Special Issue Paper

Statistics in Medicine

Received 1 December 2010, Accepted 28 September 2011 Published online 25 January 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.4439

Hierarchical Bayesian formulations for selecting variables in regression models

Veronika Rockova,[a,b*†] Emmanuel Lesaffre,[a,c] Jolanda Luime[d] and Bob Löwenberg[b]

The objective of finding a parsimonious representation of the observed data by a statistical model that is also capable of accurate prediction is commonplace in all domains of statistical applications. The parsimony of the solutions obtained by variable selection is usually counterbalanced by a limited prediction capacity. On the other hand, methodologies that assure high prediction accuracy usually lead to models that are neither simple nor easily interpretable. Regularization methodologies have proven to be useful in addressing both prediction and variable selection problems. The Bayesian approach to regularization constitutes a particularly attractive alternative as it is suitable for high-dimensional modeling, offers valid standard errors, and enables simultaneous estimation of regression coefficients and complexity parameters via computationally efficient MCMC techniques. Bayesian regularization falls within the versatile framework of Bayesian hierarchical models, which encompasses a variety of other approaches suited for variable selection such as spike and slab models and the MC^3 approach. In this article, we review these Bayesian developments and evaluate their variable selection performance in a simulation study for the classical small p large n setting. The majority of the existing Bayesian methodology for variable selection deals only with classical linear regression. Here, we present two applications in the contexts of binary and survival regression, where the Bayesian approach was applied to select markers prognostically relevant for the development of rheumatoid arthritis and for overall survival in acute myeloid leukemia patients.

Keywords: Bayesian regularization; spike and slab; MC^3; probit regression; Weibull regression

1. Introduction

The simultaneous assessment of the associations between multiple disease factors and a health outcome is an important topic in epidemiological research. The two fundamental objectives implicit in these investigations are (i) determining which predictors are prognostically or diagnostically important and (ii) selecting a combination of factors capable of accurate prediction of the disease outcome. The two goals are somewhat at odds with each other. Models that possess high prediction accuracy are usually not easily interpretable and might even contain insignificant variables, and their estimated effects may be biased [1]. When the focus shifts from prediction to explanation, usually, parsimonious models are preferred consisting of only variables that are truly influential for the outcome. Finding such a model in the regression framework can be recast as a problem of variable selection.

The customary variable selection strategies involving the sequential search (forward selection, backward elimination, or stepwise selection) or all-subset regression by using different optimization criteria have several well-acknowledged deficiencies. They become increasingly ineffective and impractical in higher dimensions and exhibit high sensitivity towards small changes in the data [2, 3]. The stepwise selection procedures are also prone to getting trapped in locally optimal models [4] and face problems in designs with complex patterns of multicollinearity [5]. Despite the drawbacks, they are still the immediate choice in routine data analysis.

[a]*Department of Biostatistics, Erasmus MC Rotterdam, The Netherlands*
[b]*Department of Hematology, Erasmus MC Rotterdam, The Netherlands*
[c]*L-BioStat, Katholieke Universiteit Leuven, Belgium*
[d]*Department of Rheumatology, Erasmus MC Rotterdam, The Netherlands*
**Correspondence to: Veronika Rockova, Department of Biostatistics, Erasmus MC Rotterdam, The Netherlands.*
[†]*E-mail: v.rockova@erasmusmc.nl*

Recently, researchers have devoted a great deal of attention to the development of different regularization methods for simultaneous variable selection and coefficient estimation [6–8]. The statistical concept of regularization can be vaguely characterized as imposing additional requirements on the regression solutions in that the more 'useful' solutions are preferred over other ones. What is meant by useful depends on the purpose. If variable selection is the ultimate goal, sparse solutions (i.e., solutions with the redundant coefficients effectively zeroed out) are more desirable. The preference requirements can take the form of restrictions on the space of the solutions (which is equivalent to imposing the frequentist penalty term to the log-likelihood being maximized) or, in a Bayesian way, putting a suitable prior on the regression coefficients.

The two regularization concepts are closely related to each other. The general principle behind the frequentist regularization is to maximize $\text{logLik}(\boldsymbol{\theta}|\mathbf{y}) - \text{pen}(\boldsymbol{\theta})$ with respect to the vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$, where $\text{logLik}(\cdot)$ denotes the logarithm of the likelihood and $\text{pen}(\cdot)$ is a regularization term, which controls the complexity of the solution. The most popular penalty terms are the l_p penalties, $l_p(\boldsymbol{\theta}) = \sum_{i=1}^{q} |\theta_i|^p$, with $p = 1$ (the LASSO penalty [6]) and $p = 2$ (the ridge penalty [9]). The solution to the penalized maximum likelihood estimation by using the l_p penalties possesses a Bayesian interpretation [6, 10]. It coincides with the mode of the joint posterior distribution of regression coefficients arising from independent individual priors of the form $p(\theta_j|\eta_j) \approx \exp(-\tau_j|\eta_j|^p)$, better known as exponential power priors [11, 12]. However, the fully Bayesian approach to regularization entails evaluation of the whole posterior distribution, rather than finding just its mode. Such exploration is most often achieved by MCMC methodology.

The Bayesian regularization constitutes only a fraction of Bayesian methodology currently available for variable selection. In the Bayesian paradigm, the task of variable selection is recast as parameter estimation in hierarchical models. In fact, the classical variable selection methods based on penalization of likelihood with a fixed multiple of model dimension (e.g., using *AIC*, C_p, and *BIC* criteria) can be regarded as special cases of hierarchical Bayesian model selection under a particular class of priors with fixed choices of hyperparameters [13]. Alternatively, George and Foster [13] proposed to estimate the hyperparameters from the data to obtain adaptive penalty criteria. The versatility of the hierarchical formulations together with the availability of numerous sophisticated MCMC techniques have lead to the development of a variety of Bayesian variable selection strategies [14–18]. The appeal of the Bayesian approach resides in several features: (i) the inference is purely probabilistic, as opposed to the frequentist hypotheses testing; (ii) it provides a natural framework for the assessment of model uncertainty and thereby creates a basis for eventual model averaging; (iii) it enables the incorporation of past external information through priors; (iv) it extends naturally to settings with multivariate responses; and (v) it is applicable for high-dimensional variable selection ('small n large p' setting).

In this article, we provide an overview of several Bayesian variable selection methods in the unified framework of Bayesian hierarchical models and highlight discrepancies and connections between them. We evaluated the empirical performance (with regard to variable selection accuracy) of the presented Bayesian methods and compared it with the classical strategies in a simulation study. The results demonstrate that Bayesian variable selection offers improved performance in detecting the true underlying model. The majority of Bayesian developments for variable selection occurred in the context of the classical linear model. The concept can be applied in other regression settings as well. To illustrate the application of Bayesian variable selection in binary and survival regression, we present an application from rheumatoid arthritis and from acute myeloid leukemia (AML). We have implemented Bayesian variable selection for probit and Weibull regression WinBUGS, which in combination with `R` provides an easy-to-use interface that is potentially attractive for users interested in applying the methodology.

The outline of the article goes as follows. In Section 2, we introduce the two data sets and describe the research questions to be addressed. Section 3 is devoted to the discussion on Bayesian hierarchical models for variable selection. We present the results of the simulation study in Section 4, present the Bayesian analysis of the data in Section 5, and wrap up with a discussion in Section 6.

2. The data

Here, we apply the Bayesian methodology for variable selection on two data sets. The goal of the first analysis was to identify markers predictive for development of rheumatoid arthritis, whereas the second analysis deals with joint assessment of prognostic capability of preselected mutation and gene expression markers for overall survival in patients with AML.

2.1. *Rotterdam Early Arthritis Cohort data*

Rheumatoid arthritis is an autoimmune disease characterized by chronic synovial inflammation and destruction of cartilage and bone in the joints. The Rotterdam Early Arthritis Cohort (REACH) study was initiated in 2004 to investigate the development of rheumatoid arthritis in patients with early manifestations of joint impairment. Information regarding basic patient characteristics, serological measurements, and patterns of disease involvement at baseline has been gathered in 681 recruited patients. It is of interest to know which of the following 12 factors are potentially associated with the development of rheumatoid arthritis considered as a binary (yes/no) outcome: *ACCP* (cyclic citrullinated peptide antibody), *age*, *ESR* (erythrocyte sedimentation rate), *DC* (duration of complaints in days), *stiffness* (duration of morning stiffness in minutes), *RF* (rheumatoid factor), *gender*, *Sym* (symmetrical pattern of joint inflammation; yes/no), *SJC* (swollen joint count), *TJC* (tender joint count), *BCPH* (bilateral compression pain in hands; yes/no), and *BCPF* (bilateral compression pain in feet; yes/no).

The standard approach to analyze these data would be to use logistic/probit regression combined with some off-the-shelf variable selection method. The F-to-out backward selection with $p = 0.05$ yields a model with the following variables: *ACCP*, *ESR*, *DC*, *Sym*, *SJC*, and *BCPH*. The model with the most favorable value of the *AIC* selected after an exhaustive model evaluation contains two extra variables: *RF* and *stiffness*. Which of these models provide the best approximation to the true underlying relationships is, if at all possible, difficult to assess. In the Bayesian approach, however, these individual models can be effectively compared using one particular measure, the posterior model probability, which quantifies the amount of confidence in each of the given models. The Bayesian analysis of the REACH data is presented in Section 4.

2.2. *Acute myeloid leukemia data*

Acute myeloid leukemia describes a group of hematopoietic disorders characterized by the expansion of immature myeloid blood cells. Risk stratification and therapy decision making is nowadays based mainly on karyotype information. However, about 45% of patients lacks any cytogenetical aberration. These patients exhibit various responses to therapy, and therefore, more targeted treatment protocols are required to improve their survival outcome. Identification of prognostic markers associated with survival in these 'intermediate-risk' patients would contribute to improved risk stratification. Recently, various markers have been individually identified as prognostically relevant. These include various mutation markers (*FLT3ITD*, *FLT3TKD*, *NPM1*, *NRAS*, *IDH1*, *IDH2*, and *CEBPA* single mutation (*SM*) and double (*DM*) mutation) as well as gene expression markers (*ABCB1*, *BCL2*, *BAALC*, *ERG*, *EVI1*, *CD34*, *MN1*, *FLT3*, *INDO*, and *WT1*). These markers were assessed and/or measured in a series of 318 AML patients with normal karyotype or a karyotype of no recognized prognostic value. Here, we focus on the joint assessment of the prognostic importance and the selection of a combination of the markers to be used for prediction/stratification.

For modeling the relationship between the markers and survival, we used a parametric Weibull model. Backward selection ($p = 0.05$) identified variables *CD34*, *ERG*, *BCL2*, and *CEBPA DM* as relevant, whereas *AIC* selection selected in addition *NPM1*, *FLT3ITD*, and *IDH2*.

In the Bayesian approach, the research question can be formulated and answered in a variable-specific way rather than model-wise. The conclusion about which variables are important for the survival outcome then again follows from posterior probabilities rather than from p-values. We present the Bayesian analysis of this data in Section 4.

3. Bayesian hierarchical formulations for variable selection

Consider an outcome random variable Y that we want to relate to the set of explanatory variables $X_1, \ldots, X_p$ by means of a regression model. The regression framework encompasses a variety of modeling platforms for different types of responses (Gaussian, time-to-event, binary), where the distribution of the response is related to the linear combination of covariates in a way which is specific for the type of outcome. Most often, only a subset of the available predictors play an important role in explaining the variability of the response, and the goal of the analysis was to identify these variables.

Each regression model is uniquely characterized by a vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$ of binary inclusion variables indicating whether or not the variable enters the model. Each model γ is then characterized by a specific linear combination of covariates of the form $\beta_0 + \boldsymbol{X}_\gamma' \boldsymbol{\beta}_\gamma$, where $\boldsymbol{X}_\gamma$ and $\boldsymbol{\beta}_\gamma$ denote subvectors of covariates and model parameters corresponding to the configuration $\boldsymbol{\gamma}$ and β_0 is the intercept.

In the Bayesian framework, we select variables based on posterior information obtained from hierarchical mixture models. Given the set of all plausible models $\{\boldsymbol{\gamma}_s : s \in S\}$, the hierarchical setup starts by assigning a prior probability $p(\boldsymbol{\gamma}_s)$ to each of the individual models, proceeds with choosing a prior distribution $p(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma} = \boldsymbol{\gamma}_s)$ over coefficients within each model, and is completed by the specification of the likelihood $p(Y|\boldsymbol{\beta}_{\gamma}, \boldsymbol{\gamma})$. The various Bayesian variable selection strategies emerge by considering different prior specifications and by choosing the actual posterior processing strategy.

3.1. *The 'model space' approach*

A natural way to compare models is by inspecting the individual posterior model probabilities

$$p(\boldsymbol{\gamma}_s|\boldsymbol{Y}) = \frac{p(\boldsymbol{Y}|\boldsymbol{\gamma}_s)p(\boldsymbol{\gamma}_s)}{\sum_{k\in S} p(\boldsymbol{Y}|\boldsymbol{\gamma}_k)p(\boldsymbol{\gamma}_k)},$$

where

$$p(\boldsymbol{Y}|\boldsymbol{\gamma}_k) = \int p(\boldsymbol{Y}|\boldsymbol{\gamma}_k, \beta_0, \boldsymbol{\beta}_{\gamma})p(\beta_0, \boldsymbol{\beta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma})\mathrm{d}(\beta_0, \boldsymbol{\beta}_{\gamma}) \tag{1}$$

denotes the marginal likelihood. The posterior model probabilities quantify the posterior evidence for selecting each particular model and as such immediately suggest models with the highest values as suitable candidates. With an increasing number of predictors, the exhaustive evaluation of the whole model space to find these models becomes impractical. As an alternative to the deterministic solutions based on stepwise search [19], stochastic alternatives have been suggested that exploit MCMC techniques to simulate a chain of models to find interesting regions of the model space with an accumulation of posterior mass. The most popular and intuitively appealing MCMC strategy adapted for this setting is MCMC model composition (MC^3) originally proposed in the context of graphical models [20]. The procedure results in a sequence $\boldsymbol{\gamma}^{(1)}, \boldsymbol{\gamma}^{(2)}, \ldots, \boldsymbol{\gamma}^{(T)}$ of visited models, which are generated according to the Metropolis–Hastings (MH) routine [21, 22]. The MH proposal distribution is concentrated at close proximity of the current state $\boldsymbol{\gamma}$, thereby restricting to models differing by an inclusion or an exclusion of just one variable. The candidate model $\boldsymbol{\gamma}^*$ sampled from the proposal distribution is then accepted with probability $\alpha = \min\left[1, \frac{p(\boldsymbol{\gamma}^*|\boldsymbol{Y})}{p(\boldsymbol{\gamma}|\boldsymbol{Y})}\right]$. The posterior model ratio is obtainable in closed form in conjugate regression designs. Otherwise, suitable approximations to the marginal likelihood in (1), for example, *BIC* approximation [23], can be used.

The prior distribution over models $p(\boldsymbol{\gamma})$ is an important ingredient in MC^3 and in other Bayesian variable selection procedures. The common choice of this prior distribution assumes independence amongst the binary inclusion indicators $\gamma_1, \ldots, \gamma_p$ and follows a product of individual Bernoulli distributions, that is, $p(\boldsymbol{\gamma}) = \prod_{j=1}^{p} w_j^{\gamma_j}(1 - w_j)^{1-\gamma_j}$, where w_j is the prior probability that the jth variable is in the model. In some hierarchical setups, the probability of inclusion w_j is assigned another prior layer. Keeping the parameters w_j fixed and equal to $1/2$, we obtain a uniform prior on the model space.

The actual variable selection can proceed in several ways. Two strategies most often applied in practice are (i) to pick a model with the highest estimated posterior probability $\widehat{p(\boldsymbol{\gamma}|y)} = \sum_{t=1}^{T} \mathrm{I}(\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma})/T$ (the highest posterior density (HPD) model) and (ii) to pick variables with estimated posterior inclusion probabilities $\widehat{p(\gamma_k|y)} = \sum_{t=1}^{T} \mathrm{I}(\gamma_k^{(t)} = 1)/T$ higher than 0.5 (the median probability model (MPM) [24]). Barbieri and Berger [24] studied the appropriateness of HPD model selection. The authors have shown that in orthogonal linear regression settings, the optimal model from a Bayesian predictive viewpoint was the MPM rather than the HPD model.

Special MCMC samplers capable of evaluating more than one model in parallel at each MCMC iteration have been developed to achieve faster and more efficient exploration of the model space, for example, the evolutionary MCMC [25] or the shotgun stochastic search [5]. The model space approach can also be implemented using some other MCMC samplers such as reversible jump MCMC [26–28] that explores simultaneously model and parameter space and dynamically adjusts for the differences in dimensionality of the sampled vectors $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. Carlin and Chib [14] proposed another strategy, on the basis of the Gibbs sampler.

3.2. *Spike and slab priors*

In many practical situations, it is desirable to estimate the values of selected coefficients after the model configuration has been chosen. In MC^3 and related strategies, the focus rests purely on variable selection,

leaving the inference about model parameters aside. The parameter estimates then can be obtained by 'post-model selection estimation' (using, e.g., posterior means or least square estimates). However, such strategy leads to biased estimates as it ignores the uncertainty in the model selection [13]. Alternatively, one could consider estimates that are not conditional on one selected model but rather averaged over all or highly probable models. Model averaging on the other hand does not provide sparse representation as it yields nonzero estimates of all coefficients regardless of how many of them are actually zero. A convenient solution would be to combine the model averaging and variable selection in one estimation process. This can be achieved in the Bayesian context by using the so-called variable selection priors.

Variable selection priors, better known as spike and slab priors [15, 17, 18], induce a positive prior probability on the hypothesis $H_0 : \beta_k = 0$. In the original formulation [16, 29], the spike and slab distribution is defined as a mixture of a Dirac measure concentrated at zero and a uniform diffuse component. Similarly, as in [30], we will slightly deviate from the original definition here. By a spike and slab prior, we understand any prior that is a mixture of two continuous distributions, implying high prior probability close to zero. These peak-shaped mixtures can be regarded as approximations to the point mass priors, which are computationally feasible for more conventional MCMC samplers. Such priors can be represented as conditionally Gaussian, that is, normal scale mixtures specified through the prior on hypervariances. The different variants of the spike and slab formulations emerge by considering different priors for the hypervariance (a two-point or continuous distribution). We now elaborate in more detail on the two most popular spike and slab priors: stochastic search variable selection (SSVS) prior [17] and normal mixture of inverse gamma (NMIG) in [15].

3.2.1. Stochastic search variable selection prior. George and McCulloch [17] proposed SSVS for variable selection in the context of linear regression. In SSVS, the model coefficients β_k are assumed to have a mixture prior of 'spike' and 'slab' Gaussian components. The spike element concentrates closely around zero, reflecting the actual absence of the variable in the model (γ_k equals zero). The slab component has a sufficiently large variance to allow the 'nonzero' coefficients to spread over larger values. The degree of separation between the two components is regulated by two tuning parameters τ_k and c_k, where $\tau_k^2 > 0$ is the variance in the spike component and $c_k^2\tau_k^2 > 0$ is the variance in the slab component. To guide the choice of τ_k and c_k, note that the two Gaussian densities intersect at the points $\pm\delta_k = \tau_k\varepsilon_k$, where $\varepsilon_k = \sqrt{2(\log c_k)c_k^2/(c_k^2-1)}$. The point δ_k can be regarded as a threshold for declaring practical significance in that all coefficients falling into the interval $[-\delta_k, \delta_k]$ can be interpreted as 'practically zero'. Given the parameter c_k, the variance τ_k^2 can be selected such that the intersection point reflects our perception of practical significance. The mathematical formulation of the SSVS hierarchical prior setup is the following:

$$\begin{aligned}
\beta_k|\lambda_k &\sim \mathrm{N}(0, \lambda_k),\\
\lambda_k|c_k, \tau_k^2, \gamma_k &\sim (1-\gamma_k)\delta_{\tau_k^2}(\cdot) + \gamma_k\delta_{c_k^2\tau_k^2}(\cdot),\\
\gamma_k|w_k &\sim \mathrm{Bernoulli}(w_k),\\
w_k &\sim \mathrm{Uniform}[0,1],
\end{aligned}$$

where $\delta_x(\cdot)$ denotes the Kronecker delta concentrated at point x. Because of the non-conjugacy, the analytical simplification of posterior distributions $p(\beta_k|\boldsymbol{y})$ and $p(\gamma_k|\boldsymbol{y})$ is not tractable. George and McCulloch [17] suggested an MCMC approximation to the posteriors by using the Gibbs sampler, which yields a chain of regression coefficients, and visited models $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\gamma}^{(1)}), \ldots, (\boldsymbol{\beta}^{(T)}, \boldsymbol{\gamma}^{(T)})$. Variable selection is then achieved through posterior model probabilities, posterior inclusion probabilities, or posterior distribution of the individual regression coefficients. Processing the MCMC information in $p(\beta_k|\boldsymbol{y})$ is complicated by the fact that the distribution can be multimodal, which makes the interpretation of posterior summary statistics less meaningful. Nevertheless, in case of strong evidence against the inclusion of the variable, the spike will dominate the posterior, which will effectively shrink the posterior mean towards zero. The decision on whether or not a variable enters the model can be carried out by hard shrinkage (HS; hard thresholding/selection shrinkage) [3, 31], where variables are included whenever the absolute value of the estimated coefficient (e.g., posterior mean) exceeds some threshold value.

Dellaportas [32] considered one particular variant of SSVS called Gibbs variable selection (GVS). The author suggested to introduce the binary inclusion indicators also in the likelihood so that only the variables that are literally present in the model contribute to the linear predictor, which now equals

$\beta_0 + \sum_{j=1}^{q} \gamma_j \beta_j X_j$. Apart from that, the prior setup for regression coefficients is analogous to SSVS. Examples of an application of SSVS priors in other linear regression settings can be found in [33] and in [34].

3.2.2. Normal mixture of inverse gamma. In the SSVS prior formulation, the variances λ_k have a discrete distribution with a support $\{\tau_k^2, c_k\tau_k^2\}$, which implies a two-point Gaussian mixture prior for the regression coefficient. In the context of linear regression, Ishwaran and Rao [15] suggested to move the spike and slab element down in the hierarchy and place it on the variances rather than on the regression coefficients. They argued that considering a continuous bimodal distribution for the variance introduces more uncertainty, which might potentially diminish the sensitivity towards the tuning of hyperparameters. In the original formulation [15, 30], the variance was parametrized as a product of two random variables, one having a two-point distribution and the second one having an inverse gamma distribution. Similarly, as in [35], we adopt a different parametrization by using a two-point mixture of inverse gammas. This yields the following hierarchical model:

$$\begin{aligned} \beta_k|\lambda_k &\sim \mathrm{N}(0, \lambda_k), \\ \lambda_k|v_0, v_1, \gamma_k, a, b &\sim (1-\gamma_k)\mathrm{IG}\left(a, \frac{v_0}{b}\right) + \gamma_k \mathrm{IG}\left(a, \frac{v_1}{b}\right), \\ \gamma_k|w_k &\sim \mathrm{Bernoulli}(w_k), \\ w_k &\sim \mathrm{Uniform}[0,1]. \end{aligned}$$

The role of τ_k^2 and c_k in SSVS is now taken by the parameters v_0 and v_1. Ishwaran and Rao [15] suggested to use $v_1 = 1$ by default for standardized covariates and rescaled responses in the linear model. Similarly, as in SSVS, the 'practical significance' argument can be applied to specify the other hyperparameters. Note that the marginal prior for the regression coefficients obtained by integrating out the variance is a two-point mixture of scaled t-distributions (with $2a$ degrees of freedom and respective scales $s_1 = \sqrt{\frac{bv_0}{a}}$ and $s_2 = \sqrt{\frac{bv_1}{a}}$). The two densities intersect at the points $\delta = \pm\sqrt{\frac{2a(1-r)}{\frac{r}{s_2^2} - \frac{1}{s_1^2}}}$, where $r = \left(\frac{s_2}{s_1}\right)^{\frac{2}{2a+1}}$. Similarly, as in SSVS, the preferred threshold for practical significance can be achieved by a suitable constellation of the hyperparameters a, b, v_0, and v_1. Konrath *et al.* (technical report University Munich) considered the extensions to non-Gaussian and hazard rate models. For an application in the context of additive regression models, see [35].

3.3. Bayesian regularization

In spike and slab hierarchies, all possible models are embodied within one hierarchical formulation, and the inference for variable selection can be carried out model-wise or from selection shrinkage. Whereas in the spike and slab formulations the peaked shape of the prior is achieved somewhat artificially by assuming a mixture distribution, it is possible to approximate the spike and slab shape with just one continuous prior component, for example, using the exponential power priors [10, 36] of the form $p(\beta_j|\eta_j) \approx \exp(-\eta_j|\beta_j|^p)$, where $p > 0$ and η_j is some variance-related parameter. The most popular powered exponential priors are the Laplace prior [6, 10] with $p = 1$ and the ridge prior with $p = 2$. If $0 < p \leqslant 1$, the prior has a singularity at origin, which promotes an intensive shrinkage towards the zero prior mean. For $0 < p \leqslant 2$, these distributions can be represented as scale mixtures of normals [37]. The class of normal scale mixtures has been recognized to generate many popular procedures for regularized regression, most notably the LASSO [6, 10], which is equivalent to the MAP estimation under normal/exponential (Laplace) prior. More recent normal scale mixture priors proposed for the shrinkage estimation in linear regression are the normal/gamma [38], the normal/Jeffreys [39, 40], or the horseshoe prior [41], where the mixing density belongs to the class of inverted beta distributions.

Unlike in the model space or spike and slab approaches, the sparsity approach avoids the specification of priors over models or individual hypotheses $H_{0k} : \beta_k = 0$. The variable selection rests purely on the inspection of the posterior behavior of the model coefficients. The posterior summary measures (mean or median) are never zero with a positive probability, and zeroing the redundant variables out then needs to be carried out through HS. Several authors augmented the shrinkage priors to include a point mass at zero [42]. Conceptually, these approaches belong to the spike and slab framework discussed in the previous section.

3.3.1. Bayesian LASSO: the Laplace prior. The Laplace (LASSO) prior arises from the scale normal mixture formulation assuming exponentially distributed variance [37, 43]. One version of the LASSO hierarchical formulation considered in the context of linear regression [10, 44, 45] is the following:

$$\beta_k|\lambda_k \sim \mathrm{N}(0, \sigma^2\lambda_k),$$
$$\lambda_k|\tau_k^2 \sim \frac{\tau_k^2}{2}\mathrm{e}^{-\lambda_k\tau_k^2/2}\mathrm{I}(\lambda_k > 0),$$

which implies a conditional Laplace prior $p(\beta_k|\sigma^2, \tau_k) = \frac{\tau_k}{2\sigma}\mathrm{e}^{-\tau_k|\beta_k|/\sigma}$. Bae and Mallick [39] used a variant of this hierarchy without the separate variance parameter σ^2 in the probit regression context. Instead of considering separate shrinkage parameters, it is customary to assume that $\tau_1^2 = \cdots = \tau_p^2 = \tau^2$. The parameter τ^2 then takes the role of the complexity parameters in the frequentist LASSO [6]. Whereas the frequentist perception of regularization assumes the shrinkage parameter fixed, the Bayesian LASSO allows to learn about the amount of shrinkage from the data by treating the parameter τ^2 as a random variable with its own prior distribution, such as gammas. The prior then combines with the evidence in the data to determine the optimal amount of shrinkage. Hans [42] complemented the LASSO prior with the point mass at zero and provided Gibbs sampling schemes alternative to the approach of Park and Casella [10] (see also [45]).

Keeping the variances λ_k equal and fixed, the MAP estimation corresponds to the frequentist ridge regression. However, such prior is not flexible enough to accommodate different shrinkage patterns for the individual coefficients. Assuming priors for the idiosyncratic variances assures more adaptivity. The fully Bayesian setup for ridge regression assumes the conjugate inverse gamma prior distribution for the variances, which implies a marginal scaled Student prior distribution for the individual regression coefficients.

Recent efforts in generalizing the penalization methodology to more complex data structures crystalized in several innovations of LASSO, which can be in turn transformed into MAP estimation in Bayesian hierarchical models. In linear regression setting, Tibshirani *et al.* [46] proposed fused LASSO for predictors that have a natural ordering, where the penalty is a linear combination on l_1 penalty on coefficients themselves and l_1 penalty on their first order differences. Such penalty induces similarity between neighboring coefficients. In case grouping among regression coefficients is suspected, but unknown, Zou and Hastie [8] suggested elastic net, which combines LASSO and ridge into one penalty and as such tends to keep the related variables in the model as a group. When the groups among predictors are known (e.g., group of dummy variables or spline coefficients), Yuan and Lin [7] proposed a grouped LASSO, which penalizes elliptical norms of the coefficients for each group. The Bayesian counterparts of these LASSO alternatives emerge by considering adequate alternations of powered exponential priors, which can be again represented as scale mixtures of normals [47, 48].

3.3.2. The elastic net prior. Bayesian elastic net, proposed by Zou and Hastie [8] and Li and Lin [48] in the context of linear regression, constitutes a compromise between the LASSO and ridge enjoying the advantages of the two. The elastic net prior inherits the sparsity property from the LASSO, because it is also not differentiable at zero, and at the same time encourages grouping as typical for the ridge prior. By a grouping effect, we refer to the ability to retain a group of strongly correlated variables in a model and keeping their estimated coefficients nearly equal (up to a change of sign for negatively correlated ones). This behavior is appreciated in modeling, for instance, gene expression data where related genes should enter the model as a group. The frequentist penalty term l_{net} for the 'naive' elastic net [8] is the linear combination of l_1 and l_2 penalties; that is, $l_{\mathrm{net}}(\boldsymbol{\beta}) = a_1\sum_{k=1}^q|\beta_k| + a_2\sum_{k=1}^q\beta_k^2$, which corresponds to the marginal MAP estimation implied by the following prior hierarchy:

$$\beta_k|\tau_k \sim \mathrm{N}\left(0, \left[\frac{a_2}{\sigma^2}\frac{\tau_k}{\tau_k - 1}\right]^{-1}\right),$$
$$\tau_k \sim \mathrm{Gamma}\left[0.5, \frac{8a_2\sigma^2}{a_1^2}, (1, \infty)\right],$$

where $\mathrm{Gamma}[a, b, (c, d)]$ refers to the truncated gamma distribution with shape a, scale b, and a support restricted to the interval (c, d). Diffuse hyperpriors for the two penalization parameters a_1 and a_2 can be added in the formulation to circumvent the uncertainty in their selection. Similarly as for the

LASSO prior, Hans (technical report University Ohio) augmented the elastic net prior to include the point mass at zero and suggested a Gibbs sampling algorithm in Gaussian regression models.

3.4. Extensions to other than linear regression settings

The variable shrinkage/selection priors outlined in the previous sections can be applied in other regression modeling settings where the response is not Gaussian. In probit regression, data augmentation strategies [49] assuming a linear model on latent continuous data greatly facilitate the implementation of efficient MCMC schemes. A similar approach can be adapted for handling ordered categorical data [49]. Data augmented Bayesian logistic regression was enabled by the introduction of Kolmogorov–Smirnov random variables [50]. Holmes and Held [50] further described variable selection approach in logistic regression by using reversible jump MCMC. Gramacy and Polson (technical report University of Chicago) dealt with regularized logistic regression. Many studies including [39, 51–53] considered variable selection in binary regression models. In survival regression context, Sha *et al.* [54] applied the variable selection priors in accelerated failure time model.

4. Simulation study

We evaluated the empirical variable selection performance of the outlined Bayesian methodology (*SSVS*, *NMIG*, *GVS*, *MC*^{3}, Bayesian LASSO, ridge, and elastic net) in a simulation study carried out to (i) compare the classical versus Bayesian variable selection, (ii) assess the sensitivity of spike and slab priors to the choice of tuning parameters, and (iii) compare the different approaches with processing of the posterior information. We performed the simulation study on data with binary responses, generated according to the latent variable probit regression scheme

$$\begin{aligned} Z_i &\overset{\text{ind}}{\sim} \mathrm{N}(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \\ Y_i &= \mathrm{I}(Z_i > 0), \quad (i = 1, \ldots, n), \end{aligned}$$

which is equivalent to assuming $Y_i \sim \text{Bernoulli}\left[\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}/\sigma)\right]$. We assume four linear regression models for the latent continuous data $\boldsymbol{Z}$, which reflect settings with different degree of sparsity, magnitude of the main effects, and pattern of collinearity among the predictors. We adopted the first three designs from the original LASSO paper of Tibshirani [6]. We drew the predictors independently from $\mathrm{N}_8(\mathbf{0}, \Sigma)$ with $\Sigma = (\sigma_{ij})_{i,j}$ and $\sigma_{ij} = \rho^{|i-j|}$. The first and third models (Designs 1 and 3) mimic rather sparse situations with relatively large values of nonzero coefficients. We chose the parameters as follows: $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ with $\rho = 0.5$ and $\sigma = 3$ in Design 1, and $\boldsymbol{\beta} = (5, 0, 0, 0, 0, 0, 0, 0)'$ with $\rho = 0.5$ and $\sigma = 2$ in Design 3. In Design 2, all the eight predictors are weakly informative, that is, $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)'$ with $\rho = 0.5$ and $\sigma = 3$. In the last design, we generated the predictors as follows: $x_{ij} = m_{ij} + z_1, j = 1, \ldots, 5$, $x_{ij} = m_{ij} + z_2, j = 6, \ldots, 10$, and $x_{ij} = m_{ij} + z_3, j = 11, \ldots, 15$, where m_i was drawn independently from $\mathrm{N}_{15}(\mathbf{0}, \mathrm{I}_{15})$, with I_{15} as the identity matrix, and z_i $(i = 1, 2, 3)$ is standard normal. Such specification induces correlations of about 0.5 within the three blocks of predictors. We chose the vector of coefficients equal to $\boldsymbol{\beta} = (3, 3, 3, 3, 3, 0, 0, 0, 0, 0, -3, -3, -3, -3, -3)'$.

For modeling the relationship between the binary response and the predictors, we used probit regression (data augmentation formulation [49], which assumes the latent normal linear model).

For each of the four models, we simulated 50 data sets, each consisting of $n = 100$ observations. To evaluate the variable selection properties, we keep track of the following quantities: (i) number of false discoveries (FDN), which is the number of coefficients falsely identified as nonzero; (ii) number of false nondiscoveries (FNN), which stands for the number of unrevealed nonzero coefficients; and (iii) dimension of the model, which is the number of nonzero coefficients.

4.1. Settings

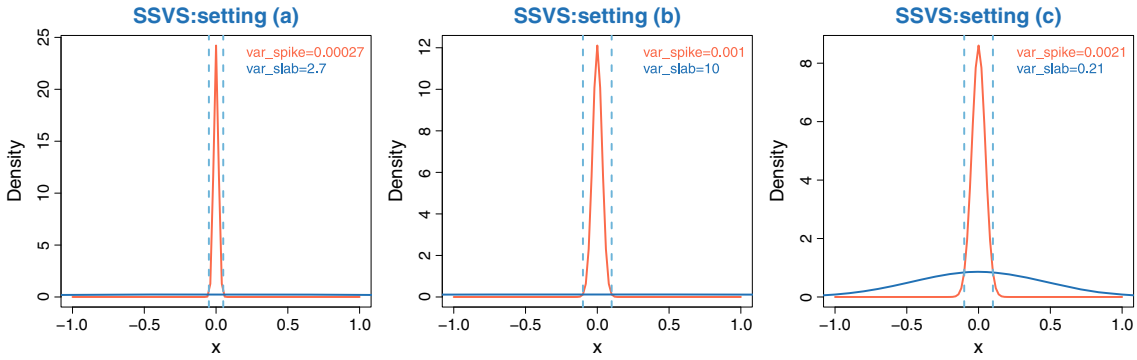
To assess the sensitivity of the spike and slab priors to the choice of tuning parameters, we considered three sets of hyperparameters. For SSVS, we selected these sets of hyperparameters, considering different values of the intersection point δ of the two normal mixture components and different ratio c^2 of the slab versus spike variance. We have the following settings: (i) $\delta = 0.05$ and $c = 100$ (spike variance $\text{var}_{\text{sp}} = 0.00027$ and slab variance $\text{var}_{\text{sl}} = 2.7$), (ii) $\delta = 0.1$ and $c = 100$ ($\text{var}_{\text{sp}} = 0.001$ and $\text{var}_{\text{sl}} = 10$), and (iii) $\delta = 0.1$ and $c = 10$ ($\text{var}_{\text{sp}} = 0.0021$ and $\text{var}_{\text{sl}} = 0.21$). We depict the three mixture

densities in Figure 1(a)–(c). We used similar settings in NMIG, where the parameters were chosen so that the intersection point of the scaled t-distributions and the ratio of the variances match to each of the three previous SSVS settings. We depict the NMIG mixture variance priors in Figure 2(a)–(c). We selected the mixture prior for GVS as in SSVS (ii). In the Bayesian LASSO and the Bayesian elastic net, the regularization parameters τ, a_1, and a_2 are assigned prior exp(0.01), where exp(μ) denotes the exponential distribution with the expectation $1/\mu$. A non-informative prior N(0, 1000) is used for the intercept term. Whenever applicable, we used the uniform prior on the model space.

For spike and slab models as well as for GVS, we investigated the highest posterior model (HPD) selection, the MPM selection, and the HS rule. Bayesian regularization (LASSO, ridge regression, and elastic net) enables only HS selection, whereas only MPM and HPD are applicable in MC^3. We based the interval decision criterion for HS on a one standard deviation interval around the posterior mean. We excluded from the final model only coefficients whose decision interval covers zero. Finally, we contrasted the Bayesian methodology with the F-to-out backward selection with $p = 0.05$ (STEP 1) and $p = 0.1$ (STEP 2) and exhaustive evaluation by using *AIC*. We based the MC^3 variable selection on the run of 1000 MCMC iterations. We estimated the remaining Bayesian hierarchical models by using 10,000 iterations with 1000 burn-in period and 10-fold thinning.

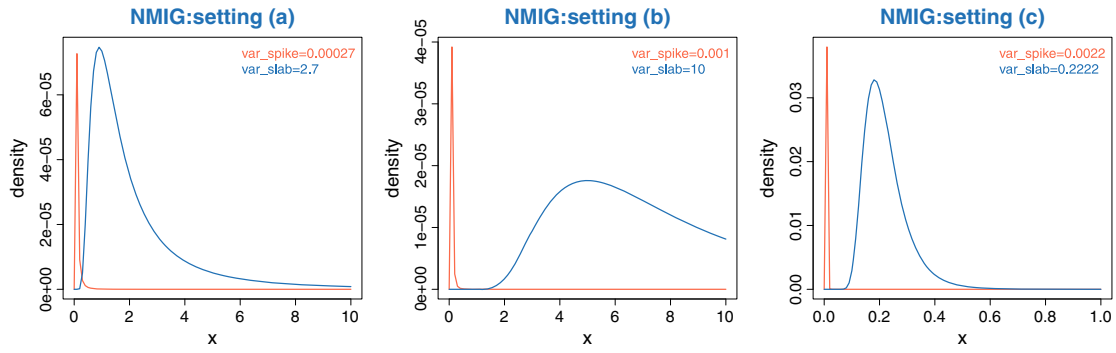
4.2. *Software*

The majority of the available software for Bayesian variable selection deals only with linear regression models. Shrinkage estimation by using sparsity priors (ridge, Laplace, normal/gamma, horseshoe) coupled with the reversible jump variable selection is obtainable through the `R` package `monomvn` of Gramacy and Pantaleo [27]. Spike and slab variable selection with NMIG priors can be found in the package `spikeSlabGAM` of Scheipl. Bayesian model averaging as well as MC^3 for linear regression



(a) $\delta = 0.05, c = 100$ (b) $\delta = 0.1, c = 100$ (c) $\delta = 0.1, c = 10$

Figure 1. Tuning parameters for the SSVS mixture priors.



(a) $a = 2, b = 2.7, v_0 = 0.0001$ (b) $a = 3, b = 20, v_0 = 0.0001$ (c) $a = 10, b = 2, v_0 = 0.01$

Figure 2. Tuning parameters for the NMIG mixture priors.

models has been implemented in the package `BMA` [19]. Bayesian regularized logistic regression applicable for high-dimensional data has been implemented in the package `reglogit` of Gramacy and Polson (technical report University of Chicago). The frequentist regularization for generalized linear models can be found in package `glmnet` [55]. A hybrid spike and slab variable selection procedure for linear (high-dimensional) regression has been made available in the package `spikeslab` of Ishwaran and Rao [30]. To implement the spike and slab models, Bayesian regularization, and GVS in the probit (Weibull) regression context, we used WinBUGS. We have adapted the code for SSVS from the BUGS code of Ntzoufras [56]. We have implemented the MC^3 for probit (Weibull) regression in `R`. Our source codes can be found in the Supporting Information.[‡]

4.3. Results

We summarize the results for MPM selection in Figure 3 and HS selection in Figure 4. For each of the methods, the figures present a triplet of bars. The left one represents the number of times a correct model has been identified (out of the 50 simulated data sets). These numbers relate to the vertical axis on the left. The middle and right bars correspond to the average FDN and FNN, respectively. These values relate to the vertical axis on the right from each graph. The average model dimension estimated by each of the methods is attached at the top of the three bars. Results for the highest posterior model selection greatly overlap with the MPM selection in first three designs (results not presented). The difference, however, emerged in Design 4, where the HPD model selection for all spike and slab models as well as MC^3 and GVS did not correctly identify the right model in any of the 50 repetitions.

Looking at the two figures, several observations can be made:

(1) The difference between HS selection and MPM selection is less apparent in the first three designs. Discrepancies again occur in Design 4, where MPM in 'properly calibrated' spike and slab models outperforms the regularization priors. In Design 4, as expected, the elastic net performed the best among the regularization priors in including the groups of correlated regressors in the model.
(2) The choice of tuning parameters is influential on spike and slab variable selection, which is particularly evident in Designs 2 and 4, where the performance increases with a decreasing variance in the slab component. This is a little at odds with the intuition that high hypervariance represents the prior belief that the coefficient can attain 'arbitrarily' large values. To explain this behavior, we note that in spike and slab models, the high slab hypervariance induces a stronger penalization on weak nonzero effects and hence expresses the prior opinion that many of the coefficients will be zero. To support this statement, we plotted the three mixture (SSVS) densities (Figure 5(a)) as well as their minus logarithms, which are proportional to the frequentist penalty functions (Figure 5(b)). Among the mixture priors, the setting with the lowest slab variance (SSVS (c)) places more prior emphasis on smaller effects. This forces the penalty to elevate more gradually with an increasing distance from origin and indeed causes the small nonzero effects to be penalized to a lesser extent. On the other hand, the shape of the penalty function arising from the 'narrow spike wide slab' prior (setting (a)) provides the closest approximation to the l_0 type of penalty, which penalizes nonzero effects equally regardless their magnitude.
(3) The distributional assumption underlying the constitution of spike and slab can influence the variable selection. With the comparison of the Gaussian and Student mixtures with matched variances and intersection points, the t-mixture implies weaker penalization of larger effects due to heavier slab tails (Figure 5(c)). The impact of this behavior is particularly evident in the non-sparse designs (Designs 2 and 4). On the other hand, the two spike and slab models exhibit similar mixing properties (evaluated by the number of visited models) in all the four designs, and their computation time in our implementation was comparable.
(4) The classical frequentist variable selection was outperformed by spike and slab variable selection (regardless of the posterior inference) in the two sparse regression designs. The Bayesian regularization, on the other hand, emerged as more accurate (compared with the backward selection and the *AIC* full subset selection) in finding the true underlying model, as long as there were many nonzero effects.

[‡]*Supporting information may be found in the online version of this article.*

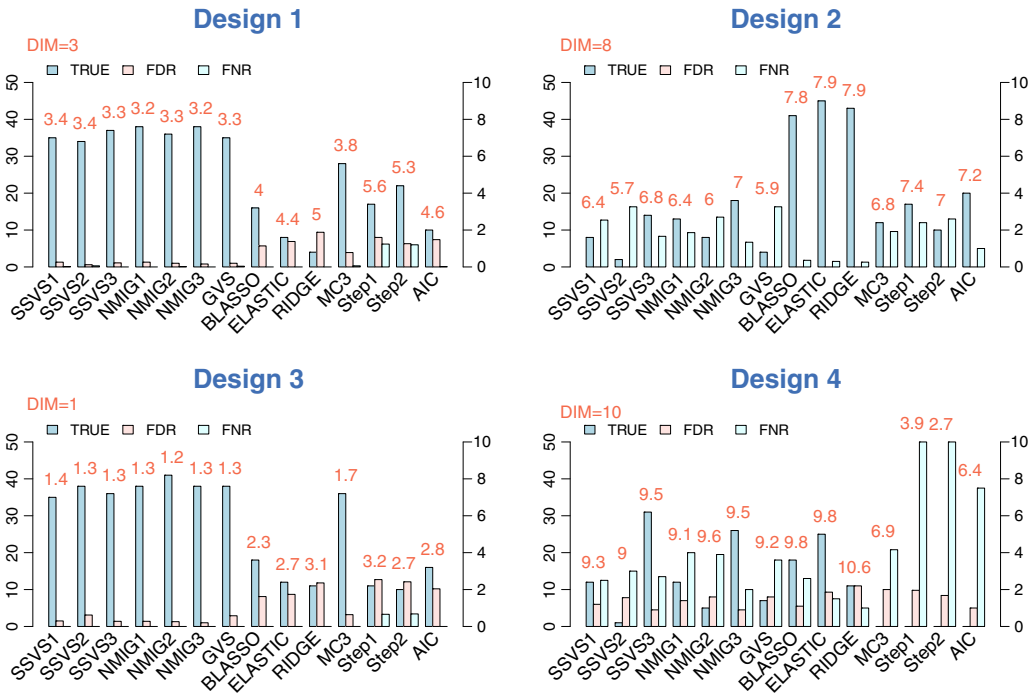


Figure 3. Simulation results: MPM selection (FDR and FNR refer to the average FDN and FNN numbers, respectively).

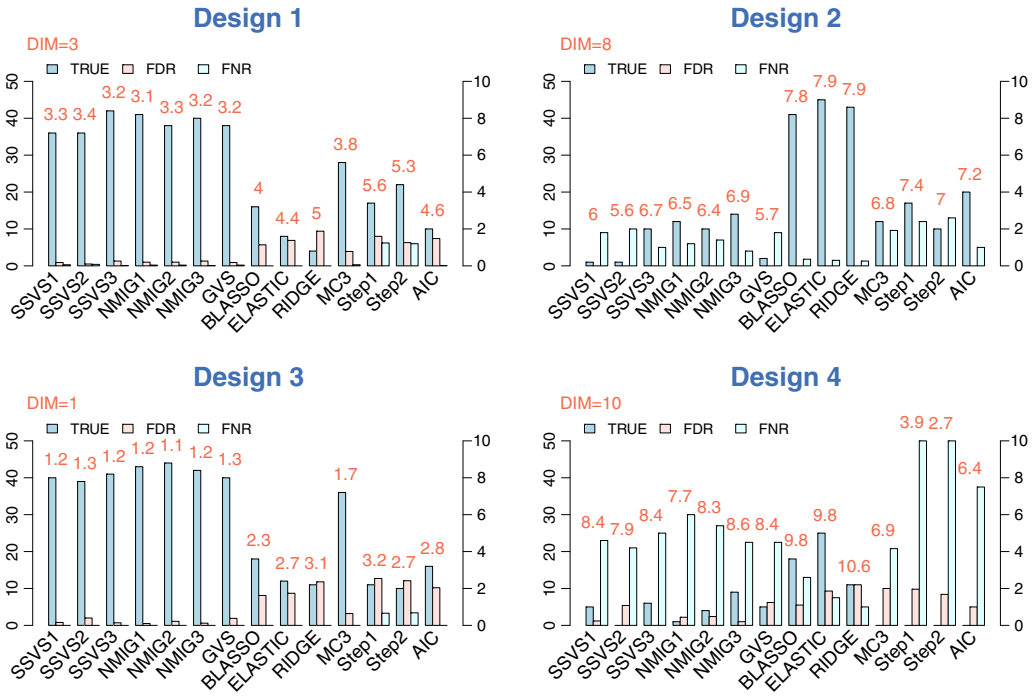
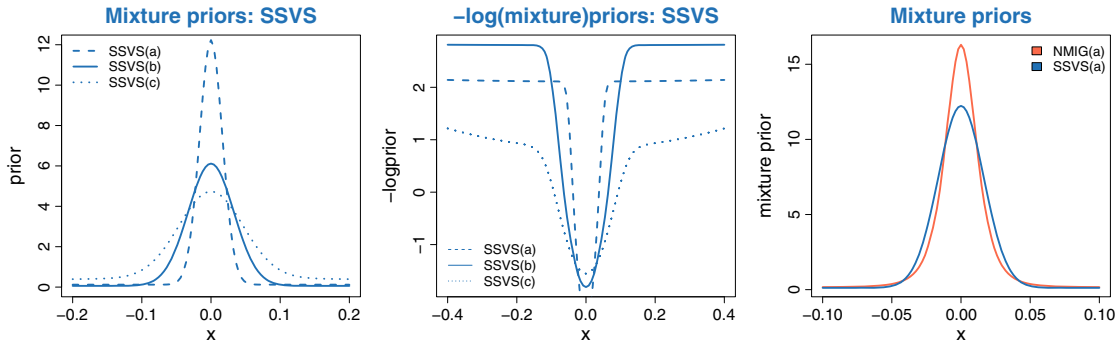


Figure 4. Simulation results: HS model selection (FDR and FNR refer to the average FDN and FNN numbers, respectively).

5. Data analysis

In Section 1, we presented the frequentist analysis of the two data sets. Whereas in the classical approach, the emphasis is often put on finding a single representation of the data by one model, the



(a) SSVS (b) -log(SSVS) (c) SSVS & NMIG

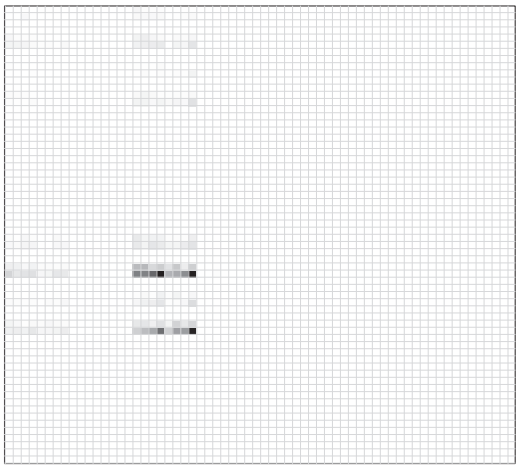
Figure 5. Left panel: minus logarithm of the three SSVS mixture priors. Middle panel: minus logarithm of SSVS(a) prior and NMIG(a) prior. Right panel: SSVS(a) and NMIG(a) priors.

Bayesian approach enables to assess uncertainty surrounding such decision and prepares grounds for the eventual model averaging. In the analysis of the REACH data, we apply the Bayesian model selection and uncertainty assessment via posterior model probabilities, as well as the shrinkage and MPM variable selection. In the AML data, we elaborate further on the shrinkage approaches and the 'model-averaged' Bayesian variable selection. In both the analyses, all continuous regressors were standardized. We based the estimation on Markov chains of the length 10,000 for MC^3 and 15,000 with a burn-in 5000 thinned by 10 for all the other Bayesian models.

5.1. The Rotterdam Early Arthritis Cohort data

In the simulation study, we have seen that the variable selection (implied by HS, posterior inclusion probabilities, or posterior model probabilities) is influenced by the prior specification, both in terms of the choice of the prior (mixture) distribution and hyperparameter calibration. To amplify this point, we applied the same prior settings on the REACH data. We present variables selected by the highest posterior model selection and HS, if applicable, for each of the Bayesian variable selection method in Figure 6(b). We do not present the MPM selection separately as it yields the same models as the HPD selection for all the methods.

Again, we see how sensitive the Bayesian variable selection can be towards the prior settings. In HPD variable selection, this 'sensitivity' connects to the mixing properties of the chain sampling individual models. There, the actual model selection depends on the chain's ability to find interesting regions of



	ACCP	Age	ESR	DC	Stiff.	RF	Sex	Sym.	SJC	TJC	BCPH	BCPF
Highest posterior model selection												
SSVS1	■	□	■	□	□	□	□	■	□	□	■	□
SSVS2	■	□	■	□	□	□	□	■	□	□	■	□
SSVS3	■	□	■	□	□	■	□	■	■	□	■	□
NMIG1	■	□	■	□	□	■	□	■	■	□	■	□
NMIG2	■	□	■	□	□	□	□	■	□	□	■	□
NMIG3	■	□	■	□	□	■	□	■	■	□	■	□
MC3	■	□	■	■	□	□	□	■	■	□	■	□
GVS	■	□	■	□	□	□	□	■	□	□	■	□
Hard shrinkage												
SSVS1	■	□	■	□	□	□	□	■	□	□	■	□
SSVS2	■	□	■	□	□	□	□	■	□	□	■	□
SSVS3	■	□	■	■	□	■	□	■	■	□	■	□
NMIG1	■	□	■	□	□	■	□	■	■	□	■	□
NMIG2	■	□	■	□	□	□	□	■	□	□	■	□
NMIG3	■	□	■	■	□	■	□	■	■	□	■	□
BLASSO	■	□	■	■	■	■	□	■	■	□	■	■
ELASTIC	■	□	■	■	■	■	□	■	■	□	■	■
RIDGE	■	□	■	■	■	■	□	■	■	□	■	■
Frequentist												
STEP1	■	□	■	■	□	□	□	■	■	□	■	□
STEP2	■	□	■	■	■	■	□	■	■	□	■	□
AIC	■	□	■	■	■	■	□	■	■	□	■	□

(a) MC^3 approximation (b) Table of models selected by BVS

Figure 6. (a) MC^3 stochastic approximation to the posterior distribution over all models for the REACH data. (b) The table of models selected by the different Bayesian variable selection methods.

the posterior model space. We depict one particular stochastic approximation to the posterior model distribution for the REACH data, which was obtained by MC^3, in Figure 6(a). We sorted the individual models in a matrix, where the simplest one (no covariates at all) is located in the lower right corner and the full model in the upper left corner. The logic of the sorting is that related models create blocks in the matrix. The darker the gray color of each squared spot (model), the more often the model was visited by the MC^3 Markov chain. The darkest spot then corresponds to the estimated HPD model (with covariates *ACCP*, *ESR*, *DC*, *Sym*, *SJC*, and *BCPH*), which the chain occupied 576 times during the run of 10,000 iterations. There are clearly more candidate models with a high number of visits. In fact, the second most frequent model was visited 571 times. Therefore, the posterior evidence contained within the estimated model probabilities from MC^3 does not vote unequivocally for the selection of just one model. Out of the 4096 possible models, the MC^3 Markov chain encountered only 229 different models. The number of visited models in GVS was only 33, whereas in the three versions of SSVS (NMIG), there were 80, 56, and 175 (126, 74, and 176) different models among the 1000 recorded MCMC iterations. The fast mixing ability of the last SSVS and NMIG specification is to be expected, because the variance of the slab is sufficiently small, allowing the chain to escape from the spike more easily. The sharpness of the prior spike together with the magnitude of the prior slab variance hence influences how many models will be visited and implicitly determines the shape of the posterior distribution of individual coefficients. If models with a particular variable included were not often encountered in the sequence of sampled models, the spike will dominate the posterior shape of the corresponding coefficient, which will result in shrinkage of the posterior mean towards zero.

The regularization priors such as ridge, elastic net, or LASSO induce rather 'soft' shrinkage, leading to many nonzero selected coefficients, whereas the shrinkage from spike and slab priors is more aggressive, especially when the slab versus spike variance ratio is sufficiently large. If the preference is to select a model with all the included variables strongly associated with the outcome, we might opt for spike and slab variable selection with a 'sharp spike and flat slab' shape. In this case, we would end up with a model with only four covariates *ACCP*, *ESR*, *Sym*, and *BCPH*. Relaxing the requirements for the model parsimony and giving preference to a model suitable rather for prediction, we might choose the model indicated by the Bayesian regularization, which contains one extra variable compared with the *AIC* model.

A similar interplay between the model complexity and practical significance of included factors can be achieved by selecting different significance thresholds in stepwise selection. However, the Bayesian variable selection (HS and MPM) in spike and slab models accounts for the uncertainty introduced by the model selection process, because the posterior distribution, on which the decision is based on, is averaged over more candidate models.

5.2. *The acute myeloid leukemia data*

Unlike in the analysis of REACH data, where we applied all three types of posterior inference for variable selection (i.e., HPD, MPM, and HS), in the AML data, we assess the variable selection only via posterior inclusion probabilities (MPM) and HS. We suspect that the approximation to the posterior model distribution provided by the spike and slab models may not be sufficiently accurate to find the highest posterior model in the setting with this many variables. Furthermore, the posterior inclusion probabilities and posterior distribution of coefficients provide model-averaged decision criterion that is potentially more reliable than just comparing individual model probabilities.

We summarize the results of Bayesian variable selection applied on the AML data in Figure 7. The upper panel depicts estimated marginal inclusion probabilities. According to the MPM selection rule, a variable is included in a model whenever the inclusion probability exceeds 0.5 (indicated by the horizontal line). The lower panel displays point estimates (posterior means) for each regression coefficient, accompanied with $\pm$SD inclusion interval. The point estimates are weighted averages of estimated posterior means arising from visited models with underlying slab (spike) prior on the present (absent) coefficients. We determine the weights from the frequencies of visits of each model. Intervals that exclude zero (again marked by the horizontal line) imply the inclusion of the variable in the model by the HS rule. The NMIG appeared to show poorer mixing (358, 23, and 118 visited models for each of the settings compared with 763, 127, and 626 models for SSVS). That is why we present the results only from SSVS spike and slab models, as we believe they are more reliable.

To compare the different shrinkage behavior of the spike and slab prior and the regularization prior, we depict the approximations to the posterior distribution of three selected coefficients (for variables

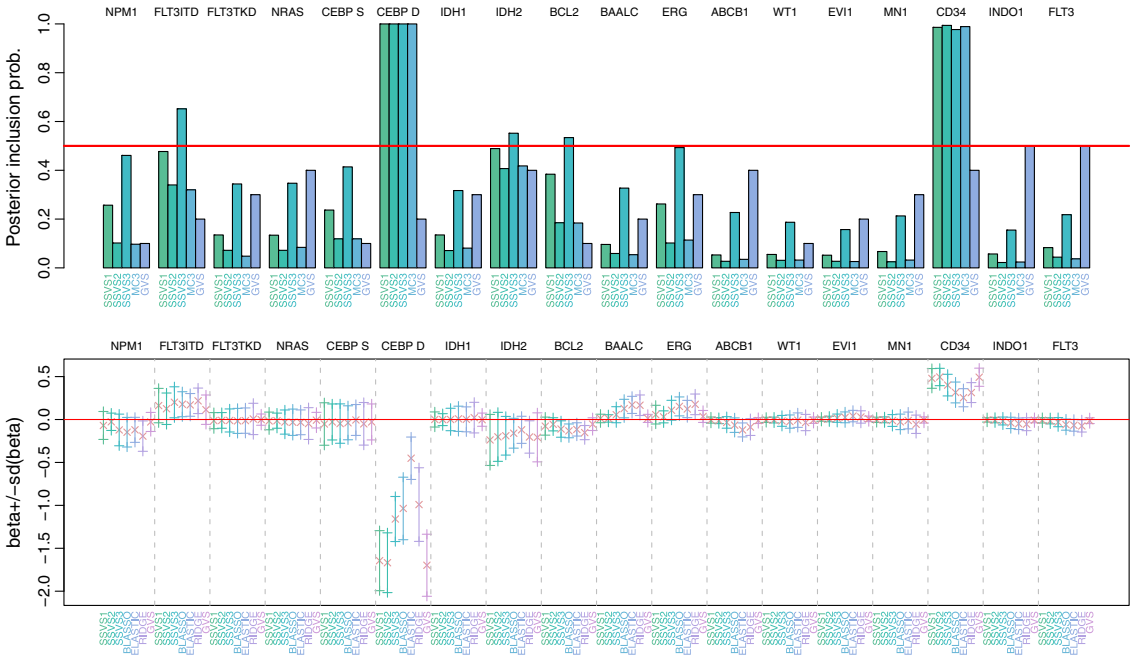


Figure 7. AML data: the upper panel depicts the estimated inclusion probabilities for each of the markers, and the lower panel depicts the estimated coefficients together with ±SD interval.

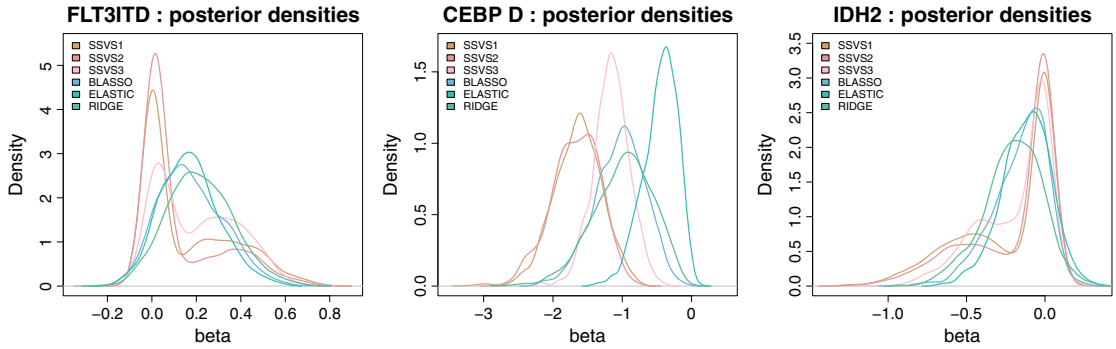


Figure 8. Approximation to the posterior distribution of three selected coefficients.

FLT3ITD, *CEBPA DM*, and *IDH2*) in Figure 8. For the first coefficient (*FLT3ITD*), the evidence for the inclusion is not strongly convincing, and therefore, the spike and slab posteriors are bimodal. We observe the sharpest posterior spike (i.e., the strongest penalization of larger values) for the SSVS with the biggest prior slab variance (setting (b)). Less stringent penalization of the SSVS setting (c) is evident from the pronounced bimodal shape of the posterior. In the case of *CEBPA DM*, the Bayesian regularization places heavier penalties on larger effects (compared with the inflated-slab-variance priors), which inevitably introduces estimation bias.

Conclusively, the Bayesian approach gives strong evidence for the two markers *CEBPA DM* and *CD34*. Some of the methods included variables *FLT3ITD*, *ERG*, or *BCL2*, but their estimated effects are rather small. In the frequentist approach, we ended up with models that are quite complex, whereas the Bayesian approach points at more parsimonious models and enables to quantify the importance of each individual marker by means of a posterior inclusion probability or posterior distribution of the coefficients. These summaries are averaged over posterior model uncertainty and therefore provide more objective quantitative assessment than p-values.

6. Discussion

The purpose of this article was to survey the evolution of Bayesian variable selection and highlight some of its recent developments. The list of the discussed methodology is surely not exhaustive as the methodology is continuously evolving and its potential has only begun to be realized. We have restricted our attention to the general discussion on the principles rather than technical details on implementation by using sophisticated MCMC techniques. We have omitted discussion on the nonparametric relaxations of considered hierarchical models by using the Dirichlet process priors [57–59] as well as application of the prior hierarchies on factor analytic models [60], additive regression models [35], and others.

The practical utility of the Bayesian methodology (regularization and spike and slab models) would be particularly appreciated in the analysis of high-dimensional data (genomics, proteomics), where the estimation in Bayesian hierarchical models constitutes a coherent alternative to approaches based on corrections of multiple testing. Here, we have confined our application to the classical regression settings and, in the simulation study, demonstrated that non-negligible practical gain can be obtained, yet less involved, also in these modeling tasks. Among the outlined methodology, the spike and slab models constitute an approach that is particularly, conceptually appealing. They are closely connected to the Bayesian regularization in the sense that they provide a Bayesian framework that gives rise to the similar type of penalties as the l_0 frequentist complexity penalty. Abramovich *et al.* [61] have pursued the Bayesian formalism for these penalties in the context of high-dimensional normal means models. The spike and slab models, however, provide a different perspective on the l_0 frequentist penalization. The connection between penalized l_0 estimation and Bayesian spike and slab models follows quite analogously as between Bayesian MAP estimates from the Laplace priors and the frequentist LASSO. The frequentist implementation of the optimization problem in l_0 penalized models is hampered by the non-singularity and discontinuity of this penalty at origin. Continuous approximations to this frequentist penalty have been suggested that facilitate the computation [62]. On the other hand, the penalty induced by the Bayesian mixture priors (which is proportional to the logarithm of the mixture prior) can be regarded as another type of continuous approximation to the l_0 type of penalty. The advantage of the Bayesian formulation is that the MCMC machinery can be used to obtain the approximation to the whole posterior distribution instead of employing some involved optimization techniques to find the posterior mode.

In this paper, we have restricted our attention to $p < n$ setting. Nevertheless, the modern applications of Bayesian variable selection deal mostly with high-dimensional data. The complexity of such problems renders several presented Bayesian variable selection methods less appealing from the computational time and storage efficiency standpoints. Kwon *et al.* [63] and Yang and Song [52] have considered adaptations of SSVS algorithm suitable for high-dimensional data that avoid sampling the individual regression coefficients. Despite the advances in high-dimensional stochastic model search [5, 25], the shrinkage approaches (eventually accompanied with the reversible jump sampling) might be preferred in such situations [27]. Alternatively, the involved MCMC computation in hierarchical shrinkage models can be avoided using EM algorithm [64].

We exemplified the Bayesian hierarchical models for variable selection in probit regression and Weibull regression. Previously, Sha *et al.* [51], Kwon *et al.* [65], Yang and Song [52], Zhou *et al.* [53], Bae and Mallick [39] have discussed the Bayesian variable selection methods in the context of probit regression models and Sha *et al.* [54] in survival models. Although our WinBUGS programs offer a working solution to fitting the hierarchical models with sparsity/variable selection priors in low-dimensional settings, customized algorithms/implementations are needed in higher dimensions. For instance, the Bayesian regularized logistic regression has been implemented in package `reglogit`. Nevertheless, the majority of the discussed hierarchical constructions are still awaited to be transferred to/implemented in other than linear regression settings.

In the simulation study, we demonstrated that Bayesian variable selection leads to improved performance in identifying the true underlying model, when compared with the frequentist methods. We used several Bayesian variable selection approaches, none of which could be postulated as the methodological ideal for all the considered simulation settings and neither it should be. The choice of the particular Bayesian approach should be context dependent as some of the discussed methodologies are customized for particular data structures (groups of correlated predictors) and inferential goals (prediction rather than variable selection). Information regarding the correlation structure and the expected dimension of the solution can be beneficial when finding the 'true' pattern of sparsity.

In the theoretical discussion, we focused mainly on absolutely continuous priors, also within the spike and slab context. The point mass spike and slab priors [16, 27] on the other hand offer a correct characterization of the model uncertainty and avoid making subjective choices on tuning hyperparameters. These facts have contributed to the fact that the point mass priors have begun to be realized as benchmark for Bayesian variable selection. Recently, point mass shrinkage priors have been made available through standard software [27] for linear regression.

Despite the conceptual appeal of Bayesian variable selection, the wide acceptance of it as the preferred variable selection strategy has been hampered by the unavailability of implementation in standard software. Catalyzed by advances in the MCMC computation, the methodology has become no longer problematic to implement in Bayesian software such as WinBUGS for the classical regression settings. However, the computational challenges increase with the dimensionality of the data, where developments in numerical approximations and/or MCMC techniques will hopefully make the methodology more approachable for more practically oriented users.

References

1. Copas JB. Regression, prediction and shrinkage. *Journal of the Royal Statistical Society. Series B (Methodological)* 1983; **45**(3):311–354.
2. Breiman L. Heuristics of instability and stabilization in model selection. *The Annals of Statistics* 1996; **24**(6):2350–2383.
3. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 2001; **96**:1348–1360.
4. Hocking R. The analysis and selection of variables in linear regression. *The Annals of Statistics* 1976; **32**(1):1–49.
5. Hans C, Dobra A, West M. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* 2007; **102**(478):507–516.
6. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 1994; **58**:267–288.
7. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B (Methodological)* 2006; **68**(1):49–67.
8. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Methodological)* 2005; **67**:301–320.
9. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 1970; **12**(1): 55–67.
10. Park T., Casella G. The Bayesian lasso. *Journal of the American Statistical Association* 2008; **103**(482):681–686.
11. Frank IE, Friedman JH. A statistical view of some chemometrics regression tools. *Technometrics* 1993; **35**(2):109–135.
12. Fu WJ. Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* 1998; **7**(3):397–416.
13. George EI, Foster DP. Calibration and empirical Bayes variable selection. *Biometrika* 2000; **87**:731–747.
14. Carlin BP, Chib S. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)* 1995; **57**(3):473–484.
15. Ishwaran H, Rao JS. Detecting differentially expressed genes in microarrays using Bayesian model selection. *Journal of the American Statistical Association* 2003; **98**(462):438–455.
16. Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. *Journal of the American Statistical Association* 1988; **83**(404):1023–1032.
17. George EI, McCulloch RE. Variable selection via Gibbs sampling. *Journal of the American Statistical Association* 1993; **88**(423):881–889.
18. George EI, McCulloch RE. Approaches for Bayesian variable selection. *Statistica Sinica* 1997; **7**(2):339–373.
19. Madigan D, Raftery A, Wermuth N, York J, Zucchini W. Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association* 1994; **89**:1535–1546.
20. Madigan D, York J, Allard D. Bayesian graphical models for discrete data. *International Statistical Review (Revue Internationale de Statistique)* 1995; **63**(2):215–232.
21. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 1953; **21**(6):1087–1092.
22. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970; **57**(1): 97–109.
23. Schwarz G. Estimating the dimension of a model. *The Annals of Statistics* 1978; **6**(2):461–464.
24. Barbieri MM, Berger JO. Optimal predictive model selection. *The Annals of Statistics* 2004; **32**(3):870–897.
25. Bottolo L, Richardson S. Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis* 2010; **5**(3):583–618.
26. Green P. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995; **82**:711–732.
27. Gramacy RB, Pantaleo E. Shrinkage regression for multivariate inference with missing data, and an application to portfolio balancing. *Bayesian Analysis* 2010; **5**:237–262.
28. Lunn DJ, Best N, Whittaker JC. Generic reversible jump MCMC using graphical models. *Journal Statistics and Computing* 2009; **19**:395–408.
29. Leamer EE. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. Wiley: New York, 1978.

30. Ishwaran H, Rao JS. Spike and slab variable selection: frequentist and Bayesian strategies. *The Annals of Statistics* 2005; **33**(2):730–773.
31. Johnstone IM, Silverman BW. Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics* 2004; **32**:1594–1649.
32. Dellaportas P, Forster JJ, Ntzoufras I. On Bayesian model and variable selection using MCMC. *Statistics and Computing* 2002; **12**(1):27–36.
33. George EI, McCulloch RE, Tsay RS. Two approaches to Bayesian model selection with applications. In *Bayesian Analysis in Statistics and Econometrics*. Wiley: New York, 1996; 339–348.
34. Ntzoufras I, Forster JJ, Dellaportas P. Stochastic search variable selection for log-linear models. *Journal of Statistical Computation and Simulation* 2000; **68**(1):23–37.
35. Fahrmeir L, Kneib T, Konrath S. Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Statistics and Computing* 2010; **20**(2):203–219.
36. Box GEP, Tiao GC. *Bayesian Inference in Statistical Analysis*. Addison-Wesley: Boston, 1973.
37. Andrews DR, Mallows CL. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)* 1974; **36**(1):99–102.
38. Griffin JE, Brown PJ. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis* 2010; **5**:17–188.
39. Bae K, Mallick BK. Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics* 2004; **20**(18): 3423–3430.
40. Figueiredo M. Adaptive sparseness using Jeffreys prior. *Advances in Neural Information Processing Systems* 2002; **14**:697–704.
41. Carvalho CM, Polson NG, Scott JG. The horseshoe estimator for sparse signals. *Biometrika* 2010; **97**:465–480.
42. Hans C. Model uncertainty and variable selection in Bayesian lasso regression. *Statistics and Computing* 2010; **20**:221–229.
43. West M. On scale mixtures of normal distributions. *Biometrika* 1987; **74**(3):646–648.
44. Carlin BP, Polson NG. Inference for nonconjugate Bayesian models using the Gibbs sampler. *The Canadian Journal of Statistics* 1991; **19**:399–405.
45. Hans C. Bayesian lasso regression. *Biometrika* 2009; **96**:835–845.
46. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 2005; **67**(1):91–108.
47. Kyung M, Gilly J, Ghoshz M, Casella G. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis* 2010; **5**(2):369–412.
48. Li Q, Lin N. The Bayesian elastic net. *Bayesian Analysis* 2010; **5**(1):151–170.
49. Albert JH, Chib S. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 1993; **88**(422):669–679.
50. Holmes CC, Held L. Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis* 2006; **1**:145–168.
51. Sha N, Vannucci M, Tadesse MG, Brown PJ, Dragoni I, Davies N, Roberts TC, Contestabile A, Salmon M, Buckley C. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics* 2004; **60**(3):812–819.
52. Yang A-J, Song X-Y. Bayesian variable selection for disease classification using gene expression data. *Bioinformatics* 2010; **26**(2):215–222.
53. Zhou X, Liu KY, Wong ST. Cancer classification and prediction using logistic regression with Bayesian gene selection. *Journal of Biomedical Informatics* 2004; **37**(4):249–259.
54. Sha N, Tadesse MG, Vannucci M. Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* 2006; **22**(18):2262–2268.
55. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 2010; **33**(1):1–22.
56. Ntzoufras I. Gibbs variable selection using BUGS. *Journal of Statistical Software* 2002; **7**(7):1–19.
57. Nott DJ. Bayesian methods for highly correlated exposure data. *Epidemiology* 2008; **28**(3):199–207.
58. Nott DJ. Predictive performance of Dirichlet process shrinkage methods in linear regression. *Computational Statistics & Data Analysis* 2008; **52**(7):3658–3669.
59. Kim S, Dahly DB, Vannucci M. Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis* 2009; **4**(4):707–732.
60. Carvalho CM, Chang J, Lucas JE, Nevins JR, Wang Q, West M. High-dimensional sparse factor modelling: applications in gene expression genomics. *Journal of the American Statistical Association* 2008; **103**(484):1438–1456.
61. Abramovich F, Angelini C, De Canditiis D. On optimality of Bayesian testimation in the normal means problem. *Annals of Statistics* 2007; **35**(5):2261–2286.
62. Liu Y, Wu Y. Variable selection via a combination of the l_0 and l_1 penalties. *Journal of Computation and Graphical Statistics* 2007; **16**(4):782–798.
63. Kwon D, Landi MT, Vannucci M, Issaw HJ, Prieto D, Pfeiffer RM. An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Computational Statistics and Data Analysis* 2011; **55**(10):2807–2818.
64. Kiiveri HT. A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations. *BMC Bioinformatics* 2008; **9**(195):1–9.
65. Kwon D, Tadesse MG, Sha N, Pfeiffer RM, Vannucci M. Identifying biomarkers from mass spectrometry data with ordinal outcome. *Cancer informatics* 2007; **3**(4):19–28.