Robust Bayesian Nonparametric Variable Selection for Linear Regression

Alberto Cabezas[*] Marco Battiston[*] Christopher Nemeth[*]

Abstract

Spike-and-slab and horseshoe regression are arguably the most popular Bayesian variable selection approaches for linear regression models. However, their performance can deteriorate if outliers and heteroskedasticity are present in the data, which are common features in many real-world statistics and machine learning applications. In this work, we propose a Bayesian nonparametric approach to linear regression that performs variable selection while accounting for outliers and heteroskedasticity. Our proposed model is an instance of a Dirichlet process scale mixture model with the advantage that we can derive the full conditional distributions of all parameters in closed form, hence producing an efficient Gibbs sampler for posterior inference. Moreover, we present how to extend the model to account for heavy-tailed response variables. The performance of the model is tested against competing algorithms on synthetic and real-world datasets.

1 Introduction

Bayesian variable selection is a popular tool in statistics and machine learning that can be used for feature selection in linear regression models. The two most popular models are arguably the spike-and-slab and horseshoe regression models. Both models rely on choosing a sparsity inducing prior over the regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. In the spike-and-slab model, the prior is a mixture between a point mass at zero and a diffuse prior, while in the horseshoe model, a continuous prior balances the local and global shrinkage (see Section 2). Compared to the estimator for the regression coefficients in standard linear regression, the posterior distribution under these priors produces a shrinkage effect toward zero in the posterior point estimates. This is especially useful in the $p >> n$ regime, in which the number of attributes p can be much larger than the number of observations n.

In this paper we propose extensions of the spike-and-slab and horseshoe regression models to deal with heteroskedasticity and outliers in the data. Specifically, we propose a Bayesian nonparametric model in which each observation is endowed with its own specific variance σ_i^2, which is sampled from an unknown distribution P. The distribution P is a Dirichlet process prior and the resulting model is a Dirichlet process scale mixture model. Due to the discreteness of the Dirichlet process, the resulting vector of variances $(\sigma_1^2, \dots, \sigma_n^2)$ will be partitioned into groups, hence also producing a corresponding clustering of the observations, in which observations belonging to the same group will have the same conditional variance. Moreover, some observations may be allocated to a single group having much larger variance than the others, hence allowing for outliers in the data.

The key features of our proposed model are:

- **Parsimonious regression construction:** Our nonparametric method performs feature selection with the posterior concentrating on the most relevant prediction coefficients. However, unlike variable coefficient models, there is no increase in the degrees-of-freedom associated with the number of regression parameters (or smoothness of parameters to control, c.f. functional regression).
- **Interpretable model structure:** The proposed linear regression model changes the structure of the marginal variance components while retaining the highly interpretable properties of a standard sparse regression formulation.
- **Posterior credible intervals:** A fully Bayesian approach gives credible intervals for the regression coefficients. We propose several ways to perform posterior inference selection using these intervals, which provide uncertainty quantification for decision makers working with high-dimensional data.
- **Efficient inference:** Under our model, full conditional distributions of all parameters can be derived in *closed form*, hence producing an effi-

[*] Department of Mathematics and Statistics, Lancaster University

cient Gibbs algorithm that gives a tuning-free and rejection-free Markov chain Monte Carlo (MCMC) sampler.

The rest of the paper is organized as follows. Section 2 briefly reviews the spike-and-slab and horseshoe regression models. Section 3 introduces the Bayesian nonparametric framework and our Dirichlet process mixture model for linear regression along with details of our MCMC sampler. In Subsection 3.3.2, we provide an extension to our model to account for heavy-tailed response variables. Finally, Sections 4 and 5 present an empirical comparison of our proposed model against popular alternative models on synthetic data and two real-world data examples.

2 Bayesian Variable Selection for Linear Regression

The Bayesian approach to variable selection for linear regression models is to introduce sparsity-inducing priors on the regression coefficients. The two most popular approaches in the literature, which we shall review here, are the discrete mixture priors known as the spike-and-slab (Mitchell and Beauchamp, 1988; George and McCulloch, 1993), and the continuous shrinkage priors, most notably the horseshoe prior (Carvalho et al., 2010).

2.1 Spike-and-slab priors

The original spike-and-slab model was initially proposed by Mitchell & Beauchamp (1988) and significantly developed by Madigan & Raftery (1994) and George & McCulloch (1997). The final adjustments to the model were completed by Ishwaran & Rao (2005) in what they refer to as the *stochastic variable selection* model. The spike-and-slab prior is intuitively simple and consists of two components. The *spike* is a delta function centered at zero indicating $\beta_j \approx 0$, and the *slab* gives probability mass to non-zero coefficients.

In our regression framework, we have data $\mathbf{y} \in \mathbb{R}^n$ that is explained by a matrix of attributes $\mathbf{X} \in \mathbb{R}^{n \times p}$ and coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$. Assuming a spike-and-slab prior for $\boldsymbol{\beta}$, we have the following model,

$$\begin{aligned}
y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2) \qquad i = 1, \ldots, n \\
\beta_j|\eta_j, \tau_j^2 &\sim \mathcal{N}(0, \eta_j \tau_j^2) \qquad j = 1, \ldots, p \\
\eta_j|\nu_0, \omega &\sim (1-\omega)\delta_{\nu_0}(\cdot) + \omega\delta_1(\cdot) \\
\tau_j^2|a_1, a_2 &\sim \mathrm{IG}(a_1, a_2) \qquad\qquad (1) \\
\omega &\sim \mathrm{Unif}[0, 1] \\
\sigma^2|b_1, b_2 &\sim \mathrm{IG}(b_1, b_2)
\end{aligned}$$

The likelihood for the data assumes a standard Gaussian regression model, and the prior for $\boldsymbol{\beta}$ is a scale mixture of Gaussians. The variable η_j is a latent indicator and $\delta_{v_0}(\cdot)$ denotes a point mass at v_0, where v_0 is a value chosen close to 0, and thus the variance of the prior on $\boldsymbol{\beta}$ is either almost zero or a broad Gaussian distribution.

2.2 Horseshoe priors

Spike-and-slab priors are intuitively appealing, but in practice their discrete nature makes posterior inference computationally difficult. A popular alternative perspective on sparse Bayesian regression is given by the horseshoe prior construction (Carvalho et al., 2009, 2010, see). The horseshoe prior is a continuous shrinkage prior which makes posterior computation more efficient when using gradient-based MCMC sampling tools such as STAN (Carpenter et al., 2017).

$$\begin{aligned}
y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2) \qquad i = 1, \ldots, n \\
\beta_j|\lambda_j^2, \tau^2 &\sim \mathcal{N}(0, \lambda_j^2 \tau^2) \qquad j = 1, \ldots, p \\
\lambda_j &\sim \mathcal{C}^+(0, 1) \\
\tau &\sim \mathcal{C}^+(0, 1) \qquad\qquad (2) \\
\sigma^2|b_1, b_2 &\sim \mathrm{IG}(b_1, b_2)
\end{aligned}$$

Under the horseshoe regression model, the parameters λ_j and τ are the *local* and *global* shrinkage parameters, respectively. Following Carvalho et al. (2010), we choose half-Cauchy priors for λ_j and τ. The intuition behind the horseshoe prior is that the global parameter τ will force the regression coefficients towards zero, while the heavy tails of the half-Cauchy prior for the local shrinkage parameters λ_j will allow for non-zero $\boldsymbol{\beta}$ coefficients.

The horseshoe model used in our experiments follows the form originally proposed by Carvalho et al. (2010). A standard Gibbs sampling approach on this parametrization of the model is difficult to implement due to the non-conjugate posterior form of the shrinkage parameters in the model. In our implementation we introduce auxiliary variables that lead to conjugate full conditionals for all parameters as suggested by Makalic and Schmidt (2015a), and allow for a straightforward implementation of Gibbs sampling.

The parameterization given by Makalic and Schmidt (2015a) makes use of an inverse gamma scale mixture representation of a random variable with a half-Cauchy distribution. It can be shown that $X \sim \mathcal{C}^+(0, A)$ if $X^2|a \sim \mathrm{IG}(1/2, 1/a)$ and $a \sim \mathrm{IG}(1/2, 1/A^2)$. Using this decomposition leads to the reparametrized horseshoe model

$$\begin{aligned}
y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2) && i = 1, \dots, n \\
\beta_j|\lambda_j^2, \tau^2 &\sim \mathcal{N}(0, \lambda_j^2 \tau^2) && j = 1, \dots, p \\
\lambda_j^2 &\sim \mathrm{IG}(1/2, 1/\nu_j) \\
\tau^2 &\sim \mathrm{IG}(1/2, 1/\xi) && (3) \\
\nu_1, \dots, \nu_p, \xi &\sim \mathrm{IG}(1/2, 1) \\
\sigma^2|b_1, b_2 &\sim \mathrm{IG}(b_1, b_2),
\end{aligned}$$

for which the sampling schemes presented in the next section easily follow.

3 Nonparametric Stochastic Variable Selection

3.1 Background to Dirichlet processes

In Bayesian nonparametric statistics, unknown parameters are infinite dimensional and cannot be parametrized by a subset of Euclidean space. The most common examples of nonparametric problems are the estimation of density functions, distribution functions and nonparametric regression (Hjort et al., 2010).

The most popular nonparametric distribution function is the Dirichlet process, introduced in Ferguson (1973). In this setting, P has a *Dirichlet process* distribution with base measure P_0 and concentration parameter α, denoted $P \sim DP(\alpha, P_0)$. For every measurable partition $(A_1, \dots, A_k)$, the random vector $(P(A_1), \dots,$
$P(A_k))$ has a Dirichlet distribution on the k-dimensional simplex with parameters $(\alpha P_0(A_1), \dots, \alpha P_0(A_k))$. The base measure P_0 is the mean of the prior (i.e. $\mathbb{E}(P) = P_0$), while the concentration parameter α regulates prior uncertainty around P_0. A large α value implies a strong belief in the prior.

There are many different representations of the Dirichlet process (see Ghosal and Van der Vaart (2017), Chapter 4). For the purpose of this paper, we shall utilize the *Chinese restaurant process representation*, which gives the marginal distribution of the data when the unknown distribution P has been marginalized out. Specifically, if $\theta_{1:n}$ is an i.i.d. sample from P, $\theta_i|P \overset{iid}{\sim} P$, and $P \sim DP(\alpha, P_0)$, then the marginal distribution of $\theta_{1:n}$, when P is integrated out, is

$$\pi(\theta_{1:n}|\alpha, P_0) = \int \prod_{i=1}^{n} \mathcal{P}(\theta_i) \mathrm{DP}(d\mathcal{P}; \alpha, P_0),$$

which can be described by the marginal of θ_1 and the conditional distributions of $\theta_i|\theta_{1:i-1}$ for $i = 2, \dots, n$, $\pi(\theta_{1:n}|\alpha, P_0) = \pi(\theta_1|\alpha, P_0) \prod_{i=2}^{n} \pi(\theta_i|\theta_{1:i-1}, \alpha, P_0)$, where $\pi(\theta_1|\alpha, P_0) = P_0$ and

$$\theta_i|\theta_{1:i-1}, \alpha, P_0 \sim \frac{1}{i-1+\alpha} \sum_{i=1}^{n} \delta_{\theta_i} + \frac{\alpha}{i-1+\alpha} P_0.$$

Thus, a sample $\theta_{1:n}$ from $\pi(\theta_{1:n}|\alpha, P_0)$ will display ties with positive probability. The parameter α regulates the number of clusters in $\theta_{1:n}$. All else being equal, larger α will lead to more distinct values within $\theta_{1:n}$.

3.2 Dirichlet process mixture model

The Dirichlet process mixture model (DPM) (Lo, 1984) assumes each observation y_i is sampled from a countable mixture model. The likelihood of each mixture component, $K(y_i; \theta_k^*)$, is parametrized by an unknown parameter vector θ_k^*. In the following, K will be a Gaussian kernel and θ_k^* is the group specific variance. Both the mixture parameters θ^* and the mixture weights w_k can be encoded into a random measure $P = \sum_{k \geq 1} w_k \delta_{\theta_k^*}$ which is endowed with a Dirichlet process prior, hence producing the following mixture model for $i = 1, \dots, n$

$$\begin{aligned}
y_i|P &\sim \sum_{k=1}^{\infty} w_k K(y_i; \theta_k^*) = \int K(y_i; \theta) P(d\theta) \\
P &\sim DP(\alpha, P_0).
\end{aligned}$$

By introducing latent class variables, $\{\theta_i\}_{i=1}^n$ s.t. $\mathbb{P}(\theta_i = \theta_k^*) = w_k$, the DPM model is equivalent to the latent variable mixture model

$$\begin{aligned}
y_i|\theta_i &\sim K(y_i; \theta_i) && i = 1, \dots, n \\
\theta_i|P &\overset{iid}{\sim} P && i = 1, \dots, n && (4) \\
P &\sim DP(\alpha, P_0)
\end{aligned}$$

The DPM model has been used in a variety of applications in statistics and machine learning for density estimation and clustering. In density estimation, the goal is to estimate the unknown density of the observations through a mixture model, while in clustering the goal is to cluster observations into groups with a similar distribution. For the latter task, the discreteness property of the Dirichlet process is very convenient, since with positive probability we will observe ties among the latent variables $\theta_{1:n}$. Two observations y_i and y_j having the same value of the latent variables θ_i and θ_j will be assigned to the same cluster and have the same distribution.

In Section 3.3 we utilize the properties of the Dirichlet process mixture model to propose a Bayesian nonparametric extension to the spike-and-slab and horseshoe regression models. We show that our new model is able to capture heteroskedasticity and outliers in the observations, as well as capturing clustering behaviour in the variance structure. Under both DP variable selection models the mixture component has a likelihood with a cluster-specific variance term which is not fixed *apriori*, but is learned from the data.

3.3 Dirichlet process variable selection

In both the spike-and-slab (1) and horsehoe (3) models, one assumes that each y_i has the same conditional variance σ^2. Under our nonparametric Dirichlet process model, we introduce an observation dependent variance σ_i^2 for each data point. The vector $\boldsymbol{\sigma^2} := \sigma^2_{1:n}$ is assumed to be sampled from an unknown discrete distribution P sampled from a Dirichlet process. Specifically, we consider the following general hierarchical model,

$$\begin{aligned} y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\sigma^2} &\sim \mathcal{N}(\boldsymbol{x}_i^\top\boldsymbol{\beta}, \sigma_i^2) && i = 1, \ldots, n \\ \boldsymbol{\beta} &\sim \pi \\ \sigma_i^2|P &\sim P(d\sigma_i^2) && i = 1, \ldots, n \\ P|\alpha &\sim \mathrm{DP}(\alpha, \mathrm{IG}(b_1, b_2)) \\ \alpha &\sim \mathrm{Gamma}(d_1, d_2), \end{aligned} \tag{5}$$

where the prior π for $\boldsymbol{\beta}$ either follows the spike-and-slab construction in lines 2-5 of eq. (1), or the horseshoe model in lines 2-4 of eq. (3). We refer to the general model in eq. (5) as the *Dirichlet process variable selection model* and recognize that the spike-and-slab and horseshoe models are special cases, which we denote as *Dirichlet process spike-and-slab* (DPSS) and *Dirichlet process horseshoe* (DPHS), respectively.

In this model, we are assuming a variance σ_i for each observation y_i, where the vector of variances $\sigma_1, \ldots, \sigma_n$ is assumed to be conditionally i.i.d. from an unknown distribution P. Specifically, we use a Dirichlet process as a nonparametric prior for P, centered at $\mathrm{IG}(b_1, b_2)$ with concentration parameter α. From the properties of the DP, P will almost surely be a discrete distribution. This implies that, with positive probability, we will observe *ties* among $\sigma_1, \ldots, \sigma_n$. In other words, some σ_i will take the same value. Let's denote by $\sigma_1^*, \ldots, \sigma_{K_n}^*$ the K_n distinct values assumed by $\sigma_1, \ldots, \sigma_n$. We will then have K_n clusters among our observations $\{y_i\}_{i=1}^n$, where observations within a cluster have the same variance, but different clusters have different variances. Specifically, if $\sigma_i = \sigma_j$, then y_i and y_j will be in the same cluster and have the same conditional variance, but with possibly different means, depending on their attributes.

3.3.1 Posterior inference

Under the DPSS and DPHS models, posterior inference can be carried out efficiently using a Markov chain Monte Carlo sampler. We propose a general Gibbs sampler to sample $\sigma^2_{1:n}$ (see Algorithm 1) based on the algorithm by Escobar and West (1995), where at each iteration, we first resample a classification vector $c_{1:n}$ assigning each σ_i^2 to a block in the partition, and then resample the distinct values $\sigma^{2*}_{1:k}$.

Algorithm 1 DP variable selection Gibbs sampler

```
for $t$ in 1: number of iterations do
    for i in 1:n do
        Sample classification value $c_i$ with probability (6).
    end for
    for k in 1:$K_n$ do
        Sample $\sigma_k^{2*}$
            $\sigma_k^{2*} \sim \mathrm{IG}\left(b_1 + \frac{n_k}{2},\, b_2 + \frac{1}{2}\sum_{i:C_k}(y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta})^2\right),$
        where $C_k = \{\sigma_i^2 \mid c_i = k\}$
    end for
    if Spike-and-Slab then
        Sample $\boldsymbol{\beta}$, $\tau_j$, $\eta_j$ and $\omega$ using Algorithm 2
    else if Horseshoe then
        Sample $\boldsymbol{\beta}$, $\lambda_j$ and $\tau$ using Algorithm 3
    end if
    Sample $\alpha$, by sampling the latent variable
        $\psi|\alpha, K_n \sim \mathrm{Beta}(\alpha, n)$ then
        $\alpha|\psi, K_n \sim \mathrm{Gamma}(d_1 + K_n, d_2 - \log\psi)$
end for
```

The partition generated by the variance value assigned to each observation is distributed, *a priori*, as a *Chinese Restaurant Process* (Aldous, 1985). This distribution can be understood intuitively as a Chinese restaurant serving an infinite amount of customers arriving in succession. Each customer needs to be seated on a unbounded number of tables, where the probability of seating a customer at an occupied table is proportional to the number of people already seated at that table. Alternatively, the customer could be seated at a new table with probability proportional to the concentration parameter α. Assuming i.i.d. variances, we can assume, *a posteriori*, that each observation is the last customer to arrive and each table represents a cluster of variances. The customer/observation will be seated at an already occupied table, or a new table with probability proportional to

$$\mathbb{P}[c_i = c \mid c_{-i}, y_{1:n}, \sigma^{2*}_{1:K_n}] \propto \begin{cases} n_{-i,c} N(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma_c^{2*}) & \text{for } 1 \le c \le K^- \\ \alpha g(y_i; \boldsymbol{x}_i, \boldsymbol{\beta}, b_1, b_2) & \text{for } c = K^- + 1, \end{cases} \tag{6}$$

where K^- are the number of distinct c_j for $j \neq i$, labeled $\{1, \dots, K^-\}$, and $n_{-i,c}$ is the number of $c_j = c$ for $j \neq i$. The likelihood of observation i being assigned a new variance, i.e. seated at a new table, is defined as

$$\begin{aligned} g(y_i; \boldsymbol{x}_i, \boldsymbol{\beta}, b_1, b_2) &:= \int \mathcal{N}(y_i; \boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma^2) \text{IG}(\sigma^2; b_1, b_2) d\sigma^2 \\ &= \frac{b_2^{b_1}}{\sqrt{2\pi}} \frac{\Gamma(b_1 + 2^{-1})}{\Gamma(b_1)} \cdot \\ &\cdot \left(\frac{(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2}{2} + b_2 \right)^{-(b_1 + \frac{1}{2})} . \end{aligned}$$

Finally, the concentration parameter α is sampled using a data augmentation trick (Ghosal and Van der Vaart, 2017, see pg.89).

We sample the regression coefficients $\boldsymbol{\beta}$ for the spike-and-slab and horseshoe models using Algorithms 2 and 3, respectively. The Gibbs sampler for the spike-and-slab model follows from Ishwaran and Rao (2005) while the Gibbs sampler for the horseshoe model is based on the data augmentation construction of Makalic and Schmidt (2015b). The full conditional distributions of all parameters in both models can be easily derived and have closed form expressions.

An advantage of our Dirichlet process model construction is that the number of clusters K_n that partition the vector of variance parameters $\sigma^2_{1:n}$ are learned by the DP model and do not need to be fixed *apriori*.

Algorithm 2 Sample β with the spike-and-slab model

```
Sample $\boldsymbol{\beta}$
    $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\beta|\cdot}, \Sigma_{\beta|\cdot})$
where $\boldsymbol{\mu}_{\beta|\cdot} = (X^T \Sigma^{-1} X + \Lambda^{-1})^{-1} X^T \Sigma^{-1} \boldsymbol{y}$ and
$\Sigma^{-1}_{\beta|\cdot} = X^T \Sigma^{-1} X + \Lambda^{-1}$, with $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$
and $\Lambda = \text{diag}(\tau_1^2 \eta_1, \dots, \tau_p^2 \eta_p)$,
for j in 1:p do
    Sample $\tau_j$
        $\tau_j^{-2} | \boldsymbol{\beta}, \eta \sim \text{Gamma}\left(a_1 + 1/2, a_2 + \beta_j^2 / 2\eta_j\right)$,
end for
for j in 1:p do
    Sample $\eta_j$
        $\eta_j | \beta, \tau, \omega \sim \frac{w_{1,j}}{w_{1,j} + w_{2,j}} \delta_{v_0}(\cdot) + \frac{w_{2,j}}{w_{1,j} + w_{2,j}} \delta_1(\cdot)$ ,
    where $w_{1,j} = (1 - \omega) v_0^{-1/2} \exp(-\beta_j^2 / 2 v_0 \tau_k^2)$ and
    $w_{2,j} = \omega \exp(-\beta_j^2 / 2\tau_j^2)$.
end for
Sample $\omega$
    $\omega | \eta, \tau \sim \text{Beta}(1 + |\{j \mid \eta_j = 1\}|, \, 1 + |\{j \mid \eta_j = v_0\}|)$
```

Clustering the variance parameters means that we can account for outlier observations if one of the variances $\sigma_1^*, \dots, \sigma_{K_n}^*$ is very large. Outlier observations belonging to that specific cluster will have much higher variance than the others and can therefore take more extreme values. In this way, the model can both simultaneously account for heteroskedasticity and outliers which are characterised by clusters with extreme variances.

Algorithm 3 Sample β with the horseshoe model

```
Sample $\boldsymbol{\beta}$
    $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\beta|\cdot}, \Sigma_{\beta|\cdot})$
where $\boldsymbol{\mu}_{\beta|\cdot} = (X^T \Sigma^{-1} X + \Lambda^{-1})^{-1} X^T \Sigma^{-1} \boldsymbol{y}$ and
$\Sigma^{-1}_{\beta|\cdot} = X^T \Sigma^{-1} X + \Lambda^{-1}$, with $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$
and $\Lambda = \text{diag}(\tau^2 \lambda_1^2, \dots, \tau^2 \lambda_p^2)$,
for j in 1:p do
    Sample the local shrinkage parameter $\lambda_j^2$ and auxiliary variable $\nu_j$,
        $\lambda_j^2 | \cdot \sim \text{IG}(1, 1/\nu_j + \beta_j^2 / 2\tau^2)$,
        $\nu_j | \cdot \sim \text{IG}(1, 1 + 1/\lambda_j^2)$;
end for
Sample the global shrinkage parameter $\tau^2$ the auxiliary variable $\xi$,
    $\tau^2 \sim \text{IG}\left( (p+1)/2, 1/\xi + \sum_{j=1}^{p} \beta_j^2 / 2\lambda_j^2 \right)$,
    $\xi \sim \text{IG}(1, 1 + 1/\tau^2)$;
```

3.3.2 Heavy Tails: Student-t extension

In many real-world datasets, the response variable $\mathbf{y}$ may contain some larger than expected values which are commonly referred to as *outliers*. A common approach to model such data is to replace the normal distribution assumption for the conditional distribution of y_i given $\boldsymbol{x}_i$ with a Student-t distribution. Compared to a normal distribution, the Student-t distribution has heavier tails and therefore permits outlier observations with a higher probability than under the normal model. We can account for outliers in our Dirichlet process variable selection model (5) by replacing the conditional distribution of y_i with

$$y_i | \boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\sigma^2} \sim \text{Student-t}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma_i^2, \nu) \quad i = 1, \dots, n,$$

for some degrees of freedom ν. The parameter ν regulates the thickness of the tails of the distribution. A small value for ν leads to thicker tails with the expectation that large or extreme values for y_i will be observed. As $\nu \to \infty$, the Student-t distribution recovers the normal distribution.

A Gibbs sampler to perform posterior inference for the Student-t model can be derived from Algorithm 1 by including an additional data augmentation step. The required conditional distributions are derived by representing the Student-t distribution as a normal distribution with an inverse gamma variance. The remainder of this section describes these procedures.

We assume our response variables are distributed as $Y \sim \text{Student-t}(\mu, \sigma^2, \nu)$ with density,

$$f(y|\mu, \sigma^2, \nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu\sigma^2}} \left(1 + \frac{(y-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \quad (7)$$

with mean $\mathbb{E}(Y) = \mu$ and variance $\text{Var}(Y) = \sigma^2 \frac{\nu}{\nu-2}$. A well-known reparameterization of the model follows that if,

$$Y|G \sim \mathcal{N}(\mu, G^{-1}) \quad \text{and} \quad G \sim \text{Gamma}(\nu/2, \nu\sigma^2/2),$$

then the marginal of Y (when integrating out G) is $Y \sim \text{Student-t}(\mu, \sigma^2, \nu)$.

The extension of the DPSS and DPHS models to a Student-t model is as follows

$$\begin{aligned} y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \sigma_i^2 &\sim \text{Student-t}(\boldsymbol{x}_i^\top \boldsymbol{\beta}, \sigma_i^2, \nu) \\ \boldsymbol{\beta} &\sim \pi_{\text{SS/HS}} \\ \sigma_i^2|P &\sim P \\ P &\sim \text{DP}(\alpha, \text{Gamma}(b_1, b_2)). \end{aligned} \quad (8)$$

Using the reparameterized Student-t model, and integrating out P from (8), the joint posterior distribution $\pi(\beta, G_{1:n}, \sigma^2_{1:n}, \alpha|Y_{1:n}, X_{1:n})$ is proportional to,

$$\prod_{i=1}^{n} \mathcal{N}(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}, G_i^{-1}) \text{Gamma}(G_i; \nu/2, \nu\sigma_i^2/2) \\ \times \pi(\sigma_{1:n}|\alpha)\pi(\alpha)\pi_{SS/HS}(\beta),$$

where as before, $\pi(\sigma_{1:n}|\alpha)$ denotes the marginal likelihood of $\sigma_{1:n}$ when P is integrated out, which can be represented using the Chinese restaurant process. The proposed Gibbs sampler is specified in Algorithm 4, where at each iteration, we first resample a class vector $c_{1:n}$ assigning each σ_i^2 and G_i to a block in the partition, and then resample the distinct values $\sigma^{2*}_{1:k}$ followed by each latent variable G_i. In the heavy tailed case the observation is related to a variance through its latent variable and hence the posterior probability of

Algorithm 4 DP variable selection Gibbs sampler, Student-t extension

```
for $t$ in 1: number of iterations do
    for i in 1:n do
        Sample classification vector $c_i$ with probability (9).
    end for
    for k in 1:$K_n$ do
        Sample $\sigma_k^{2*}$
            $$\sigma_k^{2*} \sim \text{Ga}\left(\frac{\nu}{2}n_k + b_1, \frac{\nu}{2}\sum_{i:C_k} G_i + b_2\right),$$
        where $C_k = \{G_i \mid c_i = k\}$.
    end for
    if Spike-and-Slab then
        Sample $\boldsymbol{\beta}$, $\tau_j$, $\eta_j$ and $\omega$ using Algorithm 2, replacing $\Sigma = \text{diag}(\sigma^2_{1:n})$ with $\Sigma = \text{diag}(G_{1:n})^{-1}$.
    else if Horseshoe then
        Sample $\boldsymbol{\beta}$, $\lambda_j$ and $\tau$ using Algorithm 3, replacing $\Sigma = \text{diag}(\sigma^2_{1:n})$ with $\Sigma = \text{diag}(G_{1:n})^{-1}$.
    end if
    for i in 1:$n$ do
        Sample $G_i$
            $$G_i \sim \text{Ga}\left(\frac{\nu+1}{2}, \frac{1}{2}[(Y_i - X_i^T\beta)^2 + \nu\sigma_i^2]\right).$$
    end for
    Sample $\alpha$, by sampling the latent variable
        $\psi|\alpha, K_n \sim \text{Beta}(\alpha, n)$  then
        $\alpha|\psi, K_n \sim \text{Gamma}(d_1 + K_n, d_2 - \log\psi)$
end for
```

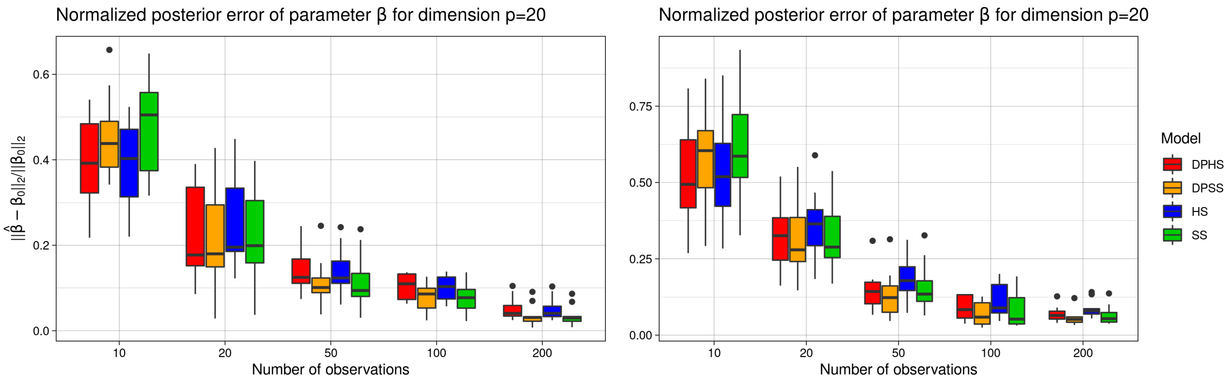

Figure 1: Estimation error $\|\hat{\beta} - \beta_0\|_2/\|\beta_0\|_2$ as a function of n for fixed $p = 20$ with homoskedastic (left) and heteroskedastic errors (right).

the classification vector is

$$\begin{aligned} &\mathbb{P}[c_i = c \mid c_{-i}, G_{1:n}, \sigma^{2*}_{1:K_n}] \propto \\ &\quad \propto \begin{cases} n_{-i,c}\mathrm{Ga}(G_i; \nu/2, \nu\sigma_c^{2*}/2) & \text{for } 1 \leq c \leq K^- \\ \alpha g(G_i; \nu, b_1, b_2) & \text{for } c = K^- + 1, \end{cases} \end{aligned} \tag{9}$$

where K^- be the number of distinct c_j for $j \neq i$, labeled $\{1, \ldots, K^-\}$, and $n_{-i,c}$ is the number of $c_j = c$ for $j \neq i$. Also, the likelihood of observation i being assigned a new variance is modified to

$$\begin{aligned} g(G_i; \nu, b_1, b_2) &:= \int \mathrm{Ga}(G_i; \nu/2, \nu\sigma^2/2)\mathrm{Ga}(\sigma^2; b_1, b_2)d\sigma^2 \\ &= \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} G_i^{\frac{\nu}{2}-1} \frac{b_2^{b_1}}{\Gamma(b_1)} \frac{\Gamma(b_1 + \frac{\nu}{2})}{(b_2 + \nu G_i/2)^{b_1 + \frac{\nu}{2}}}. \end{aligned}$$

Finally, regression coefficients $\boldsymbol{\beta}$ are sampled using Algorithms 2 and 3 with small modifications to include the augmented variable G_i.

4 Synthetic Data Experiments

In this section we compare the DPSS and DPHS models against the standard spike-and-slab and horseshoe models of Ishwaran and Rao (2005) and Carvalho et al. (2010). We begin by studying their predictive performance, their ability to model heterogeneous data and their variable selection capabilities over a range of scenarios utilizing synthetic data. In Section 5 we test the models' performance on real-world datasets. Firstly to assess the likelihood of predictions when modelling various open-source datasets and secondly on the reconstruction of a network of genomic data through variable selection on a linear model. Code for sythentic and real experiments can be found in `https://github.com/albcab/RobustVariableSelection`.

For all experiments, each algorithm is run for $J = 10,000$ iterations with a burn-in period of $J/2$. Convergence and mixing of the Markov chain is confirmed through visual diagnostics. Hyper-parameters for the models are set to $a_1 = b_1 = 2.01$, $a_2 = b_2 = d_1 = 1$ and $d_2 = 1/2$, which leads to weakly informative priors. This claim is verified with simple cross validation for all the datasets. It is worth noting that one could place additional hyper-priors on these parameters and add extra steps to the MCMC algorithm. However, in doing so we would run the risk of rejecting steps potentially leading to longer mixing times. It is worth noting that in comparable studies (e.g. Ishwaran and Rao, 2005), it is common to simply fix these parameters when assessing performance. Finally, both the dependent and independent variables are normalized to have zero mean for the testing of each model.

Performance is evaluated in the case of synthetic data over a range of regression scenarios varying the number of attributes $p = 20, 50, 100, 200$ and data lengths $n = 10, 20, 50, 100, 200$. The sparsity pattern will be kept similar across experiments according to the Briemann structure (Breiman et al., 2001). Specifically, we consider two scenarios for Gaussian linear regression: Scenario (1) - homoskedastic errors with $\epsilon \sim \mathcal{N}(0, 1)$ as the single component; Scenario (2) - heteroskedastic errors with five components $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ for $\sigma_i^2 \in \{0.5, 1, 1.5, 2, 2.5\}$ distributed evenly among the observations as well as $1, 2$ and 4 outliers $\epsilon \sim \mathcal{N}(0, 10)$ in the case of $n = 50, 100, 200$, respectively.

In each scenario and for each combination of attributes and data lengths, 10 synthetic datasets are created and parameters are estimated for each of these datasets. The results presented in the remainder of this section consider all 10 estimates for robustness.

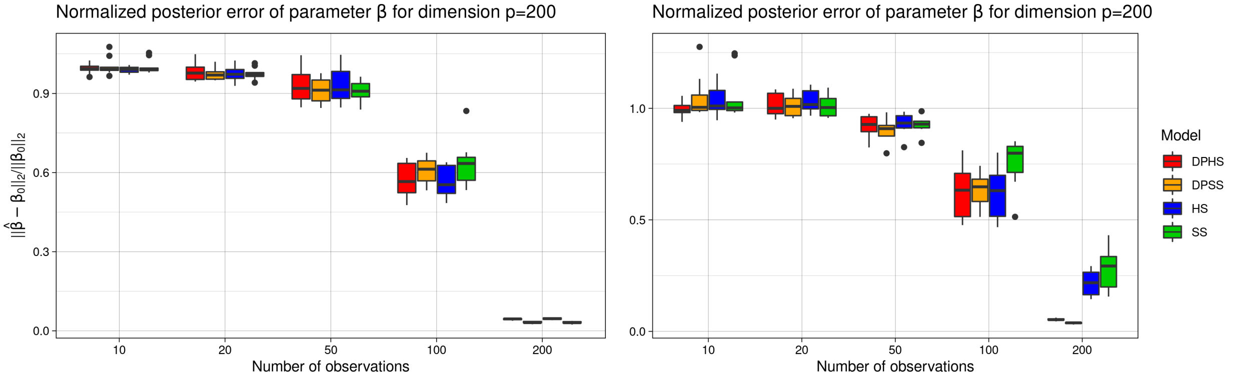

Figure 2: Estimation error $\|\hat{\beta} - \beta_0\|_2/\|\beta_0\|_2$ as a function of n for fixed $p = 200$ with homoskedastic (left) and heteroskedastic errors (right).

4.1 Homoskedastic and Heteroskedastic Errors

We compare the Dirichlet process variable selection models against the standard spike-and-slab and horseshoe models in the presence of homoskedastic (*Scenario 1*) and heteroskedastic (*Scenario 2*) noise. Figures 1 and 2 present the posterior error $\|\hat{\beta} - \beta_0\|_2/\|\beta_0\|_2$ for the low $p = 20$ and high $p = 200$ settings, respectively, as the number of observations n increases, with *Scenario 1* plotted on the left and *Scenario 2* given on the right.

Robustness - for *Scenario 2*, the robustness provided by the DPSS and DPHS models improves the predictive estimates both in the low and high dimensional settings, with improvements more noticeable in higher dimensions with sufficient data, or in lower dimensions when data is abundant. Under *Scenario 1*, as expected, the results are similar for the models allowing heterogeneous and homogeneous variances.

Coefficients - we also consider the setting of fixing the number of observations and varying the number of coefficients $\boldsymbol{\beta}$ under *Scenario 2*. Figure 3 presents the posterior errors for $n = 100$ against an increasing number of model coefficients. We see again the Dirichlet process model performs well in the presence of heterogeneity, particularly as the number of model coefficients increases. It is interesting to notice that modelling the variances nonparametrically significantly corrects the error in prediction of the horseshoe prior when observations are scarce.

Clusters - One of the key advantages of the nonparametric approach we adopt is the availability of posterior densities for the number of variance components K. This allows us to assess the uncertainty in the number of clusters identified in the model. For instance, Figure 4 plots histograms for the parameter K with $p = 50$ and $n = 50$. Heavy tails demonstrate the Dirichlet process' assumption of a number of clusters which scale logarithmically towards infinity as n grows to infinity. In the case where a finite number of clusters in the data is known, or the number of clusters K should scale faster than logarithmically, priors that generalize the Dirichlet process can be used, e.g. the Pitman-Yor process (Pitman and Yor, 1997).

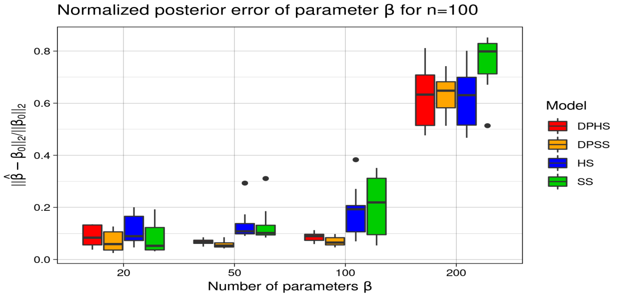

Figure 3: Normalised posterior error $\|\hat{\beta} - \beta_0\|_2/\|\beta_0\|_2$ as a function of dimension p for fixed $n = 100$.

4.2 Support Recovery

In addition to estimating the regression coefficients, a further task of interest, especially in high-dimensional regression, is to select a subset of attributes which are deemed significant for the predictive model. In a frequentist context, this is usually achieved via forward selection with sequential F-tests, or with ℓ_1 or ℓ_0 shrinkage and/or selection, for instance through the use of the AIC or BIC criteria

One advantage of the Bayesian approach is that we can sample from the joint posterior of the coefficients, thus we can construct credible intervals with relative ease. In this section, we compare two methods for estimating the support of the regression model:

Model	n	10		20		50		100		200	
SS	TP	0	[0,1]	**3**	[1.45,5.55]	**18**	[12.7,22.55]	25	[22,28]	26	[23,28]
	FP	0	[0,1]	4	[1,6.55]	20	[7,48.6]	4	[0,18.05]	1	[0,4.2]
DPSS	TP	0	[0,0.55]	1	[0,2.55]	9	[0.45,15.55]	**27**	[24.8,28]	**28**	[28,28]
	FP	0	[0,1]	**0**	[0,1.55]	**3**	[0,13.25]	**0**	[0,0]	**0**	[0,0.559]
HS	TP	0	[0,0]	0	[0,1.55]	8	[3.45,12]	25	[21.45,28]	26	[22.35,28]
	FP	0	[0,0.55]	0	[0,0]	0	[0,2.1]	2	[0.45,3.65]	2	[0.45,3]
DPHS	TP	0	[0,0]	0	[0,1.55]	8	[3.45,12]	**28**	[27,28]	**28**	[28,28]
	FP	0	[0,0.55]	0	[0,0]	0	[0,2]	**1**	[0,2]	**2**	[0,3]

Table 1: Tabulated results for FP and TP as a function of n ($p = 100$ with 28 positives) with 95% credible intervals.

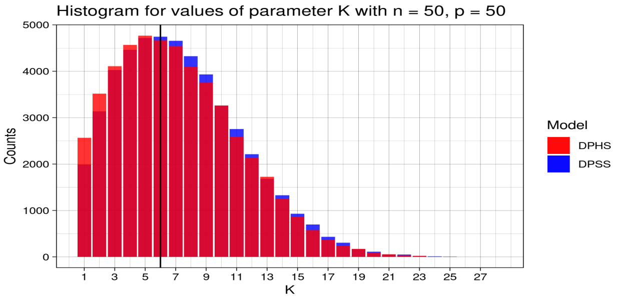

Figure 4: Histogram of the derived number of clusters K from the posterior of $\sigma_1, \ldots, \sigma_n$ for $p = 50$ and $n = 50$ for DPSS and DPHS, where the true number of clusters is six.

- The first approach is to simply look at the posterior of the inclusion parameter η_j for selecting either the spike, or slab. Specifically, if the mode of the posterior $\hat{\eta}_j$ is above level $1 - \zeta$ we add the index j to the estimated support set $\hat{\mathcal{M}}$. This method works only in the DPSS and SS models.
- The second approach is to look at the empirical posterior credible interval of $\hat{\beta}_i$ at the percentiles $\zeta/2 \times 100$ and $(1 - \zeta/2) \times 100$. If the constructed interval excludes zero, then we add the index to the estimated support set $\hat{\mathcal{M}}$. This method is used for the DPHS and HS models.

We note that the second approach is somewhat similar to the *z-cut* method discussed in Ishwaran and Rao (2005), however, they construct a set $\hat{\mathcal{M}} := \{i \mid \forall i \ s.t. \ |\bar{\beta}_i| \geq z_{(1-\zeta/2)}\}$ where $\bar{\beta}_i$ represents the posterior mean of the regression coefficients.

To summarise the performance of the two approaches, we take the number of true and false positive (TP, FP) coefficients included in the model and consider how this changes as a function of n. Clearly, these values will change as a function of ζ, and in practice this may be tuned to favour either TP or FP results. For simplicity, we report results with $\zeta = 0.05$. Table 1 summarises the performance of the posterior inclusion thresholding methods for different values of n in the heteroskedastic scenario. The benefit of a robust Dirichlet process model is highlighted through its reduced type II error. It is interesting to note that the results are more conservative for the horseshoe prior when $n < p$, while more precise for the spike-and-slab prior when $n > p$; even though results in the synthetic data seem to favour the spike-and-slab model for variable selection, real world results with scarce observations in the following section suggest otherwise.

5 Real Data Experiments

5.1 Prediction accuracy

We test the predictive accuracy of the robust models on various real world datasets appropriate for variable selection. Specifically, data on air pollution and mortality (McDonald and Schwing, 1973), physiological variables and diabetes (Efron et al., 2004), baseball salaries [1], and critical temperature of superconductors [2] (Hamidieh, 2018). The *baseball* and *critical temperature* datasets are log-transformed to allow negative values. Figure 5 plots the histograms for the real world datasets. Notice the presence of outliers with low temperatures in *critical temperature* and heavy tails in *pollution* and *diabetes*.

Table 2 presents log-predictive densities on the test dataset using 10-fold cross validation for each dataset. The robust models show significant improvements in predictive performance when observations are very heterogeneous; for example, in the case of a random sample of the US population in *pollution* vs. a sample of patients all with diabetes in *diabetes*. DPSS and DPHS perform best when data are sparse with various outliers, as in the *critical temperature* data where low temperatures are scarce but substantial enough to

[1] https://www4.stat.ncsu.edu/ boos/var.select/

[2] https://archive.ics.uci.edu/ml/datasets.php

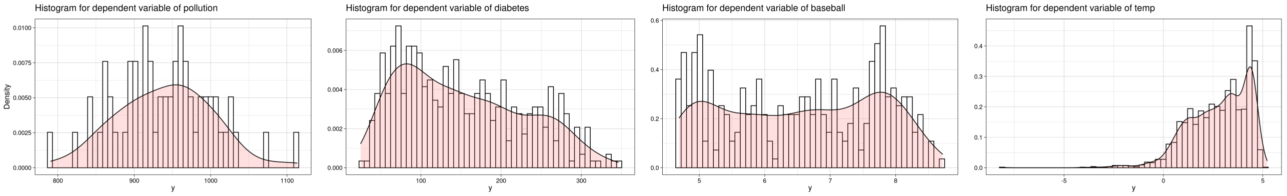

Figure 5: Dependent variables' histograms for real world datasets.

show improvements in predictive accuracy for the robust models.

The model's accuracy is also tested under the assumption that the observations are heavy tailed. The results for the *pollution* and *diabetes* datasets are presented in the bottom half of Table 2. All the models are run with fixed degrees of freedom $\nu = 2$. Relaxing the normal assumption on the observations in addition to modelling robust variances provides significant improvements on the predictive results of the model. The data augmentation trick also allows for Gibbs sampling in the case of heavy tails and the computational cost of sampling is essentially the same as under the normal assumption. This provides a very robust model for variable selection both under the spike-and-slab and horseshoe prior assumptions on the parameters.

Furthermore, we inspect the variable selection behaviour both for the models under the normal and heavy tailed assumptions. Specifically, Figure 6 plots the mean values of each model's parameters, indicating the level of inclusion of each regressor, for *diabetes*. Even though the model's parameters behave similarly under both assumptions, under the heavy tailed likelihood the non-zero values grow in magnitude and there are clearer differences between the models with homogeneous and heterogeneous variances. Under further investigation we notice that the number of components estimated under the normal assumption for both the horseshoe and spike-and-slab models is around 6 while under the t-student likelihood the number of components centers around 2. This might indicate a misspecification of the likelihood under the normal assumption that the heterogeneous models try to make up for by increasing the number of unique variances estimated. When allowing for heavy tails in the likelihood the number of components decreases to its real value, allowing for a better approximation of the mixture generating the data.

5.2 Reconstruction of transcription regulatory networks

We consider the problem of reconstructing genetic regulatory networks from gene expression data (Marbach et al., 2010). This problem can be modelled as a directed network, where each node corresponds to a different gene and each connection represents a directed interaction between two genes at the transcription level. The sparse spike-and-slab and horseshoe priors provide an attractive approach to reconstructing these networks using support recovery methods.

The models are tested on data from the challenge posed at the *The Dialogue for Reverse Engineering Assessments and Methods* (DREAM) 2009 conference, specifically the multifactorial subchallenge[3]. The challenge consists of reverse engineering five independent networks using 100 steady-state measurements for each network with 100 genes. The levels of expression of all the genes are measured under different perturbed conditions. Each gene X_i for $i = 1, ..., 100$ is treated independently and modelled through a linear relationship with the rest of the genes in the network $\boldsymbol{X}$, i.e. $X_i = \boldsymbol{X}^T \boldsymbol{\beta}_i + \epsilon$ with every $\boldsymbol{\beta}_i$ having independent spike-and-slab or horseshoe priors and assuming ϵ is a normal random variable with mean zero and homogeneous variances in the case of SS and HS or heterogeneous variances in the case of DPSS and DPHS. For each gene, the support is recovered using the method described in Section 4.2 for the case of a spike-and-slab prior with $\boldsymbol{\beta}_i$. In the case of a horseshoe prior, we follow Steinke et al. (2007) and approximate the posterior probability of a connection from gene j to gene i by the probability of the event $|(\boldsymbol{\beta}_i)_j| > 0.1$ under the posterior for $\boldsymbol{\beta}_i$.

The performance of the different approaches is evaluated using the mean of the logarithmic loss of the probabilities of connections p_{ij} from gene j to each gene i, defined as

$$\frac{1}{100} \sum_{i=1}^{100} \left\{ -\frac{1}{99} \sum_{j \neq i} [y_{ij} \log p_{ij} + (1 - y_{ij}) \log (1 - p_{ij})] \right\},$$

where $y_{ij} = 1$ if there is a directed connection from j to i and $y_{ij} = 0$ otherwise. Table 3 presents these results for each tested model. The improvement in prediction given by heterogeneous variances in the model is substantial. This supports the results from Section 4.2 on

[3] www.synapse.org/#!Synapse:syn3049712/wiki/74628

	SS	95% CI	DPSS	95% CI	HS	95% CI	DPHS	95% CI	train/test
Pollution	-26.4734	[-27.7,-24.72]	**-26.0378**	[-27.53,-24.24]	-24.9396	[-28.54,-22.78]	**-24.5585**	[-27.18,-21.69]	55/5
Diabetes	**-230.7864**	[-236.62,-225.07]	-232.3278	[-239.1,-226.42]	**-229.9841**	[-238.17,-222.8]	-231.4566	[-240.63,-223.28]	400/42
Baseball	-30.7400	[-35.06,-24.21]	**-30.2592**	[-33.58,-26.24]	-30.8921	[-35.81,-24.04]	**-30.2742**	[-33.78,-26.25]	300/37
Temperature	-243.2871	[-271.72,-219.62]	**-228.8550**	[-242.35,-209.99]	-243.1998	[-272.91,-219.51]	**-228.5229**	[-241.43,-209.6]	1800/200
Pollution	-34.5245	[-42.32,-24.35]	**-19.9301**	[-26.68,-12.99]	-24.3429	[-33.24,-16.02]	**-19.3170**	[-25.91,-10.29]	55/5
Diabetes	-362.7476	[-384,-344.69]	**-148.4859**	[-163.74,-131.94]	-343.5113	[-369.7,-322.28]	**-149.0032**	[-166.24,-131.76]	400/42

Table 2: Log predictive densities for 10-fold cross validation on various real datasets with 95% Credible Intervals.

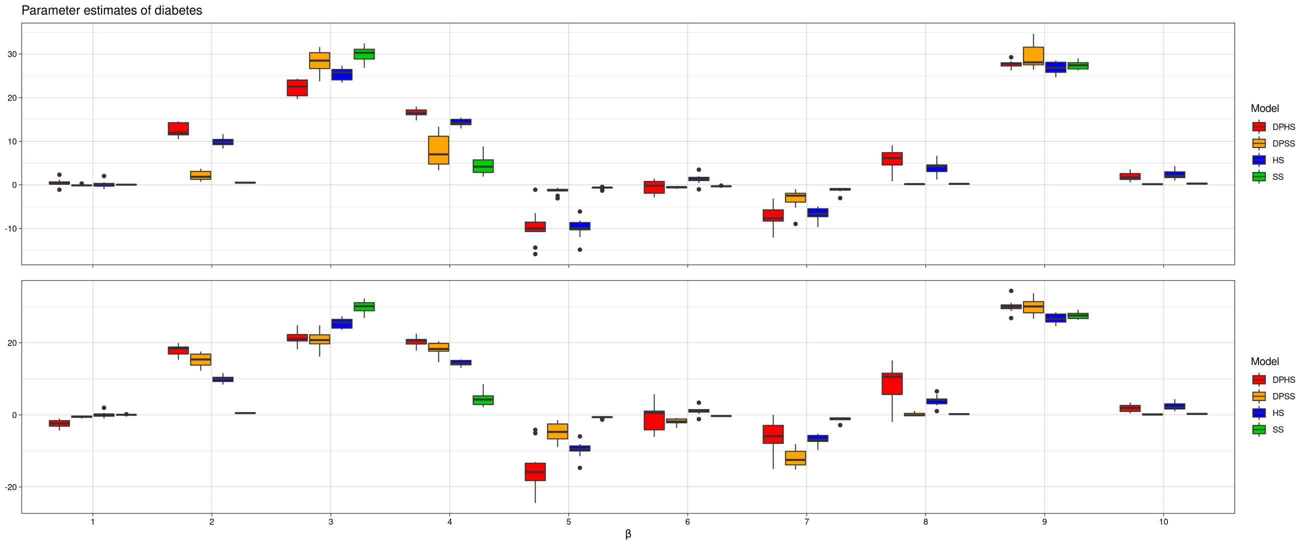

Figure 6: Estimated parameters for real world dataset diabetes for normal likelihood (top) and Student-t likelihood (bottom) models.

synthetic data, illustrating the improvement offered by robust models to correctly classify variables included in the model, specifically in their capability of reducing type II error in the predictions while correctly identifying the significant coefficients in the linear relationship. It is interesting to note that, while the DPSS model provided better support recovery results in the synthetic data example, the DPHS provides better results in the real data example of gene network reconstruction.

	SS	DPSS	HS	DPHS
N1	0.1143	**0.0928**	0.0920	**0.0873**
N2	0.1498	**0.1382**	0.1248	**0.1202**
N3	0.1610	**0.1185**	0.1196	**0.0914**
N4	0.1588	**0.1152**	0.1124	**0.0964**
N5	0.1314	**0.1079**	0.1044	**0.0939**

Table 3: Logarithmic-loss errors for the five DREAM Gene Network detection datasets (N1-N5).

6 Conclusion

This paper outlines a nonparametric extension of popular Bayesian variable selection models to account for heterogeneity, outliers and clustering effects. We also provide an extension to the model to allow for heavy-tailed data distributions. Fitting our Dirichlet process variable selection models is computationally efficient using a Gibbs sampling construction. The results presented for both synthetic and real data examples show a robust improvement in both predictive accuracy on test data and improved efficiency in identifying key model attributes.

This work could be further extended to encompass other variable selection models, including extensions of existing models, such as the regularized horseshoe model (Piironen et al., 2017). It would also be interesting to explore nonparametric priors that allow for more flexible assumptions on the expected clustering effect, such as the Pitman-Yor process (Pitman and Yor, 1997).

Bibliography

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *Annals of statistics*, 32(2):407–499.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.

Hamidieh, K. (2018). A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354.

Hjort, N. J., Holmes, C., Muller, P., and Walker, S. G. (2010). *Bayesian Nonparametrics*. Cambridge University Press.

Ishwaran, H. and Rao, J. S. (2005). Spike and Slab Variable Selection: Frequentist and Bayesian Strategies. *The Annals of Statistics*, 33(2):730–773.

Lo, A. Y. (1984). On a class of bayesian nonparametric estimates: I. density estimates. *The annals of statistics*, pages 351–357.

Makalic, E. and Schmidt, D. F. (2015a). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Makalic, E. and Schmidt, D. F. (2015b). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010). Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107(14):6286–6291.

McDonald, G. C. and Schwing, R. C. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, 15(3):463–481.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Piironen, J., Vehtari, A., et al. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051.

Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, pages 855–900.

Steinke, F., Seeger, M., and Tsuda, K. (2007). Experimental design for efficient identification of gene regulatory networks using sparse bayesian models. *BMC systems biology*, 1(1):1–15.