Computational Statistics and Data Analysis 95 (2016) 1–16

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian variable selection for finite mixture model of linear regressions

Kuo-Jung Lee[a], Ray-Bing Chen[a,*], Ying Nian Wu[b]

[a] *Department of Statistics, National Cheng Kung University, Taiwan*

[b] *Department of Statistics, University of California, Los Angeles, United States*

ARTICLE INFO

Article history:
Received 16 July 2014
Received in revised form 11 September 2015
Accepted 12 September 2015
Available online 9 October 2015

Keywords:
Gibbs sampler
Median probability criterion
Sparsity
Stochastic search variable selection

ABSTRACT

We propose a Bayesian variable selection method for fitting the finite mixture model of linear regressions. The model assumes that the observations come from a heterogeneous population which is a mixture of a finite number of sub-populations. Within each sub-population, the response variable can be explained by a linear regression on the predictor variables. If the number of predictor variables is large, it is assumed that only a small subset of variables are important for explaining the response variable. It is further assumed that for different sub-populations, different subsets of variables may be needed to explain the response variable. This gives rise to a complex variable selection problem. We propose to solve this problem within the Bayesian framework where we introduce two sets of latent variables. The first set of latent variables are membership indicators of the observations, indicating which sub-population each observation comes from. The second set of latent variables are inclusion/exclusion indicators for the predictor variables, indicating whether or not a variable is included in the regression model of a sub-population. Variable selection can then be accomplished by sampling from the posterior distributions of the indicators as well as the coefficients of the selected variables. We conduct simulation studies to demonstrate that the proposed method performs well in comparison with existing methods. We also analyze a real data set to further illustrate the usefulness of the proposed method.

1. Introduction

Variable selection is a fundamental problem in linear regression and has become increasingly important for many modern applications. During the past decade, a rich literature has been developed around this problem, especially for the case where large numbers of variables are collected and the number of variables exceeds the number of observations. The methods proposed for the problem of variable selection can be roughly classified into two categories.

One category consists of various penalized least squares methods, including the famous Lasso method (Tibshirani, 1996) based on the convex ℓ_1 penalty for regularization, as well as non-convex penalties such as SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). The Lasso approach has also been extended to more sophisticated forms such as the group Lasso and graphical Lasso; see (Tibshirani, 2011) for a review.

The other category consists of various Bayesian variable selection methods, such as stochastic search variable selection (SSVS) (George and McCulloch, 1993; Chen, 2012) and Bayesian Lasso (Park and Casella, 2008). The Bayesian method assumes

[*] Corresponding author.
E-mail addresses: kjlee@stat.ncku.edu.tw (K.-J. Lee), rbchen@stat.ncku.edu.tw (R.-B. Chen), ywu@stat.ucla.edu (Y.N. Wu).

http://dx.doi.org/10.1016/j.csda.2015.09.005

a prior distribution on the regression coefficients. A popular prior is the so-called "spike and slab" prior, which is a mixture of a point mass at zero and a diffused Gaussian distribution. Such a prior distribution assumes a binary indicator for each variable, indicating whether a variable is included in the regression or not. Variable selection can be accomplished by sampling from the posterior distributions of the latent indicators and the posterior distributions of the coefficients of the selected variables.

The two categories of methods are related in that the penalty terms correspond to specific Bayesian prior distributions. The penalized least squares approaches, especially the Lasso and its extensions, usually enjoy a computational advantage since the objective functions are convex and can be easily minimized.

Despite the wide applicability of the linear regression model powered by modern variable selection tools, a single regression model can be inadequate if the data come from a heterogeneous population that consists of a number of different sub-populations with different characteristics. In this situation, it is possible that a separate linear regression model is needed for each sub-population. Moreover, the regression models in different sub-populations may use different subsets of predictor variables (or regressors, covariates) to explain the response variable. If the memberships of the observations are unobserved, then we naturally have a finite mixture model of linear regressions, where each mixture component is a linear regression model with its own subset of predictor variables. This gives rise to a variable selection problem that is more complex than that of a single linear regression model.

To solve the variable selection problem in the mixture model, one may extend the penalized least squares methods to the penalized likelihood of mixture models (Khalili and Chen, 2007; Städler et al., 2010). However, the negative log-likelihood of the mixture model would no longer be convex, causing it to lose one of the most appealing features of the Lasso method and its various extensions. As a result, one can only draw inference based on local minima of the objective functions. This difficulty motivates us to adopt a Bayesian alternative which appears more natural.

Specifically, the Bayesian variable selection method for the mixture model involves two sets of latent variables or indicators. The first set of latent variables are membership indicators associated with the observations, indicating which sub-population each observation comes from. The second set of latent variables are inclusion/exclusion indicators for the variables, where, for each sub-population or mixture component, a binary indicator is associated with each predicator variable, indicating whether or not this variable is included in the linear regression of this mixture component. Variable selection, clustering, and parameter estimation can then be carried out by sampling from the posterior distributions of the indicators and the posterior distributions of the coefficients of the selected variables. Our simulation studies show that the Bayesian method performs well compared with existing methods and that the corresponding Markov chain Monte Carlo (MCMC) algorithm may be capable of escaping the traps of local minima. We also analyze a real data set to further illustrate the usefulness of the proposed method.

The rest of the article is organized as follows. Section 2 presents the finite mixture model of linear regressions and its Bayesian treatment, including the prior specifications and the MCMC algorithm for posterior sampling. Section 3 illustrates our method by simulation studies, where our method is compared with existing methods. Section 4 describes our analysis of a real data set to further illustrate our method. Section 5 discusses implementation issues concerning posterior inference, MCMC sampling and model selection criteria. Finally Section 6 concludes with a brief discussion.

2. Finite mixture model of linear regressions

Let $(y_i, x_i),\ i = 1, \dots, n$, be a data set of n observations that come from a heterogeneous population, where y_i is the response variable of the ith observation, and $x_i = (x_{i1}, \dots, x_{ip})'$ collects the p predictor variables or covariates of the ith observation. We assume that the heterogeneous population consists of M sub-populations or mixture components, and within each sub-population, (y_i, x_i) follows a separate linear regression model. Specifically,

$$y_i|(\rho_m, \beta_m, \sigma_m^2, m = 1, \dots, M) \sim \sum_{m=1}^{M} \rho_m \cdot N\left(x_i'\beta_m, \sigma_m^2\right), \tag{1}$$

where $\rho = (\rho_1, \dots, \rho_M)$ is the proportion vector of the M sub-populations, with $\rho_m \geq 0$ and $\sum_{m=1}^{M} \rho_m = 1$. $\beta_m = (\beta_{m1}, \dots, \beta_{mp})'$ is the coefficient vector for the linear regression in the mth sub-population. σ_m^2 is the corresponding variance of the Gaussian residual errors.

2.1. Two sets of latent variables

As is standard for the mixture model, we introduce a latent variable z_i for each observation i, so that $z_i = m$ indicates that the ith observation comes from the mth sub-population. Thus $P(z_i = m) = \rho_m$, i.e.,

$$z_i \sim \text{Multinomial}(\rho_1, \dots, \rho_M), \quad \text{and} \quad [y_i|z_i = m] \sim N\left(x_i'\beta_m, \sigma_m^2\right).$$

In modern applications of linear regression models, the number of predictors p can be large, and it is often assumed that the coefficient vector is sparse, i.e., only a small number of its components are different from zero. In other words, only a small number of predictor variables are to be included in the regression model. We assume that this is the case with the linear regression models of the sub-populations. We further assume that the sparsity patterns of the coefficient vectors of different sub-populations can be different from each other, i.e., in the linear regression models of different sub-populations, different subsets of variables may be selected to explain the response variable. The goal of inference is to infer the sparse coefficient vectors of the sub-populations while classifying the observations into the sub-populations.

To facilitate variable selection, we introduce a latent vector $r_m = (r_{m1}, \dots, r_{mp})'$ for each sub-population m. r_{mj} is an inclusion/exclusion indicator, so that $\beta_{mj} = 0$ if $r_{mj} = 0$ and $\beta_{mj} \neq 0$ if $r_{mj} = 1$. Let $\beta_m(r_m)$ collect all the non-zero elements of β_m and let $x_i(r_m)$ be the active elements of x_i corresponding to those elements of r_m that are equal to 1. The model in Eq. (1) can then be expressed as

$$y_i|(\rho_m, \beta_m, \sigma_m^2, r_m, m = 1, \dots, M) \sim \sum_{m=1}^{M} \rho_m \cdot N\left(x_i'(r_m)\beta_m(r_m), \sigma_m^2\right).$$

With the two sets of latent variables $\{z_i, i = 1, \dots, n\}$ and $\{r_m = (r_{mj}, \dots, r_{mp}), m = 1, \dots, M\}$, the Bayesian computation can be conveniently conducted. In the following subsections, we shall introduce the prior assumptions first and then present the details of the Bayesian analysis.

2.2. Prior distributions

First, consider the probability vector ρ of mixture proportions. Similar to Viele and Tong (2002), we assume a conjugate Dirichlet prior distribution on ρ,

$$\rho \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_M).$$

In each mixture component of the regression model, the prior distributions of the indicator variables r_{mj} are assumed to be independent Bernoulli(d_{mj}) for $j = 1, \dots, p$. As a result, the joint distribution of $r_m = (r_{m1}, \dots, r_{mp})'$ is

$$\pi(r_m) = \prod_{j=1}^{p} d_{mj}^{r_{mj}} (1 - d_{mj})^{1 - r_{mj}}.$$

Now consider the prior distribution of the coefficient vector β_m. Let Y_m be the response vector consisting of responses in the mth sub-population, and let $X_m(r_m)$ be the corresponding design matrix with rows consisting of $x_i'(r_m)$ where the observation i belongs to the sub-population m. Following Zellner (1996), the prior of $\beta_m(r_m)$ is assumed to be the g-prior which is given by

$$\beta_m(r_m) \sim N\left(\hat{\beta}_m(r_m), g_m\sigma_m^2 \left[X_m'(r_m)X_m(r_m)\right]^{-1}\right),$$

where $\hat{\beta}_m(r_m) = \left[X_m'(r_m)X_m(r_m)\right]^{-1} X_m'(r_m)Y_m$ and g_m is a positive number. To avoid the problem of a singular $X_m'(r_m)X_m(r_m)$, we can replace $\left[X_m'(r_m)X_m(r_m)\right]^{-1}$ by $\left[X_m'(r_m)X_m(r_m) + \lambda_m I\right]^{-1}$, where λ_m is a positive number called the ridge parameter (Gupta and Ibrahim, 2007; Baragatti and Pommeret, 2012), and I is the identity matrix. Thus we have a modified g-prior distribution given as follows:

$$\beta_m(r_m) \sim N\left(\hat{\beta}_m^{\lambda_m}(r_m), g_m\sigma_m^2 \left[X_m'(r_m)X_m(r_m) + \lambda_m I\right]^{-1}\right),$$

where $\hat{\beta}_m^{\lambda_m}(r_m) = (X_m'(r_m)X_m(r_m) + \lambda_m I)^{-1} X_m'(r_m)Y_m = w_m(r_m)X_m'(r_m)Y_m$ with $w_m(r_m) = (X_m'(r_m)X_m(r_m) + \lambda_m I)^{-1}$. Note that $\hat{\beta}_m(r_m) = \hat{\beta}_m^{\lambda_m}(r_m)$ if we set $\lambda_m = 0$. Finally, for the prior of each σ_m^2, we assume $\sigma_m^2 \sim \text{IG}\left(\frac{a_{m_0}}{2}, \frac{b_{m_0}}{2}\right)$ independently.

We assume that z_i's are independent of (r_m, β_m, σ_m), and (r_m, β_m, σ_m) are assumed to be independent of each other. For each m, r_{mj}'s are assumed to be independent of each other.

2.3. Gibbs sampler scheme for variable selection

With the prior specification described in the above subsection, the joint posterior distribution is derived as follows. The complete-data likelihood of the mixture regression model for $y = (y_1, \dots, y_n)$ and $z = (z_1, \dots, z_n)$ is

$$\ell(y, z|\theta) = \prod_{i=1}^{n} \rho_{z_i} f(y_i|\theta_{z_i}) = \prod_{m=1}^{M} \rho_m^{n_m} \left[\prod_{i \in G_m} f(y_i|\theta_{z_i})\right],$$

where $n_m = \sum_{i=1}^{n} I\{z_i = m\}$. G_m is the set containing the members in the mixture component m. $\theta_m = (\beta_m, \sigma_m^2, \rho_m, r_m)$. $f(y_i|\theta_m)$ is the probability density function of a Gaussian random variable with mean $x_i'(r_m)\beta_m(r_m)$ and variance σ_m^2. Let $\theta = (\theta_1, \dots, \theta_m)$. Combining the complete-data likelihood and the prior $\pi(\theta)$, we have the posterior distribution given by

$$p(\theta, z|y) \propto \prod_{m=1}^{M} \rho_m^{n_m} \left[\prod_{i \in G_m} \left[f(y_i|\theta_{z_i})\right]\right] \pi(\theta).$$

We use the Gibbs sampler to generate the parameters from the posterior distribution. To implement the Gibbs sampler, we need to derive the full conditional distributions of all the components of θ and z separately. Starting from an initial value, the Gibbs sampler proceeds by iteratively drawing samples from the following conditional distributions:

Step 1 The conditional distribution of z given θ and y is, for $i = 1, \dots, n$,

$$P(z_i = m|\theta, y) = \frac{\rho_m f(y_i|\theta_m)}{\sum_{m=1}^{M} \rho_m f(y_i|\theta_m)}.$$

Step 2 The conditional distribution of ρ given z is

$$\rho \sim \text{Dirichlet}(n_1 + \alpha_1, \dots, n_M + \alpha_M).$$

Step 3 The conditional distribution of σ_m^2 given β_m, r_m, z, y is $\sigma_m^2 \sim IG\left(\frac{a_m}{2}, \frac{b_m}{2}\right)$, where $q_m = \sum_{j=1}^{p} r_{mj}$; $a_m = n_m + q_m + a_{m_0}$, and

$$\begin{aligned} b_m &= \left[Y_m - X_m'(r_m)\beta_m(r_m)\right]' \left[Y_m - X_m'(r_m)\beta_m(r_m)\right] \\ &\quad + \frac{[\beta_m(r_m) - \hat{\beta}_m^{\lambda_m}(r_m)]' w_m^{-1}(r_m)[\beta_m(r_m) - \hat{\beta}_m^{\lambda_m}(r_m)]}{n_m} + b_{m_0}. \end{aligned}$$

Step 4 The conditional distribution of β_m given σ_m^2, r_m, z, y is

$$\beta_m(r_m) \sim N\left(\mu_m, \Omega_m\right),$$

where $\mu_m = \Omega_m \left(n_m X_m(r_m) Y_m' + w_m(r_m)^{-1} \hat{\beta}_m^{\lambda_m}(r_m)\right) / n_m \sigma_m^2$ and $\Omega_m^{-1} = \left[\frac{n_m X_m'(r_m) X_m(r_m) + w_m^{-1}(r_m)}{n_m \sigma_m^2}\right]$.

Step 5 The conditional distribution of r_m given $\beta_m, \sigma_m^2, z, y$ is, for each j in G_m,

$$p(r_{mj} = 1|r_{m(-j)}, z, y) = \frac{p(r_{mj} = 1|r_{m(-j)}, z, y)}{p(r_{mj} = 1|r_{m(-j)}, z, y) + p(r_{mj} = 0|r_{m(-j)}, z, y)},$$

where $r_{m(-j)}$ is the vector obtained from r_m by excluding r_{mj}.

2.4. Variable selection criterion and component assignment

After a sufficient number, say K, of samples of parameters are drawn from the posterior distribution by the Gibbs sampler, they are used for posterior inference. In order to determine the active variables of the linear regression model of each sub-population, we collect the posterior samples of r_{mj}'s and adopt the median probability criterion (Barbieri and Berger, 2004) for variable selection. Specifically, we calculate the relative frequency of the variable x_j within the sub-population m by

$$\hat{p}(r_{mj} = 1|y) = \frac{1}{K} \sum_{k=1}^{K} I\{r_{mj}^{(k)} = 1\},$$

where $r_{mj}^{(k)}$ is the kth posterior sample generated by the Gibbs sampler, and where $I\{\}$ is the indicator function. This gives us an estimate of the posterior variable inclusion probability as a measure of the relative importance of the jth predictor variable within the mth sub-population or mixture component. Specifically, based on the median probability criterion, we claim that the variable x_j is active for the mth sub-population if

$$\hat{p}(r_{mj} = 1|y) \geq 1/2. \tag{2}$$

The posterior samples of z_i's can be used for membership assignment of each observation i. Specifically, an observation y_i is assigned to the mth sub-population if

$$\hat{p}(z_i = m|y) = \max_g \hat{p}(z_i = g|y) = \max_g \frac{1}{K} \sum_{k=1}^{K} I\{z_i^{(k)} = g\}, \tag{3}$$

where $z_i^{(k)}$ is the membership of the ith observation in the kth posterior sample generated by the Gibbs sampler.

2.5. Identifiability

A common issue with mixture models is the nonidentifiability of the component parameters. This is due to the so-called "label-switching" problem (Celeux, 1998) caused by the symmetry in the likelihood of the model parameters. For a review of the identifiability issue in Bayesian mixture modeling, the interested reader can consult Jasra et al. (2005) and the references therein. For an M-component mixture, the parameter space has $M!$ regions over which the likelihood is identical, that is, the component parameters are not marginally identifiable. Thus, if $(\theta_1, \dots, \theta_M)$ is a local maximum of the likelihood function, so is $(\theta_{\omega_1}, \dots, \theta_{\omega_M})$ for every permutation ω. This makes maximization and exploration of the posterior distribution difficult.

A commonly adopted solution to this problem is to remove the symmetry by imposing an identifiability constraint on the parameters. In this paper, we arrange the mixture components in the order of increasing variances $\sigma_1^2 < \sigma_2^2 < \cdots < \sigma_M^2$. For a more detailed discussion and some examples of nonidentifiability, please see Frühwirth-Schnatter (2006).

2.6. Number of sub-populations

In the finite mixture model, the number M of the mixture components or sub-populations is usually unknown, and misspecification of the number of mixture components can cause misleading results during the variable selection process, as well as misclassified observations. Instead of pre-specifying a fixed number of mixture components, several information criteria can be adopted to determine the number of components.

Most existing approaches address the issue of model selection using criteria such as the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Deviance Information Criterion (DIC) (Celeux et al., 2006). Let M be the possible number of components in the mixture model. Let d_M be the number of parameters. Let $\hat{\vartheta}_M$ be the estimates of the parameters. The AIC, BIC and DIC for the mixture model with M components are defined as, respectively,

$$\begin{aligned}\mathrm{AIC}_M &= -2\log L(y|\hat{\vartheta}_M) + 2d_M;\\ \mathrm{BIC}_M &= -2\log L(y|\hat{\vartheta}_M) + d_M\log(n);\\ \mathrm{DIC}_M &= -2\log L(y|\hat{\vartheta}_M) + p_D,\end{aligned}$$

where p_D is the posterior mean deviance minus deviance evaluated at the posterior mean of the parameters, which is used to measure the complexity in terms of the effective number of parameters. Evidently, BIC assigns more penalty to the complexity of the selected model. However, neither BIC nor AIC takes into account the fact that in a clustering context, a finite mixture model is fitted with the hope of finding a good partition of the data. To overcome this shortcoming, the integrated classification of likelihood (ICL) criterion was proposed by Biernacki et al. (2000) with the specific aim of selecting the number of components of a finite mixture model in model-based clustering. McLachlan and Peel (2000) showed that when the number of members within each group is large enough, ICL is approximately equal to

$$\text{ICL-BIC}_M = \mathrm{BIC}_M + 2\mathrm{EN}(\hat{\vartheta}_M),$$

where $\mathrm{EN}(\hat{\vartheta}_M) = -\sum_{i=1}^n \sum_{m=1}^M P(z_i = m|y_i, \hat{\vartheta}_M)\log P(z_i = m|y_i, \hat{\vartheta}_M)$. This so-called entropy term measures the degree of the inability of the fitted M-component mixture model to provide a good partition of the data. It is close to 0 if the resulting clusters are well separated and will have a large value if this is not the case. Therefore, the ICL and its asymptotic variant penalize not only the complexity, but also the failure of classification in well-separated clusters of the fitted model. We can select the model by minimizing AIC_M, BIC_M, ICL-BIC_M, or DIC_M. The performance of these criteria will be illustrated in the next section.

3. Simulation study

In this section we conduct simulation studies to illustrate the performance of the proposed method for fitting the mixture model of linear regressions. We shall also demonstrate that our method can be easily applied to the problem of $p > n$. The settings of our simulation studies are based on those in Städler et al. (2010) and Khalili and Chen (2007). In both papers, the penalized least squares methods are used for variable selection. Specifically, given the log-likelihood function of the model (1), the penalty terms are added to induce sparsity, i.e.,

$$l_n(\theta) = \sum_{i=1}^n \log\left(\sum_{m=1}^M \rho_m \cdot \phi\left(y_i; x_i'\beta_m, \sigma_m^2\right)\right) - \sum_{m=1}^M \pi_m\left(\sum_{j=1}^p p_{nm}(\beta_{mj})\right),$$

where $\phi(y; \mu, \sigma^2)$ is the Gaussian density function, and $p_{nm}(\beta_{mj})$ is a penalty function that is nonnegative and nondecreasing in $|\beta_{mj}|$, the absolute value of β_{mj}. Both papers develop EM-type algorithms for variable selection. We shall show that our method exhibits better performance in variable selection compared to the penalized least squares methods.

3.1. Simulation 1

We first conduct a simulation study according to the setting of Khalili and Chen (2007). For simplicity, we consider the mixture model of two linear regressions, i.e. $M = 2$, given by

$$y \sim \rho N(x'\beta_1, 1) + (1-\rho)N(x'\beta_2, 1). \tag{4}$$

Table 1
Average numbers of correct and incorrect zero coefficients with $n = 100$. The numbers within parentheses are the standard errors of the corresponding estimates.

Method	$\pi = 0$				$\pi = 0.5$			
	Component 1		Component 2		Component 1		Component 2	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
$\rho = 0.5$								
Bayes	2.98$_{(0.01)}$	0.016$_{(0.01)}$	1.99$_{(0.005)}$	0.011$_{(0.001)}$	2.97$_{(0.02)}$	0.021$_{(0.01)}$	1.99$_{(0.005)}$	0.022$_{(0.002)}$
MixSCAD	2.98	0.021	1.99	0.024	2.94	0.024	1.98	0.058
MixLASSO	2.93	0.026	1.98	0.025	2.52	0.027	1.77	0.078
$\rho = 0.3$								
Bayes	2.95$_{(0.02)}$	0.051$_{(0.02)}$	1.98$_{(0.006)}$	0.016$_{(0.002)}$	2.90$_{(0.05)}$	0.055$_{(0.02)}$	1.96$_{(0.012)}$	0.025$_{(0.003)}$
MixSCAD	2.81	0.057	1.92	0.044	2.84	0.089	1.96	0.024
MixLASSO	2.74	0.060	1.94	0.028	2.54	0.133	1.78	0.042
$\rho = 0.1$								
Bayes	2.90$_{(0.04)}$	0.107$_{(0.04)}$	1.95$_{(0.010)}$	0.022$_{(0.005)}$	2.85$_{(0.10)}$	0.122$_{(0.04)}$	1.94$_{(0.018)}$	0.055$_{(0.006)}$
MixSCAD	2.10	0.246	1.88	0.023	2.40	0.577	1.99	0.026
MixLASSO	2.64	0.733	1.83	0.040	2.58	0.919	1.71	0.044

The predictor variables x in Eq. (4) are simulated from a 5-dimensional multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix Σ. Let $\Sigma(i,j)$ denote the (i,j) entry of Σ. We set $\Sigma(i,j) = \pi^{|i-j|}$ where $-1 < \pi < 1$. We consider two correlation structures with $\pi = 0$ and $\pi = 0.5$ respectively. The regression coefficients are set to be $\beta_1 = (1, 0, 0, 3, 0)$ and $\beta_2 = (-1, 2, 0, 0, 3)$ respectively, the same as in Khalili and Chen (2007). Three different choices of mixing proportions $\rho = 0.1, 0.3, 0.5$ are considered.

For the prior parameters, we first set the prior inclusion probabilities of the variables $d_{mj} = 0.5$ for all variables j and sub-populations m, to reflect the fact that we do not have any prior information about which variables are active in different sub-populations. In the inverse gamma prior for σ_m^2, the nearly non-informative priors are used by setting $a_{m_0} = b_{m_0} = 0.001$ for all m. We set $\alpha_i = 2$ in the Dirichlet prior for $i = 1$ and 2. For the value of g_m in the g-prior, we set $g_m = n_m$. Since there is no singularity problem, we set the ridge parameter $\lambda_m = 0$.

We run the Gibbs sampler 1,000,000 sweeps and collect the posterior samples. As advocated by Flegal et al. (2008), it is not necessary to have a burn-in period; thus, we keep all the samples as our posterior samples for Bayesian inference. The median probability criterion (2) is used to decide if a variable is active or not. In fact, with 100,000 iterations, the marginal posterior probabilities in our simulation are very stable over random restarts, implying convergence. We also studied the performance with the burn-in step included. Since the results are similar to those without burn-in, they are not shown here due to space limitations. We will discuss this issue further in Section 5.

To examine the performance of our method and compare our results to those of Khalili and Chen (2007), we calculate the average estimates of the number of correct zeros in the coefficient vectors and the corresponding standard errors over 100 simulated data sets with $n = 100$. The results are given in Table 1. In this table, we also show the results of MixLASSO and MixSCAD reported in Khalili and Chen (2007). Overall, our approach (labeled as "Bayes" in the table) is superior to the MixLASSO and MixSCAD methods in terms of the correct identification rates, especially in the case where $\rho = 0.1$. In addition, in most cases, the proposed approach also produces lower incorrect identification rates. This shows that the proposed Bayesian variable selection approach performs better variable selection than the methods employed by Khalili and Chen (2007).

3.2. Simulation 2

To illustrate the performance of our method in various scenarios (described below), we conduct another simulation. There are four purposes in this simulation study. First, we examine the ability of our approach to select the active variables even when the design matrix is highly correlated. Second, we investigate the influence of different values of g_m of the g-prior on the posterior estimation and identification of the active variables. Third, we demonstrate that the information criteria can be used to determine the number of sub-populations. Finally, we apply our method to the problem of large p and small n by increasing the number of inactive variables.

In this simulation study, we consider the same setting as the model M4 in Städler et al. (2010). Suppose that there are $M = 3$ sub-populations with a total of 150 observations and the corresponding β's are $\beta_1 = (3, 3, 0, 0, 0, 0)$, $\beta_2 = (0, 0, -2, -2, 0, 0)$, and $\beta_3 = (0, 0, 0, 0, -3, 2)$ in the three regression models respectively. We set $\sigma_m^2 = 0.5$ for all the sub-populations m and the mixing proportions are assumed to be $\rho = (1/3, 1/3, 1/3)$.

In the Bayesian analysis, except for the prior parameters λ_m and g_m in the g-prior, the other prior parameters are the same as those in Simulation 1. We again iterate the Gibbs sampler 1,000,000 sweeps to generate the posterior samples. The median probability criterion (2) is used to determine which variables are active. The classification of each observation is determined according to (3).

To evaluate the performance of our method, we consider the following measures: the True Classification Rate (TCR), the True Positive Rate (TPR), the False Positive Rate (FPR), and the rate of the True Classification of Observations (TCO). They are

Table 2
The measures of TCR, TPR, FPR, and TCO for different values of λ_m and g_m.

λ_m	π											
	0.25			0.5			0.75			0.995		
	$\frac{1}{2p}$	$\frac{1}{p}$	$\frac{2}{p}$	$\frac{1}{2p}$	$\frac{1}{p}$	$\frac{2}{p}$	$\frac{1}{2p}$	$\frac{1}{p}$	$\frac{2}{p}$	$\frac{1}{2p}$	$\frac{1}{p}$	$\frac{2}{p}$
TCR	0.99	0.99	0.99	0.99	0.99	0.98	0.98	0.97	0.97	0.83	0.84	0.86
TPR	1	1	1	1	1	0.98	0.98	0.97	0.97	0.67	0.58	0.58
FPR	0.008	0.010	0.011	0.009	0.010	0.010	0.011	0.012	0.012	0.12	0.13	0.14
TCO	0.92	0.91	0.91	0.92	0.91	0.90	0.92	0.91	0.90	0.92	0.92	0.92

defined as follows:

$$\text{TCR} = \frac{\text{number of correctly selected variables}}{\text{number of variables}};$$

$$\text{TPR} = \frac{\text{number of correctly selected active variables}}{\text{number of active variables}};$$

$$\text{FPR} = \frac{\text{number of falsely selected active variables}}{\text{number of inactive variables}};$$

$$\text{TCO} = \frac{\text{number of correct classifications of observations}}{\text{number of observations}}.$$

TCR is an overall evaluation of the accuracy in the identification of the active and inactive variables. TPR is the average rate of active variables identified correctly and is used to measure the power of the method. FPR is the average rate of inactive variables that are included in the regression and can be considered as the type I error rate of the selected approach. TCO is the rate that the observations are classified into the correct sub-populations. Larger values of TCR, TPR, and TCO indicate better performance than smaller values, whereas smaller values of FPR indicate better performance than larger values.

We generate the variables from 5-dimensional multivariate normal distributions with different pairwise correlations $\Sigma(i,j) = \pi^{|i-j|}$, where $\pi = 0.25, 0.5, 0.75$, and 0.995. Since we want to investigate the effect of λ_m on selecting important variables in the case of multicollinearity, we fix $g_m = n_m$ as suggested by Zellner (1996). We start with $\lambda_m = \frac{1}{p}$ as suggested by Baragatti and Pommeret (2012), and then try $\lambda_m = \frac{1}{2p}$ and $\frac{2}{p}$. The results over 100 replications are shown in Table 2. The performances under different λ_m are similar, so we choose $\lambda_m = \frac{1}{p}$ in the remaining experiments. Note that increasing the degree of collinearity between the variables corresponds to a decrease in the correct identification rate.

To demonstrate that our approach can work for the problem of large p and small n, we consider $p = 65, 85, 105, 125, 155$. We set β_{mj} at the same values as in the previous simulation except $\beta_{mj} = 0$ for $j = 7, \ldots, p$. For all m and the additional covariates, x_i's are independently generated from a multivariate normal distribution with zero mean vector and identity covariance matrix. According to the previous experiments, we fix $\lambda_m = 1/p$ for all m. In this simulation study, we find that the value g_m has an effect on the results of variable selection. This is a typical phenomenon whenever the g-prior is used. In general, smaller values of g_m tend to result in more complex models, whereas larger values of g_m tend to produce more parsimonious models. In our simulation study, we find that when $g_m = n_m$, some inactive variables are selected, especially in the large p examples (e.g. the case where $p > 100$). Thus, if the ratio of the number of variables to that of observations, p/n, is greater than or equal to 3, we recommend setting $g_m = 100 \times p \times M/n$. For each p, we repeat the simulation 100 times. The box plots of TPR, FPR and TCO for different numbers of covariates, p, are shown in Fig. 1. These box plots show that our method is quite effective at identifying the true active variables and classifying the observations into the sub-populations. Moreover, the distribution of TCO indicates that most of the time the observations are correctly classified. As for the computational cost of the proposed approach, we record the average computing time for 10,000 iterations of MCMC with $\pi = 0.5$ and different numbers of variables p in this simulation study. The computation time is reported in Table 3. The computation is done on a Linux cluster containing Xeon E5620 2.4 GHz 2 Quad-core, 16 GB × 8 = 128 GB.

Next, we evaluate the performances of AIC, BIC, ICL–BIC, DIC, and the marginal likelihood on selecting the correct number of mixture components via a simulation study. We generate 100 simulated data sets with $\pi = 0.5$ for different n and p with the true number of components equal to 3. In each simulated data set, the estimated number of components is determined respectively in terms of the minimum values of AIC, BIC, ICL–BIC, DIC, and the maximum value of the marginal likelihood. Table 4 displays the frequencies of the estimated numbers of components for each method. It demonstrates that the information criteria can be used to determine the number of components. Moreover, we find that even if the assumed number of components is larger than the true number, the Gibbs sampler still tends to select the correct number of components by assigning no observations to the extra components. Evidently, more observations result in higher accuracy in determining the number of components and classifying the observations into these components.

Finally to demonstrate the capability of the proposed approach to handle the problem of large p, we consider $p = 1000$ and $n = 500$. The values of the parameters are the same as in the setting above with $M = 3$. The value of g_m is taken to be the sample size of each group because the ratio of p to n is less than 3. Also, we take the value of $\lambda_m = 1/p$ as suggested in the experiment conducted above. For illustration, only a simulated data set with the pair correlation between covariates

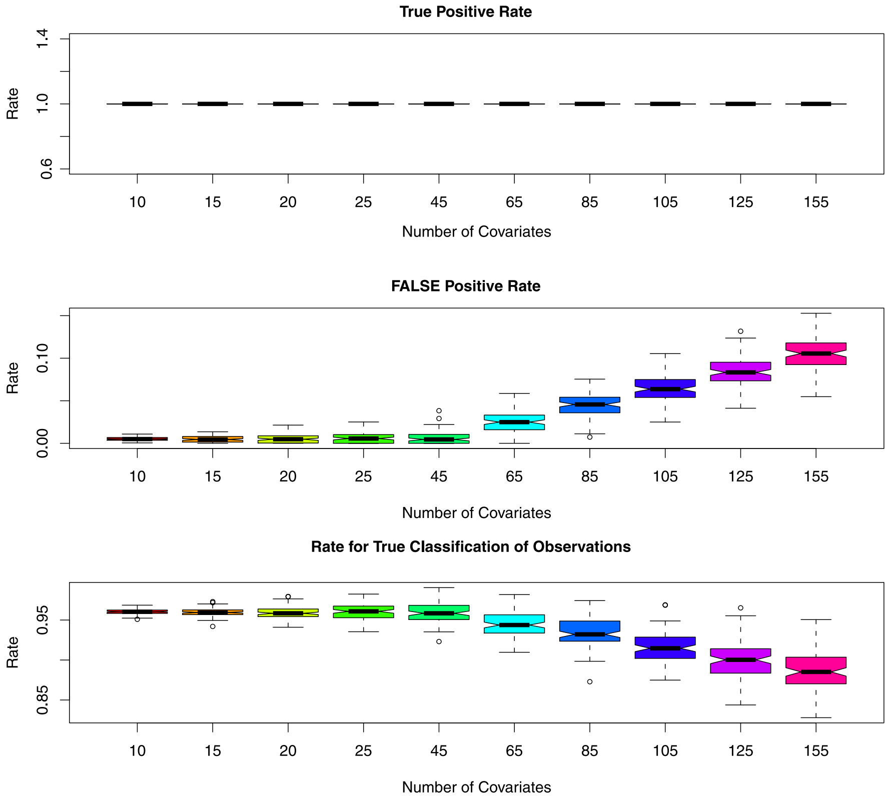


Fig. 1. Box plots for the true positive rate (TPR), false positive rate (FPR), and the rate of true classification of observations (TCO), for different numbers of covariates.

Table 3
The computing time in seconds for different scenarios.

p	25	50	100	150
CPU time (s)	1079	2829	5618	7805

Table 4
The true number of mixture components is 3. Each cell shows the frequency of the estimated number of components over 100 simulated data sets where the number of components is determined by AIC, BIC, ICL–BIC, DIC and marginal likelihood respectively. We allow the number of components to be chosen from $\{3, 4, 5\}$.

M	$n = 150, p = 50$			$n = 300, p = 100$		
	3	4	5	3	4	5
AIC	94	1	5	95	1	4
BIC	94	4	2	96	3	1
ICL–BIC	95	3	2	97	3	0
DIC	94	1	5	96	2	2
Marginal loglikelihood	92	4	4	95	3	2

Table 5
The measures of TCR, TPR, and FPR for the case of large p and small n.

TCR	TPR	FPR	TCO
0.99	1	0.008	0.93

$\pi = 0.5$ is used. For this data set, the four measures are shown in Table 5. The result shows that the proposed approach can be applied to the large p problem.

Table 6
Posterior estimates of mixing proportions and variances. "Bayes" denotes our results. MixLASSO and MixSCAD are the results from Khalili and Chen (2007).

Covariate	Bayes		MixSSAD		MixLASSO	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
ρ	0.28	0.72	0.24	0.76	0.28	0.72
σ	0.58	0.30	0.32		0.25	
R^2	0.99	0.99	0.90	0.94	0.95	0.96

4. Real example

In addition to the simulation studies, a real data set is analyzed to demonstrate the ability of the proposed approach at identifying important variables. The data set focuses on baseball salary, and the variables are highly correlated. The same data set was analyzed in Khalili and Chen (2007).

The baseball salary data contains 337 major league baseball players' salaries (in thousands) in the year 1992 with 16 performance measures from the year 1991. The players, excluding pitchers, played at least one game in both the 1991 and 1992 seasons. The main interest in this analysis is to detect which performance measures play important roles in determining the salaries.

The 16 performance measures are: batting average (x_1), on-base percentage (x_2), runs (x_3), hits (x_4), doubles (x_5), triples (x_6), home runs (x_7), runs batted in (x_8), walks (x_9), strikeouts (x_{10}), stolen bases (x_{11}), and errors (x_{12}); and indicators of free agency eligibility (x_{13}), free agent in 1991/2 (x_{14}), arbitration eligibility (x_{15}), and arbitration in 1991/2 (x_{16}). We standardize the explanatory variables, x_i, for $i = 1, \dots, 12$, in our analysis. The last four dummy variables indicate the degree of freedom for the player to move to another team. Watnik (1998) suggested that there could be potentially important interactions between the quantitative variables x_1, x_3, x_7, and x_8 and the last four variables about the freedom of a player. Among the predicator variables, x_1 and x_7 measure the performance of a player, while x_3 and x_8 measure the contributions of a player to the team. The total number of variables that may affect a player's salary is 32, including the original variables and their interactions.

Since the salaries are highly right-skewed, a log transformation is used to make log(salary) the response variable. Khalili and Chen (2007) suggested that a mixture of two normal linear models should be a proper model assumption, i.e.

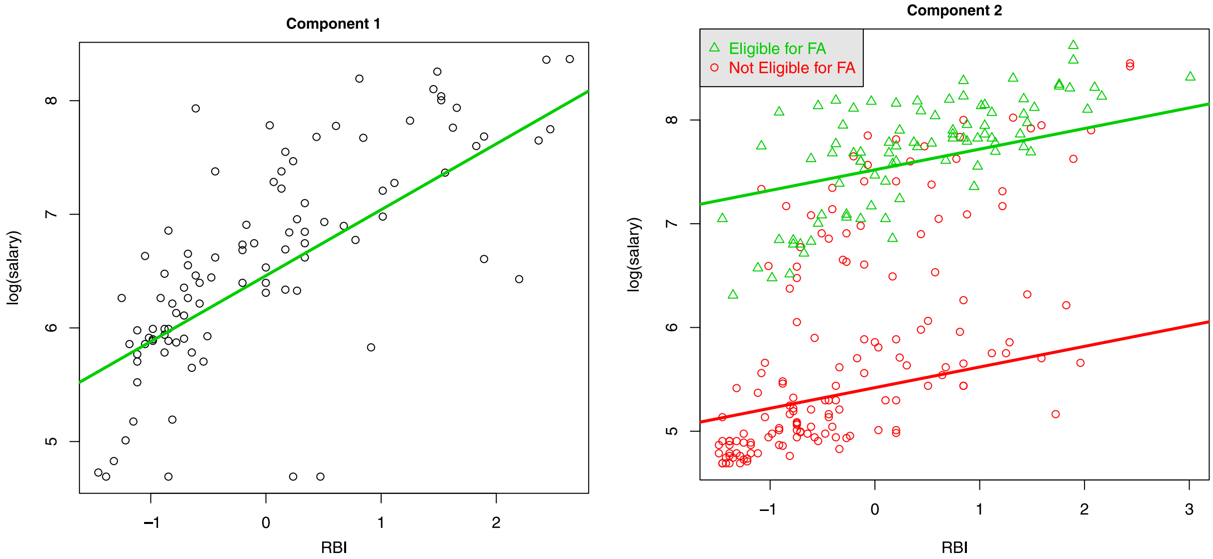
$$y = \log(\text{salary}) \sim \rho N(x'\beta_1, \sigma_1^2) + (1-\rho)N(x'\beta_2, \sigma_2^2)$$

where x is a 33×1 vector consisting of all 32 potential covariates plus an intercept. MixLASSO and MixSCAD were used by Khalili and Chen (2007) to identify the important variables. Unlike the homogeneous variance assumption in Khalili and Chen (2007), we assume different variances σ_1^2 and σ_2^2.

To analyze this baseball salary data, we set the prior parameters as follows. First, the prior inclusion probability, d_{mj}, is fixed at 0.5 for all m and j. For the inverse gamma prior of σ, we still set $a_{m_0} = b_{m_0} = 0.001$, which is a nearly non-informative prior. The α_i's in the Dirichlet prior are fixed at 2. For the parameters λ_m and g_m in the g-prior, from our pilot study, we find that the singularity problem among the variables does exist, and thus we set the ridge parameter $\lambda_m = 1/p = 1/33$ for all m. For the g_m in the g-prior, since the number of observations, n, is larger than that of the covariates, p, we choose $g_m = n = 337$. For sensitivity analysis, we also try different values of g_m, e.g., 10, 100, 500, 1000, but no significant difference is detected. 1,000,000 sweeps of Gibbs sampling is implemented to generate the posterior samples of z_i and r_{mj}.

Based on the estimates of the posterior inclusion probabilities in Fig. 5(a), we conclude that RBIs play an important role in achieving a high salary in both sub-populations, according to the median probability criterion. RBI is a measure of the contribution of a player to a team. The more RBI a player has, the more likely the team is to win a game. Fig. 2(a) evidently shows that the more RBIs a player contributes to, the higher salary he has in the first sub-population. We also find that Runs is another factor that affects a player's salary in the second population. Although RBIs and Runs are highly correlated (0.88), they both play a role in affecting a player's salary. Moreover, the most significant difference between the two sub-populations are the factors of eligibility for free agency and arbitration. Being a free agent and having the eligibility for arbitration lead to a higher salary in the second sub-population, see Fig. 2(b). On the contrary, neither variable is helpful for getting higher pay for the players in the first sub-population. The visualization of the two variables affecting a player's salary are shown in Fig. 3 for each sub-population. Though the number of variables selected by the proposed approach is less than that of Khalili and Chen (2007), our model has more potential to explain the patterns across sub-populations as will be described next.

Based on our results, we classify the observations into different sub-populations, and then fit the linear model for each sub-population using the corresponding variables selected by our method to obtain the adjusted R^2 value. Compared with the model fitting results by MixLASSO and MixSCAD in Khalili and Chen (2007), the performance of our method is better in terms of the values of R^2. These results suggest that our approach accurately classifies the observations so that observations classified into the same sub-population behave in a similar way. See Table 6. We find that the estimated proportions for the sub-populations are not significantly different for different estimation procedures such as Bayes, MixSCAD, and MixLASSO. However, the variances in different sub-populations are clearly different as illustrated in Fig. 4. Therefore, it is more reasonable to assume different variances for different sub-populations.



(a) The green line is drawn using the β value obtained by our approach.

(b) The red/green lines are drawn using the β values obtained by our approach. The green points illustrate the players with eligibility for free agency, whereas the red points are the players without such eligibility.

Fig. 2. Scatter plot of logarithm of baseball salary versus RBI. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
The estimated values of information criteria for different models.

Information	Number of components		
	2	3	4
AIC	502	579	604
BIC	540	621	658
DIC	1	43	54
Marginal loglikelihood	−241	−279	−288
ICL–BIC	760	858	901

Here we set the number of sub-populations to 2 according to the suggestion of Khalili and Chen (2007). To check this assumption, we first use the information criteria studied in Section 2.6 to determine the number of sub-populations. Table 7 shows the estimated information criteria with respect to the number of mixture components. These criteria suggest that it is appropriate to fit the 2-component model to the data.

For a more thorough investigation, we rerun our analysis by assuming 3 sub-populations or components. We classify the observations into the 3 groups, and order the groups according to the numbers of observations in each group, in descending order. The posterior inclusion probabilities of the variables under the 3-component model and the 2-component model are shown in Fig. 5. Based on the median probability criterion, the variables selected are the same for the first two groups in both the 2-component model and the 3-component model. It is worth noting that group 3 of the 3-component model turns out to be empty. This confirms that the 2-component model is the correct one.

5. Discussion

In this section, we discuss several issues about the proposed Bayesian variable selection approach. First, we describe how to estimate the regression coefficients based on the posterior samples. Next, we study the issue of the number of iterations in MCMC and the performance of the proposed algorithm with and without the burn-in step. Finally, we compare the performance of two different model selection criteria, the median posterior probability and the highest posterior probability, on variable selection via a stimulation study.

5.1. Posterior inference of regression coefficients

Since our focus is on variable selection, we first identify the active variables and then re-estimate the corresponding coefficients by setting the coefficients of the inactive variables to zero. In Step 4 of the proposed Gibbs sampler, we still generate samples of the coefficients β_m. Thus, we can estimate these active coefficients based on the posterior samples.

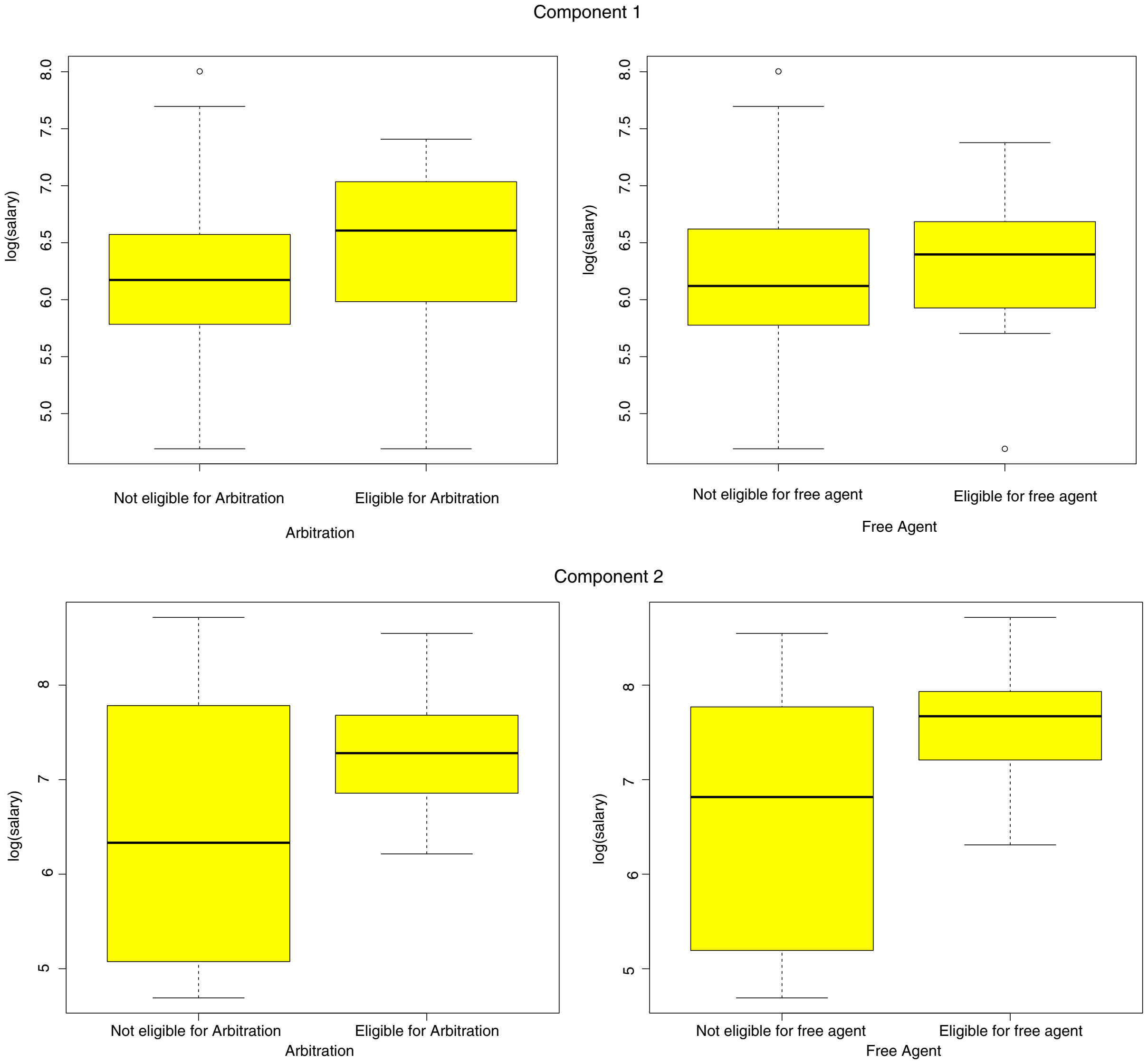


Fig. 3. Box plots of salaries based on factors of eligibility for free agency and arbitration in different sub-populations.

Given the posterior samples of the coefficients, the posterior mean $\mathrm{E}(\beta|y)$ is a natural estimate of the coefficients. In fact, the posterior mean can be approximated in a straightforward manner from simulated samples. We use Rao-Blackwellization method (Gelfand and Smith, 1990) to estimate the regression coefficient β_{mj}

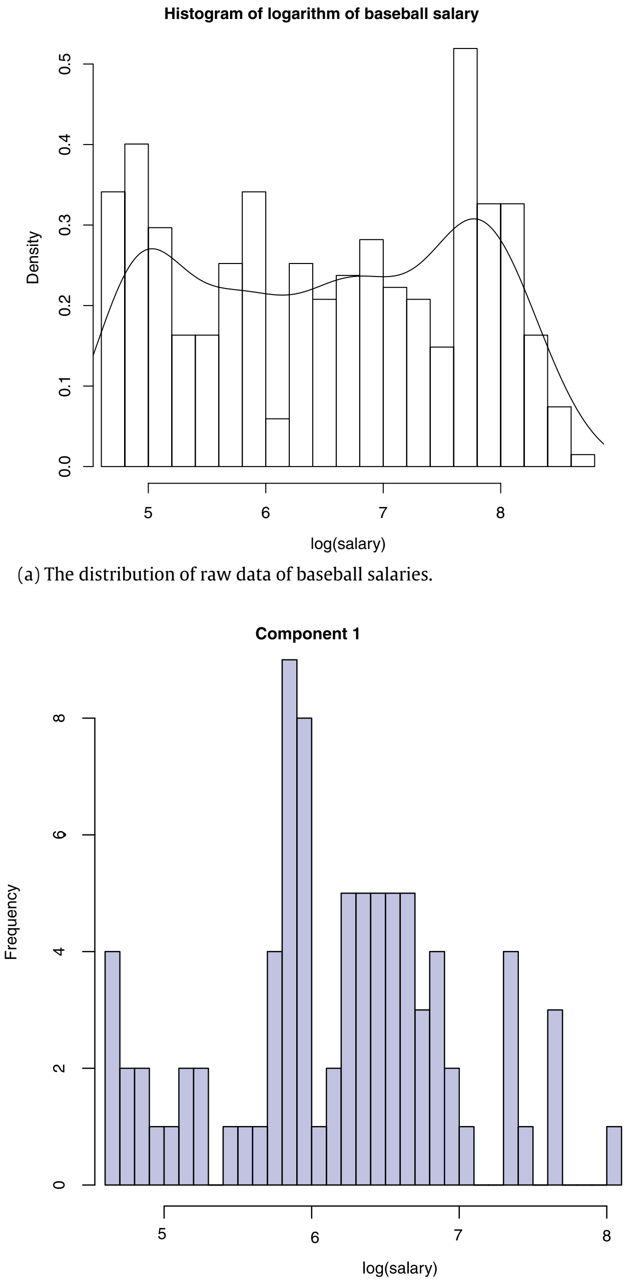
$$\mathrm{E}(\beta_{mj}|y) = \sum_{r_{mj}} E\left[\beta_{mj}|r_{mj}, y\right] p\left(r_{mj}|y\right) \approx \frac{1}{K_j}\sum_{k=1}^{K}\beta_{mj}^{[k]},$$

where $K_j = \sum_{k=1}^{K} r_{mj}^{[k]}$, and where $r_{mj}^{[k]}$ and $\beta_{mj}^{[k]}$ are the MCMC samples in the kth iteration.

We revisit Simulation 2 with $p = 50$ and $n = 150$, and the pair correlation $\pi = 0.5$. We investigate the accuracy of the posterior estimates for the parameters by comparing the estimates to the true values. Based on the same setup as in Simulation 2 (i.e. $\lambda = 1/p$ and $g_m = n_m$), 1,000,000 samples are generated by the Gibbs sampler. The estimates of the β's are given in Table 8. The estimates of the non-zero coefficients are close to the true parameter values. In addition to the coefficients, the posterior means of σ's and ρ's are also shown in Table 8.

5.2. The convergence property and burn-in step

We assess the convergence of the proposed Gibbs sampler based on the Monte Carlo standard error (MCSE) in Flegal et al. (2008). Specifically, we check the convergence of the proposed MCMC algorithm by tracing the corresponding MCSE of



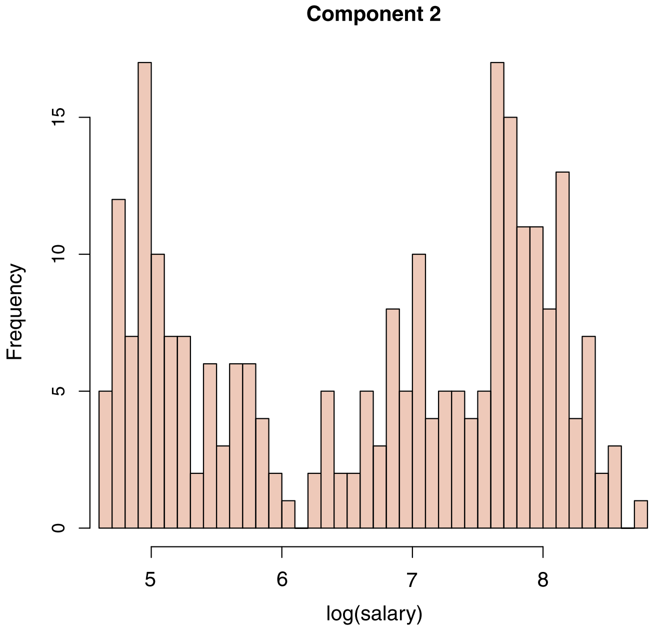
(a) The distribution of raw data of baseball salaries.

(b) The distribution of observations in the first component.

Fig. 4. Baseball salary study.

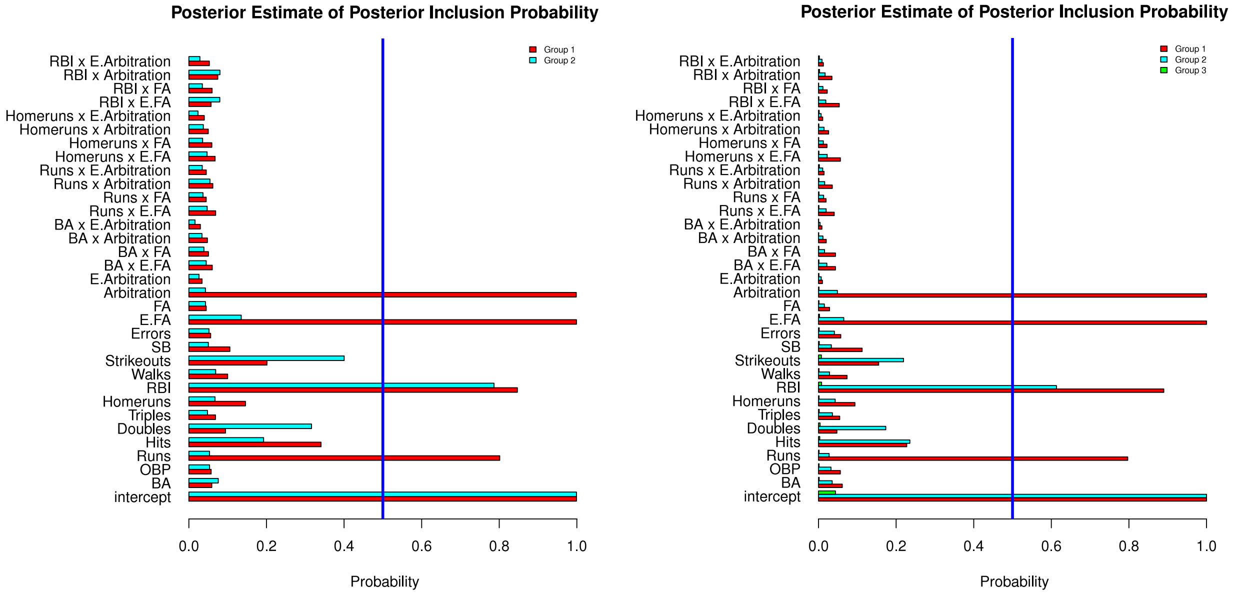
σ's. The MCMC chain stops whenever the largest MCSE of the estimates of σ's is less than 0.005, following the suggestion of Flegal et al. (2008). Once again, we revisit Simulation 2 with $p = 50$, $n = 150$ and $\pi = 0.5$ to study the issue of convergence and the burn-in step. Using the same tuning parameters as before (i.e. $g_m = n_m$ and $\lambda = 1/p$), 1,000,000 samples are generated for the simulated data. In fact, the stopping rule suggests that it is not necessary to have such a long chain. For instance, the chain would be stopped after 120,000 iterations when all the MCSEs of the estimates of σ's are less than 0.005. Putting together the findings from Table 9 and the visual assessment in Fig. 6, we can conclude that the MCMC outputs indicate convergence. Once the largest MCSE of σ's is less than 0.005, the chain stops and the samples are used to estimate the parameters of interest.

We further investigate the effect of the burn-in step on parameter estimation. Flegal et al. (2008) suggest that the burn-in step might not be necessary if the starting values are reasonable. This can be somewhat problematic, however, since there



(c) The distribution of observations in the second component.

Fig. 4. (*continued*)



(a) Two components.

(b) Three components.

Fig. 5. The estimated posterior inclusion probability of each variable in each component for 2 different models. The vertical blue line represents the probability of 0.5. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is no accurate way to assess the quality of the starting values. To overcome this problem, we first run multiple chains by randomly assigning the starting values of r's. We find that different chains produce almost the same Bayesian inference results, and therefore we believe that the null model is a reasonable initialization. To study the effect of the burn-in step, we compare the estimates of σ^2's with and without burn-in. Based on the previous study on the stopping rule, the chain is stopped after 120,000 iterations. We then burn in 10,000 and 20,000 samples for estimation. Table 10 shows the estimates of σ's by a chain with and without the burn-in step. Table 10 illustrates that the estimates of σ's in three different components are identical with or without burn-in. In fact, all of the results from the different examples in the paper are essentially the same with or without the burn-in step. Thus, we conclude that the burn-in step is not necessarily required, at least in the case of our examples.

Table 8
The estimates of non-zero coefficients and parameters.

Parameters	Component 1		Component 2		Component 3	
	True	Estimate	True	Estimate	True	Estimate
β_1	3	3.06	0	0	0	0
β_2	3	2.84	0	0	0	0
β_3	0	0	−2	−2.04	0	0
β_4	0	0	−2	2.00	0	0
β_5	0	0	0	0	−3	−2.92
β_6	0	0	0	0	2	1.96
ρ	1/3	0.32	1/3	0.34	1/3	0.34
σ	0.50	0.45	0.50	0.51	0.50	0.61

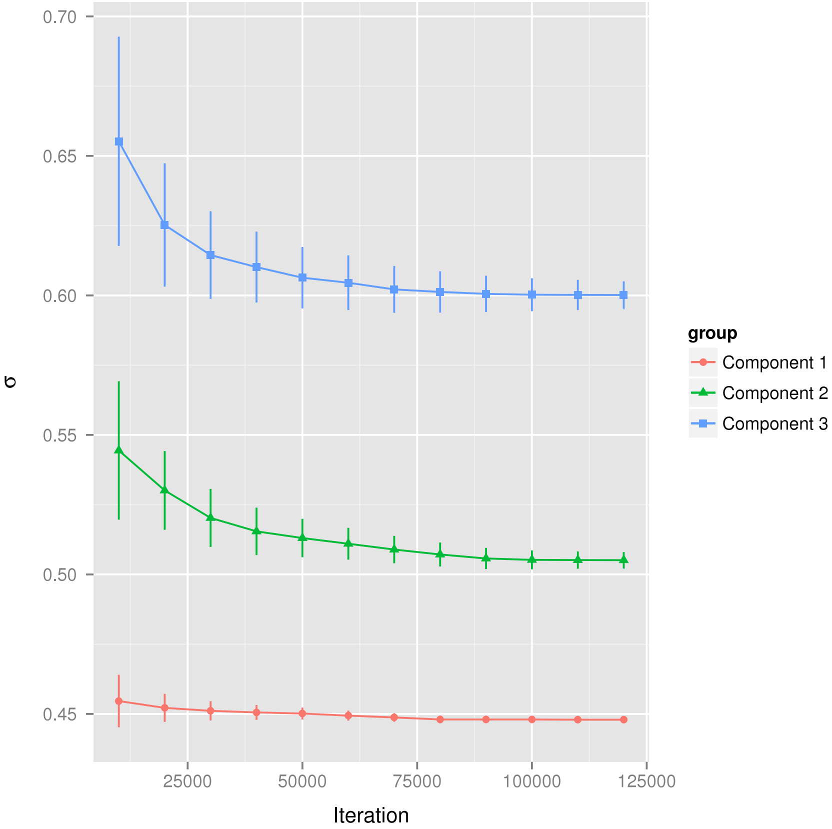


Fig. 6. The cumulative estimates of σ's for every 10,000 iterations. The error bars are the estimates of σ's plus and minus twice the MCSEs of the corresponding estimates.

5.3. Selection criteria

In this paper, the median probability criterion in (2) is used to identify the proper variables in each subgroup. The highest posterior probability criterion is another commonly used standard. According to the highest posterior probability criterion, the set of variables is selected by maximizing the posterior probability among all 2^p possible models. Barbieri and Berger (2004) have shown that under certain conditions, both criteria can identify the same model in the case of linear regression.

We compare these two selection criteria using a simulation study. The data is generated by following Simulation 2 with $M = 3$, $n = 150$, $p = 50$ and $\pi = 0.5$. Using the same tuning parameters in Simulation 2, the proposed Gibbs sampler is iterated for 1,000,000 sweeps to generate the samples. The selection results obtained by these two different criteria over 100 replications are shown in Table 11. The three performance measures indicate that the two selection approaches perform equally well, with results that are similar to each other.

In addition to these three performance measures, we also examine the ranking of the regression model in each component identified by the median probability approach in terms of its posterior probability. The average rankings for the three components over 100 simulated data sets are 1.1 (0.25), 1.1 (0.23), 1.2 (0.40), respectively, where the values in parentheses are standard deviations. This result shows that the model selected by the median probability criterion is in the top ranking

Table 9
The cumulative estimates of σ's and the corresponding MCSEs for every 10,000 iterations.

Iteration	Component 1		Component 2		Component 3	
	$\hat{\sigma}_1$	MCSE($\hat{\sigma}_1$)	$\hat{\sigma}_2$	MCSE($\hat{\sigma}_2$)	$\hat{\sigma}_3$	MCSE($\hat{\sigma}_3$)
1.0E+04	0.46	0.00942	0.54	0.02480	0.65	0.03750
2.0E+04	0.45	0.00502	0.52	0.01400	0.62	0.02210
3.0E+04	0.45	0.00345	0.52	0.01040	0.61	0.01570
4.0E+04	0.45	0.00266	0.52	0.00850	0.61	0.01270
5.0E+04	0.45	0.00214	0.51	0.00690	0.61	0.01100
6.0E+04	0.45	0.00179	0.51	0.00570	0.60	0.00980
7.0E+04	0.45	0.00154	0.51	0.00490	0.60	0.00840
8.0E+04	0.45	0.00135	0.51	0.00430	0.60	0.00740
9.0E+04	0.45	0.00121	0.51	0.00380	0.60	0.00650
1.0E+05	0.45	0.00109	0.51	0.00340	0.60	0.00590
1.1E+05	0.45	0.00099	0.51	0.00310	0.60	0.00540
1.2E+05	0.45	0.00091	0.51	0.00290	0.60	0.00490
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.0E+05	0.45	<0.00001	0.51	<0.00001	0.60	<0.00001
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.0E+06	0.45	<0.00001	0.51	<0.00001	0.60	<0.00001

Table 10
The estimates of σ's with and without burning in the samples.

Burn In	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$
0	0.45	0.51	0.60
10,000	0.45	0.51	0.60
20,000	0.45	0.51	0.60

Table 11
The measures of TCR, TPR, and FPR from two different selection criterion.

	Median probability	Highest posterior probability
TCR	0.99	0.98
TPR	1	1
FPR	0.010	0.018

in terms of the posterior probabilities. Thus, we believe that both criteria would select very similar models. Based on computational efficiency, we suggest using the median probability criterion.

6. Conclusions

In this article, we develop a Bayesian variable selection method for the finite mixture model of linear regressions. Two different sets of indicators for the observations and variables are augmented into the model, and a Gibbs sampler is proposed to generate the posterior samples so that they can be used to infer the active variables within different components or sub-populations. For the g-prior assumption, we demonstrate how to incorporate ridge regression to allow for the case where the model matrix is not of full rank. A further extension for the g-prior is the use of the mixture of g-priors (Liang et al., 2012). Our method can be applied to the situation where the number of predictor variables is larger than the number of observations, a problem that has become increasingly common in practice.

We have investigated the problem of determining the number of sub-populations or mixture components, M, using various information criteria. These criteria perform well for identifying the number of components. With a prior on the number of components, a fully Bayesian approach powered by the reversible jump MCMC algorithm (Richardson and Green, 1997; Tadesse et al., 2005) can be used to infer the number of components while selecting the variables for each component. However, such an approach may dramatically increase the computational cost since the MCMC algorithm requires extra steps such as birth and death, splitting and merging, etc. Such approaches require the Metropolis–Hastings algorithm within the Gibbs sampling procedure. A recent implementation of the reversible jump MCMC for Bayesian variable selection is Liu et al. (2014), where the performance of the weighted g-prior on variable selection was studied. We shall extend their approach to our problem in our future work.

Yau and Holmes (2011) and Chung and Dunson (2009) studied the hierarchical Bayesian nonparametric mixture model, where the number of mixture components, M, is inferred based on the posterior distribution. However, they did not consider the problems of $p > n$ and high collinearity. We also aim to extend their methods to address these problems in our future work.

In addition to handling variable selection, we can extend the methodology developed in this paper to other interesting problems. First, the proposed variable selection approach is not limited to the finite mixture of regression models. It can easily be extended to latent class models (Ghosh et al., 2011) with p larger than n. Additionally, similar to Tran et al. (2012), expert knowledge can be incorporated into the proposed model through the prior distribution of the proportion of observations in each sub-population. This may increase the accuracy of classification.

Acknowledgments

The research of Chen and Lee is supported in part by the National Science Council under grant NSC 99-2118-M-006-006-MY2 and MOST 103-2118-M-006-002-MY2 (Chen) and NSC 101-2118-M-006-007- (Lee), and the Mathematics Division of the National Center for Theoretical Sciences in Taiwan. The research of Wu is supported by NSF DMS 1310391, ONR MURI N00014-10-1-0933, and DARPA MSEE FA8650-11-1-7149.

The authors would like thank the referees, the associate editor and the editor for their insightful and constructive comments that have substantially improved this article, and thank Levon Demirdjian for polishing our English writing.

References

Baragatti, M., Pommeret, D., 2012. A study of variable selection using g-prior distribution with ridge parameter. Comput. Statist. Data Anal. 56, 1920–1934.
Barbieri, M., Berger, J.O., 2004. Optimal predictive model selection. Ann. Statist. 32, 870–897.
Biernacki, C., Celeux, G., Govaert, C., 2000. Assessing a mixture model for clustering with the integrated classification likelihood. IEEE Trans. Pattern Anal. Mach. Intell. 22, 719–725.
Celeux, G., 1998. Bayesian inference for mixture: The label switching problem. In: COMPSTAT. pp. 227–232.
Celeux, G., Forbes, F., Robert, C.P., Titterington, D.M., 2006. Deviance information criteria for missing data models. Bayesian Anal. 4, 651–674.
Chen, B., 2012. Bayesian model selection in finite mixture regression, Dissertations & Theses—Gradworks, URL: http://gradworks.umi.com/35/48/3548634.html.
Chung, Y., Dunson, D.B., 2009. Nonparametric Bayes conditional distribution modeling with variable selection. J. Amer. Statist. Assoc. 104, 1646–1660.
Fan, J., Li, R., 2001. Variable selection via non-concave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96, 1348–1360.
Flegal, J.M., Haran, M., Jones, G.L., 2008. Markov chain Monte Carlo: Can we trust the third significant figure? Statist. Sci. 23, 250–260.
Frühwirth-Schnatter, Sylvia, 2006. Finite Mixture and Markov Switching Models. Springer.
Gelfand, A.E., Smith, A.F.M., 1990. Sampling-based approaches to calculating marginal densities. J. Amer. Statist. Assoc. 85, 398–409.
George, E.I., McCulloch, R.E., 1993. Variable selection via Gibbs sampling. J. Amer. Statist. Assoc. 88, 881–889.
Ghosh, J., Herring, A.H., Siega-Riz, A.M., 2011. Bayesian variable selection for latent class models. Biometrics 67, 917–925.
Gupta, M., Ibrahim, J., 2007. Variable selection in regression mixture modeling for the discovery of gene regularory nectworks. J. Amer. Statist. Assoc. 102, 867–880.
Jasra, A., Holmes, C.C., Stephens, D.A., 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. Statist. Sci. 20, 50–67.
Khalili, A., Chen, J., 2007. Variable selection in finite mixture of regression models. J. Amer. Statist. Assoc. 102, 1025–1038.
Liang, F., Paulo, R., Molina, G., Clyde, M.A., Berger, J.O., 2012. Mixtures of g priors for Bayesian variable selection. J. Amer. Statist. Assoc. 103, 410–423.
Liu, W., Zhang, B., Zhang, Z., Tao, J., Branscum, A.J., 2014+. Model selection in finite mixture of regression models: a Bayesian approach with innovative weighted g priors and reversible jump Markov chain Monte Carlo implementation. J. Stat. Comput. Simul. 2014+.
McLachlan, G.J., Peel, D., 2000. Finite Mixture Models. Wiley.
Park, T., Casella, G., 2008. The Bayesian Lasso. J. Amer. Statist. Assoc. 103, 681–686.
Richardson, S., Green, P.J., 1997. On Bayesian analysis of mixtures with an unknown number of components. J. R. Stat. Soc. Ser. B 59, 731–792.
Städler, N., Bühlmann, P., van de Geer, S., 2010. ℓ_1-Penalization for mixture regression models. TEST 19, 209–256.
Tadesse, M.G., Sha, N., Vannucci, M., 2005. Bayesian variable selection in clustering high-dimensional data. J. Amer. Statist. Assoc. 100, 602–617.
Tibshirani, R., 1996. Regression shrinkage and selection via the LASSO. J. R. Stat. Soc. Ser. B 58, 267–288.
Tibshirani, R., 2011. Regression shrinkage and selection via the Lasso: a retrospective. J. R. Stat. Soc. Ser. B 73, 273–282.
Tran, M., Nott, D.J., Kohn, R., 2012. Simultaneous variable selection and component selection for regression density estimation with mixtures of heteroscedastic experts. Electron. J. Stat. 6, 1170–1199.
Viele, K., Tong, B., 2002. Modeling with mixtures of linear regressions. Stat. Comput. 12, 315–330.
Watnik, M.R., 1998. Pay for play: Are baseball salaries based on performance? J. Stat. Educ. 6 (2).
Yau, C., Holmes, C., 2011. Hierarchical Bayesian nonparametric mixture models for clustering with variable relevance determination. Bayesian Anal. 6, 329–352.
Zellner, A., 1996. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: Bayesian Inference and Decision Techniques: Essays in Honor of Brunode Finetti. North-Holland/Elsevier, pp. 233–243.
Zhang, C.-H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942.