume that there are K components in the population and each component has its own parameter values $(\boldsymbol{\beta}_k, \sigma_k^2)$. A common approach to handle the varying parameters is finite mixture regression (FMR), which assumes that each observation y_i is generated from $(\boldsymbol{\beta}_1, \sigma_1^2), \ldots, (\boldsymbol{\beta}_K, \sigma_K^2)$ with probabilities $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_K\}$, respectively. The likelihood density of FMR is defined as

$$\begin{aligned} P\left(\boldsymbol{y}|\boldsymbol{\psi}\right) &= \prod_{i=1}^{n}\sum_{k=1}^{K}\omega_k P\left(y_i|\boldsymbol{\beta}_k, \sigma_k^2\right) \\ &= \prod_{i=1}^{n}\sum_{k=1}^{K}\omega_k \mathcal{N}\left(y_i \mid \boldsymbol{x}_i^{'}\boldsymbol{\beta}_k, \sigma_k^2\right), \end{aligned} \tag{5}$$

where $\boldsymbol{\psi} = \{\boldsymbol{\beta}_k, \sigma_k^2, \omega_k \mid k = 1, \ldots, K\}$ is a reparameterized entity and K is the number of components.

The idea of FMR with the likelihood in (5) first came from Quandt (1972), who maximized the likelihood function using a numerical method. Hartigan (1977) attempts an EM-type algorithm to find the ML estimation, but the direct use of the EM algorithm is by DeSarbo and Cron (1988). Quandt and Ramsey (1978) use a method of moments estimator based on the moment-generating function. In Bayesian domain, the two common estimation approaches are the Gibbs sampling (Diebolt and Robert, 1990) and the Metropolis-Hastings sampling (Hurn et al., 2003 and Celeux

et al., 2000). We next discuss the Gibbs sampling with data augmentation (Dempster et al., 1977; Tanner and Wong, 1987).

The idea of data augmentation is to incorporate the latent allocation variable. Let $\boldsymbol{z}$ be the allocation vector such that $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)$, $z_i \in \{1, \ldots, K\}$, indicating the membership of each observation. Thus, we can have the complete-data likelihood density,

(6)
$$P(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{\psi}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \left[\mathrm{I}_{\{z_i=k\}} \omega_k P\left(y_i \mid \boldsymbol{\beta}_k, \sigma_k^2\right) \right] = \left\{ \prod_{k=1}^{K} \prod_{i:z_i=k} \mathcal{N}\left(y_i \mid \boldsymbol{x}_i^{'} \boldsymbol{\beta}_k, \sigma_k^2\right) \right\} \left\{ \prod_{k=1}^{K} \omega_k^{n_k} \right\}.$$

where $\mathrm{I}_{\{.\}}$ is the indicator function and $n_k = \sum_{i=1}^{n} \mathrm{I}_{\{z_i=k\}}$ is the number of observations in the k^{th} component. The likelihood in (6) is factorized into two independent parts with $(\boldsymbol{\beta}_k, \sigma_k^2)$'s and $\boldsymbol{\omega}$, respectively. A natural way of specifying the prior is that $P(\boldsymbol{\psi}) = P(\boldsymbol{\omega}) \prod_{k=1}^{K} P(\boldsymbol{\beta}_k, \sigma_k^2)$ assuming independence between $\boldsymbol{\omega}$ and $(\boldsymbol{\beta}_k, \sigma_k^2)$'s. Therefore, the complete-data posterior distribution has the same structure as (6):

$$P(\boldsymbol{\psi}|\boldsymbol{y}, \boldsymbol{z}) = \left\{ \prod_{k=1}^{K} P\left(\boldsymbol{\beta}_k, \sigma_k^2 | \boldsymbol{y}, \boldsymbol{z}\right) \right\} P(\boldsymbol{\omega}|\boldsymbol{z}),$$

where

(7)
$$P\left(\boldsymbol{\beta}_k, \sigma_k^2 | \boldsymbol{y}, \boldsymbol{z}\right) \propto P\left(\boldsymbol{\beta}_k, \sigma_k^2\right) \prod_{i:z_i=k} P\left(y_i | \boldsymbol{\beta}_k, \sigma_k^2\right),$$
$$P(\boldsymbol{\omega}|\boldsymbol{z}) \propto P(\boldsymbol{\omega}) \prod_{k=1}^{K} \omega_k^{n_k}.$$

In the above setting $(\boldsymbol{\beta}_k, \sigma_k^2)$ and $\boldsymbol{\omega}$ can be sampled independently, which makes the sampling much easier. The following sampling scheme shows that the parameters and the latent variables $\boldsymbol{z}$ can be sampled iteratively:

(8)
$$\begin{array}{ccc} \boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi} & \longrightarrow & \boldsymbol{\omega} \mid \boldsymbol{z} \\ \nwarrow & & \swarrow \\ & \boldsymbol{\beta}_k, \sigma_k^2 \mid \boldsymbol{y}, \boldsymbol{z} & \\ & \text{for } k = 1, \ldots, K & \end{array}$$

Now we are ready to discuss the goal of this paper — componentwise variable selection. The sampling scheme in (8) allows for two iterative steps: cluster observations into some components through sampling $\boldsymbol{z}$, and then sample other parameters within each component. In other words given $\boldsymbol{z}$, the parameters $(\boldsymbol{\beta}_k, \sigma_k^2, \boldsymbol{\omega})$ can be estimated the way they are treated in the non-mixture context. If we could make a further step by assigning special priors on $(\boldsymbol{\beta}_k, \sigma_k^2)$ which lead to model sparseness, variable selection can be achieved and it is within each component. Some examples of special priors for $(\boldsymbol{\beta}_k, \sigma_k^2)$ will be discussed in Section 2.2. In the rest of this section, we first incorporate an extra parameter into (8) and then work out the conditional distributions of $\boldsymbol{\omega}$ and $\boldsymbol{z}$.

We first add an extra step into (8) such that

(9)
$$\begin{array}{ccccc} \boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\psi}, \boldsymbol{\gamma}_k & \longrightarrow & \boldsymbol{\gamma}_k \mid \boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\psi} & \longrightarrow & \boldsymbol{\omega} \mid \boldsymbol{z} \\ & \nwarrow & \text{for } k = 1, \ldots, K & \swarrow & \\ & & \boldsymbol{\beta}_k, \sigma_k^2 \mid \boldsymbol{\gamma}_k, \boldsymbol{y}, \boldsymbol{z} & & \\ & & \text{for } k = 1, \ldots, K & & \end{array}$$

where $\boldsymbol{\gamma}_k$ is the indicator vector denoting which explanatory variables are included in or excluded from the k^{th} component, for $k = 1, \ldots, K$. The definition and the derivation of its posterior distribution of $\boldsymbol{\gamma}_k$ will be discussed in detail in Section 2.2.1. It is worth noting that (9) is a general scheme which may have variations depending on the variable selection technique implemented. For example, if Reversible Jump MCMC is used as the variable selection technique, $\boldsymbol{\gamma}_k$ and $\boldsymbol{\beta}_k$ are estimated in one single step instead of two separate steps.

Posterior distributions need to be worked out in order to run the Gibbs sampling. We first work on the posteriors for $\boldsymbol{\omega}$ and $\boldsymbol{z}$, while posteriors of other parameters depends on specific variable selection techniques and will be discussed in Section 2.2. First, from (7), we can see that n_k follows a multinomial distribution. The conjugate prior is thus a Dirichlet distribution denoted by $\mathcal{D}(\cdot)$. We let $\boldsymbol{\omega} \sim \mathcal{D}(\alpha_1, \ldots, \alpha_K)$, which leads to the following posterior distribution

$$P(\boldsymbol{\omega}|\boldsymbol{z}) = \mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K).$$

Second, $\boldsymbol{z}$ tells the membership of each observation. In fact, the weight parameter $\boldsymbol{\omega}$ can be viewed as the distribution of z_i without observing the data, i.e., $P(z_i = k) = \omega_k$, for $k = 1, \ldots, K$. The posterior distribution of each z_i can thus be derived following Bayes' rule:

(10)
$$P(z_i = k \mid y_i, \boldsymbol{\psi}) = \frac{\mathcal{N}\left(y_i \mid \boldsymbol{x}_i^{'} \boldsymbol{\beta}_k, \sigma_k^2\right) \omega_k}{\sum_{j=1}^{K} \mathcal{N}\left(y_i \mid \boldsymbol{x}_i^{'} \boldsymbol{\beta}_j, \sigma_j^2\right) \omega_j}.$$

We now have a general framework for variable selection within each component in finite mixture regression, as expressed in the following algorithm.

Algorithm 1. (**Data Augmentation Approach for Componentwise Variable Selection in FMR**)

Start with a random allocation $\boldsymbol{z}^{(0)}$ and repeat the following steps for $t = 1, \ldots, M$:

1. Sampling $\boldsymbol{\psi}^{(t)}$ conditional on $\boldsymbol{z}^{(t-1)}$:
   (a) Generate $\boldsymbol{\omega}^{(t)}$ from $\mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$.

(b) Generate $\boldsymbol{\beta}_k^{(t)}$, $\sigma_k^{2(t)}$ and the indicator vector $\boldsymbol{\gamma}_k^{(t)}$ if necessary for $k = 1, \ldots, K$. Appropriate priors or procedures are set up in order to impose model sparseness.

Now we have $\boldsymbol{\psi}^{(t)} = (\boldsymbol{\omega}^{(t)}, \boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_K^{(t)}, \sigma_1^{2(t)}, \ldots, \sigma_K^{2(t)}, \boldsymbol{\gamma}_1^{(t)}, \ldots, \boldsymbol{\gamma}_K^{(t)})$.

2. Sampling $\boldsymbol{z}^{(t)}$ conditional on $\boldsymbol{\psi}^{(t)}$:
Sample $z_i^{(t)}$ from $P(z_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\psi}^{(t)})$ given in (10) for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

Algorithm 1 is quite flexible so that many variable selection techniques can be easily incorporated into the algorithm.We next demonstrate how different techniques are incorporated into Algorithm 1.

2.2 Choice of variable selection methods

There are numerous studies on Bayesian variable selection methods. In Section 1, we listed some important methods in the literature. The data augmentation framework in Algorithm 1 is ready to incorporate any of them as a component to achieve variable selection. In the following we use the g-prior and SSVS to demonstrate the details of how they are embedded into our proposed approach.

2.2.1 g-prior

The general form of g-prior was presented in (4). We now introduce a latent binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ with $\gamma_j = 1$ or 0, which tells the variable is in the model if $\gamma_j = 1$, and vice versa. The natural prior of γ_j is a Bernoulli distribution, i.e., $P(\gamma_j = 1) = \pi$. The common choice is $\pi = .5$. Now by including $\boldsymbol{\gamma}$ we can have the g-prior in the context of variable selection:

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma}, \sigma^2, g \sim \mathcal{N}\left(\mathbf{0}, g\sigma^2 \left(X_{\boldsymbol{\gamma}}^{'} X_{\boldsymbol{\gamma}}\right)^{-1}\right),$$

where $X_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are the design matrix and the regression coefficients, respectively, corresponding to the model $M_{\boldsymbol{\gamma}}$ that contains explanatory variables with $\gamma_j = 1$. Coefficients of the explanatory variables not included in the model are set to zero, as done by Mitchell and Beauchamp (1988). The g-prior incorporates the relationship between explanatory variables since the term $(X_{\boldsymbol{\gamma}}' X_{\boldsymbol{\gamma}})^{-1}$ provides correlated priors for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ by borrowing the covariance structure of the design matrix $X_{\boldsymbol{\gamma}}$ in the data. The scalar g controls the amount of information in the prior relative to the data. The other advantage of the g-prior is that since the prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is conditional on σ^2, the full conditional distributions have an analytic form. Specifically, if the prior for σ^2 is

$$P\left(\sigma^2\right) \propto \sigma^{-2},$$

which is actually the limiting form of the inverse gamma distribution, the full conditional distributions can be derived as:

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \sigma^2, \boldsymbol{y}, X_{\boldsymbol{\gamma}} \sim \mathcal{N}\left(\frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}, \frac{g}{g+1}\sigma^2 \left(X_{\boldsymbol{\gamma}}' X_{\boldsymbol{\gamma}}\right)^{-1}\right),$$

(11)
$$\sigma^2 \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{y}, X_{\boldsymbol{\gamma}} \sim \mathcal{G}^{-1}\left(\frac{n}{2}, \frac{s^2}{2} + \frac{1}{2(g+1)}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}' X_{\boldsymbol{\gamma}}' X_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right),$$

where $s^2 = (\boldsymbol{y} - X_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})'(\boldsymbol{y} - X_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}})$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the ordinary least squares estimate. The posterior marginal density for $\boldsymbol{\gamma}$ is also available

$$P(\boldsymbol{\gamma} \mid \boldsymbol{y}, X) \propto \left(\frac{1}{g+1}\right)^{(q_\gamma+1)/2} \left(\boldsymbol{y}'\boldsymbol{y} - \frac{g}{g+1}\boldsymbol{y}' X_{\boldsymbol{\gamma}} \left(X_{\boldsymbol{\gamma}}' X_{\boldsymbol{\gamma}}\right)^{-1} X_{\boldsymbol{\gamma}}' \boldsymbol{y}\right),$$

where q is the number of variables selected. This is very useful for designing a stochastic search for the most likely model when there are many candidate explanatory variables. To do this, we need the full conditional distribution of each γ_j. Let $\boldsymbol{\gamma}_{-j} = \{\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots \gamma_p\}$. The full conditional distribution of γ_j, $P(\gamma_j \mid \boldsymbol{\gamma}_{-j}, \boldsymbol{y}, X)$, can be computed by evaluating the probability of $\gamma_j = 0$ and $\gamma_j = 1$ while all the other explanatory variables in $\boldsymbol{\gamma}_{-j}$ stay the same from the previous step. The most likely model can then be identified by making inference from the posterior samples of individual γ_j's. We can easily fit this setting for variable selection in Step 1 of Algorithm 1. The parameters are sampled for each component as follows.

First, we iteratively sample each γ_{kj}, which takes the values 1 or 0. Here the subscript kj denotes the j^{th} variable in the k^{th} component. The probability of $\gamma_{kj} = 1$ is

$$P\left(\gamma_{kj} = 1 \mid \boldsymbol{\gamma}_{k(-j)}, \boldsymbol{y}_k, X_k\right) = \frac{o_{kj}}{1 + o_{kj}}, \tag{12}$$

where o_{kj} is the conditional odds:

$$\begin{aligned} o_{kj} &= \frac{P\left(\gamma_{kj} = 1 \mid \boldsymbol{\gamma}_{k(-j)}, \boldsymbol{y}_k, X_{\boldsymbol{\gamma}_k}\right)}{P\left(\gamma_{kj} = 0 \mid \boldsymbol{\gamma}_{k(-j)}, \boldsymbol{y}_k, X_{\boldsymbol{\gamma}_k}\right)} \\ &= \frac{P\left(\boldsymbol{y}_k \mid \gamma_{kj} = 1, \boldsymbol{\gamma}_{k(-j)}, X_{\boldsymbol{\gamma}_k}\right)}{P\left(\boldsymbol{y}_k \mid \gamma_{kj} = 0, \boldsymbol{\gamma}_{k(-j)}, X_{\boldsymbol{\gamma}_k}\right)} \frac{P(\gamma_{kj} = 1)}{P(\gamma_{kj} = 0)}, \end{aligned} \tag{13}$$

where $P(\boldsymbol{y}_k \mid \gamma_{kj}, \boldsymbol{\gamma}_{k(-j)}, X_{\boldsymbol{\gamma}_k})$ is given in (11) and $P(\gamma_{kj})$ is the model prior.

Second, we draw samples of $\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}$ and σ_k^2 conditional on the current state of $\boldsymbol{\gamma}_k$ within each component,

$$\begin{aligned} &\boldsymbol{\beta}_{\boldsymbol{\gamma}_k} \mid \boldsymbol{\gamma}_k, \sigma_k^2, \boldsymbol{y}_k, X_{\boldsymbol{\gamma}_k} \\ &\quad \sim \mathcal{N}\left(\frac{g_k}{g_k+1}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_k}, \frac{g_k}{g_k+1}\sigma_k^2 \left(X_{\boldsymbol{\gamma}_k}' X_{\boldsymbol{\gamma}_k}\right)^{-1}\right) \\ &\sigma_k^2 \mid \boldsymbol{\gamma}_k, \boldsymbol{\beta}_{\boldsymbol{\gamma}_k}, \boldsymbol{y}_k, X_{\boldsymbol{\gamma}_k} \\ &\quad \sim \mathcal{G}^{-1}\left(\frac{n_k}{2}, \frac{s_k^2}{2} + \frac{1}{2(g_k+1)}\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_k}' X_{\boldsymbol{\gamma}_k}' X_{\boldsymbol{\gamma}_k} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_k}\right), \end{aligned} \tag{14}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_k}$ and s_k^2 are the least square estimates of $\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}$ and σ_k^2, respectively. The Gibbs sampling procedure in (12)–(14) is illustrated in the diagram below:

$$\begin{array}{ccccc} \boldsymbol{\gamma}_k^{(t)} & \longrightarrow & \sigma_k^{2(t)} & \longrightarrow & \boldsymbol{\beta}_{\boldsymbol{\gamma}_k}^{(t)} \\ \downarrow & & & & \\ \boldsymbol{\gamma}_k^{(t+1)} & \longrightarrow & \sigma_k^{2(t+1)} & \longrightarrow & \boldsymbol{\beta}_{\boldsymbol{\gamma}_k}^{(t+1)}. \end{array}$$

The above sampling scheme is different from the usual Gibbs sampling in that the drawing of $\boldsymbol{\gamma}_k$ does not depend on $(\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}, \sigma_k^2)$. This is not problematic: the Gibbs sampling of $\boldsymbol{\gamma}_k$ converges to the target distribution $P(\boldsymbol{\gamma}_k \mid \boldsymbol{y}_k, X_k)$. Since $(\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}, \sigma_k^2)$ are sampled depending on $\boldsymbol{\gamma}_k$ through $P(\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}, \sigma_k^2 \mid \boldsymbol{\gamma}_k, \boldsymbol{y}_k, X_k)$, they are also ensured to converge to the target distribution $P(\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}, \sigma_k^2 \mid \boldsymbol{y}_k, X_k)$ (see Hoff, 2009, Chapter 9).

We now summarize the Gibbs sampling method in the following algorithm.

Algorithm 2. (**Componentwise Variable Selection in FMR using Gibbs Sampling with g-prior**)

Start with a random allocation $\boldsymbol{z}^{(0)}$ and repeat the following steps for $t = 1, \ldots, M$:

1. Sampling $\boldsymbol{\psi}^{(t)}$ conditional on $\boldsymbol{z}^{(t-1)}$:
   (a) Generate $\boldsymbol{\omega}^{(t)}$ from $\mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$.
   (b) Generate $\gamma_{kj}^{(t)}$ from $\mathcal{BER}(\frac{o_{kj}}{1+o_{kj}})$ given in (12) for $j = 1, \ldots, p$ and $k = 1, \ldots, K$.
   (c) Generate $\sigma_k^{2(t)}$ from $P(\sigma_k^2 \mid \boldsymbol{\gamma}_k^{(t)}, \boldsymbol{\beta}_{\boldsymbol{\gamma}_k}^{(t-1)}, \boldsymbol{y}_k^{(t-1)}, X_{\boldsymbol{\gamma}_k}^{(t-1)})$ given in (14) for $k = 1, \ldots, K$.
   (d) Generate $\boldsymbol{\beta}_{\boldsymbol{\gamma}_k}^{(t)}$ from $P(\boldsymbol{\beta}_{\boldsymbol{\gamma}_k} \mid \boldsymbol{\gamma}_k^{(t)}, \sigma_k^{2(t)}, \boldsymbol{y}_k^{(t-1)}, X_{\boldsymbol{\gamma}_k}^{(t-1)})$ given in (14) and let other β_{kj}'s with $\gamma_{kj}^{(t)} = 0$ be zero for $k = 1, \ldots, K$.

   Now we have $\boldsymbol{\psi}^{(t)} = (\boldsymbol{\omega}^{(t)}, \boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_K^{(t)}, \sigma_1^{2(t)}, \ldots, \sigma_K^{2(t)}, \boldsymbol{\gamma}_1^{(t)}, \ldots, \boldsymbol{\gamma}_K^{(t)})$.
2. Sampling $\boldsymbol{z}^{(t)}$ conditional on $\boldsymbol{\psi}^{(t)}$:
   Sample $z_i^{(t)}$ from $P(z_i \mid y_i, \boldsymbol{x}_i, \boldsymbol{\psi}^{(t)})$ given in (10) for $i = 1, \ldots, n$ and $k = 1, \ldots, K$.

2.2.2 SSVS

Stochastic Search Variable Selection (SSVS) by George and McCulloch (1993) is another well-known method for variable selection that can be used to implement 1. Similar to the method above SSVS also relies on the binary vector $\boldsymbol{\gamma}$. The difference is that SSVS allows β_{kj} to shrink towards zero by assigning a special prior as described next. For each component, SSVS specifies the priors as follows:

$$\begin{aligned} \gamma_{kj} &\sim \mathcal{B}(\delta), \\ \beta_{kj}|\gamma_{kj} &\sim \mathrm{I}_{\{\gamma_{kj}=0\}}\mathcal{N}(0, \tau^2) + \mathrm{I}_{\{\gamma_{kj}=1\}}\mathcal{N}(0, c^2\tau^2), \\ \sigma_k^2|\boldsymbol{\gamma}_k &\sim \mathcal{G}^{-1}(n_0/2, S_0/2), \end{aligned}$$

for $j = 1, \ldots, p$, where $\mathcal{B}(\cdot)$ is the density function of a binomial distribution. We set τ small so that β_{kj} is likely to be close to zero if $\gamma_{kj} = 0$, and set c ($c > 0$) large so that β_{kj} is away from zero if $\gamma_{kj} = 1$. The priors of β_{kj}'s in can be a multivariate normal,

$$\boldsymbol{\beta}_k|\boldsymbol{\gamma}_k \sim \mathcal{N}(\mathbf{0},\ D_{\boldsymbol{\gamma}_k} R_k D_{\boldsymbol{\gamma}_k}), \tag{15}$$

where $D_{\boldsymbol{\gamma}} = diag(a_1\tau_1, \ldots, a_p\tau_p)$ with $a_j = 1$ if $\gamma_{kj} = 0$ and $a_j = c$ if $\gamma_{kj} = 1$. George and McCulloch (1993) derive the full conditional distributions for $\boldsymbol{\beta}_k$, σ_k^2, and $\boldsymbol{\gamma}_k$ as follows. Note that in the light of data augmentation the posteriors of parameters are all conditional on the allocation $\boldsymbol{z}$.

$$\begin{aligned} P(\boldsymbol{\beta}_k \mid \boldsymbol{\gamma}_k, \sigma_k^2, X_k, \boldsymbol{y}_k) &= \mathcal{N}(\sigma_k^{-2}\boldsymbol{b}_k, B_k), \\ P(\sigma_k^2 \mid \boldsymbol{\beta}_k, \boldsymbol{\gamma}_k, X_k, \boldsymbol{y}_k) &= \mathcal{G}^{-1}(N_k/2, s_k^2/2), \\ P(\gamma_{kj} \mid \boldsymbol{\gamma}_{k(-j)}, \boldsymbol{\beta}_k, \sigma_k^2) &= \mathcal{B}\left(\frac{a}{a+b}\right), \end{aligned} \tag{16}$$

where $N_k = n_0 + n_k$, $s_k^2 = S_0 + (\boldsymbol{y}_k - X_k\boldsymbol{\beta}_k)'(\boldsymbol{y}_k - X_k\boldsymbol{\beta}_k)$, $B_k = (\sigma_k^{-2}X_k'X_k + D_{\boldsymbol{\gamma}_k}^{-1}R_k^{-1}D_{\boldsymbol{\gamma}_k}^{-1})^{-1}$, $\boldsymbol{b}_k = B_kX_k'\boldsymbol{y}_k$, $a = \rho \cdot f(\boldsymbol{\beta}_k|\gamma_{kj} = 1, \boldsymbol{\gamma}_{k(-j)})$, and $b = (1-\rho)\cdot f(\boldsymbol{\beta}_k|\gamma_{kj} = 0, \boldsymbol{\gamma}_{k(-j)})$. Here ρ is the pre-specified prior probability of each variable being included in the model, $f(\cdot)$ is the prior density in (15), and $\boldsymbol{\gamma}_{k(-j)} = (\gamma_{k.1}, \ldots, \gamma_{k.(j-1)}, \gamma_{k.(j+1)}, \ldots, \gamma_{k.p})$. See George and McCulloch (1993) for details of the derivation of (16). The following algorithm shows how SSVS is embedded into the general approach in Algorithm 1.

Algorithm 3. (**Componentwise Variable Selection in FMR using SSVS**)

Start with a random allocation $\boldsymbol{z}^{(0)}$ and repeat the following steps for $t = 1, \ldots, M$:

1. Sampling $\boldsymbol{\psi}^{(t)}$ conditional on $\boldsymbol{z}^{(t-1)}$:
   (a) Generate $\boldsymbol{\omega}^{(t)}$ from $\mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$.
   (b) Generate $\sigma_k^{2(t)}$ from $P(\sigma_k^2 \mid \boldsymbol{\beta}_k^{(t-1)}, \boldsymbol{\gamma}_k^{(t-1)}, X_k^{(t-1)}, \boldsymbol{y}_k^{(t-1)})$ given in (16) for $k = 1, \ldots, K$.
   (c) Generate $\boldsymbol{\beta}_k^{(t)}$ from $P(\boldsymbol{\beta}_k \mid \boldsymbol{\gamma}_k^{(t-1)}, \sigma_k^{2(t)}, X_k^{(t-1)}, \boldsymbol{y}_k^{(t-1)})$ given in (16) for $k = 1, \ldots, K$.
   (d) Generate $\gamma_{kj}^{(t)}$ from $P(\gamma_{kj} \mid \boldsymbol{\gamma}_{k(-j)}^{(t)}, \boldsymbol{\beta}_k^{(t)}, \sigma_k^{2(t)})$ given in (16) for $j = 1, \ldots, p$ and $k = 1, \ldots, K$.

   Now we have $\boldsymbol{\psi}^{(t)} = (\boldsymbol{\omega}^{(t)}, \boldsymbol{\beta}_1^{(t)}, \ldots, \boldsymbol{\beta}_K^{(t)}, \sigma_1^{2(t)}, \ldots, \sigma_K^{2(t)}, \boldsymbol{\gamma}_1^{(t)}, \ldots, \boldsymbol{\gamma}_K^{(t)})$.
2. Sampling $\boldsymbol{z}^{(t)}$ conditional on $\boldsymbol{\psi}^{(t)}$:
   Sample $z_i^{(t)}$ from $P(z_i|y_i, \boldsymbol{x}_i, \boldsymbol{\psi}^{(t)})$ given in the following for $i = 1, \ldots, n$,

$$\begin{aligned} &P(z_i = k \mid y_i, \boldsymbol{x}_i, \boldsymbol{\psi}^{(t)}) \\ &= \frac{\mathcal{N}(y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}_k^{(t)}, \sigma_k^{2(t)})\,\omega_k^{(t)}}{\sum_{j=1}^K \mathcal{N}(y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}_j^{(t)}, \sigma_j^{2(t)})\,\omega_j^{(t)}}, \end{aligned}$$

   for $k = 1, \ldots, K$.

We have demonstrated how different variable selection techniques can be embedded in the general Algorithm 1 to accomplish componentwise variable selection in FMR. It would be helpful if we can evaluate how all these methods perform in the following simulation studies. We conducted a preliminary study to compare g-prior versus SSVS and found their performance was very close. It would be valuable if an extensive simulation study could be conducted to carefully compare their performance under various scenarios. However, running many scenarios for each of these methods would require excessive amount of coding effort and simulation work. For this reason, only the Gibbs sampling with g-prior as in Algorithm 2 was selected to be implemented in the simulation studies discussed in the next section. Comparing these techniques or even incorporating more techniques could be an interesting topic of future research.

The Gibbs sampling with g-prior as in Algorithm 2 is selected for the simulation studies for the following reasons. First, g-prior has a great number of successful applications. The Gibbs sampling is usually the search algorithm accompanying g-prior. Second, g-prior only requires the specification of g to set up for $\boldsymbol{\beta}$ and σ^2. Other techniques would need either tuning or more specifications for hyperparameters. Third, the $(X^{'}X)^{-1}$ term in the g-prior takes into account collinearity in the design matrix, which is a common problem in regression.

3. SIMULATION STUDY

In the previous section we proposed a Bayesian approach to componentwise variable selection in FMR and demonstrated how specific MCMC algorithms are embedded. In practice, there are many factors that would affect the performance of the approach. For instance, if the sample size is small relative to the number of parameters, then the posterior distribution could be multimodal and have large variance. In this section we investigate the performance of the proposed method under various scenarios via simulations.

3.1 Simulation design

Our proposed method is found to have reliable performance on simpler models with fewer components and explanatory variables (e.g., $K = 2$ and $p = 2$) in the preliminary study (not shown here). In this section, we only considered challenging cases with more components and explanatory variables. We simulated a FMR model with four components, each of which has a different setting of explanatory variables. Table 1 shows the true values of the coefficients of each component in the simulated model.

In addition, the following aspects were considered to create a wide set of simulation scenarios:

- Component weight $\boldsymbol{\omega}$. Even and uneven weights were considered. The even weight is $\boldsymbol{\omega} = (.25, .25, .25, .25)$ and the uneven one is $\boldsymbol{\omega} = (.3, .3, .3, .1)$.

Table 1. Coefficient parameters of simulated data

	$(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$
Component 1	(.3, 1, 0, 0, 3, 0)
Component 2	(.8, -4, 2, 0, 0, 3)
Component 3	(.8, -2, 1, 0, 2, 1)
Component 4	(1, 2, 0, 0, -3, 4)

Table 2. Various considerations for σ_k^2, ρ, ω, and n

	Specifications
σ_k^2	0.5, 1
ρ	0, 0.5, 0.7
ω	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $(\frac{3}{10}, \frac{3}{10}, \frac{3}{10}, \frac{1}{10})$
n	300, 600, 900

- Sample size n. Three levels of sample size were set up such that the smallest component under uneven component weight is guaranteed to have at least 30 observations.
- Variance of the error term σ_k^2. We specified two levels of variance: $\sigma_k^2 = 0.5$ and $\sigma_k^2 = 1$.
- Collinearity between variables. The explanatory variables were generated from a multivariate normal distribution $\mathcal{N}(\mathbf{0}, I)$. We followed Khalili and Chen (2007) to assign an autoregressive type of correlation between x_j's, that is, $Cov(x_i, x_j) = \rho^{|i-j|}$, where ρ controls the degree of correlation. Three levels of correlation were specified: $\rho = 0$, $\rho = 0.5$, and $\rho = 0.7$.

Table 2 summarizes the specifications of data generation by taking the above discussion into account. As a result, there are 36 simulation scenarios having different levels of noise and difficulty.

The results from the main simulations to assess performance of our proposed method will be shown in Section 3.3. Before that, we would also like to investigate other important issues such as selecting appropriate hyperparameters and determining the number of components. It is computationally costly and unnecessary to look into these issues on all the scenario combinations described in Table 2. Rather, we selected some scenarios representing various levels of difficulty as shown in Table 3: Scenario 1 is the easiest scenario, Scenario 2 has medium level of noise, and Scenarios 3–5 are very challenging ones with high levels of noise. Those selected scenarios are biased high-level noises in order for a stress test: if they behave well then we are confident that the remaining scenarios will also behave well.

All the simulations were programmed in the statistical package R. The computing time for analyzing 100 datasets for each scenario ranged from 16 to 21 hours on a 2.2 GHz dual core processor with 2 GB RAM.

Table 3. Short list of simulation scenarios

Scenario 1	$\sigma_k^2 = 0.5$, $\rho = 0$, $\omega = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $n = 600$
Scenario 2	$\sigma_k^2 = 0.5$, $\rho = 0.5$, $\omega = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $n = 300$
Scenario 3	$\sigma_k^2 = 1$, $\rho = 0.5$, $\omega = (\frac{3}{10}, \frac{3}{10}, \frac{3}{10}, \frac{1}{10})$, $n = 300$
Scenario 4	$\sigma_k^2 = 1$, $\rho = 0.7$, $\omega = (\frac{3}{10}, \frac{3}{10}, \frac{3}{10}, \frac{1}{10})$, $n = 600$
Scenario 5	$\sigma_k^2 = 1$, $\rho = 0.7$, $\omega = (\frac{3}{10}, \frac{3}{10}, \frac{3}{10}, \frac{1}{10})$, $n = 300$

3.2 Choice of priors

In order to run Algorithm 2, we set up the priors as follow: the model prior or the prior for inclusion indicators $P(\boldsymbol{\gamma})$, the scalar hyperparameter g in the g-prior, and the hyperparameter $\boldsymbol{\alpha}$ in the Dirichlet distribution for $\boldsymbol{\omega}$.

A popular setup for the model prior is an independent Bernoulli distribution on each γ_j such that

$$P(\boldsymbol{\gamma}) \propto \prod_{j=1}^{p} \pi^{\gamma_j}(1-\pi)^{1-\gamma_j},$$

where $\pi = P(\gamma = 1)$ is the prior inclusion probability of each covariate. The common choice is $\pi = 0.5$, that is, each covariate has a 50% chance of being included, corresponding to the popular uniform prior $P(\boldsymbol{\gamma}) = 0.5^p$. One can also assign other values rather than 0.5 to π. Another more complicated method is fully Bayesian, that is, putting a hyper-prior on π, for example, $\pi \sim Beta(a, b)$ (Brown et al., 1998 and Ley and Steel, 2009).

Eicher et al. (2011) conclude that the uniform prior with $\pi = 0.5$ outperforms the other values based on a series of simulation studies where the predictive performance is evaluated. Ley and Steel (2009) conduct a more comprehensive study comparing the fixed-π and the hyper-π model priors. The results show that $\pi = 0.5$ with $g = p^2$ has fairly comparable performance to other hyper-π priors. Based on the results of the two studies, we only consider $\pi = 0.5$ in our simulation studies.

An appeal of using g-prior is that the priors for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and σ^2 are set automatically if g is chosen, while the choice of g is the important question, which has attracted many studies in the literature. These studies can be classified into three types of methods: fixed g (e.g., Kass and Wasserman, 1995, Foster and George, 1994 and Fernández et al., 2001), empirical Bayes (George and Foster, 2000), and full Bayes (e.g., Liang et al., 2008, Ley and Steel, 2011 and Cui and George, 2008).

There are some comparison studies on the choice of g in the literature. Fernández et al. (2001) recommend $g = n$ and $g = \sqrt{n/p}$ as well as their benchmark prior $g = \max(n, p^2)$ after comparing a list of fixed g values. A similar study by Eicher et al. (2011) finds that $g = n$ outperforms other values of g in terms of predictive performance. A recent study by Ley and Steel (2011) compares most of the methods for specifying g discussed above. The results from their studies indicate that $g = n$ performs equally well with and in some scenarios better than other sophisticated methods. Based on these findings, we will only consider fixed values of g in our simulation studies.

Now we discuss the prior for the weight parameter $\boldsymbol{\omega}$. As a standard choice, we adopt the conjugate prior for $\boldsymbol{\omega}$ is a Dirichlet distribution $\mathcal{D}(\alpha_1, \ldots, \alpha_K)$, which leads to the posterior distribution $\mathcal{D}(\alpha_1 + n_1, \ldots, \alpha_K + n_K)$. Obviously, if we have no subjective knowledge of the component weights, it is sensible to take $\alpha_k = \alpha$. Now the question is how to choose a value for α. A common practice is to take $\alpha = 1$. Many studies (e.g., Nobile, 2004) warn that with such small α, there is a possibility of sampling small ω_k (for the reason just shown above), which in turn causes an empty component because the components with large ω_k absorb all the observations. However, this is not the case in the problem considered in this paper because $\boldsymbol{\omega}$ is drawn from the posterior distribution $\mathcal{D}(\alpha+n_1, \ldots, \alpha+n_K)$, where $\alpha + n_k \gg 1$ no matter how small α is. Apart from $\alpha = 1$, other specifications seem arbitrary in the literature. For example, $\alpha = 4$ by Frühwirth-Schnatter (2006, p. 105), and $\alpha = 5$ or 10 by Norets and Pelenis (2009). We also consider another specification based on the average sample size, i.e., $\alpha = \frac{N}{K}$.

We further conducted a sensitivity study to evaluate the impact of g and α, and determined reasonable values for the subsequent simulation studies. Following the discussion above, we narrow down the choices of g and α as follows:

Table 4. Choices of g and α

g	n, p^2, $\max\{n, p^2\}$
α	1, 4, n/K

These choices result in nine settings. For each setting, we generated 100 datasets and implemented the proposed method in Algorithm 2 under each of the scenarios listed in Table 3. The results are shown in Table 5, which report 95% credible intervals of the correction rate of variable selection (see the definitions in Section 3.3).

We can easily see that choices of g and α have no material difference on the inference of variable selection, although the setting of $(\alpha = 1, g = n)$ seems more stable than other choices and therefore will be used in the subsequent simulations.

Before showing the simulation results, we need to answer another key question in finite mixture model. In the proposed approach in Section 2 the number of components, K, is known *a priori*. However, in real-world analysis we have no knowledge of K beforehand and have to estimate it from the data. If we fail to recover the true K in the first place, the subsequent clustering and variable selection are meaningless. Finding an optimal choice of K is an inevitable step in full analysis of finite mixture modeling. Estimating K is a big topic and has been attracting numerous studies in Bayesian and frequentist literature. Important Bayesian

Table 5. Comparison of variable selection under different hyperparameter settings. Each interval is the 95% Bayesian credible set of the correction rate of variable selection.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
$\alpha = 1$, $g = n$	(.95,1)	(1,1)	(.90,1)	(.95,1)	(.90,1)
$\alpha = 1$, $g = p^2$	(1,1)	(.95,1)	(.90,1)	(.95,1)	(.90,1)
$\alpha = 1$, $g = max\{n, p^2\}$	(.95,1)	(.95,1)	(.92,1)	(.95,1)	(.90,1)
$\alpha = 4$, $g = n$	(.95,1)	(.95,1)	(.90,1)	(.95,1)	(.82,1)
$\alpha = 4$, $g = p^2$	(.95,1)	(.95,1)	(.85,1)	(.92,1)	(.85,1)
$\alpha = 4$, $g = max\{n, p^2\}$	(.95,1)	(.95,1)	(.95,1)	(.95,1)	(.90,1)
$\alpha = n/K$, $g = n$	(.95,1)	(.95,1)	(.95,1)	(.92,1)	(.90,1)
$\alpha = n/K$, $g = p^2$	(1,1)	(1,1)	(.90,1)	(.95,1)	(.85,1)
$\alpha = n/K$, $g = max\{n, p^2\}$	(.97,1)	(.95,1)	(.95,1)	(.95,1)	(.87,1)

Table 6. Recovery rate of number of components

	BIC	AIC	DIC	Marginal	PPMS
Scenario 1	1	1	1	1	.02
Scenario 2	.98	.98	.68	.95	.52
Scenario 3	.90	.96	.08	.62	.58
Scenario 4	1	.80	.04	.70	.16
Scenario 5	.60	.92	.02	.52	.62

Table 7. Accuracy of variable Selection and clustering when $\sigma^2 = 0.5$

	Variable Selection	Clustering
	$\sigma^2 = 0.5$ $\rho = 0$	
$\boldsymbol{\omega} = (.25, .25, .25, .25)$		
$n = 900$	(.95, 1)	(.76, .82)
$n = 600$	(.95, 1)	(.76, .83)
$n = 300$	(.95, 1)	(.75, .82)
$\boldsymbol{\omega} = (.3, .3, .3, .1)$		
$n = 900$	(.95, 1)	(.76, .81)
$n = 600$	(.95, 1)	(.76, .81)
$n = 300$	(.95, 1)	(.75, .83)
	$\sigma^2 = 0.5$ $\rho = 0.5$	
$\boldsymbol{\omega} = (.25, .25, .25, .25)$		
$n = 900$	(.95, 1)	(.75, .80)
$n = 600$	(.99, 1)	(.74, .80)
$n = 300$	(.95, 1)	(.72, .81)
$\boldsymbol{\omega} = (.3, .3, .3, .1)$		
$n = 900$	(.95, 1)	(.74, .79)
$n = 600$	(.99, 1)	(.74, .80)
$n = 300$	(.90, 1)	(.72, .81)
	$\sigma^2 = 0.5$ $\rho = 0.7$	
$\boldsymbol{\omega} = (.25, .25, .25, .25)$		
$n = 900$	(.95, 1)	(.73, .77)
$n = 600$	(.95, 1)	(.72, .78)
$n = 300$	(.99, 1)	(.70, .78)
$\boldsymbol{\omega} = (.3, .3, .3, .1)$		
$n = 900$	(.95, 1)	(.72, .76)
$n = 600$	(.99, 1)	(.71, .76)
$n = 300$	(.95, 1)	(.69, .78)

methods include: information criteria such as AIC (Akaike, 1973) and BIC (Schwarz, 1978), DIC (Spiegelhalter et al., 2002), marginal likelihood (Chib, 1995), posterior predictive model selection (Laud and Ibrahim, 1995), Bayes Factor (Raftery, 1995), model space search approaches (Green, 1995), and so on. Can the true$K$ in our simulated data be recovered with these methods? In fact some of these methods were applied to the simulated data and their performance was compared in our dissertation research (Chen, 2012). The results in Table 6 show that AIC and BIC outperformed other methods.

In this paper we focus on the topic of variable selection, and are not ambitious to elaborate on these methods. Detailed discussion of them as well as a proposed new method can be found in Chen (2012).

3.3 Main results

We now discuss the performance of variable selection and clustering of the proposed method under the scenarios specified above. For each of the 36 scenarios described in Table 2, 100 datasets were generated. The Gibbs sampler described in Algorithm 2 was then run for each dataset to sample posterior distributions of the parameters. The sampler was initialized by randomly assigning observations to components, and was run for 2,500 iterations with the first 1,000 iterations discarded as burn-in.

The main simulation results are given in Tables 7 and 8, which show how well the proposed approach identified the true variables (the column "Variable Selection") and recovered the grouping of the observations (the column "Clustering"). These tables report 95% credible intervals of the the correction rate which is simply:

$$\text{Correction Rate} = \frac{\#\text{ of correctly classified items}}{\text{Total }\#\text{ of items}},$$

where the items could be variables or observations. For variables, $\tilde{\gamma}$, the posterior mean of γ is first calculated for each variable in each component. A variable is determined to be in the model if $\tilde{\gamma} > .5$, and vice versa. The variable selection based on $\tilde{\gamma}$ is compared to the true model specified in Table 1 to calculate the correction rate for variable se-

Table 8. Accuracy of variable selection and clustering when $\sigma^2 = 1$

	Variable Selection	Clustering
	$\sigma^2 = 0.5$ $\rho = 0$	
$\boldsymbol{\omega} = (.25, .25, .25, .25)$		
$n = 900$	(.95, 1)	(.71, .75)
$n = 600$	(.95, 1)	(.70, .76)
$n = 300$	(.95, 1)	(.68, .76)
$\boldsymbol{\omega} = (.3, .3, .3, .1)$		
$n = 900$	(.95, 1)	(.69, .74)
$n = 600$	(.95, 1)	(.69, .75)
$n = 300$	(.90, 1)	(.67, .76)
	$\sigma^2 = 0.5$ $\rho = 0.5$	
$\boldsymbol{\omega} = (.25, .25, .25, .25)$		
$n = 900$	(.95, 1)	(.69, .73)
$n = 600$	(.99, 1)	(.67, .73)
$n = 300$	(.95, 1)	(.65, .74)
$\boldsymbol{\omega} = (.3, .3, .3, .1)$		
$n = 900$	(.95, 1)	(.67, .72)
$n = 600$	(.95, 1)	(.67, .73)
$n = 300$	(.90, 1)	(.63, .74)
	$\sigma^2 = 0.5$ $\rho = 0.7$	
$\boldsymbol{\omega} = (.25, .25, .25, .25)$		
$n = 900$	(.95, 1)	(.65, .70)
$n = 600$	(.95, 1)	(.65, .71)
$n = 300$	(.95, 1)	(.61, .71)
$\boldsymbol{\omega} = (.3, .3, .3, .1)$		
$n = 900$	(.95, 1)	(.65, .70)
$n = 600$	(.95, 1)	(.62, .70)
$n = 300$	(.90, 1)	(.60, .70)

lection. For instance, in Component 1 the true explanatory variables are x_1 and x_4. If $\tilde{\gamma}$ concludes that x_1, x_2, and x_4 are selected, "# of correctly classified items" equals 4, and the Correction Rate is $\frac{4}{5}$. To calculate the rate of observation clustering, the posterior mean of $\boldsymbol{z}$ ($\tilde{\boldsymbol{z}}$) and the majority rule (which allocates an observation to the component with the maximum membership probability) is used to determine which component an observation is allocated to. For instance, if $\tilde{\boldsymbol{z}} = (.05, .20, .60, .15)$ then that observation is allocated to Component 3. By counting how many observations are correctly allocated, we can calculate the correction rate of clustering observations.

In the followings are the findings from the simulation studies.

1. As the results show, the performance of componentwise variable selection is quite successful in general. For most of the scenarios, even the lower bound of the 95% credible interval of the correction rate is about .90, which means in worse cases only about 2–3 variables are misidentified out of total 20 variables (5 variables in each component).
2. When $\sigma^2 = 0.5$ (small level of noise), sample size, collinearity in variables and component weights do not have much impact on the performance of componentwise variable selection. However, when σ^2 is increased to 1, sample size and component weights have combined influence. When both are stressed, performance of variable selection is slightly poorer. For example, as in Table 8, when $n = 300$ and $\omega = (.3, .3, .3, .1)$, the correction rates are comparatively lower than those from easier situations.
3. The clustering of observations has mixed performance. In easy scenarios, the correction rate of clustering is about 0.70–0.80, while in the tough scenarios, the rate is about 0.60–0.70. To explain why clusterings are not recovered as well as variable selection, let us look at membership estimation in Step 2 of Algorithm 2. The allocation $\boldsymbol{z}$ is sampled from a multinomial distribution given by

$$P(z_i = k|y_i, \boldsymbol{x}_i, \boldsymbol{\psi}) = \frac{\mathcal{N}\left(y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}_k, \sigma_k^2\right)\omega_k}{\sum_{j=1}^{K}\mathcal{N}\left(y_i \mid \boldsymbol{x}_i'\boldsymbol{\beta}_j, \sigma_j^2\right)\omega_j}, \quad k = 1, \ldots, K, \tag{17}$$

which implies that clustering is influenced by three sources of variance: sample variance (i.e., the data itself), variance in parameter estimation, and randomness of multinomial allocation. For demonstration, we take an example of 30 observations from the 4th component that were generated under Scenario 5 during the simulation study. Now we calculate the membership probability defined in (17) using the true parameters (i.e., no variance in parameter estimation is included) and apply the majority rule to allocate observations. The following histogram in Figure 1 describes how the observations generated from the 4th component are allocated to the four components. We can see that there are a considerable number of observa-

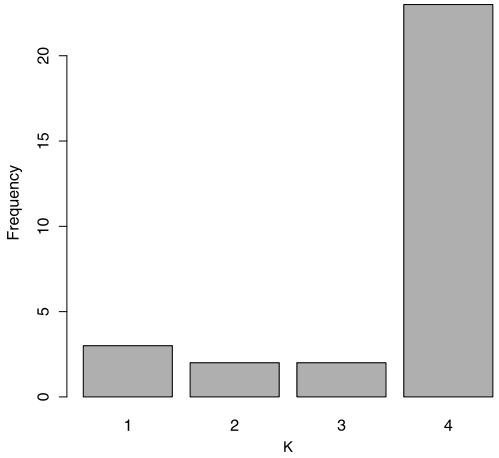


Figure 1. Allocation of observations from 4th component in Scenario 5. The histogram tells how many observations are allocated to each component given the true parameters.

Table 9. Accuracy of parameter estimation

	Scenario 1				Scenario 2				Scenario 3				Scenario 4				Scenario 5			
Component	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
									Recovery Rate											
β_0	.94	.92	.96	.92	.88	.96	1	.96	.92	1	.96	.86	.84	.98	.88	.92	.74	.94	.96	.72
β_1	.90	.88	1	.96	.96	.86	.98	.96	.86	.88	.92	.86	.94	.90	.94	.76	.80	.82	.88	.80
β_2	-	.94	.94	-	-	1	1	-	-	1	.94	-	-	.86	.86	-	-	.94	.94	-
β_3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
β_4	.86	-	.90	.92	.86	-	1	.92	1	-	.98	.82	.96	-	.92	.66	.94	-	.92	.60
β_5	-	.96	.92	.98	-	.90	.96	.88	-	.98	.96	.70	-	.92	.94	.70	-	.92	.94	.66
σ^2	.64	.02	.42	.00	.12	.00	.06	.00	.72	.24	.75	.10	.90	.78	.86	.42	.68	.92	.80	.20
$\boldsymbol{\omega}$	1	1	1	1	1	1	1	1	1	1	1	1	.98	.98	.86	.00	.96	1	1	.00
									RMSE											
β_0	.10	.07	.08	.09	.16	.14	.10	.13	.15	.17	.17	.50	.17	.10	.14	.17	.21	.20	.19	.53
β_1	.07	.09	.07	.07	.08	.15	.12	.10	.20	.19	.21	.50	.12	.14	.14	.33	.29	.23	.24	.64
β_2	-	.07	.09	-	-	.11	.10	-	-	.14	.18	-	-	.15	.14	-	-	.24	.28	-
β_3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
β_4	.12	-	.09	.06	.14	-	.11	.13	.13	-	.16	1	.10	-	.13	.40	.17	-	.20	.96
β_5	-	.06	.08	.07	-	.13	.12	.14	-	.15	.15	.85	-	.13	.14	.37	-	.19	.27	.80
σ^2	.15	.27	.16	.28	.33	.52	.38	.44	.40	.56	.43	.96	.19	.23	.20	.46	.42	.53	.45	.97
$\boldsymbol{\omega}$	.01	.01	.01	.01	.01	.01	.01	.01	.03	.03	.03	.08	.03	.02	.03	.07	.03	.03	.03	.08

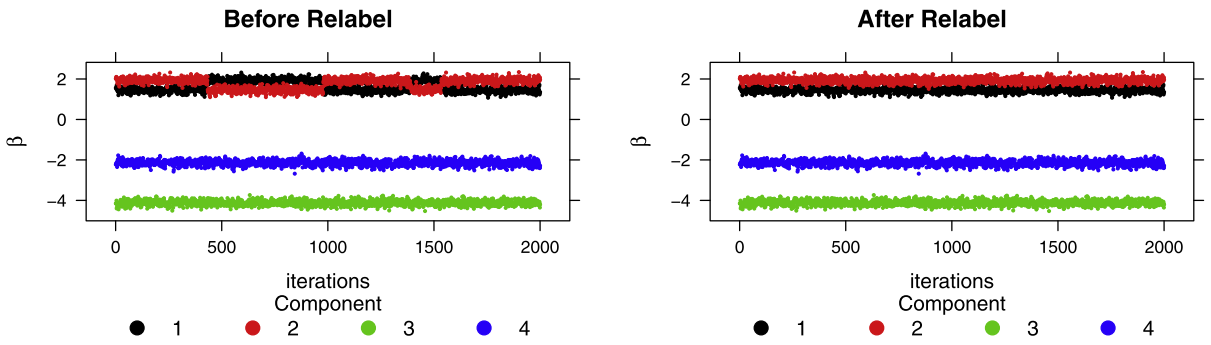
tions which are "mis-allocated" to the first three components even though only sample variance is present. When variance in the parameter estimation and randomness of multinomial allocation are both present, there will be more misclassifications, which explain the poor performance on clustering. It is worth pointing out that "misclassification" may not be an appropriate word when parameter estimation behaves well since in this case all three sources of variances are normal fluctuations in data and in the process. However, in the smallest component in Scenario 5, clustering is vulnerable since the proportion of misclassified observations from other components is fairly large due to its own small sample size. The large portion of foreign observations in return deteriorates parameter estimation and variable selection. We call this phenomenon "invasion disturbance". As shown in Table 8, the correction rate of variable selection is comparatively low for scenarios with $n = 300$ and $\omega = (0.3, 0.3, 0.3, 0.1)$. The poorer performance of parameter estimation in this situation can be found in Table 9. Performance deterioration due to mutual influence of clustering and parameter estimation for small components is inherent to the Gibbs sampling, which is a disadvantage of this approach. Resolving this shortcoming could be a topic for further research.

Parameter estimation is also evaluated for the scenarios in Table 3. For each simulation (100 datasets were generated for each scenario as stated above) the root mean square error (RMSE) and the 95% credible interval of $\boldsymbol{\beta}$, σ^2, and $\boldsymbol{\omega}$ are calculated from the Monte Carlo samples. Suppose $\theta^{(t)}$, $t = 1, \ldots, M$, are the Monte Carlo samples,

$$RMSE = \sqrt{\frac{\sum_{t=1}^{M} \left(\theta^{(t)} - \theta\right)^2}{M}},$$

where θ is the true value of the parameter that was used to simulate the data. The average of RMSE and the percentage that the true parameters are covered in the 95% credible interval (Recovery Rate) are then calculated over 100 simulations and are reported in Table 9. Here are some findings from Table 9:

1. The estimation of regression coefficients is accurate when the weights are balanced. Recovery rates are close to .95 for most of the coefficients and to 0.90 for a few of them under Scenarios 1 and 2. When the weights are unbalanced as in Scenarios 3 to 5, the estimation for the first three components are still satisfactory, but the small components have comparatively poor recovery rates. Especially, when sample size is small (Scenarios 3 and 5), the RMSE for the small component is much higher, which could be explained as stated above with Figure 1.
2. For the larger components, the estimation of the weight parameters is very accurate with almost perfect recovery rates. However, under the unbalanced weights $\boldsymbol{\omega} = (0.3, 0.3, 0.3, 0.1)$ (Scenarios 3 to 5), the weight for the small component is overestimated due to "invasion disturbance" explained in the above.
3. The performance for σ^2 has a different pattern. In general, the performance of σ^2 is not as good as that of the coefficients and the weights. The estimation deteriorates due to misclassifications. For Scenarios 1 and 2 with $\sigma^2 = 0.5$, the posterior mean of σ^2 is in the range of 0.8 and 1.2, while for Scenarios 3 to 5 with $\sigma^2 = 1$,



(a) Before Relabeling (b) After Relabeling

Figure 2. Label switch problem and effect of using relabeling algorithm.

the posterior mean was between 1.2 and 1.5. This explains why Scenarios 1 and 2 have worse recovery rates than Scenario 3 to 5 even though the former has less noise.

We have shown and discussed the performance of variable selection, clustering of observations, and parameter estimation of the proposed approach. In addition to these, an assessment of the model fit and the MCMC convergence can be found in Chen (2012).

Lastly, we want to note an extra step we implemented in the simulation studies. A notorious problem with finite mixture models is the nonidentifiability of the components and label switching in MCMC output. Label switching refers to the invariance of the likelihood under relabeling of the mixture components. As a result, during MCMC the posterior distribution can be highly symmetric and multimodal, making it meaningless to draw inference by summarizing posterior distributions. There exist many methods to resolve this problem. A classic relabeling algorithm created by Stephens (1997) has many successful applications (see, for example, Tadesse et al. (2005), Farrar (2006), and Tatarinova et al. (2008)). See Appendix A for details of the algorithm. This algorithm was adopted in our simulation studies to relabel MCMC output for meaningful posterior inference, and successfully resolved the label switching issues in almost all the scenarios. The graph in Figure 2 contains MCMC samples of a coefficient whose values in the first (black) and the second (red) component are very close. The graph demonstrates how the samples switched to each other's component during iterations and how the algorithm by Stephens (1997) successfully relabeled their membership so as to make posterior summaries meaningful.

4. REAL DATA APPLICATION

We now analyze a high-dimension real dataset in bioinformatics. Living beings depend on genes, as they specify all structures and functions of an organism through gene expression, by which information from a gene is turned into functional gene products (often proteins). Gene regulation refers to the processes that a cell uses to regulate gene expression. Knowledge of gene regulation is of fundamental importance for understanding biological processes within a cell. It is believed that a large proportion of gene regulation occurs at the transcription step, i.e., transcriptional regulation, which controls when transcription occurs and how much RNA is copied. Transcriptional regulation is known to be realized through the binding of transcription factors (TF) to specific DNA sequences (motif) that is usually located in the upstream of a gene (see the demonstration in Figure 3). Identification of binding sites or motifs of a specific TF for a certain biological process, or called motif discovery, is crucial for us understanding gene regulation.

Motif discovery attracts numerous research from many fields including statistics. One of the statistical approaches is regression models, which associate expression level of genes with candidate motifs. Various variable selection techniques could then be applied to pinpoint relevant motifs. For example, Bühlmann and Hothorn (2010) use twin boosting to select motifs. Other important studies of this type include Bussemaker et al. (2001), Conlon et al. (2003), Tadesse et al. (2004), and Zhang et al. (2008).

The regular regression model assumes that all the genes are triggered by the same set of TFs. This assumption might be challenged by the possibility of heterogeneity in the regulation process. That is, there might exist different groups of genes, each of which is regulated by a specific set of TFs. The mixture regression with componentwise variable selection offers a tool to accommodate heterogeneity in gene regulation.

The data we used is originally from a yeast cell cycle experiment (Spellman et al., 1998). There were two groups of yeast cells: the treatment and the control group, both of which were in a glucose solution. An alpha factor was added to the experiment group, but not to the control group. After a certain period of time, samples were taken and their gene expression as a function of microarry motif scores were recorded. The response variable is the log expression ratios (treatment vs. control). The dataset we use is available at ftp://ftp.stat.math.ethz.ch/Manuscripts/buhlmann/motif-

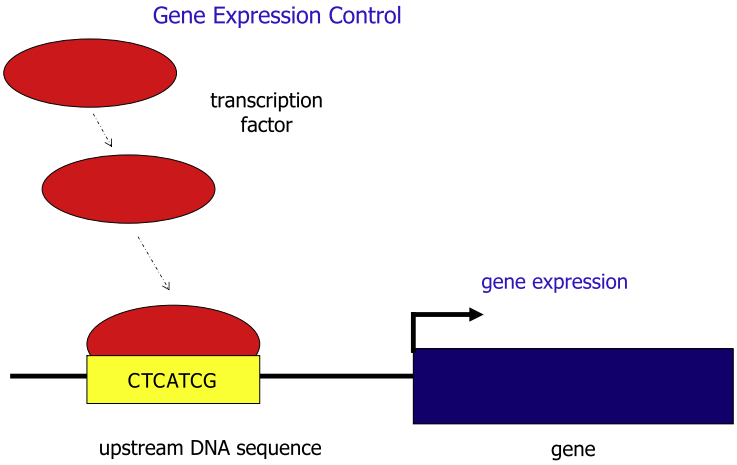


Figure 3. Transcription factor and motif in gene expression regulation (Source: http://stat.ethz.ch/events/archive/Ascona_04/Slides/conlon2.pdf).

Table 10. Criteria for finding K for Motif data

	BIC	AIC	DIC	PPMS
$K = 1$	-3804.72	-1984.13	-4675.26	-1632.34
$K = 2$	**-3605.10**	-763.90	-1341.92	-758.05
$K = 3$	-5078.90	-817.10	-1009.18	-753.78
$K = 4$	-6229.87	-547.48	-933.95	-752.76
$K = 5$	-7215.11	**-112.12**	-749.89	-751.31
$K = 6$	-8772.36	-248.78	-438.92	-750.40
$K = 7$	-10798.22	-854.04	**-407.77**	**-749.86**
$K = 8$	-12976.27	-1611.50	-500.53	-750.02
$K = 9$	-15105.65	-2320.28	-529.75	-750.01

spellman.RData. It contains a $n \times p$ matrix, $n = 4{,}443$ genes and $p = 2{,}155$ motif matching scores from candidate motifs. Our purpose is to cluster genes into several groups and identify the important FTs (motifs) for gene expression regulation within each group. This data was also analyzed by Khalili et al. (2011).

Two treatments were conducted before componentwise variable selection. First, because of the high dimension ($p = 2{,}155$) in the original data, the practical infeasibility rises due to extremely heavy computation burden and singularities in matrix calculation. To overcome this difficulty, we ran an initial screening to select 441 variables for further analysis. Second, the criteria were calculated for $K = 1$ to 9 in order to find an optimal number of components. The results in Table 10 show that the criteria except BIC suggest a range $K = 5$ to 7. With additional reviewing of posterior predictive densities we decided that $K = 6$ could be a reasonable estimate. Khalili et al. (2011) chose $K = 3$ instead, but they only tried models up to $K = 4$. Details of the treatments can be found in Chen (2012).

Given the initially selected set of variables and $K = 6$, we continued to perform componentwise variable selection described by Algorithm 2. Our strong motivation is to compare

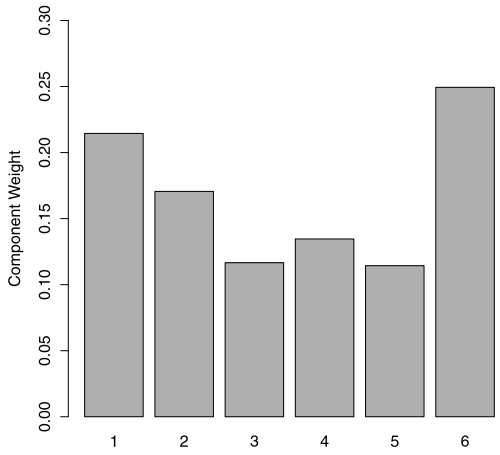


Figure 4. Size of components.

whether the mixture regression model ($K = 6$) would have a different variable selection pattern from the regular regression model ($K = 1$). For this purpose, we ran Algorithm 2 under $K = 1$ and $K = 6$. Table 11 reports the posterior inclusion probabilities, that is, the posterion mean of the inclusion indicator γ. Only the first forty motifs are included in the table to save space. It could be seen that most initially selected motifs are significant in multiple components. For example, "N1.12.2.1" plays a role in components 1–4. Some motifs have an effect in only one component. For instance, "P1.12.13.4" and "P1.10.10.2" are significant in component 2 and 4, respectively. However, a few variables selected by the non-mixture model ($K = 1$) are significant in none of the components in the mixture model ($K = 6$), for example, the motif "P1.5.5.2". Furthermore, Figure 4 shows the relative size of each component, that is, how many genes are in each component. The results show the components are quite balanced. Every component consists of at least 10% of

Table 11. Posterior inclusion probabilities for motif data (40 motifs out of 443 displayed)

	$K=1$	$K=6$					
Motif		Comp 1	Comp 2	Comp 3	Comp 4	Comp 5	Comp 6
N1.11.15	**1**	.38	.14	**1**	.12	**.53**	.11
P1.12.11	**1**	**.99**	**.66**	.12	.08	.08	**1**
P1.12.14.5	**1**	.18	**1**	**.98**	.27	**.87**	**.99**
N1.10.7.3	**1**	.12	**.60**	**1**	.13	**1**	.25
P1.5.5.2	**1**	.12	.24	.19	.37	.08	.22
N1.7.5.2	**1**	.23	**1**	.40	.10	.18	**.99**
N1.11.2	**1**	**.76**	.18	.10	.22	**1**	.10
N1.7.15.3	**1**	.30	**1**	.15	.18	.10	**1**
N1.7.4.6	**1**	.17	.15	**.52**	**1**	**1**	**.61**
N1.12.6.1	**.99**	**1**	.17	.13	**1**	.10	.13
P1.10.14.3	**.99**	.13	.11	**1**	**1**	.12	.18
N1.11.15.6	**.99**	**1**	.18	**1**	**1**	.11	.15
N1.7.14.4	**.99**	.31	.11	.31	.14	**1**	.18
P1.11.15.3	**1**	.13	.70	.17	**1**	.12	.14
P1.6.2.8	**1**	**.56**	**.86**	**1**	**.98**	.22	.16
N1.12.2.1	**.97**	**1**	**1**	**1**	**1**	.23	.10
N1.6.13.3	**1**	**1**	.27	**1**	.17	.10	**1**
N1.8.15.6	**1**	**1**	.24	**.53**	**.95**	**1**	.16
P1.11.11.1	**.97**	.12	.32	**1**	**.54**	.09	.11
P1.11.7.5	**1**	**.80**	.15	**1**	.14	**1**	**1**
N1.7.11.5	**.97**	**1**	.19	**.84**	.19	.07	.09
P1.12.13.4	**.99**	.12	**.93**	.10	.27	.40	.13
P1.10.10.2	**.99**	.10	.31	.19	**1**	.20	.18
P1.12.5.5	**.98**	.27	.17	.10	**.51**	**1**	.16
N1.11.10.3	**1**	.11	.20	.12	**1**	.12	**1**
N1.6.13.2	**.90**	.12	.17	**1**	.12	**1**	.21
P1.9.11.3	**.92**	.07	**1**	**1**	**1**	.09	.24
N1.5.13.3	**1**	.09	.38	.09	**1**	**1**	.12
P1.5.6.3	**1**	.12	.22	.22	**1**	.08	.15
P1.12.11.4	**.93**	.11	.26	.15	**1**	.07	.09
P1.11.14.1	**.92**	**.99**	.12	.09	**1**	**.85**	.11
P1.6.4	**.99**	.42	.31	**.78**	**1**	.08	.11
P1.11.13	**.96**	**1**	**.83**	.48	.12	.14	.17
N1.10.9.6	**.93**	.11	**.72**	.24	**1**	.25	.21
N1.9.8	**.97**	.28	.18	.19	**.86**	.10	.10
P1.12.3.6	**.95**	.24	.18	.15	**.91**	**1**	**.88**
N1.7.15.8	**.95**	.30	.16	**1**	**1**	**.99**	.13
N1.9.2.7	**.86**	.46	**1**	.15	.17	**1**	.11
N1.11.13.2	**.94**	.11	.10	**.64**	.11	.12	**1**
N1.12.4.8	**.91**	.15	.19	.35	**1**	**1**	.11

the genes. The smallest component (component 5) is about twice the size of the largest component (component 6).

In summary, our analysis on this motif data provides evidence to the hypothesis that there exist different patterns of gene regulation ($K = 6$ is suggestion in our analysis). Furthermore the results from componentwise variable selection suggest that there might exist different gene groups and each group is regulated by a different set of TFs. However, expert knowledge from bioinformatics is needed to justify the interpretation of our results and suggest additional studies for more insightful outcomes.

5. CONCLUSIONS AND DISCUSSIONS

We propose a general approach for componentwise variable selection in FMR, which is essentially a Gibbs sampling scheme with data augmentation that utilizes the latent membership variable. There are two iterative steps in the approach: estimation of parameters for each component and allocation of observations to each component. In the step of estimating parameters, specially designed priors (such as spike-slab prior) or processes (such as Reversible Jump MCMC) can then be used to achieve the purpose of variable selection. Componentwise variable selection is

achieved since parameter estimation is carried out individually in each component. We next illustrate how specific Bayesian variable selection techniques can be embedded in the general approach including g-prior and SSVS. We further choose the g-prior method as an example and proceed with discussing prior settings and assessing performance via a series of simulation studies.

The first part of the simulation studies is a sensitivity study on the priors of the g-prior method. The results show that the three common prior settings have no obvious effect on the performance of variable selection. As the main part, we evaluate the performance of the proposed approach for componentwise variable selection and clustering of observations in various scenarios. The results show excellent performance on variable selection even in challenging scenarios with high-level noise and small sample size. The clustering of observations does not seem to have been recovered very well with correction rates dropping to 60–70% in challenging scenarios, which, however, is caused largely by noise in the data itself. It is worth noting that the simulation studies done so far are for evaluating the g-prior method. The performance of the proposed approach embedded with other variable selection techniques such as RJMCMC is an open question for future study.

As for the Gibbs sampling method, the proposed approach provides fast convergence. For example, the sampler converged to the true state in less than 200 iterations in easy scenarios and in about 500 iterations in challenging scenarios in the simulation studies. Another advantage of the approach is that the user has the freedom to choose a specific variable selection technique. On the other hand, the proposed approach has relatively poor performance for small components in the case of unbalanced component weights (i.e., some components have considerably smaller size than the others). As explained earlier, this problem is caused by invasion of observations from large components into the small components. We thus call it "invasion disturbance". It is important to note that this drawback is different from the common problematic issue with the Gibbs sampling called "trapping states", that is, the Gibbs sampler is stuck in a local maximum and needs an enormous number of iterations to escape from it so as to converge to the true state. Rather "invasion disturbance" could not be cured by running more iterations as long as parameter estimation depends on allocation, which is unfortunately the case in the proposed approach. The Metropolis-Hastings sampling without allocation is a possible resolution, but designing a feasible proposal density is challenging. Solutions to "invasion disturbance" is an interesting topic of future research.

Lastly, we compare our method to Khalili and Chen (2007). They achieved variable selection by maximizing a penalized likelihood function using the EM algorithm. Though apparently this method is frequentist and ours belongs to the Bayesian class, there are some similarities between their method and ours: both rely on the data augmentation framework (Dempster et al., 1977) which includes the latent membership variable $\boldsymbol{z}$ in the iterative process, and both are flexible by allowing choosing from various variable selection techniques. Both are iterative methods. But as a general difference, their method is to search for MLE or maximum a posterior (MAP), while our method generates the entire posterior distribution. Having the posterior distribution gives us the potential to do many things beyond MAP, which might be viewed as an advantage of our method. From our point of view, our proposed approach has two additional advantages. First, our method allows for the latent indicator of variable inclusion $\boldsymbol{\gamma}$ so that variable selection could be treated as less independent of the value of the coefficient. This might be beneficial when a variable has a small coefficient value but its contribution is significant, which sometime is the case when variables are not normalized. Second, our method is ready to have model space search techniques embedded such as Reversible Jump MCMC so that the sampler could moves between mixture models with different K. We can then accomplish the three goals — estimating K, selecting variable, and clustering observations — simultaneously, which is a challenging task to Khalili and Chen (2007). On the other hand, since finding MLE is usually faster than sampling a distribution, their method might converge faster than our method and thus needs less computation. We have just laid out some theoretical comparisons between the two methods. As for simulation studies, we generated more challenging simulated data (4-component mixture versus 2-component mixture in their study). Our method performed satisfactory componentwise variable selection even in the most challenging scenarios, but we are not sure if their method would have comparable performance in the same situation. It will be an interesting topic of future study to look further into theoretical differences in depth and conduct more careful simulation studies to compare the performance between our method and Khalili and Chen (2007) on the same simulated data.

APPENDIX A. LABEL SWITCHING AND RELABELING

Stephens (1997, 2000) develops relabeling techniques based on minimizing the posterior expectation of a loss function. Let $\mathcal{L}(a, \theta)$ be the loss function with the action a and the true parameter θ. In his papers Stephens uses the overall classification matrix as a and the iteration-wise classification probability matrix representing θ. Specifically, let $Q = (q_{ij})$ be an $n \times K$ matrix of overall classification probabilities. Each row of Q represents the probabilities that each observation i is assigned to the K components such that $\sum_{j=1}^{K} q_{ij} = 1$.

We denote the matrix of classification probabilities at each iteration by $P(\boldsymbol{\psi}) = (p_{ij}(\boldsymbol{\psi}))$, where

$$p_{ij}(\boldsymbol{\psi}) = \frac{P\left(y_i|\boldsymbol{\theta}_j\right)\omega_j}{\sum_{k=1}^{K} P\left(y_i|\boldsymbol{\theta}_k\right)\omega_k}.$$

As for the loss function, Stephens (1997) suggests the Kullback-Leibler distance between the "true" distribution corresponding to $P(\boldsymbol{\psi})$ and the distribution corresponding to Q:

$$\mathcal{L}(Q;\boldsymbol{\psi}) = \sum_{i=1}^{n}\sum_{j=1}^{K} p_{ij}(\boldsymbol{\psi}) \log \frac{p_{ij}(\boldsymbol{\psi})}{q_{ij}},$$

which is minimized iteratively until some tolerance is reached as described in the following algorithm.

Algorithm 4. (**Kullback-Leibler Relabeling Algorithm**)

Start with initial permutations of $\{1,\ldots,K\}$ for all iterations $\nu_1,\ldots,\nu_t,\ldots,\nu_M$ and denote by $\nu_t(\boldsymbol{\psi}^{(t)})$ the rearrangement of the parameters at the tth iteration. Iterate the following steps until convergence.

1. Choose $\hat{Q} = (\hat{q}_{ij})$ to minimize the loss

$$\sum_{t=1}^{M} \mathcal{L}\left(Q;\nu_t\left(\hat{\boldsymbol{\psi}}^{(t)}\right)\right) = \sum_{t=1}^{M}\sum_{i=1}^{n}\sum_{j=1}^{K}\{p_{ij}\left(\nu_t\left(\boldsymbol{\psi}^{(t)}\right)\right) \log \frac{p_{ij}\left(\nu_t\left(\boldsymbol{\psi}^{(t)}\right)\right)}{q_{ij}}\}.$$

2. It can be shown that this is achieved by averaging MCMC samples of $p_{ij}(\boldsymbol{\psi})$,

$$\hat{q}_{ij} = \frac{1}{M}\sum_{t=1}^{M} p_{ij}\left(\nu_t\left(\boldsymbol{\psi}^{(t)}\right)\right).$$

3. For $t = 1,\ldots,M$, choose a label permutation ν_t to minimize

$$\mathcal{L}\left(\hat{Q};\nu_t\left(\boldsymbol{\psi}^{(t)}\right)\right) = \sum_{i=1}^{n}\sum_{j=1}^{K} p_{ij}\left(\nu_t\left(\boldsymbol{\psi}^{(t)}\right)\right) \log \frac{p_{ij}\left(\nu_t\left(\boldsymbol{\psi}^{(t)}\right)\right)}{\hat{q}_{ij}}.$$

Received 11 November 2013

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, volume 1, pages 267–281. Springer Verlag. MR0483125

Box, G. and Tiao, G. (1973). *Bayesian inference in statistical analysis*. Addison-Wesley series in behavioral science: quantitative methods. Addison-Wesley Pub. Co. MR0418321

Brown, P., Vannucci, M., and Fearn, T. (1998). Bayesian wavelength selection in multicomponent analysis. *Journal of Chemometrics*, 12(3):173–182.

Bühlmann, P. and Hothorn, T. (2010). Twin boosting: improved feature selection and prediction. *Statistics and Computing*, 20(2):119–138. MR2610767

Bussemaker, H., Li, H., and Siggia, E., et al. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, 27(2):167–171.

Celeux, G., Hurn, M., and Robert, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 957–970. MR1804450

Chen, B. (2012). Bayesian model selection in finite mixture regression. Ph.D. dissertation, University of Texas at San Antonio. MR3130888

Chib, S. (1995). Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 1313–1321. MR1379473

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*, 100(6):3339–3344.

Cui, W. and George, E. (2008). Empirical bayes vs. fully bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900. MR2416869

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38. MR0501537

DeSarbo, W. and Cron, W. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification*, 5(2):249–282. MR0971156

Diebolt, J. and Robert, C. (1990). Bayesian estimation of finite mixture distributions, part i: Theoretical aspects. Technical Report 111, LSTA, Université Paris VI, Paris.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499. MR2060166

Eicher, T., Papageorgiou, C., and Raftery, A. (2011). Default priors and predictive performance in bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, 26(1):30–55. MR2759908

Farrar, D. (2006). Approaches to the label-switching problem of classification, based on partition-space relabeling and label-invariant visualization. Technical report, Citeseer.

Fernández, C., Ley, E., and Steel, M. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics*, 16(5):563–576.

Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 1947–1975. MR1329177

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Verlag. MR2265601

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2009). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*, 2(3):1360–1383.

George, E. and Foster, D. (2000). Calibration and empirical bayes variable selection. *Biometrika*, 87(4):731–747. MR1813972

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88:881–889.

Green, P. (1995). Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82(4):711. MR1380810

Gupta, M. and Ibrahim, J. (2007). Variable selection in regression mixture modeling for the discovery of gene regulatory networks. *Journal of the American Statistical Association*, 102(479):867–880. MR2411650

Hartigan, J. (1977). Distribution problems in clustering. *Classification and Clustering*, 45–72.

Hoff, P. (2009). *A first course in bayesian statistical methods*. Springer Verlag. MR2648134

Hurn, M., Justel, A., and Robert, C. (2003). Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics*, 12(1):55–79. MR1977206

Kass, R. and Wasserman, L. (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 928–934. MR1354008

Khalili, A. and Chen, J. (2007). Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038. MR2411662

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(1):156–172.

Laud, P. and Ibrahim, J. (1995). Predictive model selection. *Journal of the Royal Statistical Society. Series B (Methodological)*, 247–262. MR1325389

Ley, E. and Steel, M. (2009). On the effect of prior assumptions in bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674. MR2675199

Ley, E. and Steel, M. (2011). Mixtures of g-priors for bayesian model averaging with economic application. MR2991863

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423. MR2420243

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA. MR0214227

Marin, J.-M. and Robert, C. P. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*. Springer, 1st edition. MR2289769

McLachlan, G. and Peel, D. (2000). *Finite mixture models*, volume 299. Wiley-Interscience. MR1789474

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032. MR0997578

Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *The Annals of Statistics*, 32(5):2044–2073. MR2102502

Norets, A. and Pelenis, J. (2009). Bayesian modeling of joint and conditional distributions. *Unpublished manuscript*, Princeton Univ.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103:681–686. MR2524001

Quandt, R. (1972). A new approach to estimating switching regressions. *Journal of the American Statistical Association*, 306–310.

Quandt, R. and Ramsey, J. (1978). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association*, 730–738. MR0521324

Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–164.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464. MR0468014

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297.

Spiegelhalter, D., Best, N., Carlin, B., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639. MR1979380

Stephens, M. (1997). *Bayesian methods for mixtures of normal distributions*. PhD thesis, University of Oxford.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809. MR1796293

Tadesse, M., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617. MR2160563

Tadesse, M., Vannucci, M., and Liò, P. (2004). Identification of dna regulatory motifs using bayesian variable selection. *Bioinformatics*, 20(16):2553–2561.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, pages 528–540. MR0898357

Tatarinova, T., Bouck, J., and Schumitzky, A. (2008). Kullback-leibler markov chain monte carlo–a new algorithm for finite mixture analysis and its application to gene expression data. *Journal of Bioinformatics and Computational Biology*, 6(4):727–746.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Ser. B*, 58:267–288. MR1379242

Wang, P., Puterman, M., Cockburn, I., and Le, N. (1996). Mixed poisson regression models with covariate dependent rates. *Biometrics*, pages 381–400.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g-prior distributions. In Goel, P. K. and Zellner, A., editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti*, volume 6 of *Studis in Bayesian Econometrics and Statistics*, pages 233–243. North-Holland, Amsterdam. MR0881437

Zhang, N., Wildermuth, M., and Speed, T. (2008). Transcription factor binding site prediction with multivariate gene expression data. *The Annals of Applied Statistics*, pages 332–365. MR2415606

Bin Chen
Federal Home Loan Bank of Dallas
Irving, TX
USA
E-mail address: chenbin_osu@yahoo.com

Keying Ye
The University of Texas at San Antonio
San Antonio, TX
USA