Statistical Methods and Applications manuscript No.
(will be inserted by the editor)

Bayesian Variable Selection in Linear Regression Models with non-normal Errors

Received: date / Accepted: date

Abstract This paper addresses two crucial issues in multiple linear regression analysis: (i) error terms whose distribution is non-normal because of the presence of asymmetry of the response variable and/or data coming from heterogeneous populations; (ii) selection of the regressors that effectively contribute to explaining patterns in the observations and are relevant for predicting the dependent variable. A solution to the first issue can be obtained through an approach in which the distribution of the error terms is modelled using a finite mixture of Gaussian distributions. In this paper we use this approach to specify a Bayesian linear regression model with non-normal errors; furthermore, by embedding Bayesian variable selection techniques in the specification of the model, we simultaneously perform estimation and variable selection. These tasks are accomplished by sampling from the posterior distributions associated with the model. The performances of the proposed methodology are evaluated through the analysis of simulated datasets in comparison with other approaches. The results of an analysis based on a real dataset are also provided. The methods developed in this paper result to perform well when the distribution of the error terms is characterised by heavy tails, skewness and/or multimodality.

Keywords Gaussian mixture model · g-prior · MCMC algorithm · median probability criterion

1 Introduction

The multiple linear regression model is a popular statistical tool used by many practitioners to discover existing relationships between a dependent variable Y and a vector of potential predictors $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$ that could describe patterns in the observed outcome. A staple assumption of this model requires

the distribution of the error terms to be Gaussian. However, in practical applications this distribution can depart from normality because of the presence of heavy tails and/or skewness. In order to deal with these issues, several authors have developed approaches that make use of more flexible parametric distributions for modelling the errors, such as the Student-t (Zellner, 1976; Sutradhar and Ali, 1986; Lange et al, 1989; Breusch et al, 1997; Fernandez and Steel, 1999), elliptical distributions (Chib et al, 1988; Galea et al, 1997; Liu, 2002; Diaz-Garcia et al, 2013), scale mixtures of Gaussians (Fernandez and Steel, 2000; Rubio and Yu, 2017), skew-elliptical distributions (Azzalini and Capitanio, 2003; Sahu et al, 2003; Azzalini and Genton, 2008) and skew-symmetric scale mixtures of Gaussians (Rubio and Genton, 2016). Some of these approaches have been proposed in the multivariate linear regression setting; inference procedures have been developed both in the frequentist and in the Bayesian frameworks.

Departures from normality can also be observed when a relevant categorical predictor is omitted from the model (Hosmer, 1974). Such an omission induces a form of unobserved heterogeneity in the model given by the presence of unknown subgroups of observations, each of which is characterized by a different regression model. The resulting distribution of the errors will be typically characterised by multimodality and/or asymmetry. For the latter issue, one can refer to the literature on skew-normal and skew-t regression models (Azzalini and Capitanio, 2003; Sahu et al, 2003; Azzalini and Genton, 2008). A way to jointly deal with heavy tails, skewness and multimodality relies on the use of finite mixture models (McLachlan and Peel, 2000; Frühwirth-Schnatter, 2006). Through an appropriate choice of its components, a finite mixture model is able to reproduce complex unknown distributions as well as local variations in the observed data. Thus, a flexible approach for modelling the distribution of the error terms in a linear regression analysis is represented by finite mixture models. This approach has been investigated using different types of mixture models, such as mixtures of Gaussians (Bartolucci and Scaccia, 2005; Soffritti and Galimberti, 2011; Dang and McNicholas, 2015), mixtures of Student-t distributions (Galimberti and Soffritti, 2014; Yao et al, 2014), mixtures of Laplace distributions (Song et al, 2014) and mixtures of scale mixtures of skew-normal distributions (Basso et al, 2010). Inference procedures under these models have been developed in the frequentist framework.

In this paper, we build upon the approach that makes use of Gaussian mixture models to develop a Bayesian specification and estimation of the multiple linear regression model. The model resulting from such an approach is equivalent to a mixture of Gaussian linear regressions with common regression coefficients among the components but different intercepts and variances. The model estimation is carried out by also considering the problem of variable selection. Several solutions to this problem have been proposed in the literature, such as penalized least squares methods (Tibshirani, 1996; Fan and Li, 2001; Zhang, 2010; Tibshirani, 2011) and Bayesian methods (George and McCulloch, 1993; Park and Casella, 2008). For a review see, for example, O'Hara and Sillanpää (2009). Some of these methods have been extended to be used

in finite mixtures of regression models (Khalili and Chen, 2007; Städler et al, 2010; Chen, 2012; Liu et al, 2015; Lee et al, 2016). In this paper, we perform variable selection by resorting to a stochastic search method via both indicators for active predictors and penalizing prior distributions for the regression coefficients. A similar approach has been employed in Lee et al (2016) where, however, the emphasis is mostly put into capturing the unobserved heterogeneity in the observations rather than the non-normality of the errors; furthermore, in the finite mixture of Gaussian linear regression models examined by Lee et al (2016) each linear regression has both component-specific linear predictors and variances.

The main contributions of this paper consist of:

- providing a Bayesian hierarchical formulation of the mixture of linear regression models approach proposed by Bartolucci and Scaccia (2005), in order to deal with regression settings where the normality assumption is not met;
- tackling explicitly the task of variable selection, by introducing a layer of latent variables that identify predictors actually contributing to explain patterns in the data;
- eliciting a weakly informative modified g-prior for the regression coefficients, which: (i) is a conjugate prior for the likelihood of mixture model's component; (ii) aids the variable selection process; (iii) induces a form of penalization, thus overcoming potential problems of overfitting;
- providing a hierarchical formulation that employs conjugate prior distributions for the parameters in the model, resulting in closed forms conditional posterior distributions;
- detailing a Monte Carlo Markov Chain (MCMC) implementation of the sampling procedure, consisting of Gibbs samplers steps based on full conditionals for the parameters;
- exploring, in a simulation study, the impact on variable selection due to different sources of deviation from the normality, such as multi-modality, heavy-tailedness, asymmetry; also, the simulation study allows us to evaluate the effects of varying quantities like sample size and number of predictors in the regression model.

The remainder of the paper is organized as follows. In Section 2, we describe the Bayesian specification of the linear regression model with a Gaussian mixture for the error terms and with variable selection, including information about the adopted prior distributions and the criteria for including a predictor and choosing the number of mixture components. In Section 3, we illustrate the main results of the simulation study, in which datasets have been generated from the proposed model and from models with other non-normal distributions for the error terms. Finally, Section 4 shows the results obtained from the analysis of a real dataset providing information about plasma beta-carotene and some of its potential determinants (Nierenberg et al, 1989; Stukel, 2008). Details about the joint posterior distribution, full conditional distributions comprising the Gibbs samplers, and MCMC algorithm for posterior sampling are reported in an Appendix, together with additional results from the simulation study. The

algorithm is scripted for the R software (R Core Team, 2017), and code files are available as supplementary material.

2 Bayesian model specification, estimation and variable selection

2.1 Model specification

Let the data be a set of n observations $\{y_i\}$ of the dependent variable Y, with $i = 1, \ldots, n$, and their associated vectors $\{\mathbf{x}_i = (x_{i1}, \ldots, x_{ij}, \ldots, x_{ip})'\}$ of p predictors. As in Bartolucci and Scaccia (2005), we consider a linear regression model with an error term distributed according to a mixture of Gaussian distributions:

$$y_i = \mathbf{x}_i'\boldsymbol{\eta} + \varepsilon_i, \quad \text{with} \quad \varepsilon_i \sim \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mu_k, \sigma_k^2),$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)'$ is the vector of regression coefficients. This is equivalent to assuming the following finite mixture of Gaussian linear regressions for y_i with common regression coefficients:

$$y_i|(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2, \mathbf{x}_i, K) \sim \sum_{k=1}^{K} w_k \cdot \mathcal{N}\big(\mu_k + \mathbf{x}_i'\boldsymbol{\eta}, \sigma_k^2\big), \tag{1}$$

where $\mathbf{w} = (w_1, \ldots, w_K)'$ is the vector of components' proportions, satisfying the conditions $w_k > 0$, for each k, and $\sum_{k=1}^{K} w_k = 1$; μ_k and σ_k^2 are, respectively, the component-specific intercept and variance of k-th linear regression; $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)'$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_K^2)'$. In the following, we consider the predictors as fixed, together with the number of components K. Thus, we do not specify any distribution for these quantities. In addition, we shall omit $\{\mathbf{x}_i\}$ as well as K from the conditioning set in the notation.

2.2 Bayesian model specification and estimation

In order to provide a Bayesian specification of the model in Equation (1) and to simultaneously embed Bayesian variable selection techniques, two different types of vectors of latent variables are required. The first type is unit-specific and is defined as $\mathbf{z}_i = (z_{i1}, \ldots, z_{iK})'$. If unit i belongs to component k then such vector will be full of zeros except for the k-th element, which will be $z_{ik} = 1$. Furthermore, $\pi(z_{ik} = 1) = w_k$, $k = 1, \ldots, K$. The overall number of such vectors in the model will be equal to n. The vector of latent variables of the second type is employed to inject variable selection into the model. Let this vector be denoted as $\mathbf{u} = (u_1, u_2, \ldots, u_p)'$. Its elements serve as indicators for active predictors: if $u_j = 0$ the j-th predictor will have no contribution to the regression, that is $\eta_j = 0$; on the contrary, if $u_j = 1$ the corresponding

predictor X_j will be considered active and will participate with a coefficient $\eta_j \neq 0$. Thus, $h = \sum_{j=1}^{p} u_j$ is the number of active predictors in the model.

The Bayesian specification of model (1) can be obtained through the following hierarchical representation of its elements:

$$\mathbf{w}|(\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K) \sim \text{Dir}\big(\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K\big), \tag{2}$$

$$\sigma_k^2|(a_{1,k}, a_{2,k}) \sim \text{IG}\big(a_{1k}, a_{2k}\big),\ k = 1, \ldots, K, \tag{3}$$

$$\mathbf{z}_i|\mathbf{w} \sim \text{Multinom}\big(w_1, \ldots, w_K\big),\ i = 1, \ldots, n, \tag{4}$$

$$u_j|v_j \sim \text{Bern}\big(v_j\big),\ j = 1, \ldots, p, \tag{5}$$

$$\boldsymbol{\eta}|(\boldsymbol{\sigma}^2, \mathbf{u}) \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{e}_0 \\ \mathbf{0}_{(p-h)\times 1} \end{bmatrix}, \begin{bmatrix} E_0 & \mathbf{0}_{h\times(p-h)} \\ \mathbf{0}_{(p-h)\times h} & \delta\, I_{p-h} \end{bmatrix}\right), \tag{6}$$

$$\mu_k|\sigma_k^2 \sim \mathcal{N}\big(m_{0,k}, M_{0,k}\big),\ k = 1, \ldots, K, \tag{7}$$

$$y_i|(\boldsymbol{\eta}, \mu_k, \sigma_k^2, \mathbf{u}, \mathbf{z}_i) \sim \mathcal{N}\big(\mu_k + \mathbf{x}_i'(\mathbf{u})\boldsymbol{\eta}(\mathbf{u}), \sigma_k^2\big),\ i = 1, \ldots, n. \tag{8}$$

For the vector of the components' proportions $\mathbf{w}$ and the k-th variance σ_k^2, we assume a Dirichlet distribution and a weekly informative inverse gamma distribution, respectively (see Equations (2) and (3)). In addition, the *a priori* conditional distribution of $\mathbf{z}_i$ for a given $\mathbf{w}$ is modelled using the multinomial distribution defined in Equation (4). These are standard choices in any mixture model (see, for example, Lee et al (2016)). Furthermore, we assume that the K variances are independent:

$$\pi(\boldsymbol{\sigma}^{\mathbf{2}}|a_{1,1}, a_{2,1}, \ldots, a_{1,K}, a_{2,K}) = \prod_{k=1}^{K} \pi(\sigma_k^2|a_{1,k}, a_{2,k}).$$

As far as the indicators for active/inactive predictors are concerned, as in George and McCulloch (1993), we assign marginal Bernoulli prior distributions (see Equation (5), where $v_j = \pi(u_j = 1)$) and assume that they are independent, so that

$$\pi(\mathbf{u}|v_1, \ldots, v_p) = \prod_{j=1}^{p} v_j^{u_j}(1 - v_j)^{1-u_j}.$$

The description of the remaining part of the hierarchical representation requires the preliminary definition of some additional sub-vectors and sub-matrices. Let $\mathbf{y}$ be the $n \times 1$ column vector of observations of the dependent variable Y, and $X(\mathbf{u})$ be the $n \times h$ matrix containing their corresponding active predictors. Furthermore, $\mathbf{y}_{[k]} = \{y_i : z_{ik} = 1\}$ contains the observations for units in k-th component, and $X_{[k]}(\mathbf{u}) = \{\mathbf{x}_i(\mathbf{u}) : z_{ik} = 1\}$ their corresponding active predictors' values. In addition, for a given vector $\mathbf{u}$, we select the elements of $\mathbf{x}_i$ which correspond to the active predictors and we use it to define the vector $\mathbf{x}_i(\mathbf{u})$; in the same way, we also define $\boldsymbol{\eta}(\mathbf{u})$ as the sub-vector of $\boldsymbol{\eta}$ containing the non-zero regression coefficients. The length of both vectors obtained in this way is equal to h. Finally, we define $\tilde{\mathbf{y}}_{[k]} = \mathbf{y}_{[k]} - X_{[k]}(\mathbf{u})\boldsymbol{\eta}(\mathbf{u})$

and $\ddot{\mathbf{y}}_{[k]} = \mathbf{y}_{[k]} - \mu_k \mathbf{1}_{n_k}$, where $\mathbf{1}_{n_k}$ is an n_k-dimensional unitary column vector, $n_k = \sum_{i=1}^{n} z_{ik}$ is the number of units allocated in component k, and $\ddot{\mathbf{y}}$ is $n \times 1$ column vector where the generic i-th element is $\ddot{y}_i = y_i - \mu_{z_{ik=1}}$.

The prior distribution on the regression coefficients $\boldsymbol{\eta}$ is specified so as to depend on $\mathbf{u}$, which falls under the category of Gibbs variable selection (Dellaportas et al, 1997; Carlin and Chib, 1995). In particular, in Equation (6) we assume that the prior distribution of $\boldsymbol{\eta}|(\boldsymbol{\sigma}^2, \mathbf{u})$ is a modified g-prior (Zellner, 1986; Lee et al, 2016), where the vector $\boldsymbol{\eta}$ is rearranged such that we have non-zero elements $\boldsymbol{\eta}(\mathbf{u})$ in the first h positions and zero elements in the remaining $p - h$ positions. The expected value and covariance matrix of the normal distribution in Equation (6) involve some vectors and matrices: I_h is identity matrix of order h; $\mathbf{0}_{p \times h}$ is a null matrix with p rows and h columns;

$$\mathbf{e}_0 = \left(X'(\mathbf{u})\Sigma^{-1}X(\mathbf{u}) + \delta I_h\right)^{-1} X'(\mathbf{u})\Sigma^{-1}\ddot{\mathbf{y}} \approx \boldsymbol{\eta}(\mathbf{u})_{WLS},$$
$$E_0 = \left(g^{-1}X'(\mathbf{u})\Sigma^{-1}X(\mathbf{u}) + \lambda I_h\right)^{-1}.$$

Σ is an $n \times n$ diagonal matrix with generic element $\Sigma_{ii} = \sigma^2_{z_{ik=1}}$; $\boldsymbol{\eta}(\mathbf{u})_{WLS}$ is the weighted least squares (WLS) solution (Carroll and Ruppert, 1988). Equation (6) also involves δ, which is a user-specified small positive constant. The positive quantity λ is called ridge parameter; together with the identity matrix I_h, it projects a form of penalization into the model. This regularizing term also avoids problematic inversion of $X'(\mathbf{u})\Sigma^{-1}X(\mathbf{u})$, especially in the case of highly correlated predictors. The prior distribution for $\boldsymbol{\eta}$ just described induces the estimation procedure to seek for sparse solutions among the possible ones.

Through Equation (7) we let the intercepts and variances within each component to be dependent, and we use normal g-prior distributions for modelling the *a priori* conditional distributions of μ_k given σ^2_k, $k = 1, \ldots, K$. The expected value and variance of μ_k given σ^2_k in Equation (7) are assumed to be defined as follows:

$$m_{0,k} = \sum_{i:z_{ik}=1} \tilde{y}_{i,[k]}/n_k = \bar{y}_{[k]},$$
$$M_{0,k} = \sigma^2_k/g_k.$$

Furthermore, we assume that the K component-specific intercepts are conditionally independent, so that $\pi(\boldsymbol{\mu}|\sigma^2_1, \ldots, \sigma^2_K) = \prod_{k=1}^{K} \pi(\mu_k|\sigma^2_k)$. The choice of the uninformative priors for the regression coefficients and the intercepts illustrated above is motivated by their properties: $\pi(\boldsymbol{\eta}|\boldsymbol{\sigma}^2, \mathbf{u})$ and $\pi(\boldsymbol{\mu}|\boldsymbol{\sigma}^2)$ lead to posterior distributions centered - respectively - around (modified) WLS and ordinary least squares (OLS) solutions.

The final part of the hierarchical representation of the proposed model concerns the conditional distribution of y_i, which is assumed to be affected by $\boldsymbol{\eta}$, μ_k, σ^2_k and $\mathbf{u}, \mathbf{z}_i$ (see Equation (8)). In particular, when the unit i belongs to the k-th component ($z_{ik} = 1$), the assumed conditional density function of y_i is $\pi(y_i|\boldsymbol{\eta}, \mu_k, \sigma^2_k, \mathbf{u}, z_{ik} = 1) = \phi\left(y_i; \mu_k + \mathbf{x}'_i(\mathbf{u})\boldsymbol{\eta}(\mathbf{u}), \sigma^2_k\right)$, where $\phi(\cdot; \mu, \sigma^2)$ is

the density function of a univariate Gaussian random variable with parameters μ, σ^2.

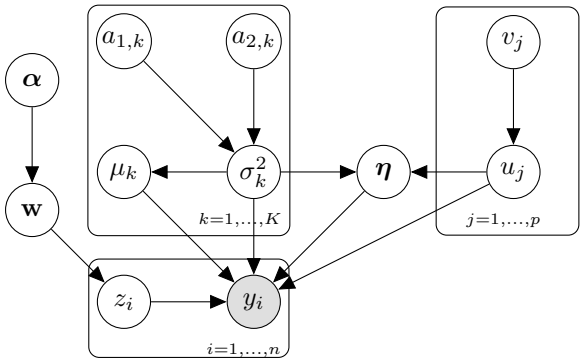

Fig. 1 Graphical representation of the hierarchical structure of the proposed model.

A graphical representation of the hierarchical structure of the resulting Bayesian model is provided in Figure 1. The complete-data likelihood of such model is

$$L(\mathbf{w}, \boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2; \mathbf{y}, \mathbf{z}, \mathbf{u}) = \prod_{k=1}^{K} w_k^{n_k} \left[\prod_{i:z_{ik}=1} \pi\left(y_i | \boldsymbol{\eta}, \mu_k, \sigma_k^2, \mathbf{u}, \mathbf{z}_i\right) \right]. \tag{9}$$

We shall remark we are assuming independence among the n units, as well as between $\boldsymbol{\eta}$ and each μ_k. Combining Equation (9) with the previously discussed priors we obtain the posterior distribution of our model: we have derived the appropriate full conditionals and implemented an MCMC algorithm with sequential Gibbs samplers steps. A sketch of the algorithm and Gibbs samplers are provided in full details in the Appendix A. Values for the hyper-parameters and the quantities λ, g, g_k, and δ are provided in Section 3.1.

2.3 Dealing with label switching

A well-known issue with mixture models is the multimodality nature of the likelihood: any arbitrary permutation of the labels of the components yields exactly the same evidence, thus producing identifiability problems (Frühwirth-Schnatter, 2006). Many approaches have been suggested, such as random permutations at each iteration of the estimating algorithm, or identifiability constraints on the components' parameters. In order to break this invariance of the likelihood, we impose a non-decreasing order among the K intercepts $\mu_1 > \mu_2 > \cdots > \mu_K$, sorting all the related quantities at each sweep of the MCMC algorithm. Other sortings can be defined, for example considering other parameters such as components' proportions $\mathbf{w}$ or components' variances $\boldsymbol{\sigma}^2$. We remark that, in general, label switching is a desirable feature of

samplers in an MCMC algorithm, as it highlights the capability of chains to explore different regions of the posterior distribution of the parameters in the model. However, label switching proves to be an issue when computing non-invariant posterior quantities from the sampled values, such as posterior means and posterior standard deviations (i.e., for the regression coefficients). Some techniques to post-process chains exhibiting label switching are discussed in Papastamoulis (2016).

2.4 Bayesian variable selection and allocation criterion for the sample observations

To decide whether a predictor should be included or not, we adopt the median probability criterion (Barbieri et al, 2004) on the posterior probability of including a predictor. Denoting with T the effective number of draws from the posterior distribution, and indexing with superscript (t) the MCMC iterations after burn-in, if $\bar{\pi}(u_j|\mathbf{y}) = \bar{\pi}(u_j = 1|\mathbf{y}) \approx T^{-1}\sum_{t=1}^{T} u_j^{(t)}$ is greater than 0.5 then predictor X_j is in the model, and discarded otherwise. After predictors are selected, posterior quantities (mean, standard deviation, credible interval) for the regression coefficients are computed through Rao-Blackwellization method (Gelfand and Smith, 1990). While allocating the sample observations into the K components of the mixture is not the primary focus of our model, partitioning units into groups might highlight features of the dataset being analyzed. This task can be accomplished through the maximum-a-posteriori (MAP) rule: unit i is allocated into the cluster having the highest average posterior probability among the K components, where each average posterior probability is computed as

$$\bar{\pi}(z_{ik}|\mathbf{y}) = T^{-1}\sum_{t=1}^{T}\pi\big(z_{ik} = 1|\mathbf{w}^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\eta}^{(t)}, \boldsymbol{\sigma}^{2^{(t)}}, \mathbf{y}\big).$$

This allocation strategy is similar to the one adopted in Lee et al (2016), and it is marginal with respect to the predictors selected in the final model.

2.5 Choosing the number of components

Deviance Information Criterion (Spiegelhalter et al, 2002) is a natural tool for Bayesian model selection, where goodness-of-fit and complexity of the model are jointly estimated through the *deviance* random variable. The conventional definition of the DIC is not adequate to deal with the presence of latent variables and/or parameters with non-continuous distributions. As it is the case with our mixture modeling context, we overcome this limitation by relying on one of the alternative versions of DIC proposed in Celeux et al (2006) (labeled, in the same manuscript, as DIC_3):

$$\text{DIC} = -4\text{E}[\log\pi(\mathbf{y}|\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2)] + 2\log\hat{\pi}(\mathbf{y}),$$

where both terms can be computed with values sampled at each iteration of the MCMC algorithm. More specifically, we have:

$$\mathrm{E}[\log \pi(\mathbf{y}|\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\eta}, \boldsymbol{\sigma}^2)] = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{n}\log\left\{\sum_{k=1}^{K} w_k^{(t)} \cdot \phi\big(y_i; \mu_k^{(t)} + \mathbf{x}_i'\boldsymbol{\eta}^{(t)}, \sigma_k^{2^{(t)}}\big)\right\},$$

and

$$\hat{\pi}(\mathbf{y}) = \prod_{i=1}^{n}\bar{\pi}(y_i), \text{ where } \bar{\pi}(y_i) = \frac{1}{T}\sum_{t=1}^{T}\left\{\sum_{k=1}^{K} w_k^{(t)} \cdot \phi\big(y_i; \mu_k^{(t)} + \mathbf{x}_i'\boldsymbol{\eta}^{(t)}, \sigma_k^{2^{(t)}}\big)\right\}.$$

The model selected is then the one with associated lowest DIC value among the possible candidates.

3 Simulation study

3.1 Design for the simulation study

In order to assess the performance of our model and compare it with similar approaches under different circumstances, we select four main factors to be altered during the generating process of the simulated datasets. In particular, we consider:

- three sample sizes: $n \in \{100, 250, 500\}$;
- three increasing numbers $(p - h)$ of non-active predictors to be appended to the $h = 5$ active predictors, leading to three totals $p \in \{10, 25, 50\}$;
- four probability distributions for the error term, where:
  Scenario 1: a mixture of $K_{\text{true}} = 2$ normal distributions, with: $\boldsymbol{\mu} = (20, -18)$, $\boldsymbol{\sigma}^2 = (0.8, 0.4)$, $\mathbf{w} = (0.7, 0.3)$;
  Scenario 2: a Student-t distribution with degrees of freedom $\nu = 1$;
  Scenario 3: a skew-normal distribution (Azzalini, 2013) with location, scale, and skewness parameters $\xi = 0, \omega = 1, \alpha = 25$;
  Scenario 4: a skew-t distribution (Azzalini, 2013) with location, scale, skewness, and degrees of freedom parameters $\xi = 0, \omega = 1, \alpha = 25, \nu = 5$;
- two correlation intensities for the predictors $(X_1, \ldots, X_p)'$, that is: $\mathrm{corr}(X_i, X_j) = \rho^{|i-j|}$ with $\rho \in \{0, 0.5\}$.

We set the active predictors' coefficients as $\boldsymbol{\eta}(\mathbf{u}) = (2.5, -5, 3, -2, 4)'$ for all scenarios. When considering **Scenarios 2-3-4**, we set the model intercept to be 8.6. The data generating process (DGP) for the simulated datasets is the following:

1. set all the parameters according to the **Scenario**;
2. generate the predictors' matrix X from a p-variate normal distribution with zero mean vector and correlation structure: $\mathrm{corr}(X_i, X_j) = \rho^{|i-j|}$;
3. compute the linear predictor $X\boldsymbol{\eta}$, with $\boldsymbol{\eta} = \big(\boldsymbol{\eta}(\mathbf{u})', \mathbf{0}_{1\times(p-h)}\big)'$;

4. simulate an $n \times 1$ vector $\boldsymbol{\varepsilon}$ of errors, distributed according to the **Scenario**;
5. append the error term $\boldsymbol{\varepsilon}$ such that: $\mathbf{y} = X\boldsymbol{\eta} + \boldsymbol{\varepsilon}$.

We remark we are assuming an additive noise model when generating each dataset. For each **Scenario 1-2-3-4**, we generate 100 independent datasets for each of the six combinations: $\{(n = 100, p = 50, \rho = 0.0), (n = 100, p = 50, \rho = 0.5), (n = 250, p = 25, \rho = 0.0), (n = 250, p = 25, \rho = 0.5), (n = 500, p = 10, \rho = 0.0), (n = 500, p = 10, \rho = 0.5)\}$. We set $\lambda = p^{-1}$, $g = n$, $\delta = 10^{-6}$, and $g_k = 1$ for every k. As far as hyper-parameters are concerned, we set: $\alpha_k = 1$, $a_{1,k} = a_{2,k} = 1$, for every k; $v_j = 0.5$ for every j.

Models with $K \in \{1, 2, 3\}$ are considered in each scenario. For $K = 1$, our model coincides with a multiple penalized ridge regression. The other competitor is the LASSO (Tibshirani, 1996) approach for variable selection. As we are interested in choosing predictors, we focus on the following quantities computed from the number of True Positive (TP), True Negative (TN) and False Positive (FP):

$$\text{TSR} = \frac{\text{TP} + \text{TN}}{p} \qquad \text{TPR} = \frac{\text{TP}}{h} \qquad \text{FPR} = \frac{\text{FP}}{p - h}$$

where TSR is the True Selection Rate (also known as *accuracy*), TPR is the True Positive Rate (*sensitivity* or *recall*), and FPR is the False Positive Rate (*fall-out*). In our context: True Positive is an active predictor (u_j=1) that is correctly selected as active ($\bar{\pi}(u_j|\mathbf{y}) > 0.5$); True Negative is an inactive predictor ($u_j = 0$) that is correctly not selected ($\bar{\pi}(u_j|\mathbf{y}) \leq 0.5$); False Positive is an inactive predictor ($u_j = 0$) that is wrongly selected as active ($\bar{\pi}(u_j|\mathbf{y}) > 0.5$). These quantities are computed in each independent dataset and then averaged across the 100 replications: we thus report average TSR ($\overline{\text{TSR}}$), average TPR ($\overline{\text{TPR}}$), and average FPR ($\overline{\text{FPR}}$). Additional information about the performances of the proposed models are available in Tables in Appendix B.

3.2 Finite mixture of normal distributions

First batch of simulations is used to assess: (i) the performance of our proposed methodology when the error structure is a mixture of normal distributions (that is, when the model is correctly specified); (ii) the validity of the adopted selection criterion discussed in Section 2.5.

Table 1 provides a summary of the results obtained when $\rho = 0.0$. In particular, for each examined setting, we report $\overline{\text{TSR}}$, $\overline{\text{TPR}}$ and $\overline{\text{FPR}}$ for the LASSO, ridge regression ($K = 1$), and our proposed model with $K = 2$ and $K = 3$. Rows with K_{DIC} refer to average TSR, average TPR and average FPR across the 100 replicated datasets with varying K, selected among the values $K \in \{1, 2, 3\}$ according to the lowest DIC value for each dataset. Finally, as a measure of reliability of the selection criterion, Table 1 contains the frequency distributions for the values of K selected by DIC, that is the number of datasets in which $K_{\text{DIC}} = \{1, 2, 3\}$.

Setting ($\rho = 0.0$) n	p	Model	$\overline{\text{TSR}}$	$\overline{\text{TPR}}$	$\overline{\text{FPR}}$	K_{DIC} 1	2	3
100	50	LASSO	0.84	0.44	0.11			
		$K = 1$	0.90	0.01	0.00			
		$K = 2$	0.95	0.94	0.05	0	59	41
		$K = 3$	0.89	0.95	0.12			
		K_{DIC}	0.93	0.99	0.07			
250	25	LASSO	0.76	0.88	0.27			
		$K = 1$	0.85	0.28	0.00			
		$K = 2$	0.98	0.99	0.02	0	87	13
		$K = 3$	0.97	1.00	0.04			
		K_{DIC}	0.98	1.00	0.03			
500	10	LASSO	0.73	1.00	0.54			
		$K = 1$	0.92	0.84	0.01			
		$K = 2$	0.99	1.00	0.02	0	90	10
		$K = 3$	0.98	1.00	0.02			
		K_{DIC}	0.99	1.00	0.03			

Table 1 Results for **Scenario 1** of simulation study in terms of TSR, TPR, and FPR, averaged across the 100 independent datasets; last column provides the frequency distribution of the selected values of K according to DIC. Error: mixture of two normal distributions; correlation: $\rho = 0.0$.

When comparing the different methods, we see that our proposed approach tends to outperform LASSO and penalized ridge regression ($K = 1$), both when the correct number of component is considered ($K = 2$) and when K is selected according to DIC. It is worth mentioning that LASSO seems to be characterized by the largest values for $\overline{\text{FPR}}$ for all the examined values of the n-to-p ratio, thus suggesting a tendency to include inactive predictors. On the contrary, the smallest values for $\overline{\text{TPR}}$ appear to be associated with penalized ridge regression, even though this method can achieve appreciable $\overline{\text{TSR}}$. This could be explained by the fact that in this first batch of simulations penalized ridge regression tends to select no predictor, irrespective of its role (see Table 7 in Appendix B reporting the fraction of datasets in which each active predictor is actually selected). Given the number of truly non-active predictors $p - h$ considered in this simulation study, this behavior inflates the corresponding $\overline{\text{TSR}}$. As far as the behavior of the proposed approach is concerned, as expected, having a higher n-to-p ratio in the samples improves the quality of the results, producing slightly higher $\overline{\text{TPR}}$ and smaller $\overline{\text{FPR}}$. Furthermore, correlation among predictors seems to have a limited impact on the performance of the proposed method: comparing Table 1 with Table 9 in Appendix B, we can note only slight differences in *accuracy*, *recall* and *fall-out* both when $K = 2$ and when K is selected according to DIC.

Some additional information about the posterior distributions of the regression coefficients obtained using the proposed Bayesian strategy are reported in Appendix B (Tables 8 and 11). According to them, posterior distributions obtained using mixtures appear to have lower variability than the ones produced by penalized ridge regression. Finally, in this scenario the reliability of DIC as model selection criterion can be evaluated by considering the number

of samples for which $K_{\text{DIC}} = 2$. It is possible to note that this reliability tends to decrease as the n-to-p ratio decreases, or when predictors are correlated.

3.3 Heavy-tailed distribution for the errors

Second batch of simulations serves us to explore the behavior of the compared methodologies when the non-normality of the errors is declined into heavy-tailed distribution and, in particular, a Student-t distribution with $\nu = 1$ degrees of freedom. In general, the results obtained on data generated according to **Scenario 2** seem to confirm those presented is Section 3.2. In particular, our proposed approach seems superior to the others in terms of $\overline{\text{TPR}}$ in all settings (see Table 2 and Table 14 in Appendix B). However, it is worth mentioning the large values of $\overline{\text{FPR}}$ obtained using mixture models ($K = 2, 3$ and K_{DIC}) when $n = 100$ and $p = 50$, providing some evidence of a tendency to select many inactive predictors when the n-to-p ratio is low. Furthermore, the behavior of LASSO and penalized regression is consistent with that observed on data generated according to **Scenario 1**.

As in the previous batch of simulations, the posterior distributions for the regression coefficients obtained considering K_{DIC} seem to show a lower variability than those obtained using penalized ridge regression. When DIC is considered, most of the times it leads to select $K = 3$, thus suggesting that a 3 component mixture is needed to approximate a Student-t distribution with $\nu = 1$ degrees of freedom.

Setting ($\rho = 0.0$) n	p	Model	$\overline{\text{TSR}}$	$\overline{\text{TPR}}$	$\overline{\text{FPR}}$	K_{DIC} 1	2	3
100	50	LASSO	0.84	0.54	0.13			
		$K = 1$	0.93	0.32	0.00			
		$K = 2$	0.69	0.85	0.32	0	31	69
		$K = 3$	0.54	0.92	0.51			
		K_{DIC}	0.63	0.94	0.40			
250	25	LASSO	0.74	0.77	0.27			
		$K = 1$	0.89	0.47	0.00			
		$K = 2$	0.99	0.99	0.02	0	10	90
		$K = 3$	0.96	0.99	0.05			
		K_{DIC}	0.96	1.00	0.05			
500	10	LASSO	0.66	0.80	0.48			
		$K = 1$	0.75	0.50	0.00			
		$K = 2$	0.98	0.98	0.01	0	3	97
		$K = 3$	0.97	0.99	0.04			
		K_{DIC}	0.98	0.99	0.04			

Table 2 Results for **Scenario 2** of simulation study in terms of TSR, TPR, and FPR, averaged across the 100 independent datasets; last column provides the frequency distribution of the selected values of K according to DIC. Error: Student-t distribution; correlation: $\rho = 0.0$.

3.4 Asymmetrically distributed errors

While the previous analyses are conducted setting the prior expected value of $\boldsymbol{\eta}$ equal to $\boldsymbol{\eta}_{\text{WLS}}$, data generated according to **Scenario 3** are analyzed considering $\mathbf{e}_0 = \mathbf{0}$. This change is motivated by some preliminary results (not shown here) highlighting a poor performance of the proposed method in terms of $\overline{\text{FPR}}$. The idea that motivates the choice of $\mathbf{e}_0 = \mathbf{0}$ is to strengthen the shrinkage effect of the prior and to knock-down predictors which are truly not active predictors in the DGP. As seen in Table 3, when $\rho = 0$ there are only slight differences in $\overline{\text{TPR}}$ among the methods considered in this batch of simulations. Similar results are obtained with a higher correlation intensity (see Table 19, Appendix B). Nevertheless, as Tables 18 and 21 in Appendix B show, mixtures seem to lead to a higher precision in the posterior distributions.

Setting ($\rho = 0.0$)		Model	$\overline{\text{TSR}}$	$\overline{\text{TPR}}$	$\overline{\text{FPR}}$	K_{DIC}		
n	p					1	2	3
100	50	LASSO	0.78	0.98	0.24	0	4	96
		$K = 1$	0.98	0.89	0.01			
		$K = 2$	0.94	0.96	0.06			
		$K = 3$	0.73	0.94	0.29			
		K_{DIC}	0.73	0.94	0.29			
250	25	LASSO	0.73	1.00	0.34	0	9	91
		$K = 1$	0.99	1.00	0.01			
		$K = 2$	0.99	1.00	0.02			
		$K = 3$	0.97	1.00	0.04			
		K_{DIC}	0.97	1.00	0.04			
500	10	LASSO	0.69	1.00	0.63	0	0	100
		$K = 1$	1.00	1.00	0.01			
		$K = 2$	0.99	1.00	0.02			
		$K = 3$	0.99	1.00	0.03			
		K_{DIC}	0.99	1.00	0.03			

Table 3 Results for **Scenario 3** of simulation study in terms of TSR, TPR, and FPR, averaged across the 100 independent datasets; last column provides the frequency distribution of the selected values of K according to DIC. Error: skew-normal distribution; correlation: $\rho = 0.0$.

3.5 Heavy-tailed and asymmetrically distributed errors

There seems to be a ranking of performance based on which is the source of departure from normality. Multimodality is, as expected, the most obvious case where our model is preferable instead of simple penalized ridge regression. Heavy-tailed distributed errors is another situation where the flexibility of our model produces better performance in terms of variable selection. Asymmetrically distributed errors, on the other hand, seems to not affect too much the properties of ridge regression, especially if there is only a mild

asymmetry in the DGP additive noise term. We test the competing methodologies when the distribution of the errors possesses both heavy-tailedness and asymmetry to assess whether a combined detrimental effect might show up as degraded $\overline{\text{TSR}}$, $\overline{\text{TPR}}$ and $\overline{\text{FPR}}$. As in Section 3.4, we consider $\mathbf{e}_0 = \mathbf{0}$.

According to Table 4, the main differences between penalized ridge regression and the method proposed in this paper seem to emerge when considering $n = 100$ and $p = 50$. In this setting, our method outperforms penalized ridge regression in terms of $\overline{\text{TPR}}$, but continues to suffer from high $\overline{\text{FPR}}$. In the other two settings, apparently *accuracy*, *recall* and *fall-out* are only slightly affected by the choice of K. However, this choice has an impact on the precision of the posterior distribution, as shown in Table 23 in Appendix B, where using a mixture with either three or K_{DIC} components leads to a slightly better performance in terms of accuracy of the estimated posterior distribution of the regression coefficients. Similar conclusions holds when $\rho = 0.5$ (see Tables 24 and 26, Appendix B).

Setting ($\rho = 0.0$)		Model	$\overline{\text{TSR}}$	$\overline{\text{TPR}}$	$\overline{\text{FPR}}$	K_{DIC}		
n	p					1	2	3
100	50	LASSO	0.80	0.91	0.22	0	4	96
		$K = 1$	0.96	0.59	0.00			
		$K = 2$	0.95	0.90	0.04			
		$K = 3$	0.77	0.92	0.25			
		K_{DIC}	0.77	0.93	0.25			
250	25	LASSO	0.70	1.00	0.37	0	6	94
		$K = 1$	0.99	0.96	0.01			
		$K = 2$	0.99	1.00	0.01			
		$K = 3$	0.97	1.00	0.04			
		K_{DIC}	0.97	1.00	0.04			
500	10	LASSO	0.69	1.00	0.61	0	0	100
		$K = 1$	0.99	0.99	0.01			
		$K = 2$	0.99	1.00	0.01			
		$K = 3$	0.99	1.00	0.02			
		K_{DIC}	0.99	1.00	0.02			

Table 4 Results for **Scenario 4** of simulation study in terms of TSR, TPR, and FPR, averaged across the 100 independent datasets; last column provides the frequency distribution of the selected values of K according to DIC. Error: skew-t distribution; correlation: $\rho = 0.0$.

4 Real world dataset: plasma beta-carotene and its potential determinants

Beta-carotene, along with other micronutrients such as retinol, is believed to be potentially protective against some diseases, especially cancer. Although numerous observational and retrospective studies tried to capture this beneficial effect, it is yet controversial to claim that evidence was found (see Stukel (2008) and references therein), but interest in which determinants of a person

lifestyle might alter this nutrient's concentration levels is still prominent. We use data originally collected in Nierenberg et al (1989) and further analyzed in Stukel (2008), where the dependent variable we focus on is the level of beta-carotene in the plasma, measured as nanograms per milliliters (ng/ml) in the original scale (Y). We select this dataset because of the heterogeneity of the observed beta-carotene plasma levels in the dataset, which was already mentioned in Schlattmann (2009), hinting at a rewarding use of models that can account for deviation from the normality assumption of the errors' distribution. In addition, a set of potential determinants that could explain these beta-carotene levels are available, and employable as predictors in a regression setting. In particular, we have at our disposal plasma beta-carotene concentrations $\mathbf{y}$ for $n = 315$ subjects, and the following additional subject-specific information: age of the subject (`age`); calories consumed per day (`calories`); grams of fat consumed per day (`fat`); grams of fiber consumed per day (`fiber`); number of alcoholic drinks consumed per week (`alcohol`); grams of cholesterol consumed per day (`chol`); micrograms (per day) of beta-carotene provided by diet (`betadiet`); micrograms (per day) of retinol provided by diet (`retdiet`). Furthermore, we have:

- `smokestat`: smoking status of the subject; we dichotomized the original predictor by collapsing "Never" and "Former" into 0 and "Current smoker" into 1;
- `vituse`: vitamine use of the subject; we dichotomized the original predictor by collapsing "Yes" and "Yes (fairly often)" into 1 and "No" into 0;
- `bmi`: body-mass index, also known as Quetelet index, measured as weight (in kg) divided by the square of height (in meters).

Previous analysis on this very same dataset usually dealt with a logarithmic transformation of the data (Nierenberg et al, 1989; Stukel, 2008), as the resulting marginal distribution becomes more symmetrical in comparison with the distribution on the original scale.

Instead of resorting on this transformation, with our model we are able to frame the regression with Y as our dependent variable; furthermore, we work with the standardized form of the design matrix X, so that for each continuous predictor we have $\mathrm{E}[X_j] = 0$ and $\mathrm{Var}[X_j] = 1$. From the 315 observations, we remove three units as potentially bad records from the data entry step: for example, one removed observation is a woman who reported 209 alcoholic drinks per week. These units were also removed in Stukel (2008). We want to point out that our model can accommodate for deviations related to the dependent variable Y, while the three removed units are anomalous observations with respect to the predictors.

An exploratory analysis of the predictors shows a set of correlated predictors (`fat`, `calories`, `alcohol`, `chol`), while `bmi` is highlighted as correlated with the dependent variable but not with the aforementioned set. We observe a p-value of 0.0374 for the Pearson's χ^2 test, with Yates' continuity correction, between the smoking status and the use of vitamin, which suggests the two categorical predictors to be significantly associated. The competing models we

fit are: LASSO; penalized ridge regression; our model with $K \in \{2, 3, 4\}$. For LASSO, the penalty parameter is selected via cross-validation, with the functionality provided in the package `R glmnet` (Simon et al, 2011). We provide results for two possible values of $\mathbf{e}_0$, which are $\mathbf{e}_0 = \mathbf{0}$ and $\mathbf{e}_0 = \boldsymbol{\eta}_{\text{WLS}}$, to look for sensitivity issues with respect to prior specification. In addition, we explore models with and without the two predictors `vituse` and `smokestat`, as they were dichotomized by an arbitrary choice of collapsing of the original categories.

Models without (`vituse`, `smokestat`) as potential predictors			
Prior specification	K	DIC	Predictors selected ($\bar{\pi}(u_j\|\mathbf{y})$)
$\mathbf{e}_0 = \mathbf{0}$	1	4139.52	`bmi` (0.65)
	2	3867.93	`fiber` (0.88)
	3	3795.08	`bmi` (0.77)
	4	3797.76	`bmi` (0.90)
$\mathbf{e}_0 = \boldsymbol{\eta}_{\text{WLS}}$	1	4133.15	`bmi` (0.66), `fiber` (0.67)
	2	3861.94	`bmi` (0.53), `fiber` (0.98)
	3	**3794.85**	`bmi` (0.80)
	4	3796.48	`bmi` (0.96)
		LASSO	`age`, `bmi`, `fat`, `fiber`, `chol`, `betadiet`

Table 5 Predictors selected by each compared model; best model in bold font; posterior probabilities of inclusion are reported in brackets.

When the analysis is carried out without `vituse` and `smokestat`, most of the models select `bmi` as the only relevant predictor (see Table 5). Sometimes, `bmi` is accompanied by `fiber`, and the best fit - controlling also for complexity - is associated to our model with $K = 3$ and $\mathbf{e}_0 = \boldsymbol{\eta}_{\text{WLS}}$. The results suggest that, once the heterogeneity in the data is controlled by the mixture layer of the model, the only relevant predicting information is the Quetelet index (the BMI), which is interpretable as a proxy of one person's lifestyle with respect to quality of their diet. The Rao-Blackwellized posterior mean of $\eta_{\texttt{bmi}}$ is equal to -16.49 with a posterior standard deviation of 4.22: an increasing body-mass index, usually associated with low-quality diet habits and low physical activity levels, is reflected by a decreasing plasma beta-carotene concentration. When selected, $\eta_{\texttt{fiber}}$ shows a positive sign, as expected because of the nature of beta-carotene as a micronutrient prevalently found in vegetables and fruits, both high in fiber quantities.

When `vituse` and `smokestat` are also used as potential predictors in the analysis, `bmi` seems to be the only predictor useful to explain patterns in the data once heterogeneity and non-normality are accounted for by the 3-components mixture (see Table 6). The magnitude and variability of $\eta_{\texttt{bmi}}$ are coherent with the model already discussed (posterior mean equal to -16.61, posterior standard deviation of 4.14). As far the LASSO approach is concerned, we see that LASSO tends to select most of the regressors in an attempt to capture as much as possible of the extra-variability and non-normality of the errors. In Figure 2, we visualize the posterior distribution of the parameters

Models with (`vituse`, `smokestat`) as potential predictors			
Prior specification	K	DIC	Predictors selected ($\bar{\pi}(u_j\|\mathbf{y})$)
$\mathbf{e}_0 = \mathbf{0}$	1	4136.43	`bmi` (0.65)
	2	3880.36	-
	3	3802.18	-
	4	3811.17	-
$\mathbf{e}_0 = \boldsymbol{\eta}_{\text{WLS}}$	1	4138.84	`bmi` (0.59)
	2	3888.49	-
	3	**3799.82**	`bmi` (0.54)
	4	3811.01	-
		LASSO	`age`, `bmi`, `fat`, `fiber`, `alcohol`, `chol`, `betadiet`, `retdiet`, `vituse`, `smokestat`

Table 6 Predictors selected by each compared model; best model in bold font; posterior probabilities of inclusion are reported in brackets.

associated to the selected predictor (`bmi`), under the model $K = 3$, with prior specification $\mathbf{e}_0 = \boldsymbol{\eta}_{\text{WLS}}$ and (`vituse`, `smokestat`) included among the potential predictors. For comparison, we also provide the posterior distribution for $K = 1$ and same prior specification. The posterior distribution of $\eta_{\texttt{bmi}}$ under our model with $K = 3$ (*light-blue* curve) is more concentrated and exhibits less uncertainty when compared to the posterior distribution of the same parameter under penalized ridge regression model (*red* curve). The posterior distributions for the other parameters in the model (see Figure B.4 in Appendix B) show that there is clear separation among the components of the mixture, with respect to the proportions, intercepts, and variances. We also performed the analysis using the log-scale of the dependent variable Y without `vituse` and `smokestat` as predictors: in this situation, the number of components according to DIC is equal to $K_{\text{DIC}} = 2$, thus suggesting that the transformation is not sufficient to account for the non-Gaussianity of the conditional distribution of Y. Moreover, the selected predictors according to this model are `age`, `bmi`, `fiber`, `chol`, and `betadiet`.

5 Conclusions

We have provided a Bayesian variable selection procedure within the framework of a multiple linear regression model with non-normal errors. A flexible distribution for the error terms has been structured through a Gaussian mixture model. This has led to a mixture of Gaussian regression models with common regression coefficients among the components but specific intercepts and variances. Variable selection has been achieved by introducing a layer of latent variables that serve as indicators for active predictors.

We have explored some simulation environments to study the performances of our proposed approach. Appreciable results have been obtained in the presence of heavy tailed or multimodal errors, both in terms of variable selection and in posterior summary quantities.

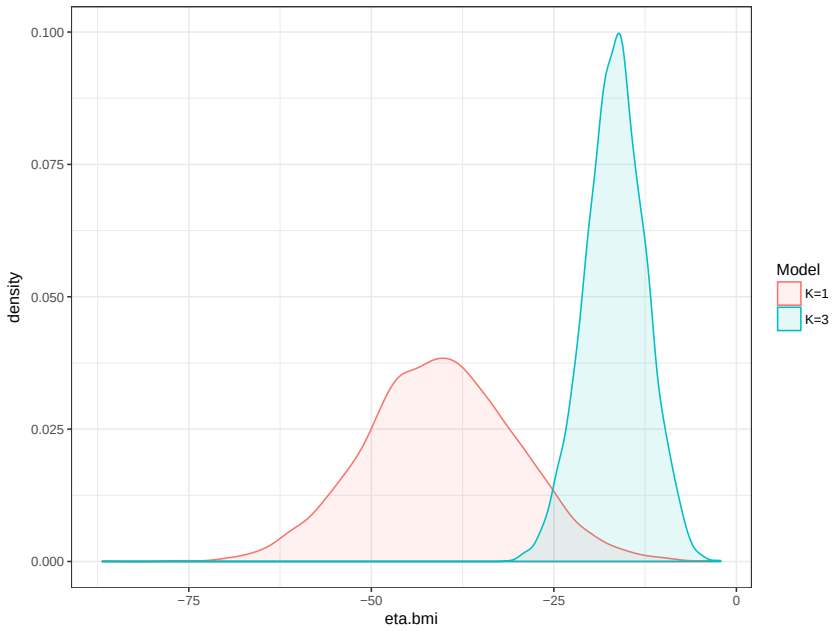

Fig. 2 Posterior distributions of $\eta_{\texttt{bmi}}$, sampled under: penalized ridge regression model with $K = 1$, red curve; our model with $K = 3$ and $\mathbf{e}_0 = \boldsymbol{\eta}_{\text{WLS}}$, light-blue curve.

It is worth remarking that K (the number of mixture components) has been kept fixed in the MCMC algorithm developed in this paper. In order to select an optimal value for K, we have suggested a *post hoc* choice based on the Deviance Information Criterion. However, we feel that a fully Bayesian approach might improve the overall appeal of the methodology. In order to do this, the number of components should be treated as an unknown model parameter. Similarly to Liu et al (2015), a reversible jump MCMC algorithm (Richardson and Green, 1997) could be developed, by introducing a prior distribution for K. It would be interesting to evaluate whether the increase in the computational costs due to the extra steps related to birth-and-death and splitting-and-merging of components is counterbalanced by substantial improvements in the performances.

As a further development, a multivariate extension of the proposed variable selection procedure could be investigated, which considers a vector of D correlated observed outcomes $\mathbf{Y} = (Y_1, \ldots, Y_D)'$. Departures from a multivariate Gaussian distribution for the error terms are suitably accounted for by using a multivariate Gaussian mixture model, with possibly non-diagonal component covariance matrices (see, for example, Soffritti and Galimberti, 2011; Galimberti et al, 2016). In this scenario, one could examine either the specific case of multivariate regression models (Srivastava, 2002) or the general framework of seemingly unrelated linear regression models (Srivastava and Giles, 1987). In the first situation, an active predictor is assumed to affect all the D outcomes,

while in the second one specific sets of active predictors can be considered for each observed outcome.

Acknowledgments

We would like to thank the two anonymous referees, who provided helpful suggestions and comments to improve the overall quality of the manuscript.

References

Azzalini A (2013) The skew-normal and related families, vol 3. Cambridge University Press

Azzalini A, Capitanio A (2003) Distributions generated by perturbation of symmetry with emphasis on a multivariate skew-t distribution. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65:367–389

Azzalini A, Genton M (2008) Robust likelihood methods based on the skew-t and related distributions. International Statistical Review 76:106–129

Barbieri MM, Berger JO, et al (2004) Optimal predictive model selection. The Annals of Statistics 32(3):870–897

Bartolucci F, Scaccia L (2005) The use of mixtures for dealing with non-normal regression errors. Computational Statistics & Data Analysis 48(4):821–834

Basso R, Lachos V, Cabral C, Ghosh P (2010) Robust mixture modeling based on scale mixtures of skew-normal distributions. Computational Statistics & Data Analysis 54:2926–2941

Breusch T, Robertson J, Welsh A (1997) The emperor's new clothes: a critique of the multivariate t regression model. Statistica Neerlandica 51:269–286

Carlin BP, Chib S (1995) Bayesian model choice via Markov chain Monte Carlo methods. Journal of the Royal Statistical Society Series B (Methodological) 57(3):473–484

Carroll RJ, Ruppert D (1988) Transformation and weighting in regression, vol 30. CRC Press

Celeux G, Forbes F, Robert CP, Titterington DM (2006) Deviance information criteria for missing data models. Bayesian Anal 1(4):651–673, DOI 10.1214/06-BA122, URL https://doi.org/10.1214/06-BA122

Chen B (2012) Bayesian model selection in finite mixture regression. Dissertations & Theses - Gradworks.

Chib S, Tiwari R, Jammalamadaka S (1988) Bayes prediction in regressions with elliptical errors. Journal of Econometrics 38:349–360

Dang UJ, McNicholas PD (2015) Families of parsimonious finite mixtures of regression models. In: Advances in Statistical Models for Data Analysis, Springer, pp 73–84

Dellaportas P, Forster J, Ntzoufras I (1997) On Bayesian model and variable selection using MCMC. Technical report. Technical report, Department of Statistics, Athens University of Economics and Business, Athens Greece

Diaz-Garcia J, Rojas M, Leiva-Sanchez V (2013) Influence diagnostics for elliptical multivariate linear regression models. Communications in Statistics - Theory and Methods 32:625–642

Fan J, Li R (2001) Variable selection via non-concave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96:1348–1360

Fernandez C, Steel M (1999) Multivariate Student-t regression models: pitfalls and inference. Biometrika 86:153–167

Fernandez C, Steel M (2000) Bayesian regression analysis with scale mixtures of normals. Econometric Theory 80:80–101

Frühwirth-Schnatter S (2006) Finite Mixture and Markov Switching Models. Springer Science & Business Media, New York

Galea M, Paula G, Bolfarine H (1997) Local influence in elliptical linear regression models. Statistician 46:71–79

Galimberti G, Soffritti G (2014) A multivariate linear regression analysis using finite mixtures of t distributions. Computational Statistics & Data Analysis 71:138–150

Galimberti G, Scardovi E, Soffritti G (2016) Using mixtures in seemingly unrelated linear regression models with non-normal errors. Statistics and Computing 26(5):1025–1038

Gelfand AE, Smith AF (1990) Sampling-based approaches to calculating marginal densities. Journal of the American Statistical Association 85(410):398–409

George EI, McCulloch RE (1993) Variable selection via Gibbs sampling. Journal of the American Statistical Association 88(423):881–889

Hosmer D (1974) Maximum likelihood estimates of the parameters of a mixture of two regression lines. Communications in Statistics - Theory and Methods 3:995–1006

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

Lange K, Little R, Taylor J (1989) Robust statistical modeling using the t distribution. Journal of the American Statistical Association 84:881–896

Lee K, Chen R, Wu Y (2016) Bayesian variable selection for finite mixture model of linear regressions. Computational Statistics & Data Analysis 95:1–16

Liu S (2002) Local influence in multivariate elliptical linear regression models. Linear Algebra and its Applications 354:159–174

Liu W, Zhang B, Zhang Z, Tao J, Branscum A (2015) Model selection in finite mixture of regression models: a Bayesian approach with innovative weighted g priors and reversible jump Markov chain Monte Carlo implementation. Journal of Statistical Computation and Simulation 85:2456–2478

McLachlan G, Peel D (2000) Finite mixture models. Wiley, Chichester.

Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg R, Group SCPS (1989) Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology 130(3):511–521

O'Hara RB, Sillanpää MJ (2009) A review of Bayesian variable selection methods: what, how and which. Bayesian Analysis 4(1):85–117

Papastamoulis P (2016) label.switching: An `R` Package for Dealing with the Label Switching Problem in MCMC Outputs. Journal of Statistical Software, Code Snippets 69(1):1–24, DOI 10.18637/jss.v069.c01, URL https://www.jstatsoft.org/v069/c01

Park T, Casella G (2008) The Bayesian Lasso. Journal of the American Statistical Association 103:681–686

R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Richardson S, Green PJ (1997) On Bayesian analysis of mixtures with unknown number of components (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology) 59(4):731–792

Rubio F, Genton M (2016) Bayesian linear regression with skew-symmetric error distributions with applications to survival analysis. Statistics in Medicine 35:2441–2454

Rubio F, Yu K (2017) Flexible objective Bayesian linear regression with applications in survival analysis. Journal of Applied Statistics 44:798–810

Sahu S, Dey D, Branco M (2003) A new class of multivariate skew distributions with applications to Bayesian regression models. Canadian Journal of Statistics 31:129–150

Schlattmann P (2009) Medical applications of finite mixture models. Springer, Berlin.

Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. Journal of Statistical Software 39(5):1–13, URL http://www.jstatsoft.org/v39/i05/

Soffritti G, Galimberti G (2011) Multivariate linear regression with non-normal errors: a solution based on mixture models. Statistics and Computing 21:523–536

Song W, Yao W, Xing Y (2014) Robust mixture regression model fitting by Laplace distribution. Computational Statistics & Data Analysis 71:128–137

Spiegelhalter DJ, Best NG, Carlin BP, 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

Srivastava MS (2002) Methods of Multivariate Statistics. Wiley

Srivastava VK, Giles DEA (1987) Seemingly Unrelated Regression Equations Models: Estimation and Inference. CRC Press

Städler N, Bühlmann P, van de Geer S (2010) l_1-penalization for mixture regression models. Test 19:209–256

Stukel T (2008) Determinants of plasma retinol and beta-carotene levels. StatLib Datasets Archive URL http://libstatcmuedu/datasets/Plasma_Retinol

Sutradhar B, Ali M (1986) Estimation of the parameters of a regression model with a multivariate t error variable. Communications in Statistics - Theory and Methods 15:429–450

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological) pp 267–288

Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73:273–282

Yao W, Wei Y, Yu C (2014) Robust mixture regression using the t-distribution. Computational Statistics & Data Analysis 71:116–127

Zellner A (1976) Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms. Journal of the American Statistical Association 71:400–405

Zellner A (1986) On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti 6:233–243

Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics 38:894–942

A Posterior distribution, full conditionals, and sketch of the algorithm

The posterior distribution is obtained by combining the complete-data likelihood in Equation 9 with the priors discussed in the Section 2.2. Omitting the hyper-parameters, the result is

$$\pi(\mathbf{w},\boldsymbol{\eta},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\mathbf{z},\mathbf{u}|\mathbf{y}) \propto L(\mathbf{w},\boldsymbol{\eta},\boldsymbol{\mu},\boldsymbol{\sigma}^2;\mathbf{y},\mathbf{z},\mathbf{u})\pi(\mathbf{u})\pi(\mathbf{w})\pi(\boldsymbol{\eta}|\boldsymbol{\sigma}^2,\mathbf{u})\pi(\boldsymbol{\mu}|\boldsymbol{\sigma}^2)\pi(\boldsymbol{\sigma}^2)$$

which can be fully detailed as

$$\begin{aligned}
\pi(\mathbf{w},\boldsymbol{\eta},\boldsymbol{\mu},\boldsymbol{\sigma}^2,\mathbf{z},\mathbf{u}|\mathbf{y}) &\propto \prod_{k=1}^{K} w_k^{n_k}\Bigg[\prod_{i:z_{ik}=1}(\sigma^2_{z_{ik}=1})^{-\frac{1}{2}}\exp\Bigg\{-\frac{1}{2\sigma^2_{z_{ik}=1}}\Big(y_i-\mu_{z_{ik}=1}-\mathbf{x}_i'(\mathbf{u})\boldsymbol{\eta}(\mathbf{u})\Big)^2\Bigg\}\Bigg]\times\\
&\times\prod_{j=1}^{p} v_j^{u_j}(1-v_j)^{1-u_j}\prod_{k=1}^{K} w_k^{\alpha_k-1}\Bigg(|E_0|^{-\frac{1}{2}}\exp\Bigg\{\frac{1}{2}\Big[\boldsymbol{\eta}(\mathbf{u})-\mathbf{e}_0\Big]'E_0^{-1}\Big[\boldsymbol{\eta}(\mathbf{u})-\mathbf{e}_0\Big]\Bigg\}\Bigg)\times\\
&\times\prod_{k=1}^{K}\Bigg[\Big(\frac{\sigma_k^2}{g_k}\Big)^{-\frac{1}{2}}\exp\Bigg\{-\frac{g_k}{2\sigma_k^2}\Big(\mu_k-m_{0,k}\Big)^2\Bigg\}\Bigg]\prod_{k=1}^{K}\Big(\sigma_k^2\Big)^{-a_{1,k}-1}\exp\Bigg\{-\frac{a_{2,k}}{\sigma^2}\Bigg\},
\end{aligned}$$

where $|E_0|$ is the determinant of a matrix E_0. From the complete-data likelihood we can also derive the distribution of each $\tilde{\mathbf{y}}_{[k]}$ and $\ddot{\mathbf{y}}$. The former is

$$\begin{aligned}
\pi(\tilde{\mathbf{y}}_{[k]}|\mu_k,\boldsymbol{\eta},\boldsymbol{\sigma}^2,\mathbf{u},\mathbf{z}) &\propto \Big(\sigma_k^2\Big)^{-\frac{n_k}{2}}\exp\Bigg\{-\frac{1}{2\sigma_k^2}\sum_{i=1}^{n_k}\Big(\tilde{y}_{i,[k]}-\mu_k\Big)^2\Bigg\}\propto\\
&\propto \Big(\sigma_k^2\Big)^{-\frac{n_k}{2}}\exp\Bigg\{-\frac{1}{2\sigma_k^2}\Big[(n_k-1)\tilde{s}_k^2+n_k\Big(\mu_k-\bar{y}_{[k]}\Big)^2\Big]\Bigg\},
\end{aligned}$$

for every k, while the latter is

$$\begin{aligned}
\pi(\ddot{\mathbf{y}}|\boldsymbol{\mu},\boldsymbol{\eta},\boldsymbol{\sigma}^2,\mathbf{u},\mathbf{z}) &\propto |\Sigma|^{-\frac{1}{2}}\exp\Bigg\{-\frac{1}{2}\Big[\ddot{\mathbf{y}}-X(\mathbf{u})\boldsymbol{\eta}(\mathbf{u})\Big]'\Sigma^{-1}\Big[\ddot{\mathbf{y}}-X(\mathbf{u})\boldsymbol{\eta}(\mathbf{u})\Big]\Bigg\}\propto\\
\propto |\Sigma|^{-\frac{1}{2}}\exp\Bigg\{-\frac{1}{2}\Big[(n-p)\ddot{s}^2+\Big(\boldsymbol{\eta}(\mathbf{u})-\boldsymbol{\eta}_{\mathrm{WLS}}(\mathbf{u})\Big)'X(\mathbf{u})'\Sigma^{-1}X(\mathbf{u})\Big(\boldsymbol{\eta}(\mathbf{u})-\boldsymbol{\eta}_{\mathrm{WLS}}(\mathbf{u})\Big)\Bigg\},
\end{aligned}$$

where $\tilde{s}_k^2 = (n_k - 1)^{-1} \sum_{i=1}^{n_k} \left(\tilde{y}_{i,[k]} - \bar{y}_{[k]} \right)^2$, and

$$\ddot{s}^2 = (n-p)^{-1} \left[\ddot{\mathbf{y}} - X(\mathbf{u}) \boldsymbol{\eta}_{\mathrm{WLS}}(\mathbf{u}) \right]' \Sigma^{-1} \left[\ddot{\mathbf{y}} - X(\mathbf{u}) \boldsymbol{\eta}_{\mathrm{WLS}}(\mathbf{u}) \right].$$

Thanks to conjugacy between priors and the likelihood, and the properties of the normal distribution, we can derive Gibbs samplers for all the parameters and latent variables in our model. A sketch of the algorithm, with the full conditionals used within it, is given below. After initializing all the parameters and randomly assigning units to the K components, for each MCMC iteration $t = 0, \ldots, T-1$, at $t+1$:

1. sample latent variables $\mathbf{z}_i^{(t+1)}$, for each $i = 1, \ldots, n$, according to:

$$\pi(z_{ik}^{(t+1)} = 1 | w_k^{(t)}, \mu_k^{(t)}, \sigma_k^{2^{(t)}}, \boldsymbol{\eta}^{(t)}, \mathbf{u}^{(t)}, y_i) = \frac{w_k^{(t)} \cdot \phi\big(y_i; \mu_k^{(t)} + \mathbf{x}_i'(\mathbf{u}^{(t)}) \boldsymbol{\eta}^{(t)}(\mathbf{u}^{(t)}), \sigma_k^{2^{(t)}}\big)}{\sum_{k'=1}^{K} w_{k'}^{(t)} \cdot \phi\big(y_i; \mu_{k'}^{(t)} + \mathbf{x}_i'(\mathbf{u}^{(t)}) \boldsymbol{\eta}(\mathbf{u}^{(t)}), \sigma_{k'}^{2^{(t)}}\big)};$$

2. sample component proportions $\mathbf{w}^{(t+1)}$ from:

$$\mathbf{w}^{(t+1)} | (\mathbf{z}^{(t)}, \mathbf{y}) \sim \mathrm{Dir}(\alpha_1 + n_1^{(t)}, \alpha_2 + n_2^{(t)}, \ldots, \alpha_K + n_K^{(t)});$$

3. for each $k = 1, \ldots, K$, sample $\mu_k^{(t+1)}$ from

$$\mu_k^{(t+1)} | (\sigma_k^{2^{(t)}}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \mathbf{u}^{(t)}, \mathbf{y}) \sim \mathcal{N}(m_k^{(t)}, M_k^{(t)}),$$

with $m_k^{(t)} = {n_k^{(t)}}^{-1} \sum_{i=1}^{n_k^{(t)}} \tilde{y}_{i,[k]}$ and $M_k^{(t)} = (n_k^{(t)} + 1)^{-1} \sigma_k^{2^{(t)}}$;

4. for each $k = 1, \ldots, K$ sample $\sigma_k^{2^{(t+1)}}$ from

$$\sigma_k^{2^{(t+1)}} | (\mu_k^{(t)}, \boldsymbol{\eta}^{(t)}, \mathbf{z}^{(t)}, \mathbf{u}^{(t)}, \mathbf{y}) \sim \mathrm{IG}\left(\frac{b_{1,k}^{(t)}}{2}, \frac{b_{2,k}^{(t)}}{2} \right),$$

with $b_{1,k}^{(t)} = n_k^{(t)} + 1 + h^{(t)} + a_{1,k}$ and

$$b_{2,k}^{(t)} = \left(\mathbf{y}_{[k]} - \mu_k^{(t)} \mathbf{1}_{n_k} - X_{[k]}(\mathbf{u}^{(t)}) \boldsymbol{\eta}^{(t)}(\mathbf{u}^{(t)}) \right)' \left(\mathbf{y}_{[k]} - \mu_k^{(t)} \mathbf{1}_{n_k} - X_{[k]}(\mathbf{u}^{(t)}) \boldsymbol{\eta}^{(t)}(\mathbf{u}^{(t)}) \right) +$$
$$+ (\mu_k^{(t)} - m_{0,k})^2 + \left(\boldsymbol{\eta}^{(t)}(\mathbf{u}^{(t)}) - \mathbf{e}_0 \right)' E_{0,[k]}^{-1} \left(\boldsymbol{\eta}^{(t)}(\mathbf{u}^{(t)}) - \mathbf{e}_0 \right),$$

where $E_{0,[k]}$ is obtained by selecting rows of E_0 relative only to units in component k;

5. sample the regression coefficients $\boldsymbol{\eta}^{(t+1)}$ according to:

$$\boldsymbol{\eta}^{(t+1)} | (\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\sigma}^{2^{(t)}}, \mathbf{z}^{(t)}, \mathbf{u}^{(t)}, \mathbf{y}) \sim \mathcal{N}(\mathbf{e}^{(t)}, E^{(t)}),$$

where

$$E^{(t)} = \left[X'(\mathbf{u}^{(t)}) \Sigma^{-1^{(t)}} X(\mathbf{u}^{(t)}) + E_0^{-1} \right]^{-1}$$
$$\mathbf{e}^{(t)} = E^{(t)} \left[E_0^{-1} \mathbf{e}_0 + X'(\mathbf{u}^{(t)}) \Sigma^{-1} \ddot{\mathbf{y}} \right];$$

6. for each $j = 1, \ldots, p$, sample $\mathbf{u}^{(t+1)}$ from

$$\pi(u_j^{(t+1)} = 1 | \ldots, \mathbf{u}_{[-j]}^{(t)}, \mathbf{y}) = \frac{\pi(u_j^{(t)} = 1 | \ldots, \mathbf{u}_{[-j]}^{(t)}, \mathbf{y})}{\pi(u_j^{(t)} = 1 | \ldots, \mathbf{u}_{[-j]}^{(t)}, \mathbf{y}) + \pi(u_j^{(t)} = 0 | \ldots, \mathbf{u}_{[-j]}^{(t)}, \mathbf{y})},$$

where $\mathbf{u}_{[-j]}^{(t)}$ is the vector $\mathbf{u}^{(t)}$ without its j-th element;

7. relabel components according to the ordered component-specific intercepts $\mu_1^{(t)} > \mu_2^{(t)} > \cdots > \mu_K^{(t)}$.

B Additional results from the simulation study

Results are reported in terms of:

- posterior mean and posterior standard deviation of the 5 active parameters $\eta_1, \ldots, \eta_5$, averaged across the 100 datasets, for correlation $\rho = 0.0$ and $\rho = 0.5$;
- percentage of datasets in which the corresponding predictors are selected, for correlation $\rho = 0.0$ and $\rho = 0.5$;
- TSR, TPR, and FPR, averaged across the 100 independent datasets, with correlation $\rho = 0.5$;

Values are rounded to the second decimal position. Tables are divided according to the simulation scenario.

B.1 Scenario 1, error: mixture of two normal distributions

Setting		Predictor	Fraction of datasets where X_j is selected			
n	p		$K = 1$	$K = 2$	$K = 3$	K_{DIC}
100	50	X_1	0.00	0.85	0.90	0.97
		X_2	0.02	0.99	1.00	1.00
		X_3	0.02	0.96	0.97	0.99
		X_4	0.00	0.94	0.92	0.97
		X_5	0.00	0.97	0.98	1.00
250	25	X_1	0.17	0.99	1.00	1.00
		X_2	0.46	1.00	1.00	1.00
		X_3	0.30	1.00	1.00	1.00
		X_4	0.05	0.99	1.00	1.00
		X_5	0.04	0.99	1.00	1.00
500	10	X_1	0.00	0.85	0.90	0.97
		X_2	0.02	0.99	1.00	1.00
		X_3	0.02	0.96	0.97	0.99
		X_4	0.00	0.94	0.92	0.97
		X_5	0.00	0.97	0.98	1.00

Table 7 Scenario 1, accuracy in selecting active predictors, according to the compared models; correlation: $\rho = 0.0$.

Setting n	p	Parameter	Average posterior mean (*average posterior st.dev.*) $K=1$		$K=2$		$K=3$		K_{DIC}	
		$\eta_1 = 2.5$	2.77	*(0.08)*	2.53	*(0.09)*	2.53	*(0.09)*	2.50	*(0.08)*
		$\eta_2 = -5.0$	-6.67	*(0.06)*	-5.02	*(0.13)*	-4.98	*(0.10)*	-4.99	*(0.08)*
100	50	$\eta_3 = 3.0$	5.51	*(0.08)*	2.97	*(0.12)*	3.00	*(0.11)*	3.01	*(0.08)*
		$\eta_4 = -2.0$	-0.20	*(0.08)*	-2.04	*(0.12)*	-2.01	*(0.09)*	-2.00	*(0.08)*
		$\eta_5 = 4.0$	4.43	*(0.10)*	3.98	*(0.10)*	3.98	*(0.10)*	4.01	*(0.08)*
		$\eta_1 = 2.5$	2.54	*(1.12)*	2.50	*(0.05)*	2.50	*(0.05)*	2.50	*(0.05)*
		$\eta_2 = -5.0$	-5.52	*(0.81)*	-5.00	*(0.05)*	-5.01	*(0.05)*	-5.01	*(0.05)*
250	25	$\eta_3 = 3.0$	3.19	*(1.02)*	3.01	*(0.05)*	3.00	*(0.05)*	3.00	*(0.05)*
		$\eta_4 = -2.0$	-1.91	*(1.23)*	-1.99	*(0.05)*	-2.00	*(0.05)*	-2.00	*(0.05)*
		$\eta_5 = 4.0$	4.19	*(0.97)*	4.00	*(0.05)*	4.00	*(0.05)*	4.00	*(0.05)*
		$\eta_1 = 2.5$	2.56	*(0.02)*	2.50	*(0.04)*	2.51	*(0.04)*	2.50	*(0.04)*
		$\eta_2 = -5.0$	-4.97	*(0.76)*	-5.00	*(0.04)*	-4.99	*(0.04)*	-5.00	*(0.04)*
500	10	$\eta_3 = 3.0$	2.98	*(0.94)*	3.00	*(0.04)*	3.00	*(0.04)*	3.00	*(0.04)*
		$\eta_4 = -2.0$	-2.09	*(1.20)*	-2.00	*(0.04)*	-2.00	*(0.04)*	-2.00	*(0.04)*
		$\eta_5 = 4.0$	4.09	*(0.77)*	4.00	*(0.04)*	4.01	*(0.04)*	4.01	*(0.04)*

Table 8 Scenario 1, average posterior quantities of regression coefficients for the compared models; correlation: $\rho = 0.0$.

Setting ($\rho = 0.5$) n	p	Model	$\overline{\text{TSR}}$	$\overline{\text{TPR}}$	$\overline{\text{FPR}}$	K_{DIC} 1	2	3
		LASSO	0.85	0.14	0.07			
		$K=1$	0.90	0.00	0.00			
100	50	$K=2$	0.95	0.88	0.04	0	56	44
		$K=3$	0.87	0.88	0.13			
		K_{DIC}	0.93	0.96	0.07			
		LASSO	0.70	0.62	0.28			
		$K=1$	0.80	0.02	0.00			
250	25	$K=2$	0.97	0.96	0.02	0	74	26
		$K=3$	0.96	0.97	0.04			
		K_{DIC}	0.97	0.99	0.03			
		LASSO	0.67	0.95	0.60			
		$K=1$	0.78	0.58	0.01			
500	10	$K=2$	0.98	0.98	0.02	0	78	22
		$K=3$	0.97	0.96	0.02			
		K_{DIC}	0.99	1.00	0.02			

Table 9 Scenario 1, accuracy in selecting/non-selecting predictors, according to the compared models; correlation: $\rho = 0.5$.

Setting		Predictor	Fraction of datasets where X_j is selected			
n	p		$K=1$	$K=2$	$K=3$	K_{DIC}
100	50	X_1	0.00	0.82	0.84	0.96
		X_2	0.00	0.95	0.94	0.99
		X_3	0.00	0.90	0.90	0.96
		X_4	0.00	0.79	0.74	0.87
		X_5	0.00	0.96	0.99	1.00
250	25	X_1	0.01	0.99	1.00	1.00
		X_2	0.03	1.00	1.00	1.00
		X_3	0.02	1.00	1.00	1.00
		X_4	0.00	0.80	0.85	0.95
		X_5	0.03	1.00	1.00	1.00
500	10	X_1	0.50	0.97	1.00	1.00
		X_2	0.92	1.00	1.00	1.00
		X_3	0.36	0.99	1.00	1.00
		X_4	0.14	0.94	0.82	0.98
		X_5	0.96	1.00	1.00	1.00

Table 10 Scenario 1, accuracy in selecting active predictors, according to the compared models; correlation: $\rho = 0.5$.

Setting		Parameter	Average posterior mean (*average posterior st.dev.*)							
n	p		$K=1$		$K=2$		$K=3$		K_{DIC}	
100	50	$\eta_1 = 2.5$	.	*(0.00)*	2.48	*(0.11)*	2.39	*(0.13)*	2.48	*(0.11)*
		$\eta_2 = -5.0$	.	*(0.00)*	-4.76	*(0.14)*	-4.82	*(0.14)*	-4.95	*(0.12)*
		$\eta_3 = 3.0$	.	*(0.00)*	2.89	*(0.12)*	2.67	*(0.15)*	2.88	*(0.12)*
		$\eta_4 = -2.0$	.	*(0.00)*	-1.93	*(0.13)*	-1.92	*(0.11)*	-1.99	*(0.09)*
		$\eta_5 = 4.0$	.	*(0.00)*	3.81	*(0.15)*	3.84	*(0.15)*	3.92	*(0.12)*
250	25	$\eta_1 = 2.5$	3.02	*(0.12)*	2.50	*(0.07)*	2.50	*(0.07)*	2.50	*(0.06)*
		$\eta_2 = -5.0$	-6.22	*(0.09)*	-5.00	*(0.08)*	-5.01	*(0.08)*	-5.01	*(0.07)*
		$\eta_3 = 3.0$	3.70	*(0.11)*	2.84	*(0.08)*	2.89	*(0.07)*	2.96	*(0.07)*
		$\eta_4 = -2.0$	-2.25	*(0.12)*	-1.99	*(0.06)*	-2.00	*(0.05)*	-2.00	*(0.06)*
		$\eta_5 = 4.0$	4.45	*(0.08)*	3.84	*(0.07)*	3.88	*(0.07)*	3.95	*(0.06)*
500	10	$\eta_1 = 2.5$	2.48	*(1.38)*	2.50	*(0.04)*	2.51	*(0.05)*	2.50	*(0.04)*
		$\eta_2 = -5.0$	-4.06	*(1.19)*	-4.94	*(0.05)*	-5.00	*(0.06)*	-4.99	*(0.05)*
		$\eta_3 = 3.0$	2.40	*(1.43)*	2.96	*(0.05)*	2.85	*(0.05)*	2.98	*(0.05)*
		$\eta_4 = -2.0$	-1.52	*(1.51)*	-2.00	*(0.04)*	-2.00	*(0.04)*	-2.00	*(0.04)*
		$\eta_5 = 4.0$	3.69	*(0.95)*	3.96	*(0.05)*	3.86	*(0.05)*	3.99	*(0.04)*

Table 11 Scenario 1, average posterior quantities of regression coefficients for the compared models; correlation: $\rho = 0.5$.

B.2 Scenario 2, error: Student-t distribution

Setting		Predictor	Fraction of datasets where X_j is selected			
n	p		$K=1$	$K=2$	$K=3$	K_{DIC}
100	50	X_1	0.30	0.79	0.92	0.90
		X_2	0.39	0.87	0.95	0.97
		X_3	0.33	0.86	0.94	0.95
		X_4	0.21	0.81	0.89	0.93
		X_5	0.39	0.93	0.92	0.95
250	25	X_1	0.43	0.95	0.96	0.98
		X_2	0.56	0.99	0.98	1.00
		X_3	0.50	1.00	1.00	1.00
		X_4	0.29	1.00	0.99	1.00
		X_5	0.56	1.00	1.00	1.00
500	10	X_1	0.39	0.91	0.97	0.99
		X_2	0.67	0.98	0.98	0.99
		X_3	0.50	1.00	0.99	1.00
		X_4	0.35	1.00	1.00	1.00
		X_5	0.58	1.00	1.00	1.00

Table 12 Scenario 2, accuracy in selecting active predictors, according to the compared models; correlation: $\rho = 0.0$.

Setting		Parameter	Average posterior mean (*average posterior st.dev.*)							
n	p		$K=1$		$K=2$		$K=3$		K_{DIC}	
100	50	$\eta_1 = 2.5$	2.40	*(0.36)*	3.67	*(0.57)*	2.67	*(0.31)*	2.74	*(0.25)*
		$\eta_2 = -5.0$	-5.24	*(0.27)*	-4.94	*(0.29)*	-4.96	*(0.22)*	-4.99	*(0.18)*
		$\eta_3 = 3.0$	2.87	*(0.38)*	3.17	*(0.40)*	2.68	*(0.29)*	2.63	*(0.23)*
		$\eta_4 = -2.0$	-1.96	*(0.45)*	-1.81	*(0.31)*	-1.80	*(0.22)*	-1.94	*(0.20)*
		$\eta_5 = 4.0$	4.15	*(0.28)*	4.26	*(0.38)*	3.77	*(0.21)*	4.00	*(0.17)*
250	25	$\eta_1 = 2.5$	2.60	*(0.56)*	2.49	*(0.12)*	2.49	*(0.10)*	2.49	*(0.09)*
		$\eta_2 = -5.0$	-5.08	*(0.45)*	-4.99	*(0.13)*	-5.00	*(0.10)*	-5.00	*(0.10)*
		$\eta_3 = 3.0$	3.20	*(0.49)*	3.04	*(0.14)*	3.01	*(0.11)*	3.01	*(0.10)*
		$\eta_4 = -2.0$	-1.91	*(0.66)*	-2.00	*(0.14)*	-2.00	*(0.11)*	-2.01	*(0.10)*
		$\eta_5 = 4.0$	3.96	*(0.46)*	4.01	*(0.14)*	3.99	*(0.11)*	4.00	*(0.10)*
500	10	$\eta_1 = 2.5$	3.34	*(2.11)*	2.50	*(0.09)*	2.50	*(0.07)*	2.50	*(0.07)*
		$\eta_2 = -5.0$	-5.14	*(1.96)*	-5.01	*(0.10)*	-5.01	*(0.07)*	-5.01	*(0.07)*
		$\eta_3 = 3.0$	2.89	*(1.70)*	3.01	*(0.10)*	3.02	*(0.07)*	3.01	*(0.07)*
		$\eta_4 = -2.0$	-1.90	*(1.81)*	-2.00	*(0.10)*	-2.00	*(0.08)*	-2.00	*(0.07)*
		$\eta_5 = 4.0$	3.67	*(1.89)*	3.99	*(0.10)*	4.00	*(0.08)*	4.00	*(0.07)*

Table 13 Scenario 2, average posterior quantities of regression coefficients for the compared models; correlation: $\rho = 0.0$.

Setting ($\rho = 0.5$) n	p	Model	$\overline{\text{TSR}}$	$\overline{\text{TPR}}$	$\overline{\text{FPR}}$	K_{DIC} 1	2	3
100	50	LASSO	0.82	0.36	0.12	0	26	74
		$K = 1$	0.92	0.21	0.00			
		$K = 2$	0.67	0.77	0.34			
		$K = 3$	0.53	0.88	0.51			
		K_{DIC}	0.59	0.90	0.45			
250	25	LASSO	0.69	0.59	0.28	0	19	81
		$K = 1$	0.85	0.26	0.00			
		$K = 2$	0.96	0.90	0.02			
		$K = 3$	0.95	0.95	0.05			
		K_{DIC}	0.96	0.99	0.04			
500	10	LASSO	0.62	0.67	0.43	0	13	87
		$K = 1$	0.66	0.33	0.00			
		$K = 2$	0.97	0.95	0.01			
		$K = 3$	0.97	0.96	0.02			
		K_{DIC}	0.98	0.99	0.02			

Table 14 Scenario 2, accuracy in selecting/non-selecting predictors, according to the compared models; correlation: $\rho = 0.5$.

Setting n	p	Predictor	Fraction of datasets where X_j is selected $K = 1$	$K = 2$	$K = 3$	K_{DIC}
100	50	X_1	0.23	0.75	0.85	0.87
		X_2	0.28	0.85	0.90	0.92
		X_3	0.18	0.80	0.88	0.88
		X_4	0.10	0.60	0.83	0.88
		X_5	0.27	0.84	0.93	0.94
250	25	X_1	0.26	0.91	0.95	0.97
		X_2	0.34	0.91	0.98	0.99
		X_3	0.27	0.95	0.99	1.00
		X_4	0.14	0.93	0.84	0.98
		X_5	0.31	0.74	0.99	1.00
500	10	X_1	0.29	0.90	0.93	0.97
		X_2	0.46	0.98	0.98	1.00
		X_3	0.26	1.00	1.00	1.00
		X_4	0.16	0.89	0.89	0.96
		X_5	0.48	1.00	1.00	1.00

Table 15 Scenario 2, accuracy in selecting active predictors, according to the compared models; correlation: $\rho = 0.5$.

Setting n	Setting p	Parameter	Average posterior mean (*average posterior st.dev.*) $K=1$		$K=2$		$K=3$		K_{DIC}	
100	50	$\eta_1 = 2.5$	2.54	*(0.24)*	2.66	*(0.37)*	2.82	*(0.25)*	2.78	*(0.24)*
		$\eta_2 = -5.0$	-3.31	*(0.37)*	-5.15	*(0.34)*	-4.74	*(0.24)*	-5.16	*(0.22)*
		$\eta_3 = 3.0$	2.42	*(0.28)*	2.60	*(0.48)*	2.48	*(0.26)*	2.66	*(0.23)*
		$\eta_4 = -2.0$	-1.47	*(0.25)*	-1.15	*(0.34)*	-1.80	*(0.25)*	-1.86	*(0.24)*
		$\eta_5 = 4.0$	4.49	*(0.38)*	3.82	*(0.44)*	3.68	*(0.26)*	3.84	*(0.25)*
250	25	$\eta_1 = 2.5$	2.61	*(0.34)*	2.45	*(0.15)*	2.47	*(0.12)*	2.46	*(0.12)*
		$\eta_2 = -5.0$	-4.63	*(0.29)*	-4.82	*(0.18)*	-4.97	*(0.14)*	-4.98	*(0.13)*
		$\eta_3 = 3.0$	2.86	*(0.34)*	2.84	*(0.18)*	2.89	*(0.14)*	2.99	*(0.13)*
		$\eta_4 = -2.0$	-1.62	*(0.38)*	-1.78	*(0.17)*	-2.04	*(0.11)*	-2.04	*(0.13)*
		$\eta_5 = 4.0$	3.60	*(0.29)*	3.76	*(0.17)*	3.90	*(0.13)*	4.00	*(0.12)*
500	10	$\eta_1 = 2.5$	2.42	*(2.02)*	2.48	*(0.11)*	2.48	*(0.09)*	2.50	*(0.08)*
		$\eta_2 = -5.0$	-3.25	*(2.02)*	-4.91	*(0.13)*	-4.95	*(0.10)*	-4.97	*(0.10)*
		$\eta_3 = 3.0$	2.12	*(2.21)*	2.89	*(0.10)*	2.88	*(0.07)*	2.98	*(0.10)*
		$\eta_4 = -2.0$	-1.05	*(2.22)*	-2.01	*(0.12)*	-2.01	*(0.09)*	-2.01	*(0.09)*
		$\eta_5 = 4.0$	3.28	*(1.90)*	3.91	*(0.12)*	3.91	*(0.10)*	3.96	*(0.09)*

Table 16 Scenario 2, average posterior quantities of regression coefficients for the compared models; correlation: $\rho = 0.5$.

B.3 Scenario 3, error: skew-normal distribution

Setting		Predictor	Fraction of datasets where X_j is selected			
n	p		$K=1$	$K=2$	$K=3$	K_{DIC}
100	50	X_1	0.90	0.97	0.96	0.96
		X_2	1.00	1.00	1.00	1.00
		X_3	0.98	1.00	0.97	0.97
		X_4	0.57	0.85	0.78	0.78
		X_5	1.00	1.00	1.00	1.00
250	25	X_1	1.00	1.00	1.00	1.00
		X_2	1.00	1.00	1.00	1.00
		X_3	1.00	1.00	1.00	1.00
		X_4	1.00	1.00	0.99	1.00
		X_5	1.00	1.00	1.00	1.00
500	10	X_1	1.00	1.00	1.00	1.00
		X_2	1.00	1.00	1.00	1.00
		X_3	1.00	1.00	1.00	1.00
		X_4	1.00	1.00	1.00	1.00
		X_5	1.00	1.00	1.00	1.00

Table 17 Scenario 3, accuracy in selecting active predictors, according to the compared models; correlation: $\rho = 0.0$.

Setting		Parameter	Average posterior mean (*average posterior st.dev.*)							
			$K=1$		$K=2$		$K=3$		K_{DIC}	
n	p									
100	50	$\eta_1 = 2.5$	2.48	*(0.76)*	2.46	*(0.48)*	2.22	*(0.28)*	2.23	*(0.28)*
		$\eta_2 = -5.0$	-4.92	*(0.62)*	-4.75	*(0.46)*	-4.30	*(0.29)*	-4.30	*(0.28)*
		$\eta_3 = 3.0$	2.99	*(0.68)*	2.92	*(0.47)*	2.62	*(0.30)*	2.64	*(0.30)*
		$\eta_4 = -2.0$	-1.85	*(0.93)*	-1.84	*(0.50)*	-1.55	*(0.30)*	-1.56	*(0.30)*
		$\eta_5 = 4.0$	3.94	*(0.63)*	3.82	*(0.48)*	3.50	*(0.29)*	3.51	*(0.28)*
250	25	$\eta_1 = 2.5$	2.47	*(0.38)*	2.48	*(0.27)*	2.45	*(0.23)*	2.45	*(0.23)*
		$\eta_2 = -5.0$	-4.99	*(0.38)*	-5.00	*(0.26)*	-4.95	*(0.24)*	-4.95	*(0.23)*
		$\eta_3 = 3.0$	2.92	*(0.38)*	2.92	*(0.26)*	2.91	*(0.23)*	2.92	*(0.23)*
		$\eta_4 = -2.0$	-2.05	*(0.39)*	-2.01	*(0.26)*	-1.97	*(0.23)*	-1.98	*(0.23)*
		$\eta_5 = 4.0$	3.96	*(0.38)*	3.97	*(0.27)*	3.94	*(0.24)*	3.95	*(0.24)*
500	10	$\eta_1 = 2.5$	2.50	*(0.27)*	2.49	*(0.18)*	2.49	*(0.14)*	2.49	*(0.14)*
		$\eta_2 = -5.0$	-4.90	*(0.27)*	-4.92	*(0.18)*	-4.95	*(0.15)*	-4.95	*(0.15)*
		$\eta_3 = 3.0$	2.94	*(0.27)*	2.96	*(0.18)*	2.96	*(0.14)*	2.96	*(0.14)*
		$\eta_4 = -2.0$	-1.97	*(0.27)*	-1.99	*(0.18)*	-2.00	*(0.15)*	-2.00	*(0.15)*
		$\eta_5 = 4.0$	3.94	*(0.27)*	3.99	*(0.18)*	3.99	*(0.15)*	3.99	*(0.15)*

Table 18 Scenario 3, average posterior quantities of regression coefficients for the compared models; correlation: $\rho = 0.0$.

Setting ($\rho = 0.5$)		Model	$\overline{\text{TSR}}$	$\overline{\text{TPR}}$	$\overline{\text{FPR}}$	K_{DIC}		
n	p					1	2	3
100	50	LASSO	0.73	0.83	0.28	0	3	97
		$K = 1$	0.95	0.56	0.00			
		$K = 2$	0.93	0.87	0.06			
		$K = 3$	0.69	0.89	0.33			
		K_{DIC}	0.69	0.89	0.33			
250	25	LASSO	0.61	1.00	0.48	0	15	85
		$K = 1$	0.99	0.96	0.01			
		$K = 2$	0.98	0.97	0.02			
		$K = 3$	0.96	0.98	0.05			
		K_{DIC}	0.96	1.00	0.05			
500	10	LASSO	0.65	1.00	0.71	0	14	86
		$K = 1$	0.99	0.99	0.01			
		$K = 2$	0.97	0.96	0.02			
		$K = 3$	0.97	0.97	0.03			
		K_{DIC}	0.98	0.99	0.02			

Table 19 Scenario 3, accuracy in selecting/non-selecting predictors, according to the compared models; correlation: $\rho = 0.5$.

Setting		Predictor	Fraction of datasets where X_j is selected			
n	p		$K = 1$	$K = 2$	$K = 3$	K_{DIC}
100	50	X_1	0.60	0.90	0.96	0.96
		X_2	0.75	0.97	1.00	1.00
		X_3	0.51	0.86	0.85	0.85
		X_4	0.18	0.65	0.68	0.67
		X_5	0.75	0.98	0.96	0.96
250	25	X_1	1.00	1.00	1.00	1.00
		X_2	1.00	1.00	1.00	1.00
		X_3	1.00	1.00	1.00	1.00
		X_4	0.81	0.87	0.91	0.98
		X_5	1.00	1.00	1.00	1.00
500	10	X_1	1.00	1.00	1.00	1.00
		X_2	1.00	1.00	1.00	1.00
		X_3	1.00	1.00	1.00	1.00
		X_4	0.94	0.82	0.83	0.96
		X_5	1.00	1.00	1.00	1.00

Table 20 Scenario 3, accuracy in selecting active predictors, according to the compared models; correlation: $\rho = 0.5$.

Setting n	Setting p	Parameter	Average posterior mean (*average posterior st.dev.*) $K=1$		$K=2$		$K=3$		K_{DIC}	
100	50	$\eta_1 = 2.5$	2.48	*(0.74)*	2.37	*(0.54)*	2.21	*(0.28)*	2.21	*(0.28)*
		$\eta_2 = -5.0$	-4.56	*(0.70)*	-4.60	*(0.61)*	-4.25	*(0.33)*	-4.25	*(0.33)*
		$\eta_3 = 3.0$	2.60	*(0.83)*	2.68	*(0.67)*	2.38	*(0.33)*	2.38	*(0.33)*
		$\eta_4 = -2.0$	-1.55	*(0.35)*	-1.75	*(0.67)*	-1.30	*(0.35)*	-1.29	*(0.30)*
		$\eta_5 = 4.0$	3.64	*(0.56)*	3.62	*(0.53)*	3.15	*(0.32)*	3.14	*(0.32)*
250	25	$\eta_1 = 2.5$	2.45	*(0.45)*	2.48	*(0.31)*	2.45	*(0.27)*	2.47	*(0.26)*
		$\eta_2 = -5.0$	-4.92	*(0.50)*	-4.97	*(0.34)*	-4.91	*(0.30)*	-4.92	*(0.29)*
		$\eta_3 = 3.0$	2.77	*(0.51)*	2.83	*(0.34)*	2.84	*(0.29)*	2.88	*(0.29)*
		$\eta_4 = -2.0$	-2.05	*(0.48)*	-1.98	*(0.29)*	-1.97	*(0.28)*	-1.98	*(0.29)*
		$\eta_5 = 4.0$	3.83	*(0.46)*	3.86	*(0.32)*	3.88	*(0.29)*	3.94	*(0.28)*
500	10	$\eta_1 = 2.5$	2.46	*(0.31)*	2.49	*(0.22)*	2.48	*(0.17)*	2.47	*(0.17)*
		$\eta_2 = -5.0$	-4.84	*(0.35)*	-4.92	*(0.24)*	-4.92	*(0.19)*	-4.92	*(0.19)*
		$\eta_3 = 3.0$	2.84	*(0.35)*	2.80	*(0.24)*	2.80	*(0.19)*	2.90	*(0.19)*
		$\eta_4 = -2.0$	-1.94	*(0.33)*	-1.98	*(0.20)*	-2.00	*(0.15)*	-1.98	*(0.18)*
		$\eta_5 = 4.0$	3.87	*(0.32)*	3.84	*(0.22)*	3.86	*(0.18)*	3.96	*(0.18)*

Table 21 Scenario 3, average posterior quantities of regression coefficients for the compared models; correlation: $\rho = 0.5$.

B.4 Scenario 4, error: skew-t distribution

Setting		Predictor	Fraction of datasets where X_j is selected			
n	p		$K=1$	$K=2$	$K=3$	K_{DIC}
100	50	X_1	0.51	0.85	0.94	0.94
		X_2	0.80	0.99	1.00	1.00
		X_3	0.65	0.95	0.93	0.93
		X_4	0.24	0.71	0.78	0.80
		X_5	0.75	0.98	0.97	0.97
250	25	X_1	0.96	1.00	1.00	1.00
		X_2	1.00	1.00	1.00	1.00
		X_3	1.00	1.00	1.00	1.00
		X_4	0.83	1.00	0.99	1.00
		X_5	1.00	1.00	1.00	1.00
500	10	X_1	0.99	1.00	1.00	1.00
		X_2	1.00	1.00	1.00	1.00
		X_3	1.00	1.00	1.00	1.00
		X_4	0.98	1.00	0.99	0.99
		X_5	1.00	1.00	1.00	1.00

Table 22 Scenario 4, accuracy in selecting active predictors, according to the compared models; correlation: $\rho = 0.0$.

Setting		Parameter	Average posterior mean (*average posterior st.dev.*)							
n	p		$K=1$		$K=2$		$K=3$		K_{DIC}	
100	50	$\eta_1 = 2.5$	2.39	*(1.03)*	2.44	*(0.65)*	2.23	*(0.40)*	2.24	*(0.41)*
		$\eta_2 = -5.0$	-4.93	*(0.69)*	-4.77	*(0.58)*	-4.54	*(0.38)*	-4.54	*(0.38)*
		$\eta_3 = 3.0$	3.08	*(0.91)*	3.00	*(0.60)*	2.58	*(0.39)*	2.60	*(0.39)*
		$\eta_4 = -2.0$	-1.80	*(1.08)*	-1.84	*(0.68)*	-1.64	*(0.43)*	-1.66	*(0.44)*
		$\eta_5 = 4.0$	4.07	*(0.78)*	3.88	*(0.59)*	3.50	*(0.39)*	3.51	*(0.40)*
250	25	$\eta_1 = 2.5$	2.44	*(0.63)*	2.45	*(0.33)*	2.44	*(0.26)*	2.45	*(0.26)*
		$\eta_2 = -5.0$	-5.03	*(0.55)*	-5.00	*(0.33)*	-4.96	*(0.30)*	-4.97	*(0.26)*
		$\eta_3 = 3.0$	2.91	*(0.57)*	2.90	*0.33*	2.89	*(0.29)*	2.89	*(0.26)*
		$\eta_4 = -2.0$	-2.07	*(0.73)*	-2.03	*(0.33)*	-1.99	*(0.28)*	-2.00	*(0.26)*
		$\eta_5 = 4.0$	3.95	*(0.55)*	3.99	*(0.33)*	3.96	*(0.29)*	3.96	*(0.26)*
500	10	$\eta_1 = 2.5$	2.49	*(0.39)*	2.50	*(0.24)*	2.48	*(0.17)*	2.48	*(0.17)*
		$\eta_2 = -5.0$	-4.86	*(0.39)*	-4.92	*(0.24)*	-4.92	*(0.18)*	-4.92	*(0.18)*
		$\eta_3 = 3.0$	2.90	*(0.39)*	2.98	*(0.24)*	2.95	*(0.17)*	2.95	*(0.17)*
		$\eta_4 = -2.0$	-1.96	*(0.41)*	-1.99	*(0.24)*	-2.00	*(0.17)*	-2.00	*(0.17)*
		$\eta_5 = 4.0$	3.88	*(0.39)*	3.98	*(0.24)*	4.00	*(0.17)*	4.00	*(0.17)*

Table 23 Scenario 4, average posterior quantities of regression coefficients for the compared models; correlation: $\rho = 0.0$.

Setting ($\rho = 0.5$) n	p	Model	$\overline{\text{TSR}}$	$\overline{\text{TPR}}$	$\overline{\text{FPR}}$	K_{DIC} 1	2	3
100	50	LASSO	0.76	0.59	0.22			
		$K = 1$	0.91	0.15	0.00			
		$K = 2$	0.92	0.67	0.05	0	9	91
		$K = 3$	0.74	0.82	0.27			
		K_{DIC}	0.74	0.82	0.27			
250	25	LASSO	0.63	0.96	0.45			
		$K = 1$	0.95	0.77	0.01			
		$K = 2$	0.98	0.97	0.01	0	15	85
		$K = 3$	0.96	0.99	0.04			
		K_{DIC}	0.97	0.99	0.04			
500	10	LASSO	0.64	1.00	0.71			
		$K = 1$	0.97	0.96	0.01			
		$K = 2$	0.98	0.97	0.01	0	11	89
		$K = 3$	0.97	0.96	0.01			
		K_{DIC}	0.98	0.98	0.01			

Table 24 Scenario 4, accuracy in selecting/non-selecting predictors, according to the compared models; correlation: $\rho = 0.5$.

Setting n	p	Predictor	Fraction of datasets where X_j is selected $K = 1$	$K = 2$	$K = 3$	K_{DIC}
100	50	X_1	0.13	0.63	0.82	0.80
		X_2	0.23	0.83	0.96	0.95
		X_3	0.12	0.68	0.83	0.83
		X_4	0.05	0.39	0.57	0.58
		X_5	0.22	0.82	0.93	0.93
250	25	X_1	0.80	1.00	1.00	1.00
		X_2	0.94	1.00	1.00	1.00
		X_3	0.72	0.99	1.00	0.99
		X_4	0.46	0.84	0.94	0.96
		X_5	0.94	1.00	1.00	1.00
500	10	X_1	0.98	0.99	1.00	1.00
		X_2	1.00	1.00	1.00	1.00
		X_3	1.00	1.00	1.00	1.00
		X_4	0.82	0.87	0.79	0.90
		X_5	1.00	1.00	1.00	1.00

Table 25 Scenario 4, accuracy in selecting active predictors, according to the compared models; correlation: $\rho = 0.5$.

Setting		Parameter	Average posterior mean (*average posterior st.dev.*)							
n	p		$K=1$		$K=2$		$K=3$		K_{DIC}	
100	50	$\eta_1 = 2.5$	2.34	*(0.31)*	2.26	*(0.61)*	2.15	*(0.49)*	2.16	*(0.49)*
		$\eta_2 = -5.0$	-4.98	*(0.23)*	-4.43	*(0.69)*	-4.11	*(0.52)*	-4.16	*(0.52)*
		$\eta_3 = 3.0$	2.83	*(0.33)*	2.69	*(0.66)*	2.42	*(0.51)*	2.44	*(0.48)*
		$\eta_4 = -2.0$	-1.34	*(0.44)*	-1.54	*(0.70)*	-1.28	*(0.52)*	-1.24	*(0.51)*
		$\eta_5 = 4.0$	3.77	*(0.24)*	3.66	*(0.58)*	3.26	*(0.51)*	3.27	*(0.49)*
250	25	$\eta_1 = 2.5$	2.44	*(0.74)*	2.45	*(0.38)*	2.46	*(0.30)*	2.45	*(0.30)*
		$\eta_2 = -5.0$	-4.65	*(0.76)*	-4.96	*(0.44)*	-4.91	*(0.34)*	-4.97	*(0.34)*
		$\eta_3 = 3.0$	2.55	*(0.94)*	2.77	*0.44*	2.79	*(0.34)*	2.89	*(0.35)*
		$\eta_4 = -2.0$	-1.83	*(1.07)*	-2.01	*(0.41)*	-1.98	*(0.32)*	-2.00	*(0.34)*
		$\eta_5 = 4.0$	3.70	*(0.64)*	3.86	*(0.40)*	3.94	*(0.31)*	3.96	*(0.31)*
500	10	$\eta_1 = 2.5$	2.42	*(0.46)*	2.49	*(0.28)*	2.48	*(0.21)*	2.47	*(0.21)*
		$\eta_2 = -5.0$	-4.72	*(0.50)*	-4.89	*(0.31)*	-4.91	*(0.24)*	-4.91	*(0.24)*
		$\eta_3 = 3.0$	2.68	*(0.52)*	2.85	*(0.31)*	2.77	*(0.24)*	2.87	*(0.24)*
		$\eta_4 = -2.0$	-1.87	*(0.57)*	-1.97	*(0.29)*	-2.00	*(0.21)*	-2.01	*(0.22)*
		$\eta_5 = 4.0$	3.73	*(0.47)*	3.86	*(0.29)*	3.83	*(0.22)*	3.92	*(0.22)*

Table 26 Scenario 4, average posterior quantities of regression coefficients for the compared models; correlation: $\rho = 0.5$.

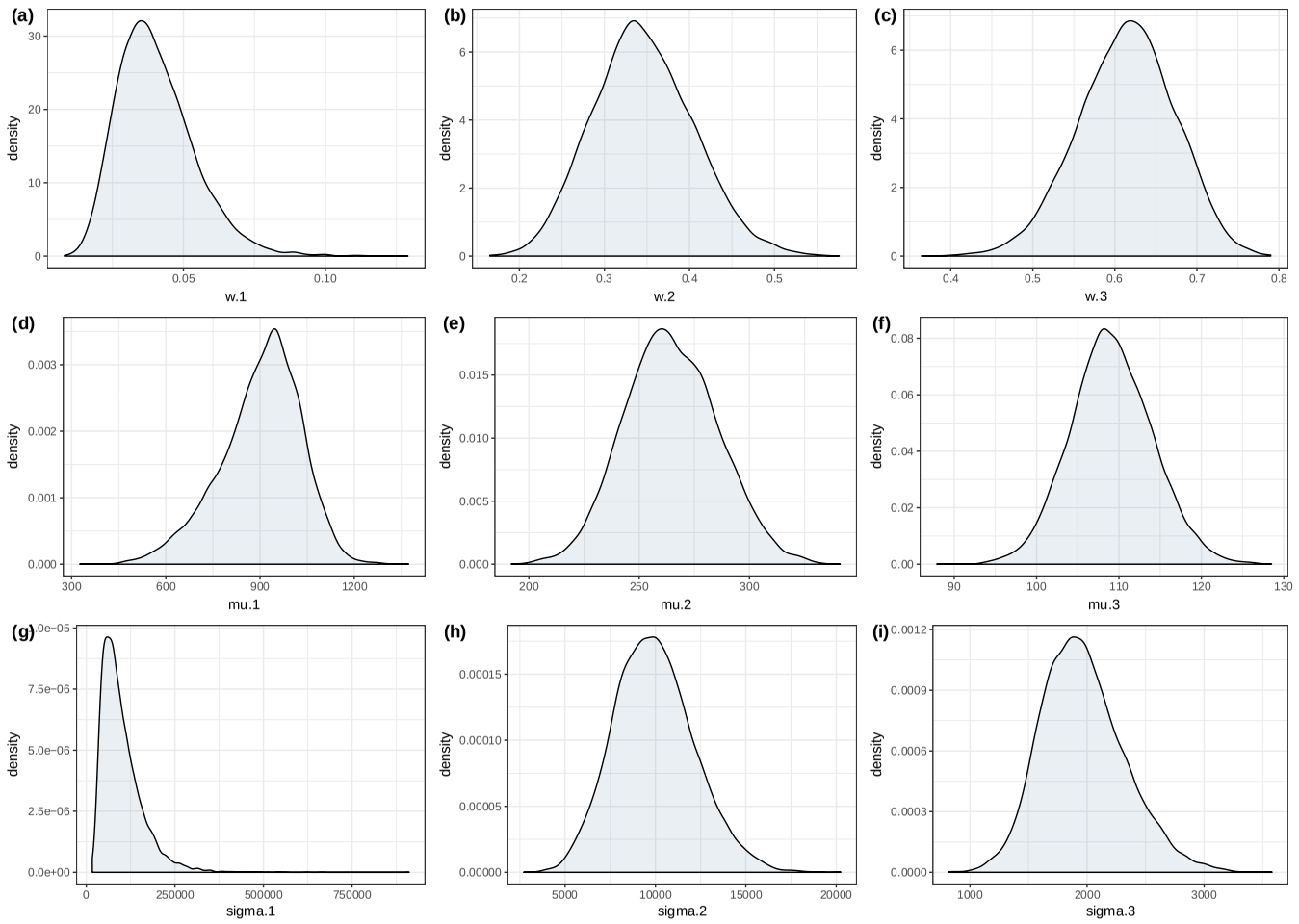

Fig. 3 Posterior distributions of components' proportions, intercepts, and variances for the mixture model with $K = 3$, $\mathbf{e}_0 = \boldsymbol{\eta}_{\text{WLS}}$, and (`vituse`, `smokestat`) included among the potential predictors.