Chapter 5
Bayesian Variable Selection

Matthew Sutton

Abstract In this chapter we survey Bayesian approaches for variable selection and model choice in regression models. We explore the methodological developments and computational approaches for these methods. In conclusion we note the available software for their implementation.

5.1 Introduction

Bayesian variable selection methodology has been progressing rapidly in recent years. While the seminal work of the Bayesian spike and slab prior [1] remains the main approach, continuous shrinkage priors have received a large amount of attention. There is growing interest in speeding up inference with these sparse priors using modern Bayesian computational approaches. Moreover, the subject of inference for these sparse models has become an increasingly important area of discussion among statisticians. A common theme among Bayesian variable selection methods is that they aim to select variables while also quantifying uncertainty through selection probabilities and variability of the estimates. This chapter gives a survey of relevant methodological and computational approaches in this area, along with some descriptions of available software.

5.2 Preliminaries

5.2.1 The Variable Selection Problem

In the context of variable selection for a regression model we consider the following canonical problem in Bayesian analysis. Suppose we want to model a sample of n

M. Sutton (✉)
Queensland University of Technology, Brisbane, QLD, Australia
e-mail: m.sutton5@lancaster.ac.uk

K. L. Mengersen et al. (eds.), *Case Studies in Applied Bayesian Data Science*,
Lecture Notes in Mathematics 2259, https://doi.org/10.1007/978-3-030-42553-1_5

observations of a response variable $Y \in \mathbb{R}^n$ and a set of p potential explanatory variables $X_1, \ldots, X_p$, where $X_j \in \mathbb{R}^n$. The variable selection problem is to find the 'best' model between the response Y and a subset of $X_1, \ldots, X_p$ where there is uncertainty in which subset to use. Throughout this chapter, we index each of the possible 2^p subset choices by the vector

$$\gamma = (\gamma_1, \ldots, \gamma_p)^T,$$

where $\gamma_j = 1$ if variable X_j is included in the model, and $\gamma_j = 0$ otherwise. We let $s_\gamma = \sum_{j=1}^{p} \gamma_j$ denote the number of selected variables for a model indexed by γ. Given γ, suppose that Y has density $p(Y \mid \beta_\gamma, \gamma)$ where β_γ is a vector of unknown parameters corresponding to the variables indexed by γ. The Bayesian approach assigns a prior probability to the space of models $p(\gamma)$, and a prior to the parameters of each model $p(\beta_\gamma \mid \gamma)$.

The probability for the model with the selected variables γ conditional on having observed Y, is the posterior model probability

$$p(\gamma \mid Y) = \frac{p(Y \mid \gamma)p(\gamma)}{\sum_{\gamma' \in \{0,1\}^p} p(Y|\gamma')p(\gamma')},$$

where

$$p(Y \mid \gamma) = \int p(Y \mid \gamma, \beta_\gamma)p(\beta_\gamma \mid \gamma)\mathrm{d}(\beta_\gamma),$$

is the marginal likelihood of Y. The priors $p(\beta_\gamma \mid \gamma)$ and $p(\gamma)$ provide an initial representation of model uncertainty and the posterior adjusts for the information in Y, allowing us to quantify the uncertainty of the variable selection. The actual variable selection in a Bayesian analysis can proceed in several ways. Two common approaches are:

1. Select the variables with the highest estimated posterior probability $p(\gamma \mid Y)$, also known as the highest posterior density model (HPD),
2. Select variables with estimated posterior probability of inclusion $p(\gamma_j = 1 \mid Y)$ greater than 0.5, also known as the median probability model (MPM).

The appropriateness of the HPD and MPM model have been studied in detail [2, 3]. It has been shown that for orthogonal linear regression, the optimal model from a Bayesian predictive objective is the MPM rather than the HPD.

In a Bayesian framework, the accuracy of the variable selection method depends on the specification of the priors for the model space and parameters. In this section, we survey priors which fall into one of four possible categories, priors on the model space, spike and slab priors, shrinkage priors and projection methods.

5.2.2 Model Space Priors

We begin by considering priors on the model space $p(\gamma)$. A common prior on the model space assumes that the γ_j are independent and Bernoulli distributed,

$$p(\gamma) = \prod_{j=1}^{p} w_j^{\gamma_j} (1 - w_j)^{1-\gamma_j}, \tag{5.1}$$

is computationally inexpensive and has been found to give sensible results in practice [4–7]. Under this prior, each variable X_j will enter the model with probability $p(\gamma_j = 1) = w_j$. A common variant of this method is to place a Beta prior on $w \sim Beta(a, b)$ which yields

$$p(\gamma) = \frac{B(a + s_\gamma,\ b + p - s_\gamma)}{B(a,\ b)},$$

where $B(a, b)$ is the beta function with hyper-parameters a and b. The choice of $a = b = 1$ corresponds to an uninformative prior on the model space. This type of prior is also recommended in [8], where the choice of hyper-parameters is considered asymptotically. More generally, one can put a prior $h(s_\gamma)$ on the model dimension and let

$$p(\gamma) = \binom{p}{s_\gamma}^{-1} h(s_\gamma),$$

which allows for the belief that the optimal models are sparse [16]. Priors of this form are considered generally by Scott in [9]. The priors described so far are useful when there is no structural information about the predictors.

Structured priors have also been considered, for example [10] propose a model space prior which incorporates known correlation in the predictors. They assume that the covariates have an underlying graphical structure and use an Ising prior to incorporate the structural information (see [11] for a survey on the Ising model). This structural information is used to capture underlying biological processes in the modelling.

5.2.3 Spike and Slab Priors

We now consider the specification of the prior for the parameters $p(\beta_\gamma \mid \gamma)$. Arguably, one of the simplest and most natural classes of prior distributions is given by the spike and slab type priors. In the original formulation [1, 12] the spike and slab distribution was defined as a mixture of a Dirac measure concentrated at zero and a uniform diffuse component. Similar to [13], we use a more general version

of the prior. In this chapter we refer to a spike and slab as any mixture of two distributions where one component is peaked at zero and the other is diffuse. More specifically, we define a spike and slab to have the form,

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)G_0(\beta_j) + \gamma_j G_1(\beta_j),$$

for $j = 1, \ldots, p$ where G_0 and G_1 are probability measures on $\mathbb{R}$ and $\gamma \sim p(\gamma)$, where $p(\gamma)$ is a prior on the model space. This framework naturally extends the model space prior discussed in the previous section. The original spike and slab (Mitchell et al. [1]) corresponds to a Dirac mass at zero δ_0 for G_0 and a uniform slab distribution for G_1.

For this section, we will assume an independent Bernoulli prior for γ_j, where $\gamma_j \sim Bernoulli(w_j)$, and $w_j \in [0, 1]$ for $j = 1, \ldots, p$. Using this prior on the model space the spike and slab can be written as the mixture

$$\beta_j \mid w_j \sim (1 - w_j)G_0(\beta_j) + w_j G_1(\beta_j),$$

where we have marginalised over the binary term γ_j. There are a number of prior specifications which use this hierarchical setup but differ in the distributions chosen for G_0 and G_1 [14]:

Kuo and Mallick The Bernoulli–Gaussian or Binary Mask model is due to [15]. This prior takes a Dirac for the spike $G_0 = \delta_0$ and a Gaussian for the slab G_1,

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)\delta_0 + \gamma_j N(0, \sigma_\beta^2),$$

where $N(\mu_\beta, \sigma_\beta^2)$ denotes a Normal distribution with mean μ_β and standard deviation σ_β. The slab distribution is chosen with sufficiently large variance to allow the non-zero coefficients to spread over large values. As noted by O'Hara and Sillanpää [14] this method can suffer poor mixing in an MCMC implementation due to the sharp shrinkage properties of the Dirac measure.

Stochastic Search Variable Selection (SSVS) A related method for variable selection is the stochastic search variable selection (SSVS) or Normal-Normal formulation proposed by George and McCulloch [6]. This prior has the aim of excluding variable β_j from the model whenever $|\beta_j| < \epsilon_j$ given $\epsilon_j > 0$ and where $|\cdot|$ denotes the absolute value. The idea is that ϵ_j is a practical threshold that can aid the identification of variables with effect size larger than some specified value. The prior has the form,

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j\tau_j^2),$$

where the separation between the two components is controlled through the tuning parameters τ_j and $c_j > 0$ which control the variance of the spike τ_j^2 and the variance of the slab $\tau_j^2 c$. To help guide the choice of these tuning parameters, [6] and [16]

note that the two Gaussians intersect at the points $\pm\epsilon_j$ where

$$\epsilon_j = \tau_j \sqrt{2\log(c_j)c_j^2/(c_j-1)}.$$

Thus posterior coefficients within the interval $[-\epsilon_j, \epsilon_j]$ can be considered "practically zero". They suggest using this to aid in the selection of the hyper-parameters τ_j and c_j. A variant of this prior is called the Gibbs variable selection (GVS) method suggested by Dellaportas et al. [17] and Carlin and Chib [18]. This method was motivated to improve convergence in MCMC implementations by reducing the sharp shrinkage of the Dirac. Their method suggests that the distribution G_1 corresponding to $\gamma_j = 0$ should be chosen so that it has no effect on the posterior. When the likelihood is Normal this method follows a similar form as the SSVS method where G_1 is a normal distribution with mean and variance chosen to minimise the effect on the posterior. This method can have good mixing properties but is difficult to tune in practice [14].

A recent extension of the SSVS type of prior was proposed by Narisetty and He [19] who propose a spike and slab priors that are Normal, but where the prior parameters depend explicitly on the sample size to achieve appropriate shrinkage. They establish model selection consistency in a high-dimensional setting, where p can grow nearly exponentially with n.

Normal Mixture of Inverse Gamma (NMIG) For linear regression, [20] proposed to move the spike and slab to the variance term rather than placing a prior on the parameter itself. The form of their prior parameterised the variance as a product of random variables with inverse gamma distribution (IG) and a Dirac. We state the equivalent parameterisation of this spike and slab model [21]

$$\beta_j \mid \tau_j^2 \sim N(0, \tau_j^2) \tag{5.2}$$

$$\tau_j^2 \mid \gamma_j \sim (1-\gamma_j) IG(a, \frac{d_0}{b}) + \gamma_j IG(a, \frac{d_1}{b}) \tag{5.3}$$

where d_0 and d_1 now have the role of τ_j^2 and c_j from the SSVS prior. Integrating over the variance terms the prior on β_j can be seen as a mixture of two scaled t-distributions. A similar argument based on the desired "practical effect" can be made for this prior to assist in the choice of hyper-parameters (see [20] and [21]).

Spike and Slab Lasso More recently priors with thicker tails have been considered for the distributions of the spike and slab. In particular, [22] propose a version of the spike and slab distribution,

$$\beta_j \mid \gamma_j \sim (1-\gamma_j) Lap(\lambda_0) + \gamma_j Lap(\lambda_1),$$

where $Lap(\lambda) = \frac{\lambda}{2} e^{-\lambda|\beta|}$ denotes a Laplace (double exponential) distribution. Taking λ_1 small and λ_0 large enables the distribution to mimic the original [1] prior with Dirac spike and diffuse slab. Taking instead $\lambda_0 = \lambda_1 = \lambda$, the prior

is equivalent to a single Laplace with parameter λ. This method provides a bridge between the weak shrinkage of the Laplace distribution and the harsh shrinkage of the original spike and slab. Additional computational advantages for mode detection are also possible due to the choice of Laplace shrinkage.

Heavy Tailed Spike and Slab Recent work of [13], have considered using distributions with heavier tails than the Laplace distribution. They advocate the use of priors of the form

$$\beta_j \mid \gamma_j \sim (1 - \gamma_j)\delta_0 + \gamma_j Cauchy(1),$$

where $Cauchy(1)$ denotes a standard Cauchy distribution. In particular they find that for the prior $\gamma_j \sim Bernoulli(w)$ for all $j = 1, \ldots, p$, if the hyper parameter w is calibrated via marginal maximum likelihood empirical Bayes, the Laplace slab is shown to lead to a suboptimal rate for the empirical Bayes posterior [13]. Heavier tailed distributions are required in order to make the empirical posterior contract at the optimal rate.

Nonlocal Priors Each of the priors considered so far places local prior densities on regression coefficients in the model. That is, the slab G_1 distributions all have positive prior density at the origin 0, which can make it more difficult to distinguish between models with small coefficients. Johnson and Rossell [23] proposed two new classes of priors which are zero at and around the origin. These priors are motivated from a Bayesian model averaging perspective and assign a lower weight to more complex models [24, 25].

5.2.4 Shrinkage Priors

Due to high computational costs spike and slab methods are often not able to scale to very high dimensional problems. This is due largely to the discrete γ variable and the large model space. Consequently, this has motivated the development of a wealth of priors that aim to provide continuous alternatives to the spike and slab. One of the earliest methods that received attention for this purpose is the Bayesian Lasso (least absolute shrinkage and selection) [26]. This method was motivated largely by the Lasso penalisation approach which has been celebrated in the statistics community for its computational efficiency and variable selection performance. For a detailed survey of the lasso and related Penalised regression methods see [27]. The Bayesian Lasso corresponds to the use of a Laplace prior on the regression coefficient. The resulting posterior mode for the Bayesian lasso is equivalent to the solution for the Lasso regression problem. While the Lasso estimate has been shown to have good variable selection properties, the Bayesian Lasso does not. Castillo et al. [8] show that the Bayesian Lasso does not make the posterior concentrate near the true value in large samples.

In recent years, continuous Bayesian priors with good shrinkage properties have been introduced to the literature. One broad class of priors is referred to as global-local shrinkage priors [28] which have the hierarchical form,

$$\beta_j \mid \eta_j, w \sim \mathcal{N}(0, w\eta_j), \tag{5.4}$$

$$\eta_j \sim \pi(\eta_j), \tag{5.5}$$

$$w \sim \pi(w) \tag{5.6}$$

where η_js are known as the local shrinkage parameters and control the degree of shrinkage for each individual coefficient β_j, while the global parameter w causes an overall shrinkage. If the prior $\pi(\eta_j)$ is appropriately heavy-tailed, then the coefficients of nonzero variables will not incur a strong shrinkage effect. This hierarchical formulation essentially places a scale mixture of Normal distributions using (5.5) and (5.6) and is found frequently in the Bayesian literature. This includes the normal-gamma [29], Horseshoe prior [30], generalised double Pareto [31], Dirichlet-Laplace (DL) prior [32] and the Horseshoe+ prior [33]. These priors all contain a significant amount of mass at zero so that coefficients are shrunk to zero.

Ghosh et al. [34] observed that for a large number of global-local shrinkage priors, the parameter η_j has a distribution that can be written as,

$$\pi(\eta_j) = K\eta_j^{-a-1}L(\eta_j), \tag{5.7}$$

where $K > 0$ and $a > 0$ are positive constants, and L is a positive measureable function. Table 1 from [35] provides a list of the more well known global-local shrinkage priors that fall into this form, their corresponding density for η_j, and the component $L(\eta_j)$. Theoretical properties and uncertainty quantification has also been considered for these types of shrinkage priors [36]. Importantly, point estimates using only shrinkage priors on the regression coefficients are not able to produce exact zeros. Quantification of the selected variables is often achieved using the estimated credible intervals. Additional inference on the regression coefficients may also be achieved using the decoupling shrinkage and selection (DSS) framework developed by Hahn and Carvalho [37].

5.3 Computational Methods

In this section we survey some of the standard methods used in computational Bayesian statistics to compute posterior inference in the Bayesian variable selection methods. For each method we outline the general implementation details. For illustrative purposes, we show how these methods may be used for a linear

regression analysis with the following hierarchical framework:

$$Y \mid \beta_\gamma, \gamma, \sigma \sim N_n(X_\gamma \beta_\gamma, \sigma^2 I) \tag{5.8}$$

$$\beta_\gamma \mid \sigma, \gamma \sim N_{s_\gamma}(\mu_\beta, \sigma^2 \Sigma_\gamma), \tag{5.9}$$

$$\sigma^2 \sim IG(d/2, d\lambda/2), \tag{5.10}$$

$$\gamma_j \overset{iid}{\sim} Bern(w) \text{ for} j = 1, \ldots, p, \tag{5.11}$$

where X_γ and β_γ denote subvectors of the covariates and regression parameters corresponding to the selected indices in γ and $\Sigma_\gamma \in \mathbb{R}^{s_\gamma \times s_\gamma}$ is the $s_\gamma \times s_\gamma$ prior covariance matrix for the selected regressors. Since $\gamma_j \overset{iid}{\sim} Bern(w)$ with w fixed, this prior on the model space favours models with wp selected variables. This prior specification for $\beta \mid \gamma$ corresponds to the Normal-Binomial or Kuo and Mallick spike and slab.

5.3.1 Markov Chain Monte Carlo Methods

The most widely used tool for fitting Bayesian models are sampling techniques based on Markov chain Monte Carlo (MCMC), in which a Markov chains is designed with stationary distribution that matches the desired posterior. In Bayesian variable selection, MCMC procedures are used to generate a sequence

$$\gamma^{(1)}, \gamma^{(2)}, \ldots \tag{5.12}$$

from a Markov chain with stationary distribution $p(\gamma \mid Y)$. In situations where there is no closed form expression for $p(\gamma \mid Y)$ we can attain a sequence of the form

$$\gamma^{(1)}, \beta^{(1)}, \sigma^{(1)}, \gamma^{(2)}, \beta^{(2)}, \sigma^{(2)} \ldots \tag{5.13}$$

from a Markov chain with distribution $p(\beta, \sigma, \gamma \mid Y)$. In the next two subsections we described various MCMC algorithms which may be used for simulating from (5.12) and (5.13). These algorithms are variants of the Metropolis–Hastings (MH) and Gibbs sampler algorithms, respectively. For more information on these algorithms and other MCMC methods for variable selection see the lecture notes [16].

5.3.2 Metropolis–Hastings

Algorithm 1 gives a generic description of an iteration of a Hastings–Metropolis algorithm that samples from $p(\gamma \mid Y)$. The MH algorithm works by sampling from

an arbitrary probability transition kernel $q(\gamma^* \mid \gamma)$ (the distribution of the proposal γ^*) and imposing a random rejection step.

Input: γ
Output: γ'
1. Sample $\gamma^* \sim q(\gamma^*|\gamma)$
2. With Probability

$$\alpha = \min\left(1, \frac{q(\gamma \mid \gamma^*)p(\gamma^* \mid Y)}{q(\gamma^* \mid \gamma)p(\gamma \mid Y)}\right)$$

Set $\gamma' \leftarrow \gamma^*$, otherwise $\gamma' \leftarrow \gamma$.

Algorithm 1: Metropolis–Hastings (MH) algorithm

The simplest transition kernel would be to take $q(\gamma^* \mid \gamma) = 1/p$ if a single component of γ is changed. This yields a Metropolis algorithm which simulates a new proposal by randomly changing one component of γ. This algorithm was originally proposed for graphical model selection by Madigan et al. [38] and is named MC^3 (Markov chain Monte Carlo model composition). Alternative transition kernels could be constructed to propose changes in d components of γ, or more generally to change a random number of components in γ. We note that the MH approach for variable selection has inspired a number of methods that are able to effectively explore a large model space. The stochastic search methods developed by Hans et al. [39] explores multiple candidate models in parallel at each iteration and moves more aggressively toward regions of higher probability. Parallel tempering together with genetic algorithms have also been adapted to help assist the exploration of the large feature space in a method called Evolutionary MCMC (EMC) [40]. This was later adapted to Bayesian variable selection by Bottolo and Richardson [41]. For variable selection problems where $p(\gamma \mid Y)$ is not easily attained, MH methods will need to sample both β_γ and γ, so care must be taken in choosing the appropriate transition kernel.

Example Details A valuable feature of the prior in (5.8) is that, due to conjugacy of the priors [16], the parameters β_γ and σ can be eliminated from $p(Y, \beta_\gamma, \sigma \mid \gamma)$ to yield,

$$p(Y \mid \gamma) \propto |X_\gamma^T X_\gamma + \Sigma_\gamma^{-1}|^{-1/2}|\Sigma_\gamma|^{-1/2}(d\lambda + S_\gamma^2)^{(-(n+d)/2)}$$

where,

$$S_\gamma^2 = Y^T Y - Y^T X_\gamma (X_\gamma^T X_\gamma + \Sigma_\gamma^{-1})^{-1} X_\gamma^T Y.$$

Thus, for the model prior $p(\gamma) = w^{s_\gamma}(1-w)^{p-s_\gamma}$ the posterior is proportional to

$$p(\gamma \mid Y) \propto p(Y \mid \gamma)p(\gamma) = g(\gamma).$$

Taking the previously defined transition kernel $q(\gamma^* \mid \gamma)$ and making use of the fact that $g(\gamma)/g(\gamma') = p(\gamma \mid Y)/p(\gamma' \mid Y)$, the MH algorithm follows the steps in Algorithm 1.

5.3.3 Gibbs Sampling

A well known MCMC approach to variable selection when the conditional distributions of the parameters are known is to apply Gibbs sampling. Unfortunately a drawback of Gibbs sampling is that it is not very generic and implementation depends strongly on the prior and model. When the prior is analytically tractable and a function $g(\gamma) \propto p(\gamma \mid Y)$ is available, the standard way to draw samples from the posterior $p(\gamma \mid Y)$ is by sampling the p components $(\gamma_1, \ldots, \gamma_p)$ as,

$$\gamma_j \sim p(\gamma_j \mid Y, \gamma_{(-j)}), \quad j = 1, \ldots, p,$$

where $\gamma_{(-j)} = (\gamma_1, \ldots, \gamma_{j-1}, \gamma_{j+1}, \ldots, \gamma_p)$ and where components γ_j may be drawn in fixed or random order. By computing the ratios

$$\frac{p(\gamma_j = 1, \gamma_{(-j)} \mid Y)}{p(\gamma_j = 0, \gamma_{(-j)} \mid Y)} = \frac{g(\gamma_j = 1, \gamma_{(-j)})}{g(\gamma_j = 0, \gamma_{(-j)})},$$

we can make use of the following [16]

$$p(\gamma_j = 1 \mid Y, \gamma_{(-j)}) = \frac{p(\gamma_j = 1, \gamma_{(-j)} \mid Y)}{p(\gamma_j = 0, \gamma_{(-j)} \mid Y)} \left(1 + \frac{p(\gamma_j = 1, \gamma_{(-j)} \mid Y)}{p(\gamma_j = 0, \gamma_{(-j)})}\right)^{-1}.$$

It is worth noting the recent work of Zanella and Roberts [42] which proposes an importance sampling version of the Gibbs sampling method with application to Bayesian variable selection. Additional computational advantages may be possible by drawing the components of γ in groups rather than one at a time. In this case the potential advantage of group updates would perform best if correlated variables are jointly updated.

Example Details As before, we have the function $g(\gamma)$

$$p(Y \mid \gamma) \propto g(\gamma) = |X_\gamma^T X_\gamma + \Sigma_\gamma^{-1}|^{-1/2} |\Sigma_\gamma|^{-1/2} (d\lambda + S_\gamma^2)^{(-(n+d)/2)}$$

where,

$$S_\gamma^2 = Y^T Y - Y^T X_\gamma (X_\gamma^T X_\gamma + \Sigma_\gamma^{-1})^{-1} X_\gamma^T Y.$$

The Bayesian update for $\gamma_j \mid Y, \gamma_{(-j)}$ is a Bernoulli draw with probability

$$p(\gamma_j = 1 \mid Y, \gamma_{(-j)}) = \frac{g(\gamma_j = 1, \gamma_{(-j)})}{g(\gamma_j = 0, \gamma_{(-j)})} \left(1 + \frac{g(\gamma_j = 1, \gamma_{(-j)})}{g(\gamma_j = 0, \gamma_{(-j)})}\right)^{-1}.$$

5.4 Software Implementations

There is a vast supply of software available to perform Bayesian variable selection. For this survey we restrict the scope to packages built for the R programming language [43]. These packages are free and available on the comprehensive R archive network CRAN (cran.r-project.org).

We start by noting that computational implementation of the priors and models described can be easily implemented in a number of generic Bayesian software. Ntzoufras [44] provide interesting examples of variable selection for the programs WinBUGS [45] and JAGS [46]. Code has also been made available for JAGS implementations of variable selection priors in the tutorial [14]. General purpose Bayesian software such as STAN [47] is not able to model discrete parameters so the spike and slab priors cannot be implemented. However, a large range of shrinkage priors such as the Horseshoe and Horseshoe+ are available. Practical examples for the analysis of variable selection has been proposed using STAN [48] (Table 5.1).

In addition to the general probabilistic programming languages, there are a large number of specific variable selection R packages. A survey of available R packages for variable selection has compared and contrasted popular software available as recent as February 17, 2017 [49]. In this chapter, we note some recent packages which were found using the PKGSEARCH R package [50]. The key words searched were *Bayesian variable selection*, *Bayesian model averaging* and *Bayesian feature selection*. From this search we note the following packages: *EMVS, basad, varbvs, BAS, spikeSlabGAM, BVSNLP, BayesS5, mombf, BoomSpikeSlab, R2GUESS, BMA, SSLASSO*.

BoomSpikeslab [51] implements a fast Gibbs sampling procedure for Bayesian modelling using a variant of the SSVS spike and slab prior. BMA implements a Metropolis Hastings (MC^3) algorithm for linear and some nonlinear sparse Bayesian models. BAS is similar to BMA in that it provides Bayesian model averaging methods. However, the sampler in BAS makes use of adaptive MCMC methods to give more efficient estimates. The mombf package provides a Gibbs sampler for the non-local and local priors (see Sect. 5.2.3). spikeSlabGAM implements a Gibbs sampler using a variant of the SSVS prior for generalised additive mixed models. Varbvs [52] implements a variational Bayesian variable selection method. As an alternative to MCMC, this package returns approximate estimates of posterior probabilities. These methods can scale much better with the dimension of the data than MCMC methods but suffer an approximation bias. R2GUESS provides an evolutionary stochastic search algorithm for both single and multiple response linear

Table 5.1 Recent packages for variable selection found using the R package PKGSEARCH

Package	Last release	Downloads	Description
BoomSpikeSlab	2019	214, 663	MCMC for Spike and Slab regression
BMA	2018	159, 652	Bayesian model averaging
BAS	2018	80, 286	Bayesian variable selection and model averaging using Bayesian adaptive sampling
mombf	2019	39, 764	Bayesian model selection and averaging for non-local and local priors
spikeSlabGAM	2018	21, 332	Bayesian variable selection and model choice for generalized additive mixed models
Varbvs	2019	14, 781	Large-scale Bayesian variable selection using variational methods
R2GUESS	2018	14, 595	A graphics processing unit-based R package for Bayesian variable selection regression of multivariate responses
BayesS5	2018	11, 295	Bayesian variable selection using simplified shotgun stochastic search with screening (S5)
BVSNLP	2019	10, 985	Bayesian variable selection in high dimensional settings using nonlocal priors
basad	2017	6187	Bayesian variable selection with Shrinking and diffusing priors
SSLASSO	2018	4407	The Spike and Slab LASSO
EMVS	2018	3816	The expectation-maximization approach to Bayesian variable selection

Year of the last release of the package, number of package downloads (calculated using CRANLOGS as of 28th July 2019)

models. BayesS5 is an efficient algorithm based on a variation of the stochastic search method and screening steps to improve computation time in high dimensions. The package BVSNLP implements considers local and nonlocal priors (similar to mombf) for binary and survival data [53]. The package basad implements variable selection with shrinking and diffusing spike and slab priors [19]. SSLASSO provides an implementation of the spike and slab lasso [22] for fast variable selection with Laplacian distributions for both the spike and slab. Finally, EMVS provides an expectation maximisation approach for Bayesian variable selection. The method provides a deterministic alternative to the stochastic search methods in order to find posterior modes.

Acknowledgement The author would like to acknowledge the Australian Research Council Centre of Excellence in Mathematical and Statistical Frontiers for funding.

References

1. T.J. Mitchell, J.J. Beauchamp, Bayesian variable selection in linear regression. J. Am. Stat. Assoc. **83**(404), 1023–1032 (1988)
2. M. Barbieri, J.O. Berger, E.I. George, V. Rockova, The median probability model and correlated variables. arXiv:1807.08336 (2020)
3. M.M. Barbieri, J.O. Berger, Optimal predictive model selection. Ann. Stat. **32**(3), 870–897 (2004)
4. F. Liang, Q. Song, K. Yu, Bayesian subset modeling for high-dimensional generalized linear models. J. Am. Stat. Assoc. **108**(502), 589–606 (2013)
5. W. Jiang, Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. Ann. Stat. **35**(4), 1487–1511 (2007)
6. E.I. George, R.E. McCulloch, Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**(423), 881–889 (1993)
7. M. Smith, R. Kohn, A Bayesian approach to nonparametric bivariate regression. J. Am. Stat. Assoc. **92**(440), 1522–1535 (1997)
8. I. Castillo, J. Schmidt-Hieber, A. van der Vaart, Bayesian linear regression with sparse priors. Ann. Stat. **43**(5), 1986–2018 (2015)
9. J.G. Scott, J.O. Berger, Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. Ann. Stat. **38**(5), 2587–2619 (2010)
10. F. Li, N.R. Zhang, Bayesian variable selection in structured high-dimensional covariate spaces with applications in genomics. J. Am. Stat. Assoc. **105**(491), 1202–1214 (2010)
11. M.E.J. Newman, G.T. Barkema, *Monte Carlo Methods in Statistical Physics* (Clarendon Press, Oxford, 1999)
12. E.E. Leamer, *Specification Searches: Ad hoc Inference with Nonexperimental Data*, vol. 53 (Wiley, Hoboken, 1978)
13. I. Castillo, R. Mismer, Empirical bayes analysis of spike and slab posterior distributions. Electron. J. Stat. **12**, 3953–4001 (2018)
14. R.B. O'Hara, M.J. Sillanpää, A review of Bayesian variable selection methods: what, how and which. Bayesian Anal. **4**(1), 85–117 (2009)
15. L. Kuo, B. Mallick, Variable selection for regression models. Sankhyā Indian J. Stat. Ser. B (1960–2002) **60**(1), 65–81 (1998)
16. H. Chipman, E.I. George, R.E. McCulloch, *The Practical Implementation of Bayesian Model Selection*. Lecture Notes–Monograph Series, vol. 38 (Institute of Mathematical Statistics, Beachwood, 2001), pp. 65–116. https://doi.org/10.1214/lnms/1215540964
17. P. Dellaportas, J.J. Forster, I. Ntzoufras, Bayesian variable selection using the Gibbs sampler. BIOSTATISTICS-BASEL- **5**, 273–286 (2000)
18. B.P. Carlin, S. Chib, Bayesian model choice via Markov chain Monte Carlo methods. J. R. Stat. Soc. Ser. B Stat. Methodol. **57**(3), 473–484 (1995)
19. N.N. Narisetty, X. He, Bayesian variable selection with shrinking and diffusing priors. Ann. Stat. **42**(2), 789–817 (2014)
20. H. Ishwaran, J.S. Rao, Detecting differentially expressed genes in microarrays using Bayesian model selection. J. Am. Stat. Assoc. **98**(462), 438–455 (2003)
21. L. Fahrmeir, T. Kneib, S. Konrath, Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. Stat. Comput. **20**(2), 203–219 (2010)
22. V. Ročková, E.I. George, The spike-and-slab lasso. J. Am. Stat. Assoc. **113**(521), 431–444 (2018)
23. V.E. Johnson, D. Rossell, Bayesian model selection in high-dimensional settings. J. Am. Stat. Assoc. **107**(498), 649–660 (2012)
24. D. Rossell, D. Telesca, Non-local priors for high-dimensional estimation. J. Am. Stat. Assoc. **112**(517), 254–265 (2017)

25. A. Nikooienejad, W. Wang, V.E. Johnson, Bayesian variable selection for binary outcomes in high-dimensional genomic studies using non-local priors. Bioinformatics **32**(9), 1338–1345 (2016)
26. R. Tibshirani, Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. **58**(1), 267–288 (1996)
27. J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space. Stat. Sin. **20**(1), 101–148 (2010)
28. N.G. Polson, J.G. Scott, Local shrinkage rules, lévy processes and regularized regression. J. R. Stat. Soc. Ser. B Stat. Methodol. **74**(2), 287–311 (2012)
29. J.E. Griffin, P.J. Brown, Inference with normal-gamma prior distributions in regression problems. Bayesian Anal. **5**(1), 171–188 (2010)
30. C.M. Carvalho, N.G. Polson, J.G. Scott, The horseshoe estimator for sparse signals. Biometrika **97**(2), 465–480 (2010)
31. A. Armagan, D.B. Dunson, J. Lee, Generalized double pareto shrinkage. Stat. Sin. **23**(1), 119–143 (2013)
32. A. Bhattacharya, D. Pati, N.S. Pillai, D.B. Dunson, Dirichlet–laplace priors for optimal shrinkage. J. Am. Stat. Assoc. **110**(512), 1479–1490 (2015)
33. A. Bhadra, J. Datta, N.G. Polson, B. Willard, The horseshoe+ estimator of ultra-sparse signals. Bayesian Anal. **12**(4), 1105–1131 (2017)
34. P. Ghosh, X. Tang, M. Ghosh, A. Chakrabarti, Asymptotic properties of bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. Bayesian Anal. **11**(3), 753–796 (2016)
35. R. Bai, M. Ghosh, High-dimensional multivariate posterior consistency under global–local shrinkage priors. J. Multivar. Anal. **167**, 157–170 (2018)
36. S. van der Pas, B. Szabó, A. van der Vaart, Uncertainty quantification for the horseshoe (with discussion). Bayesian Anal. **12**(4), 1221–1274 (2017)
37. P.R. Hahn, C.M. Carvalho, Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. J. Am. Stat. Assoc. **110**(509), 435–448 (2015)
38. D. Madigan, J. York, D. Allard, Bayesian graphical models for discrete data. Int. Stat. Rev./Rev. Int. de Stat. **63**(2), 215–232 (1995)
39. C. Hans, A. Dobra, M. West, Shotgun stochastic search for "large p" regression. J. Am. Stat. Assoc. **102**(478), 507–516 (2007)
40. F. Liang, W.H. Wong, Evolutionary monte carlo: applications to C p model sampling and change point problem. Stat. Sin. **10**(2), 317–342 (2000)
41. L. Bottolo, S. Richardson, Evolutionary stochastic search for Bayesian model exploration. Bayesian Anal. **5**(3), 583–618 (2010)
42. G. Zanella, G. Roberts, Scalable importance tempering and Bayesian variable selection. J. R. Statist. Soc. B **81**, 489–517 (2019)
43. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2013)
44. I. Ntzoufras, Gibbs variable selection usingbugs. J. Stat. Softw. **7**(7), 1–19 (2002)
45. D.J. Lunn, A. Thomas, N. Best, D. Spiegelhalter, Winbugs-a Bayesian modelling framework: concepts, structure, and extensibility. Stat. Comput. **10**(4), 325–337 (2000)
46. M. Plummer, et al., JAGS: A program for analysis of Bayesian graphical models using gibbs sampling, in *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, vol. 124 (2003)
47. B. Carpenter, A. Gelman, M.D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan: A probabilistic programming language. J. Stat. Softw. **76**(1), 1–32 (2017)
48. J. Piironen, A. Vehtari, Projection predictive model selection for gaussian processes, in *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)* (2016), pp. 1–6

49. A. Forte, G. Garcia-Donato, M. Steel, Methods and tools for Bayesian variable selection and model averaging in normal linear regression. Int. Stat. Rev./Rev. Int. de Stat. **86**(2), 237–258 (2018)
50. G. Csárdi, *pkgsearch: Search CRAN R Packages*. R package version 2.0.1. (2018). https://CRAN.R-project.org/package=pkgsearch
51. H. Ishwaran, U.B. Kogalur, J.S. Rao, spikeslab: prediction and variable selection using spike and slab regression. R J. **2**, 68–73 (2010)
52. P. Carbonetto, M. Stephens, Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. Bayesian Anal. **7**, 73–108 (2012)
53. D. Rossell, J.D. Cook, D. Telesca, P. Roebuck, mombf: moment and inverse moment bayes factors. R Package Version 1. 0, vol. 3 (2008)