Note: References highlighted in color on pg. 2-3-4 are incorrect in the printed version of the chapter

Contents

1 Discrete Spike-and-Slab Priors: Models and Computational Aspects **1**
Marina Vannucci
1.1 Introduction . . . 2
1.2 Spike-and-slab priors for linear regression models . . . 3
1.2.1 Stochastic search MCMC . . . 5
1.2.2 Prediction via Bayesian model averaging . . . 7
1.3 Spike-and-slab priors for non-Gaussian data . . . 8
1.3.1 Compositional count data . . . 8
1.4 Structured spike-and-slab priors for biomedical studies . . . 11
1.4.1 Network priors . . . 13
1.4.2 Spiked nonparametric priors . . . 15
1.5 Scalable Bayesian variable selection . . . 16
1.5.1 Variational inference . . . 18
1.6 Conclusion . . . 20

Bibliography **21**

1

Discrete Spike-and-Slab Priors: Models and Computational Aspects

Marina Vannucci

Rice University, Houston, TX (USA)

CONTENTS

1.1 Introduction 2
1.2 Spike-and-slab priors for linear regression models 3
1.2.1 Stochastic search MCMC 5
1.2.2 Prediction via Bayesian model averaging 7
1.3 Spike-and-slab priors for non-Gaussian data 8
1.3.1 Compositional count data 8
1.4 Structured spike-and-slab priors for biomedical studies 11
1.4.1 Network priors 13
1.4.2 Spiked nonparametric priors 15
1.5 Scalable Bayesian variable selection 16
1.5.1 Variational inference 16
1.6 Conclusion 19
Software 20
Acknowledgements 20

A large body of research has been devoted to variable selection in recent years. Bayesian methods have been successful in applications, particularly in settings where the amount of measured variables can be much greater than the number of observations. This chapter reviews mixture priors that employ a point mass distribution at zero for variable selection in regression settings. The popular stochastic search Markov chain Monte Carlo (MCMC) algorithm with add-delete-swap moves is described. Posterior inference and prediction via Bayesian model averaging are briefly discussed. Regression models for non-Gaussian data, including binary, multinomial, survival and compositional count data, are also addressed. Prior constructions that take into account specific structures in the covariates are reviewed. These have been particularly successful in applications as they allow the integration of different sources of external information into the analysis. A discussion of variational algorithms for scalable inference concludes the chapter. Throughout, some emphasis is given to the author's contributions.

1.1 Introduction

Variable selection, also known as feature selection, has been an important topic in the statistical literature for the past several decades, with numerous papers published in both theory and practice. Finding a subset of features that best explain an outcome of interest is often an important aspect of the data analysis, as it allows for simpler interpretation, avoids overfitting and multicollinearity, and provides insights into the mechanisms generating the data. Variable selection is especially important when the number of potential predictors is substantially larger than the sample size.

In linear regression settings, modern approaches to variable selection include criteria-based methods, such as AIC/BIC [36], penalized likelihood methods which shrink to zero coefficients of unimportant covariates [65], and Bayesian approaches that use shrinkage priors to induce sparsity, such as mixtures of two distributions (spike-and-slab priors) [6, 17, 18, 23, 37] and unimodal continuous shrinkage priors [10, 43, 46]. With spike-and-slab priors, a latent binary vector is introduced to index the possible subsets of predictors and used to induce mixture priors of two components on the regression coefficients, one peaked at zero (spike) and the other one a diffuse distribution (slab). Posterior inference is carried out via stochastic search MCMC techniques to identify the high- probability models, and variable selection is performed based on the posterior model probabilities. This chapter is devoted in particular to *discrete* spike-and-slab constructions, which employ a point mass distribution at zero as the spike component.

Bayesian methods for variable selection have several appealing features. They allow rich modeling via MCMC stochastic search strategies and incorporate optimal model averaging prediction; they extend naturally to multivariate responses and many linear and nonlinear settings; they can handle the "small n - large p" setting, i.e., situations where the number of covariates is larger than the sample size; they allow the use of priors that incorporate past and collateral information into the model.

This chapter is organized as follows. Section 1.2 briefly reviews discrete spike-and-slab priors for variable selection in linear regression, including the popular stochastic search MCMC algorithm with add-delete-swap moves, for posterior inference, and a brief discussion of Bayesian model averaging, for prediction purposes. Section 1.3 addresses regression models for non-Gaussian data, including binary, multinomial and survival outcomes. It also covers model settings for compositional count data. Section 1.4 reviews prior constructions that take into account specific structures in the covariates, together with examples of modern biomedical studies in genomics and neuroimaging that have motivated those constructions. Section 1.5 discusses variational inference strategies for scalable inference. Final remarks are given in Section 1.6.

1.2 Spike-and-slab priors for linear regression models

In the classical multiple linear regression model, a continuous response, y_i, is modeled via a linear combination of p covariates, $\mathbf{x}_i = (x_1, \ldots, x_p) \in \mathbb{R}^p$, as

$$y_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ the vector of regression coefficients and α the baseline or intercept. The variable selection problem arises when it is believed that not all p covariates

are important in explaining changes of the response and identification of the important predictors is one of the goals of the analysis. Clearly, setting to zero some of the regression coefficients in (1.1) is equivalent to excluding the corresponding subset of predictors from the model. In the Bayesian paradigm this can be achieved by imposing sparsity-inducing mixture priors, known as *spike-and-slab* priors, on the β_j coefficients [17, 18, 23, 37]. This formulation introduces a latent vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ of binary indicators

$$\gamma_j = \begin{cases} 1 & \text{if variable } j \text{ is included in model,} \\ 0 & \text{otherwise.} \end{cases}$$

Two prior constructions have been developed in parallel in the statistical literature. This chapter focuses on the *discrete* construction, which employs a mixture prior distribution on β_j with a point mass at zero, see Figure 1.1, as

$$\beta_j|\sigma^2, \gamma_j \sim (1-\gamma_j)\delta_0(\beta_j) + \gamma_j \mathcal{N}(0, h_j\sigma^2), \tag{1.2}$$

for $j = 1, \ldots, p$, with $\delta_0(\cdot)$ the Dirac function at $\beta_j = 0$ and the h_j's a set of hyperparameters. Here, $\gamma_j = 0$ excludes the j-th variable from the model since the prior on the corresponding coefficient β_j is a point mass distribution at 0, while $\gamma_j = 1$ includes the predictor into the model, leading to a normal prior on β_j. Mixture priors of type (1.2) for the linear regression setting were originally proposed by [37] and made popular by [18]. The prior formulation is completed with an independent conjugate inverse-gamma priors on σ^2 and a Gaussian prior on the intercept α,

$$\alpha|\sigma^2 \sim \mathcal{N}(\alpha_0, h_0\sigma^2), \quad \sigma^2 \sim IG(\nu/2, \lambda/2), \tag{1.3}$$

with α_0, h_0, ν and λ hyperparameters to be chosen. Setting $\alpha_0 = 0$ and taking $h_0 \to \infty$ results in a vague prior on the intercept, so that mean centering the predictors sets the posterior mean for α at $\bar{\mathbf{y}}$.

Common choices for the hyperparameters h_j's in (1.2) assume that the β_j's are *a priori* independent given $\boldsymbol{\gamma}$, for example, by choosing $h_j = c$ for every j. Generally speaking, small values of c induce shrinkage towards smaller models, while larger values favor the selection of larger models. Dependent priors that use the Zellner's g-prior of [74] have also been considered [33, 60],

$$\boldsymbol{\beta}_{(\boldsymbol{\gamma})}|\sigma^2 \sim \mathcal{N}(0, c(\mathbf{X}'_{(\boldsymbol{\gamma})}\mathbf{X}_{(\boldsymbol{\gamma})})^{-1}\sigma^2), \tag{1.4}$$

with $\boldsymbol{\beta}_{(\boldsymbol{\gamma})}$ the subset of coefficients corresponding to the elements of $\boldsymbol{\gamma}$ equal to 1 and $\mathbf{X}_{(\boldsymbol{\gamma})}$ the selected covariates. The range of values $c \in [10, 100]$ is suggested in [60]. The Zellner's prior is appealing because of its intuitive interpretation. It can, however, induce mixing problems in the MCMC, particularly with subsets of highly correlated predictors. [6] investigated prior (1.2) with h_j proportional to the j-th diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$, to alleviate this problem. Prior constructions described so far are conjugate to the Gaussian likelihood. Non-conjugate constructions, that assume independent priors on $\boldsymbol{\beta}$ and σ^2, are also possible, see for example [18]. We will revisit these constructions in Section 1.2.1.

The discrete construction (1.2) differs from the *continuous* spike-and-slab prior, which instead employs a mixture of two continuous components, typically two Gaussian distributions, one concentrated around zero and the second one more spread out over plausible large values [17, 23], see Figure 1.1. Unlike with the continuous construction, discrete priors of type (1.2) effectively exclude non-selected variables from the calculation of the likelihood. While optimality properties of continuous spike-and-slab priors have been studied fairly extensively [24, 38, 53], theoretical guarantees for the discrete construction in the linear

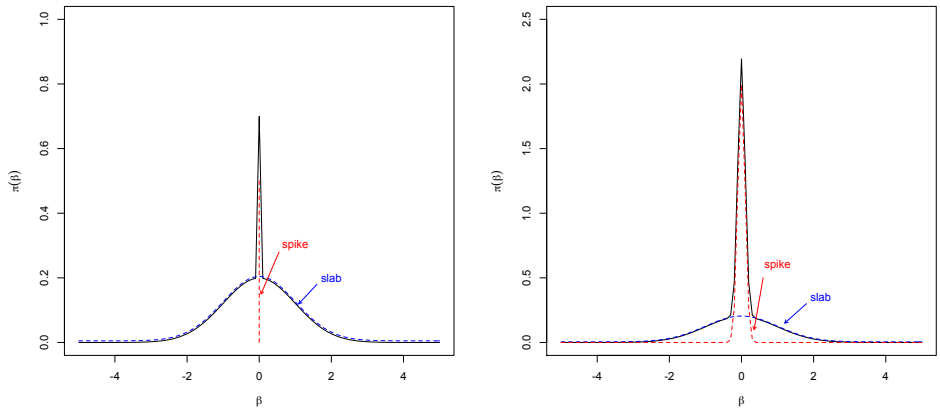

FIGURE 1.1
Spike-and-slab mixture priors for Bayesian variable selection. *Left:* The discrete construction (solid line) is a mixture of a point mass at zero (spike; dashed line) and a diffuse prior (slab; dotted line). *Right:* The continuous construction (solid line) is a mixture of two normal distributions, one peaked around zero (dashed line) and the other with a large variance (dotted line).

regression setting (1.1) have become available only recently, due to the seminal work of [13], and include optimality results for the Zellner g-prior construction (1.4), see [73].

Prior construction (1.2) requires the choice of a prior distribution on $\boldsymbol{\gamma}$. The simplest and most common choice adopted in the literature is a product of independent Bernoulli's with common parameter ω as

$$p(\boldsymbol{\gamma}|\omega) = \prod_{j=1}^{p} \omega^{\gamma_j} (1-\omega)^{1-\gamma_j}, \tag{1.5}$$

that leads to $p\omega$ being the number of variables expected *a priori* to be included in the model. Uncertainty on ω can be modeled by imposing a Beta hyperprior, $\omega \sim Beta(a, b)$, with a, b to be chosen. If inference on ω is not of interest, it can be integrated out to simplify the MCMC implementation. A weakly-informative prior can be obtained by setting $a = b = 1$, resulting in the prior expected mean value to be $m = a/(a+b) = .5$. An attractive feature of the Beta-Binomial prior construction is that it imposes an *a priori* multiplicity penalty, as argued in [57]. The intuition behind this is that the marginal prior on $\boldsymbol{\gamma}$ contains a non-linear penalty which is a function of p and therefore, as p grows, with the number of true variables remaining fixed, the posterior distribution of ω concentrates near 0. A limitation of the Beta-Binomial construction is that it assumes that the inclusion indicators are stochastically independent. Alternative priors, that exploit complex dependence structures between covariates, as induced by underlying biological processes and/or networks, have been motivated by specific applications to data from biomedical studies. Some of these prior constructions will be described in Section 1.4.

Conjugate discrete spike-and-slab prior constructions have been extended by [6, 7] to multivariate linear regression models with q response outcomes. Their construction selects variables as relevant to either all or none of the q responses. [31] proposed multivariate constructions based on partition models that allow each covariate to be relevant for subsets and/or individual response variables. Other flexible multivariate prior formulations, that

allow to select covariates for individual responses, were proposed by [52, 63]. See Section 1.3 for an example of such construction in a multivariate count data model setting.

1.2.1 Stochastic search MCMC

Let us consider the linear setting (1.1) with the discrete spike-and-slab prior construction described in the previous section. The choice of conjugate priors makes it possible to integrate out the model parameters and obtain the relative posterior distribution of $\boldsymbol{\gamma}$ as

$$p(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}) \propto p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X})p(\boldsymbol{\gamma}). \tag{1.6}$$

This distribution allows to identify the "best" models as those with highest posterior probabilities. When a large number of predictors makes the full exploration of the model space unfeasible, MCMC methods can be used as stochastic searches to explore the posterior distribution and identify models with high posterior probability. Marginalization (1.6), jointly with a *QR* deletion-addition algorithm for fast updating in the calculation of the marginal likelihood, leads to efficient MCMC schemes for posterior inference, see [18] for the univariate regression setting and [7] for multivariate regression. A commonly used algorithm is a Metropolis-Hastings scheme readapted from the MC^3 algorithm proposed by [34] in the context of model selection for discrete graphical models. It consists of *add-delete-swap* moves that allow the exploration of the posterior space by visiting a sequence of models where, at each step, the new model differs from the previously visited one by the inclusion and/or exclusion of one or two variables. More specifically, given a randomly chosen starting value, $\boldsymbol{\gamma}^0$, at a generic iteration the new model is generated from the previous one by randomly choosing one of the following transition moves:

1. (Adding or deleting) Randomly pick one of the p indices in $\boldsymbol{\gamma}^{old}$ and change its value. This results in either including a new variable in the model or in deleting a variable currently included.
2. (Swapping) Draw independently and at random a 0 and a 1 in $\boldsymbol{\gamma}^{old}$ and switch their values. This results in both the inclusion of a new variable in the model and the deletion of a currently included one.

By indicating with $\boldsymbol{\gamma}^{new}$ the candidate model, the acceptance probability is calculated as

$$\min\left[\frac{p(\boldsymbol{\gamma}^{new}|\mathbf{y}, \mathbf{X})}{p(\boldsymbol{\gamma}^{old}|\mathbf{y}, \mathbf{X})}, 1\right]. \tag{1.7}$$

Therefore, if the new candidate model has a higher probability than the current one, the chain moves to the new configuration. If not, then the move is still possible, but now only with a certain probability. Note that the acceptance probability (1.7) depends on an "exact" ratio, since the constants of proportionality from (1.6) cancel out. This allows the search to quickly move towards better models. The stochastic search results in a list of visited models, $\boldsymbol{\gamma}^{(0)}, \ldots, \boldsymbol{\gamma}^{(T)}$, and their corresponding relative posterior probabilities. Variable selection can then be achieved either by looking at the $\boldsymbol{\gamma}$ vectors with largest joint posterior probabilities among the visited models or, marginally, by calculating frequencies of inclusion for each γ_j and then choosing those γ_j's with frequencies exceeding a given cut-off value. A common choice is a cut-off value of 0.5, which results in the median probability model [1]. Methods based on expected false discovery rates can also be employed, as suggested in [39].

Gibbs sampling schemes are also possible, see for example [18]. However, these schemes typically sample all variable indicators γ_j's at each iteration, unlike Metropolis schemes that allow a more efficient exploration of the space of only the relevant variables. This

is particularly important in situations of sparsity of the true model. Improved stochastic MCMC schemes have been proposed, to achieve a faster exploration of the posterior space. See, for example, the *shotgun* algorithm of [22] and the evolutionary Monte Carlo schemes, combined with a parallel tempering step that prevents the chain from getting stuck in local modes, proposed by [5]. A correlation-based stochastic search method, the hybrid-CBS algorithm, which comprises add-delete-swap moves specifically designed to accommodate correlations among the covariates, was proposed by [29]. Adaptive schemes that specifically aim at improving the mixing of the MCMC chain have been investigated by [21, 30].

When non-conjugate priors are used, the marginalization of the model parameters is no longer possible and those parameters need to be sampled as part of the MCMC algorithm. Initial attempts employed the reversible jump algorithm of [20], to handle the varying dimensionality of the parameter vector, see for example [18]. Later, [19] showed that the reversible jump can be formulated in terms of a mixture of singular distributions, implying that the algorithm is the same as an MCMC algorithm that jointly samples parameters and binary indicators. This is key in designing efficient MCMC algorithms for variable selection in non conjugate settings, particularly for the case of non-Gaussian data and, more generally, complex models for which conjugate prior formulations may not be available. For example, this idea was used by [56] to design add-delete-swap algorithms that jointly update parameters and selection indicators in a variable selection approach that incorporates Gaussian processes within a generalized linear model framework. We will see an example of a joint sampler for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ in Section 1.3, within a model setting for multivariate count data.

1.2.2 Prediction via Bayesian model averaging

Prediction is an important aspect of the inference in linear regression settings. Given the list of models visited by the stochastic search, $\boldsymbol{\gamma}^{(0)}, \ldots, \boldsymbol{\gamma}^{(T)}$, prediction of a future observation y^f can be done based on the selected models, either via least squares on single models, or by *Bayesian model averaging* (BMA) [50], which accounts for the uncertainty in the selection process by averaging over a set of *a posteriori* likely models. For example, for model (1.1) with prior (1.2) and posterior (1.6), BMA calculates the expected value of the predictive distribution $p(y^f|\mathbf{y}, \mathbf{X}^f)$, averaging over a set of configurations of $\boldsymbol{\gamma}$, with weights given by the posterior probabilities of these configurations, as

$$\hat{y}^f = \sum_{\boldsymbol{\gamma}} \left(\hat{\alpha} + \mathbf{X}^f_{(\boldsymbol{\gamma})} \hat{\boldsymbol{\beta}}_{(\boldsymbol{\gamma})} \right) p(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}), \tag{1.8}$$

with $\mathbf{X}^f_{(\boldsymbol{\gamma})}$ the covariates corresponding to the elements of $\boldsymbol{\gamma}$ equal to 1, $\hat{\alpha} = \bar{\mathbf{y}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} = (\mathbf{X}'_{(\boldsymbol{\gamma})}\mathbf{X}_{(\boldsymbol{\gamma})} + \mathbf{H}^{-1}_{(\boldsymbol{\gamma})})^{-1}\mathbf{X}'_{(\boldsymbol{\gamma})}\mathbf{y}$, with $\mathbf{H}$ a diagonal matrix with diagonal elements the hyperparameters h_j's of the slab component in (1.2). Typically, only the best k configurations among those visited by the MCMC, according to their posterior probabilities, are used in the summation.

1.3 Spike-and-slab priors for non-Gaussian data

Spike-and-slab mixture priors for variable selection have been extended beyond Gaussian data to other model settings that express a response variable as a linear combination of the predictors. A unified treatment of the class of generalized linear models (GLM) of [35] presents some challenges. Conditional densities in the general GLM framework cannot be

obtained directly and the resulting mixture posterior may be difficult to sample from using standard MCMC methods due to multimodality. Some attempts were done by [49], who proposed approximate Bayes factors, and by [40], who developed a method to jointly select variables and the link function.

Several contributions exist on extending spike-and-slab priors to specific models in the GLM class, in particular models for binary and multinomial outcomes and parametric accelerated failure time (AFT) models for survival outcomes. For example, probit models with multinomial outcomes were considered by [59] and AFT models by [58]. In these settings, data augmentation approaches allow to express the model in a linear framework, with latent responses $\mathbf{z}$, and conjugate priors are used to integrate the regression coefficients out, obtaining the marginal likelihood $p(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{X}, \mathbf{z})$, and facilitating the implementation of MCMC schemes that update $\boldsymbol{\gamma}$ conditional upon $\mathbf{z}$. For other settings, where marginalization of the regression coefficients is not possible, joint updates of coefficients and selection indicators can be performed and, whenever possible, coupled with data augmentation schemes for more efficient samplers. Examples include logistic and negative binomial regression models, for which the Pólya-Gamma (PG) data augmentation schemes developed by [45, 47] can be used to implement Gibbs samplers with PG updates on the latent variables followed by Gaussian updates on the regression coefficients. See [69] for recent work that combines these augmentation schemes with the add-delete-swap scheme of [56], as part of a variable selection approach to non-homogeneous hidden Markov models. Also, adaptive MCMC schemes for variable selection in logistic and AFT regression models were investigated in [68].

1.3.1 Compositional count data

[67] considered a Dirichlet-multinomial (DM) regression framework for compositional count data and demonstrated how to embed spike-and-slab priors for variable selection. Compositional count data $\mathbf{y}_i = (y_{i1}, \ldots, y_{iq})$ sum up to a fixed amount. A suitable distribution for this data is the multinomial

$$\mathbf{y}_i \sim \text{Multinomial}(\dot{y}_i|\mathbf{p}_i), \tag{1.9}$$

with $\dot{y}_i = \sum_{k=1}^{q} y_{ik}$, and $\mathbf{p}_i$ defined on the q-dimensional simplex

$$S^{q-1} = \{(p_{i1}, \ldots, p_{iq}) : p_{ik} \geq 0, \forall k, \sum_{k=1}^{q} p_{ik} = 1\}.$$

A Dirichlet conjugate prior can be imposed on the probability parameter vector, $\mathbf{p}_i \sim \text{Dirichlet}(\boldsymbol{\phi}_i)$, with q-dimensional vector $\boldsymbol{\phi_i} = (\phi_{ik} > 0, \forall k)$, and then $\mathbf{p}_i$ can be integrated out to obtain the DM model $\mathbf{y}_i \sim \text{DM}(\boldsymbol{\phi}_i)$. The DM model allows more flexibility than the multinomial when encountering overdispersion, as it induces an increase in the variance by a factor of $(\dot{y}_i + \dot{\phi}_i)/(1 + \dot{\phi}_i)$.

Covariate effects can be incorporated into the DM model via a log-linear link on the concentration parameters $\boldsymbol{\phi}_i$'s, by setting $\lambda_{ik} = \log(\phi_{ik})$ and assuming

$$\lambda_{ik} = \alpha_k + \sum_{j=1}^{p} \beta_{jk}\, x_{ij}, \tag{1.10}$$

where $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{pk})^T$ represents the covariates' potential relations with the k-th compositional outcome, and α_k is a outcome-specific intercept term. Exponentiating (1.10) ensures positive hyperparmeters for the Dirichlet distribution. For variable selection purposes, the number of potential models to choose from is 2^{pq}, and grows quickly even

for a small number of covariates. [67] introduced a set of q latent p-dimensional vectors $\boldsymbol{\gamma_k} = (\gamma_{1k}, \ldots, \gamma_{pk})$ of inclusion indicators. Thus, $\gamma_{jk} = 1$ indicates that the j-th covariate is associated with the k-th compositional outcome, and 0 otherwise. A discrete spike-and-slab prior on β_{jk} is then written as

$$\beta_{jk}|\gamma_{jk} \sim (1-\gamma_{jk})\delta_0(\beta_{jk}) + \gamma_{jk}\mathcal{N}(0, r_k^2), \tag{1.11}$$

where the hyperparameters r_k^2 can be set large to impose a diffuse prior for the regression coefficients included in the model. This multivariate spike-and-slab prior, that allows to identify covariates associated with individual responses, is similar to constructions used by [52, 63] in linear regression settings for Gaussian data with multiple responses.

Posterior inference is carried out via stochastic search MCMC algorithms. Here, the regression coefficients β_{jk}'s cannot be integrated out and, therefore, need to be jointly updated with the inclusion indicators, following [56]. [67] employed this strategy within a Gibbs scheme that scans through the γ_{jk}'s and uses adaptive sampling on the β_{jk}'s. [28] incorporated the joint update within an add-delete Metropolis-Hastings within Gibbs sampling scheme that updates each α_k and a randomly selected $(\gamma_{jk}, \beta_{jk})$ at every iteration. The joint update works as follows:

- *Between-Model Step* - Randomly select a γ_{jk}.
  - Add: If the covariate is currently excluded ($\gamma_{jk} = 0$), change it to $\gamma'_{jk} = 1$. Then sample a new $\beta'_{jk} \sim \mathcal{N}(\beta_{jk}, c)$ with c fixed to a chosen value. Accept proposal with probability
  $$\min\left\{\frac{f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{X})p(\beta'_{jk}|\gamma'_{jk})p(\gamma'_{jk})}{f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{X})p(\gamma_{jk})}, 1\right\}.$$
  - Delete: If the covariate is currently included ($\gamma_{jk} = 1$), change it to $\gamma'_{jk} = 0$ and set $\beta'_{jk} = 0$. Accept proposal with probability
  $$\min\left\{\frac{f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}', \boldsymbol{\gamma}', \mathbf{X})p(\gamma'_{jk})}{f(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{X})p(\beta_{jk}|\gamma_{jk})p(\gamma_{jk})}, 1\right\}.$$
- *Within-Model Step* - Propose a $\beta'_{jk} \sim \mathcal{N}(\beta_{jk}, c)$ for each covariate currently selected in the model ($\gamma_{jk} = 1$). Accept each proposal with probability
$$\min\left\{\frac{p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}', \boldsymbol{\gamma}, \mathbf{X})p(\beta'_{jk}|\gamma_{jk})}{p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{X})p(\beta_{jk}|\gamma_{jk})}, 1\right\}.$$
This within-model step is not required for ergodicity but allows to perform a refinement of the parameter space within the existing model, for faster convergence.

As customary with spike-and-slab priors, variable selection is performed based on the marginal posterior probabilities of inclusion (PPIs).

Recently, there has been a renewed interest in the biomedical community on statistical models for compositional count data, in particular due to the availability of high-throughput sequencing technologies that have enabled researchers to characterize the composition of the microbiome by quantifying its richness, diversity and abundances. Human microbiome research aims to understand how microbiome communities interact with their host, respond to their environment, and influence disease. DM regression models allow to appropriately handle the compositional structure of the data and accommodate the overdispersion induced by sample heterogeneity and varying proportions among samples. While model formulation (1.9) assumes that counts are negatively correlated, extensions exist that allow more general correlation structures between counts, such as the Dirichlet-tree multinomial model,

FIGURE 1.2
Analysis of data on dietary intake and microbiome via a regression model for compositional count data with variable selection [28]. Plot of PPIs of associations between microbial taxa and dietary factors, with each of the $p = 97$ dietary factors having a unique inclusion indicator for each of the $q = 28$ taxa.

that deconstructs the model into the product of multinomial distributions for each of the subtrees in a tree [71]. The `R` package `MicroBVS`, accompanying [28], comprises of a suite of regression models for compositional data, including DM and Dirichlet-tree multinomial regression models. It also implements the joint model of [27] that includes a phenotypical outcome to investigate how the microbiome may affect the relation between covariates and phenotypic responses. MCMC algorithms are written in C++, to increase performance time, and accessed through `R` wrapper functions. The package includes a vignette with worked out examples using simulated data and access to open-source data used in the papers. As an example, let us consider the analysis of a benchmark data set collected to study the relation between dietary intake and the human gut microbiome [72]. Briefly, the data used consist of counts on $q = 28$ microbial taxa obtained from 16S rRNA sequencing and a corresponding set of $p = 97$ dietary intake covariates derived from diet information collected using a food frequency questionnaire on $n = 98$ subjects. In this analysis, the DM regression model was fit to the data assuming a Beta-Binomial prior with $a = 1$ and $b = 999$ on the inclusion indicators γ_{jk} of prior (1.11). At convergence of the MCMC, about 398 of the roughly 2700 terms would be selected with a threshold of .5 on the PPIs, see Figure 1.2. Heatmaps of the selected positive and negative associations are shown in Figure 1.2. For this application, knowledge of the identified relations between microbial composition and nutrients may help researchers design tailored interventions to help maintain a healthy microbiome community [72].

1.4 Structured spike-and-slab priors for biomedical studies

Spike-and-slab variable selection priors have found important applications in biomedical studies. In high-throughput genomic, for example, linear models are routinely used to relate large sets of biomarkers to disease-related outcomes, and variable selection methods are employed to identify the significant predictors. In neuroimaging, as another example, functional magnetic resonance imaging (fMRI) is used to measure blood flow changes across the whole brain and linear models are employed to detect (i.e, select) brain regions that activate in response to external stimuli. For these applications, extensions of spike-and-slab prior constructions have been motivated by specific characteristics of the data. For example,

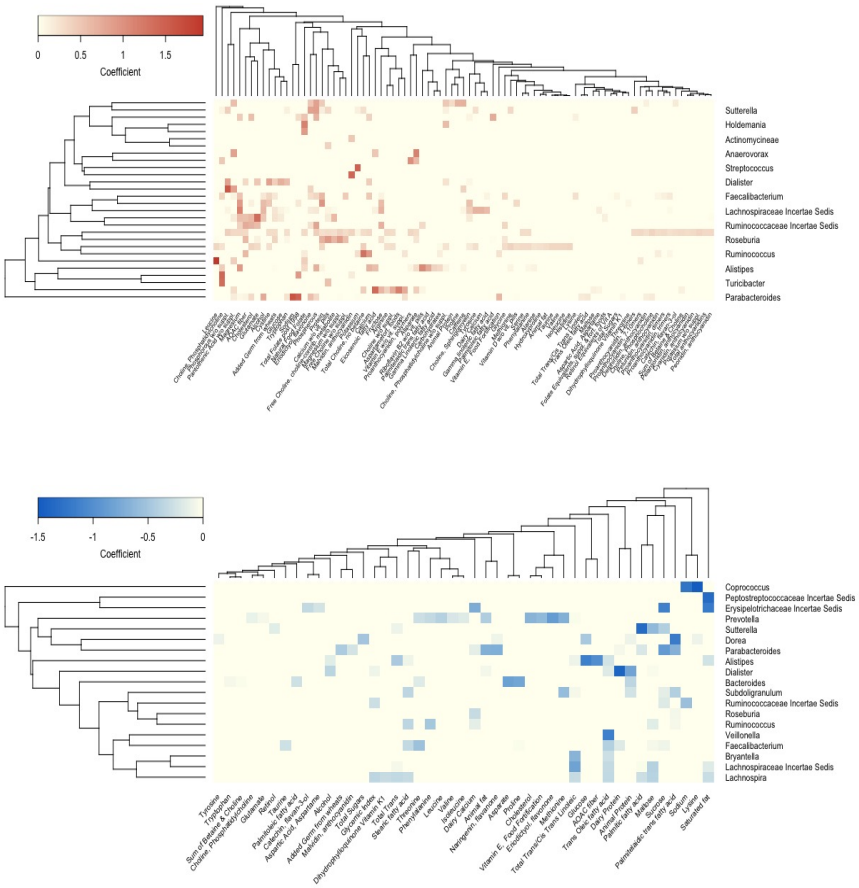

FIGURE 1.3
Analysis of data on dietary intake and microbiome via a regression model for compositional count data with variable selection [28]. Heatmaps of selected positive (upper plot) and negative (lower plot) associations.

[63] put forward a graphical model formulation of a multivariate regression model where target genes (the outcomes) are regulated by microRNAs (the covariates), which are small RNA sequences located upstream of the genes. In the proposed model formulation, spike-and-slab priors allow to identify gene-microRNA interactions, therefore inferring a biological network. In place of the independent Bernoulli priors of type (1.5), the authors assumed a logistic prior construction on $\boldsymbol{\gamma}$ of the form

$$P(\gamma_{jk} = 1|\tau, \eta) = \frac{\exp(\eta + \tau s_{jk})}{1 + \exp(\eta + \tau s_{jk})}, \tag{1.12}$$

that incorporates a set of available scores s_{jk} of possible association between gene-microRNA pairs, as obtained from external sequence/structure information. The prior assumes that the γ_{jk}'s are stochastically independent given τ and reduces to $p(\gamma_{jk} = 1) = \exp(\eta)/(1+\exp(\eta))$ when all $s_{jk} = 0$. Probit constructions that incorporate external information can also be used and have been investigated in other integrative settings [11, 48]. These constructions, while accounting for external information, still assume independence between inclusion indicators. However, in many practical applications, researchers may be interested in incorporating prior information on the dependence structure between covariates, as captured by an underlying biological process and/or a correlation network. Below, prior constructions on $\boldsymbol{\gamma}$ that account for such information are briefly described.

1.4.1 Network priors

Network priors account for known relations among covariates in the form of a graph. For example, in genomics, when covariates are chosen as gene expression levels, a network of gene-gene interactions may be known based on biological information on known pathways (i.e., groups of genes). Here, individual genes are represented by nodes in the network and relations between them by edges. This network structure can be captured via a Markov random field (MRF) prior [3], also known as Ising prior, on the binary indicator vector $\boldsymbol{\gamma}$ as

$$P(\boldsymbol{\gamma}|d, e) \propto \exp\left(d\mathbf{1}'\boldsymbol{\gamma} + e\boldsymbol{\gamma}'G\boldsymbol{\gamma}\right), \tag{1.13}$$

with $\mathbf{G}$ a $p \times p$ adjacency matrix that represents the relations between covariates, that is, with elements $g_{jj'} = 1$ if variables j and j' have a direct link in the network, and $g_{jj'} = 0$ otherwise. Hyperparameters $d \in R$ and $e > 0$ control the global probability of inclusion and the influence of neighbors' inclusion on a covariate's inclusion, respectively. According to parametrization (1.13), a covariate's inclusion probability will increase if neighboring covariates in the known network are also included in the model. The prior simplifies to the independent Bernoulli$(\exp(d)/(1 + \exp(d)))$ for $e = 0$. MRF priors of type (1.13) have been employed in linear models for genomic applications to aid the identification of predictive genes by [32]. Also, [62] considered a linear model that predicts a phenotype based on predictors synthesizing the activity of genes belonging to same pathways. The prior model encodes information on gene-gene networks via a MRF prior, as retrieved from available databases, and inference results in the identification of both relevant pathways and subsets of genes. Among more recent applications, [31] considered a linear model with multivariate responses to identify the joint effect of pollutants on DNA methylation outcomes via structured spike-and-slab priors that leverage the dependence across markers. In all these papers, authors show how small increments of the parameter e in (1.13) can drastically increase the number of selected covariates and provide guidelines on how to select suitable values and/or prior distributions for this parameter.

In situations where the network structure $\mathbf{G}$ in prior (1.13) is unknown, it can be inferred

from the data using priors and learning algorithms for undirected graphical models. [44] used this strategy to obtain a Bayesian modeling approach for linear regression settings that simultaneously performs variable selection while learning the dependence relations between covariates. In this setting, the matrix of covariates $\mathbf{X}$ is treated as random and the joint distribution of $(\mathbf{y}, \mathbf{X})$ is factorized as

$$f(\mathbf{y}, \mathbf{X}) = f(\mathbf{y}|\mathbf{X})f(\mathbf{X}), \tag{1.14}$$

with $f(\mathbf{y}|\mathbf{X})$ defining the linear regression model. Assuming Gaussianity of the $\mathbf{x}_i$'s, we have $\mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Omega})$, where $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ is a $p \times p$ precision matrix. Thus, the presence of edge $g_{jj'} = 1$ in graph $\mathbf{G}$ corresponds to $\omega_{jj'} \neq 0$ in $\mathbf{\Omega}$. [70] proposed a prior for $\mathbf{\Omega}$ that assumes continuous spike-and-slab distributions on the off-diagonal elements and exponential distributions for the diagonal components as

$$p(\mathbf{\Omega}|\mathbf{G}, v_0, v_1, \theta) \propto \prod_{j<j'} \mathcal{N}(\omega_{jj'}|0, v^2_{jj'}) \prod_j \mathcal{E}xp\,(\omega_{jj}|\theta/2)\, I_{\{\mathbf{\Omega} \in M^+\}}, \tag{1.15}$$

with $v^2_{jj'} = v_1$ if $g_{jj'} = 1$, and $v^2_{jj'} = v_0$ if $g_{jj'} = 0$, with $v_0 << v_1$, and where $\mathcal{E}xp(\cdot|\theta/2)$ represents an exponential distribution with mean $2/\theta$. The term $I_{\{\mathbf{\Omega} \in M^+\}}$ restricts the prior to the space of symmetric-positive definite matrices. The model is completed with a prior for $\mathbf{G}$, for example as a simple product of independent Bernoulli's on the $g_{jj'}$'s elements, with a common parameter π to represent the prior probability of inclusion for an individual edge. A specification of π that reflects prior beliefs of sparsity is recommended by [70]. Also, setting $\theta = 1$ implies a relatively vague prior for ω_{jj} when the data are standardized prior to analysis. For posterior inference, [44] incorporated two additional steps, two update $\mathbf{\Omega}$ and $\mathbf{G}$, following [70], within a stochastic search MCMC scheme for linear settings, to obtain simultaneous variable selection and estimation of a graph between covariates. Also, [28] extended these methods to the regression models for compositional count data discussed in Section 1.3.1. Both prior options, (1.13) with $\mathbf{G}$ known and (1.13) with prior (1.15) on $\mathbf{\Omega}|\mathbf{G}$, are available in the `R` package `MicroBVS`.

1.4.2 Spiked nonparametric priors

Other extensions of spike-and-slab priors include constructions that employ nonparametric priors [14, 25, 55]. One construction uses a mixture of a point mass at zero and a nonparametric slab, typically a Dirichlet process (DP) prior [15] with a continuous distribution as its centering distribution. Such construction clusters parameters together when information in the data provides evidence of a common effect. This, in turn, allows to borrow information across covariates, resulting in improved selection and estimation [14]. In [8], this construction is referred to as an "outer" spike-and-slab nonparametric prior, as opposed to the "inner" prior of [25], which is a DP prior where the base measure is modeled as a mixture of a point mass at zero and a diffuse measure. The inner prior formulation does not share information across covariates, but rather clusters vectors of regression coefficients across observations. Recent applications of outer discrete nonparametric constructions include covariate dependent random partition models [2] and dynamic extensions for spatio-temporal dynamic models with random effects [12].

Let us illustrate the outer construction via an application to functional magnetic resonance imaging (fMRI) data, another area of successful applications of models that employ spike-and-slab priors. In a typical task-based fMRI experiment, the whole brain is scanned at multiple time points while the subject is presented with a series of stimuli. Each scan is arranged as a 3D array of volume elements (or "voxels"), and the experiment returns time series data acquired at each voxel. Let $\mathbf{y}_{i\nu} = (y_{i\nu 1}, \ldots, y_{i\nu T})^T$ be the vector of the time

series data at voxel ν, with $\nu = 1, \ldots, V$, for subject i. Common modeling approaches for the analysis of task-based fMRI data rely on the general linear model formulation originally proposed by [16]

$$\mathbf{y}_{i\nu} = \mathbf{X}_{i\nu}\boldsymbol{\beta}_{i\nu} + \varepsilon_{i\nu},\ \varepsilon_{i\nu} \sim \mathcal{N}_T(0, \boldsymbol{\Sigma}_{i\nu}), \tag{1.16}$$

where $\mathbf{X}_{i\nu}$ is a known $T \times K$ design matrix (for K stimuli) modeled as the convolution of the stimulus patterns with a hemodynamic response function that accounts for the delay of the response with respect to the stimulus onset. The task of the inference is to detect those voxels that activate in response to the stimuli, which is equivalent to inferring the non-zero regression coefficients in (1.16). Spatial correlation among brain voxels can be accounted for in the prior construction. Examples include spike-and-slab priors that incorporate MRF priors on the selection indicators, to account for neighboring correlation among voxels, and nonparametric slabs that capture spatial correlation among possibly distant voxels [61, 76]. Let us consider the simpler case $K = 1$, i.e. one stimulus. A spiked nonparametric prior on $\beta_{i\nu}$ can be written as

$$\beta_{i\nu}|\gamma_{i\nu}, G_0 \sim (1 - \gamma_{i\nu})\delta_0(\beta_{i\nu}) + \gamma_{i\nu}G_0, \tag{1.17}$$

where G_0 denotes a Dirichlet process prior with $\mathcal{N}(0, \tau)$ as the centering distribution. With multiple subjects, a hierarchical Dirichlet Process (HDP) prior can be specified as the nonparametric slab, inducing clustering among voxels within a subject on one level of the hierarchy and between subjects on the second level, as

$$\begin{aligned} \beta_{i\nu}|\gamma_{i\nu}, G_i &\sim (1 - \gamma_{i\nu})\delta_0(\beta_{i\nu}) + \gamma_{i\nu}G_i \\ G_i|\eta_1, G_0 &\sim DP(\eta_1, G_0) \\ G_0|\eta_2, P_0 &\sim DP(\eta_2, P_0) \\ P_0 &= \mathcal{N}(0, \tau), \end{aligned} \tag{1.18}$$

with τ fixed, η_1, η_2 the mass parameters and P_0 the base measure. With this prior formulation, the subject-specific distribution G_i varies around a population-based distribution G_0, which is centered around a known parametric model P_0. The mass parameters η_1 and η_2 control the variability of the distribution of the coefficients at the subject and population level, respectively. This construction enables the model to borrow information from subjects exhibiting similar activation patterns while also capturing spatial correlation among distant voxels.

For model (1.16) with prior (1.18), [76] implemented an MCMC algorithm that combines add-delete-swap moves with sampling algorithms for HDP models that use auxiliary parameters for cluster allocation [64]. To ensure scalability, the authors also investigated an alternative approach that uses variational inference combined with an importance sampling procedure [9]. For inference, spatial maps of the activated brain regions for each subject can be produced by thresholding the PPIs of the $\gamma_{i\nu}$'s and corresponding posterior β-maps can be obtained based on the estimated regression coefficients. As an additional feature, the use of the nonparametric HDP prior construction can be exploited to obtain a clustering of the subjects for possible discovery of differential activations. The methods have been implemented in the user-friendly Matlab GUI `NPBayes-fMRI` [26], see Figures 1.4 and 1.5 for some of the available features.

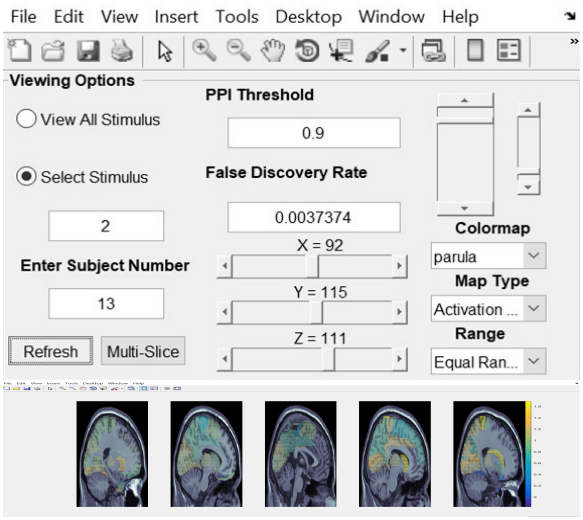

FIGURE 1.4
Matlab GUI `NPBayes-fMRI`: User friendly software that implements a nonparametric Bayesian spatio-temporal general linear model for task-based multi-subject fMRI data. Subject-level visualization interface and corresponding activation β-maps, for subject 13, stimulus 2 and PPI threshold of .9 (adapted from Kook *et al.* [26]).

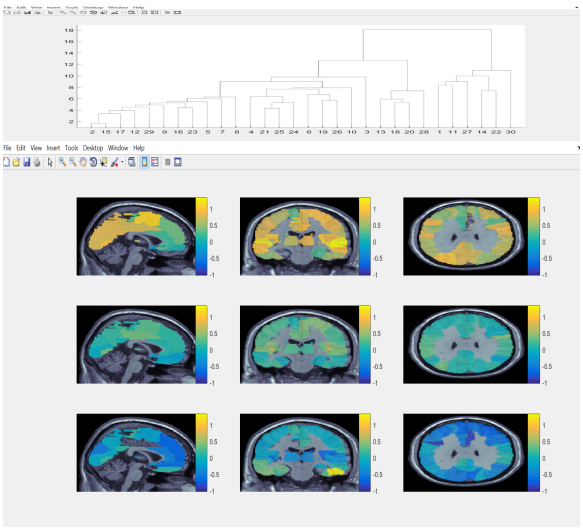

FIGURE 1.5
Matlab GUI `NPBayes-fMRI`: User friendly software that implements a nonparametric Bayesian spatio-temporal general linear model for task-based multi-subject fMRI data. Dendrogram and cluster-level β-maps obtained with three clusters (adapted from Kook *et al.* [26]).

1.5 Scalable Bayesian variable selection

Despite the flexibility offered by spike-and-slab priors, and the availability of clever data augmentation schemes, computational algorithms for posterior inference in regression models remain a challenge, particularly for model settings with a large number of predictors. Below we review alternative strategies to sampling algorithms given by variational inference methods.

1.5.1 Variational inference

Variational Bayes approaches turn inference into an optimization problem, making posterior inference scalable and computationally faster than sampling-based MCMC methods [4]. Typically, variational approaches provide good estimates of mean parameters; however, they tend to underestimate posterior variances and the correlation structure of the data. This shortcoming can be an acceptable trade-off in variable selection problems. For example, [76] performed a comparative study of an MCMC and a variational algorithm for a same linear model and show on simulated data that the variational scheme reduces the computational cost without compromising accuracy in both the detection and the estimation of the non-zero coefficients.

Variational inference (VI) works by specifying a family of approximate distributions $\mathcal{Q}$, which are densities over model parameters and latent variables that depend on free parameters $\boldsymbol{\xi}$, and then using gradient descent to find the values of $\boldsymbol{\xi}$ that minimize the Kullback-Leibler (KL) divergence between the approximate distribution and the true posterior. Let us indicate with W the set of model parameters and latent variables. As discussed in [4], minimizing the KL divergence is equivalent to maximizing the Evidence Lower BOund (ELBO), which is defined as

$$\text{ELBO} = E_{\boldsymbol{\xi}}[\log p(\mathbf{y}, W)] - E_{\boldsymbol{\xi}}[\log q(W)], \tag{1.19}$$

with $p(\mathbf{y}, W)$ the joint distribution of the observed data and the latent variables and parameters, and $q(W)$ the variational distribution of the variables in W. Clearly, the complexity of the approximating class $q(W)$ determines the complexity of the optimization procedure.

The most common approach to obtain an approximating distribution within a VI scheme is mean field approximation, which assumes that the approximating distribution factorizes over some partition of the parameters. This approach is widely used with spike-and-slab priors [9, 66, 76]. In particular, the variational distribution for $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is assumed to factorize as

$$q\left(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \boldsymbol{\xi}\right) = \prod_{j=1}^{p} q\left(\beta_j, \gamma_j; \xi_j\right), \tag{1.20}$$

with

$$q\left(\beta_j, \gamma_j; \xi_j\right) = \begin{cases} \psi_j \mathcal{N}(\beta_j \mid \mu_j, s_j^2) & \text{if } \gamma_j = 1 \\ (1 - \psi_j)\, \delta_0\left(\beta_j\right) & \text{otherwise,} \end{cases} \tag{1.21}$$

and free parameters $\xi_j = \left(\psi_j, \mu_j, s_j^2\right)$. A coordinate ascent algorithm can then be implemented to maximize the ELBO by setting the partial derivatives equal to zero. After initializing the free parameters, the algorithm updates each component of ξ_j given all the others, iteratively, until convergence of the ELBO is met. The ELBO is further maximized

by finding optimal values for the remainder of the model parameters. VI schemes can be combined with importance sampling procedures, to integrate over some of the model parameters, and/or data augmentation schemes, to implement efficient closed-form VI updates that exploit the conditional conjugacy of latent parameters [9, 41, 75]. At convergence, PPIs are approximated via variational distribution values and thresholded to select covariates. Corresponding regression coefficients are estimated as the variational distribution values at convergence. Variational approaches are only suitable for point estimation and do not allow to assess uncertainty about the estimates. Additionally, in situations with correlated covariates, performances can be sensitive to initialization and can result in poor estimation [51].

Recently, hybrid schemes that combine VI steps on $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ with expectation-maximization (EM) estimation steps on latent variables and other model parameters have also been investigated [42]. As noted by [4], the first term of the ELBO is the object to optimize in EM. One could therefore consider EM approaches as a special case of variational inference, where the variational distributions are point masses.

1.6 Conclusion

Bayesian approaches offer a coherent and flexible framework for variable selection. This chapter has reviewed discrete spike-and-slab priors for linear settings, with a focus on prior constructions and computational aspects. Theoretical properties of these priors will be discussed in the next chapter. This will be followed by treatments of the continuous spike-and-slab priors, which employ mixtures of two unimodal distributions and require careful choices of the variance parameters that separate important variables from noisy ones.

Spike-and-slab priors are sometimes referred to as two-group priors, in contrast to the one-group unimodal continuous shrinkage priors, which will be covered in the second part of the handbook. An advantage of spike-and-slab priors over continuous shrinkage priors is that, in addition to the sparse estimation of the regression coefficients, they produce PPIs for each covariate. Another advantage is that the flexibility of the constructions allows to incorporate structural information among the covariates via the prior choice on the latent indicators, for example as the network priors described in this chapter. The disadvantages are obviously in the computations, particularly in high dimensions, as stochastic search MCMC algorithms need to explore a large posterior space of possible models. Some solutions are offered by optimization procedures, such as the EMVS of [54], for continuous spike-and-slab priors, and by the variational algorithms reviewed in this chapter. These methods, however, can only produce point estimates.

Spike-and-slab priors have been extended to a wide variety of modeling frameworks, such as multivariate regression models, state-space and time-varying coefficient models, as well as to edge selection in graphical models. These topics will be discussed in the third part of the handbook.

Software

The `R` package `MicroBVS`, written by Matthew Koslovsky, is available at `https://github.com/mkoslovsky/MicroBVS`. The user-friendly Matlab GUI `NPBayes-fMRI`, written by Eric Kook, is available at `https://github.com/marinavannucci/NPBayes_fMRI`.

Acknowledgements

Many thanks to Matthew Koslovsky for producing Figures 1.2 and 1.3.

Bibliography

[1] M.M. Barbieri and J.O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32:870–897, 2004.

[2] W. Barcella, M. De Iorio, G. Baio, and J. Malone-Lee. Variable selection in covariate dependent random partition models: An application to urinary tract infection. *Statistics in Medicine*, 35:1373–89, 2016.

[3] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

[4] D.M. Blei, A. Kucukelbir, and J.D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[5] L. Bottolo and S. Richardson. Evolutionary stochastic search. *Bayesian Analysis*, 5(3):583–618, 2010.

[6] P.J. Brown, M. Vannucci, and T. Fearn. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, 60(3):627–641, 1998.

[7] P.J. Brown, M. Vannucci, and T. Fearn. Bayes model averaging with selection of regressors. *Journal of the Royal Statistical Society, Series B*, 64:519–536, 2002.

[8] A. Canale, A. Lijoi, B. Nipoti, and I. Prünster. On the Pitman-Yor process with spike and slab base measure. *Biometrika*, 104(3):681–697, 2017.

[9] P. Carbonetto and M. Stephens. Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108, 2012.

[10] C.M. Carvalho, N.G. Polson, and J.G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97:465–480, 2010.

[11] A. Cassese, A. Guindani, and M. Vannucci. A Bayesian integrative model for genetical genomics with spatially informed variable selection. *Cancer Informatics*, 13(S2):29–37, 2014.

[12] A. Cassese, W. Zhu, M. Guindani, and M. Vannucci. A Bayesian nonparametric spiked process prior for dynamic model selection. *Bayesian Analysis*, 14(2):553–572, 2019.

[13] I. Castillo, J. Schmidt-Hieber, and A. Van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, 2015.

[14] D.B. Dunson, A.H. Herring, and S.M. Engel. Bayesian selection and clustering of polymorphisms in functionally-related genes. *Journal of the American Statistical Association*, 103:534–546, 2008.

[15] T.S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, pages 209–230, 1973.

[16] K.J. Friston, P. Jezzard, and R. Turner. Analysis of functional MRI time-series. *Human Brain Mapping*, 1(2):153–171, 1994.

[17] E.I. George and R.E. McCulloch. Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 85:398–409, 1993.

[18] E.I George and R.E. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373, 1997.

[19] R. Gottardo and A.E. Raftery. Markov chain Monte Carlo with mixtures of mutually singular distributions. *Journal of Computational and Graphical Statistics*, 17:949–975, 2008.

[20] P.J. Green. Reversible jump Markov chain Monte Carlo computations and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[21] J.E. Griffin, K. Latuszynski, and M.F.J. Steel. In search of lost (mixing) time: Adaptive Markov chain Monte Carlo schemes for Bayesian variable selection with very large p. *Biometrika*, to appear, 2020.

[22] C. Hans, A. Dobra, and M. West. Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102 (478):507–516, 2007.

[23] H. Ishwaran and J.S. Rao. Spike and slab variable selection : frequentist and Bayesian strategies. *The Annals of Statistics*, 33:730–773, 2005.

[24] H. Ishwaran and J.S. Rao. Consistency of spike and slab regression. *Statistics and Probability Letters*, 81:1920–1928, 2011.

[25] S. Kim, D.B. Dahl, and M. Vannucci. Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis*, 4(4):707–732, 2009.

[26] J.H. Kook, M. Guindani, L. Zhang, and M. Vannucci. NPBayes-fMRI: Nonparametric Bayesian general linear models for single- and multi-subject fMRI data. *Statistics in Biosciences*, 11(1):3–21, 2019.

[27] M.D. Koslovsky, K.L. Hoffman, C.R. Daniel, and M. Vannucci. A Bayesian model of microbiome data for simultaneous identification of covariate associations and prediction of phenotypic outcomes. *Annals of Applied Statistics*, 14(3):1471–1492, 2020.

[28] M.D. Koslovsky and M. Vannucci. MicroBVS: Dirichlet-tree multinomial regression models with Bayesian variable selection - an R package. *BMC Bioinformatics*, 21:301:DOI 10.1186/s12859–020–03640–0, 2020.

[29] D.W. Kwon, M.T. Landi, M. Vannucci, H.J. Issaq, D. Prieto, and R.M. Pfeiffer. An efficient stochastic search for Bayesian variable selection with high-dimensional correlated predictors. *Computational Statistics and Data Analysis*, 55(10):2807–2818, 2011.

[30] D.S. Lamnisos, J.E. Griffin, and M.F.J. Steel. Adaptive Monte Carlo for Bayesian variable selection in regression models. *Journal of Computational and Graphical Statistics*, 22:729–748, 2013.

[31] K.H. Lee, M.G. Tadesse, A.A. Baccarelli, J. Schwartz, and B.A. Coull. Multivariate Bayesian variable selection exploiting dependence structure among outcomes: Application to air pollution effects on DNA methylation. *Biometrics*, 73:232–241, 2017.

[32] F. Li and N. Zhang. Bayesian variable selection in structured high-dimensional covariate space with application in genomics. *Journal of the American Statistical Association*, 105:1978–2002, 2010.

[33] F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixture of g-priors for Bayes variable selection. *Journal of the American Statistical Association*, 103:410–423, 2008.

[34] D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215 – 232, 1995.

[35] P. McCullagh and J.A. Nelder. *Generalized Linear Models, second edition.* Chapman & Hall, London, 1989.

[36] A. Miller. *Subset Selection in Regression.* Chapman & Hall/CRC, Bocan Raton: Florida, 2002.

[37] T.J. Mitchell and J.J. Beauchamp. Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1036, 1988.

[38] N.N. Narisetty and X. He. Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42:789–817, 2014.

[39] M.A. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biostatistics*, 5(2):155–176, 2004.

[40] I. Ntzoufras, P. Dellaportas, and J.J. Forster. Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference*, 111:165–180, 2003.

[41] J.T. Ormerod, C. You, and S. Müller. A variational Bayes approach to variable selection. *Electronic Journal of Statistics*, 11(2):3549 – 3594, 2017.

[42] N. Osborne, C.B. Peterson, and M. Vannucci. Latent network estimation and variable selection for compositional data via variational EM. *Journal of Computational and Graphical Statistics*, in press, 2021.

[43] T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[44] C.B. Peterson, F.C. Stingo, and M. Vannucci. Joint Bayesian variable and graph selection for regression models with network-structured predictors. *Statistics in Medicine*, 35(7):1017–1031, 2016.

[45] J.W. Pillow and J. Scott. Fully Bayesian inference for neural models with negative-binomial spiking. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1898–1906. 2012.

[46] N.G. Polson and J.G. Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010.

[47] N.G. Polson, J.G. Scott, and J. Windle. Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

[48] M.A. Quintana and D.V. Conti. Integrative variable selection via Bayesian model uncertainty. *Statistics in Medicine*, 32:4938–4953, 2013.

[49] A.E. Raftery. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, 83:251 – 266, 1996.

[50] A.E. Raftery, D. Madigan, and J.A. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179 – 191, 1997.

[51] K. Ray and B. Szabó. Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 0(0):1–12, 2021.

[52] S. Richardson, L. Bottolo, and J.S. Rosenthal. Bayesian models for sparse regression analysis of high dimensional data. In *Bayesian Statistics 9*, pages 539–569, 2010.

[53] V. Rockova. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics*, 46(1):401–437, 2018.

[54] V. Rockova and E.I. George. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109:828–846, 2014.

[55] T. Savitsky and M. Vannucci. Spiked Dirichlet process priors for Gaussian process models. *Journal of Probability and Statistics*, 2010, 2010.

[56] T. Savitsky, M. Vannucci, and N. Sha. Variable selection for nonparametric Gaussian process priors: Models and computational strategies. *Statistical Science*, 26(1):130–149, 2011.

[57] J.G. Scott and J.O. Berger. Bayes and empirical-Bayes multiplicity adjustment in the variable selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2008.

[58] N. Sha, M.G. Tadesse, and M. Vannucci. Bayesian variable selection for the analysis of microarray data with censored outcome. *Bioinformatics*, 22(18):2262–2268, 2006.

[59] N. Sha, M. Vannucci, M.G. Tadesse, P.J. Brown, I. Dragoni, N. Davies, T.C. Roberts, A. Contestabile, N. Salmon, C. Buckley, and F. Falciani. Bayesian variable selection in multinomial probit models to identify molecular signatures of disease stage. *Biometrics*, 60(3):812–819, 2004.

[60] M. Smith and R. Kohn. Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–343, 1996.

[61] M. Smith, B. Putz, B.D. Auer, and L.D. Fahrmeir. Assessing brain activity through spatial Bayesian variable selection. *Neuroimage*, 20:802–815, 2003.

[62] F.C. Stingo, Y.A. Chen, M.G. Tadesse, and M. Vannucci. Incorporating biological information in Bayesian models for the selection of pathways and genes. *Annals of Applied Statistics*, 5(3):1978–2002, 2011.

[63] F.C. Stingo, Y.A. Chen, M. Vannucci, M. Barrier, and P.E. Mirkes. A Bayesian graphical modeling approach to microRNA regulatory network inference. *Annals of Applied Statistics*, 4(4):2024–2048, 2010.

[64] Y.W. Teh, M.I. Jordan, M.J. Beal, and D.M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[65] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, pages 267–288, 1996.

[66] M.K. Titsias and M. Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems*, pages 2339–2347, 2011.

[67] W.D. Wadsworth, R. Argiento, M. Guindani, J. Galloway-Pena, S.A. Shelburne, and M. Vannucci. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. *BMC Bioinformatics*, 18(1):94, 2017.

[68] K.Y.Y. Wan and J.E. Griffin. An adaptive MCMC method for Bayesian variable selection in logistic and accelerated failure time regression models. *Statistics and Computing*, 31(6), 2021.

[69] E.T. Wang, S. Chiang, Z. Haneef, V.R. Rao, R. Moss, and M. Vannucci. Bayesian non-homogeneous hidden Markov model with variable selection for investigating drivers of seizure risk cycling. *Annals of Applied Statistics*, submitted, 2021.

[70] H. Wang. Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377, 2015.

[71] T. Wang and H. Zhao. A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. *Biometrics*, 73(3):792–801, 2017.

[72] G.D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S.A. Keilbaugh, M. Bewtra, D. Knights, W.A. Walters, and R. Knight. Linking long-term dietary patterns with gut microbial enterotypes. *Science*, 334(6052):105–108, 2011.

[73] Y. Yang, M.J. Wainwright, and M.I. Jordan. On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532, 2016.

[74] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P.K. Goel and A. Zellner, editors, *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pages 233–243. North-Holland/Elsevier, 1986.

[75] C.-X. Zhang, S. Xu, and J.-S. Zhang. A novel variational Bayesian method for variable selection in logistic regression models. *Computational Statistics & Data Analysis*, 133:1–19, 2019.

[76] L. Zhang, M. Guindani, F. Versace, J.M. Engelmann, and M. Vannucci. A spatio-temporal nonparametric Bayesian model of multi-subject fMRI data. *Annals of Applied Statistics*, 10(2):638–666, 2016.