entropy

MDPI

Article

A Dirichlet Process Prior Approach for Covariate Selection

Stefano Cabras †

Department of Statistics, Universidad Carlos III de Madrid, 28903 Madrid, Spain; stefano.cabras@uc3m.es
† Current address: Department of Statistics, Universidad Carlos III de Madrid C/Madrid, 126-28903 Getafe, Spain.

Received: 21 July 2020; Accepted: 26 August 2020; Published: 28 August 2020

Abstract: The variable selection problem in general, and specifically for the ordinary linear regression model, is considered in the setup in which the number of covariates is large enough to prevent the exploration of all possible models. In this context, Gibbs-sampling is needed to perform stochastic model exploration to estimate, for instance, the model inclusion probability. We show that under a Bayesian non-parametric prior model for analyzing Gibbs-sampling output, the usual empirical estimator is just the asymptotic version of the expected posterior inclusion probability given the simulation output from Gibbs-sampling. Other posterior conditional estimators of inclusion probabilities can also be considered as related to the latent probabilities distributions on the model space which can be sampled given the observed Gibbs-sampling output. This paper will also compare, in this large model space setup the conventional prior approach against the non-local prior approach used to define the Bayes Factors for model selection. The approach is exposed along with simulation samples and also an application of modeling the Travel and Tourism factors all over the world.

Keywords: covariate inclusion probability; conventional priors; Dirichlet process prior; non-local prior; ordinary linear regression; variable selection

1. Introduction

The variable selection problem in regression analysis consists of finding a suitable set of predictors for the response variable y from the columns p of a fixed design matrix $X_{n\times p}$ with n rows (observations). In a Bayesian setting we have to evaluate the posterior probability $\pi(\mathcal{M}_\gamma|y)$ of each regression model $\mathcal{M}_\gamma$, where with the usual notation $\gamma = (\gamma_1, \ldots, \gamma_p)$, with $\gamma_j = 1$ if column X_j, $j = 1, \ldots, p$ is in model $\mathcal{M}_\gamma$. The aim of this paper is to estimate $\pi(\mathcal{M}_\gamma|y)$ for all $\gamma \in \Gamma$. The main point is that the cardinality of Γ is of order 2^p, huge even for moderate values of $p \approx 1000$ even if it were $p < n$. In what follows, whatever p is, we only evaluate the evidence for models in which the size is no larger than n including the intercept and the regression the error variance.

Each $\pi(\mathcal{M}_\gamma|y)$ can be calculated exactly (not just estimated) by means of all 2^p Bayes Factors (BF) $BF_{\gamma 0}$, $\pi(\mathcal{M}_\gamma|y) = \pi(\mathcal{M}_\gamma)BF_{\gamma 0}(1 + \sum_{\forall \gamma \neq 0} BF_{\gamma 0})^{-1}$, where the marginal distribution of model $\mathcal{M}_\gamma$ is compared against that of a common null nested model $\mathcal{M}_0$. The reference null model considered here is the regression model with the intercept only. To obtain $\pi(\mathcal{M}_\gamma|y)$, a prior on model space Γ, $\pi(\mathcal{M}_\gamma)$ must be chosen.

Different priors on Γ have been proposed, starting from the discrete uniform $\pi(\mathcal{M}_\gamma) = 2^{-p}$ to fulfill the insufficient reasonable principle requirements, up to the hierarchical uniform prior [1] $\pi(\mathcal{M}_\gamma) \propto \binom{p}{p_\gamma}^{-1}$ in which p_γ indicates the size of model γ. Under the hierarchical uniform, prior models of the same sizes receive the same prior probability and this also controls for the false discovery rate in declaring a covariate important when it is not important. This is relevant for large p and when the model is sparse [1].

More than focusing on comparing different definitions of $\pi(\mathcal{M}_\gamma)$, we consider different definitions of $BF_{\gamma 0}$ given by the prior probability of model parameters. In particular, we center on the semi-conjugate multivariate normal on regression coefficients, which leads to closed-form expression of $BF_{\gamma 0}$, which is important for computational feasibility. For instance, the conventional prior approach in [2] has suitable properties (or *desiderata*). An important one is the so-called *predictive matching property*, which assures that $BF_{\gamma 0} = 1$ if there is not enough information in the data to distinguish between $\mathcal{M}_\gamma$ and $\mathcal{M}_0$. This is obtained by using the concept of effective sample size, which is the number used to rescale the prior covariance matrix to obtain unit information prior, which is a prior that bears the same information as one sample. The other approach is that of the non-local priors detailed in [3]. These priors are non-local in the sense that the minimum prior density is at the null hypothesis in contrast to other usual approaches. The effect is that such a set of priors, asymptotically in n, has a larger increment in recognizing the true set of important covariates (the so-called learning rate) more than the conventional prior approach, although they do not meet the predictive matching desiderata. In the sequel, we refer to these two sets of priors as conventional and non-local prior approaches and these are the only two definitions of BFs considered in this work.

Beyond choices of model parameters prior and/or model prior, the point of this work is that even under closed-form expression of BFs an exhaustive exploration of Γ is not feasible and thus a stochastic model search must be employed to obtain an estimation of $\pi(\mathcal{M}_\gamma|y)$. By estimating $\pi(\mathcal{M}_\gamma|y)$ and intending to interpret the nature of the true underlying model, a summary of $\pi(\mathcal{M}_\gamma|y)$ which is of particular interest here is the inclusion probability of a covariate j, denoted by τ and defined by marginalizing $\pi(\mathcal{M}_\gamma|y)$ overall γ such that $\gamma_j = 1$. τ is necessary to define the *median probability model* [4] defined as the model in which covariates have $\tau > 0.5$. If such a model exists, it is proved to be very near to the true model even under strong collinearity [5].

When Γ cannot be fully explored, a stochastic model search is employed in place of heuristic algorithms as they allow to approximate $\pi(\mathcal{M}_\gamma|y)$. In the case of the normal linear regression model, a popular stochastic approach is that of the well known Gibbs-sampling in [6]. Such an algorithm implements a Markov chain by jumping from one model to another based on the marginal distribution of the data (the same that appears in the Bayes Factors). This algorithm returns a dependent sequence of $\mathcal{M}^{(S)} = \left(\gamma^{(1)} \in \Gamma, \ldots, \gamma^{(S)} \in \Gamma\right)$ assumed to be a sample of size S from $\pi(\mathcal{M}_\gamma|y)$. The Gibbs-sampling operates on the space Γ using the fact that a closed-form expression of the marginal distributions is available for the normal regression model with the above-mentioned set of conjugate priors (conventional and non-local). The main advantage of Gibbs-sampling over heuristic approaches is that the theory assures that when the number of steps $S \to \infty$ all Γ has been properly explored, whereas heuristic methods (like step-wise methods) may stop at local maxima and may not bear properly posterior model uncertainty.

The aim of this paper is to use $\mathcal{M}^{(S)}$ to obtain an estimation of τ, namely $\widehat{\tau}$, and in particular the new estimators $\widehat{\tau}_{\mathrm{b}}$ and $\widehat{\tau}_f$, later defined. A different definition of $\widehat{\tau}$ derived from the Gibbs-sampling algorithm has been studied in [7] applying the known concept of sampling theory. In particular, two estimators have been fully characterized:

(*i*) The intuitive and popular *empirical* proportion of the sampled models in $\mathcal{M}^{(S)}$ containing covariate j, $\widehat{\tau}_{\mathrm{e}} = S^{-1} \sum_{\gamma \in \mathcal{M}^{(S)}}^{S} \mathbf{1}(\gamma)_{\gamma_j=1}$, where $\mathbf{1}(\gamma)_{\gamma_j=1}$ is the usual indicator function for the scalar $\gamma_j = 1$ in the vector γ. This estimator is called the empirical estimator;

(*ii*) The *renormalized* proportion of the sampled models containing covariate j, $\widehat{\tau}_{\mathrm{r}} = \frac{\sum_{\gamma \in \mathcal{M}^{(S)}} \mathbf{1}(\gamma)_{\gamma_j=1} \pi(\mathcal{M}_\gamma) BF_{\gamma^{(s)}0}}{\sum_{\gamma \in \mathcal{M}^{(S)}} \pi(\mathcal{M}_\gamma) BF_{\gamma^{(s)}0}}$, called the renormalized estimator.

According to [7] both are consistent estimators of τ for $S \to \infty$ and that the error of $\widehat{\tau}_{\mathrm{r}}$ is the sum of two components: the error of $\widehat{\tau}_{\mathrm{e}}$ plus (or minus) a term that depends on the correlation between posterior probability of a $\mathcal{M}_\gamma$ and its probability to be a visited model for the Gibbs-sampling algorithm. Basically, if the Gibbs-sampling moves very little around a model $\mathcal{M}_\gamma$ just because it has

the highest posterior probability (but not necessarily the largest one), $\widehat{\tau}_{\mathfrak{r}}$ will be more biased than $\widehat{\tau}_{\mathfrak{e}}$ [7] otherwise it will be more precise than $\widehat{\tau}_{\mathfrak{e}}$. Moreover, if the model is sparse, the visited models, for finite S depends on the initial state of the chain (for instance: the null model $\mathcal{M}_0$ or the full model for $p < n$).

The idea of this paper is to rethink about all these estimators of τ and use a proper Bayesian approach to analyze the sequence $\mathcal{M}^{(S)}$. In practice, $\mathcal{M}^{(S)}$ can be viewed as a sequence of non-ordinal categorical variables with two levels, which gives rise to representing Γ as the set of a *very sparse* contingency table of dimension 2^p cells, which are the probabilities $\pi(\mathcal{M}_\gamma|y)$. Therefore, estimating cell probabilities is equivalent to the estimate of the posterior model probabilities. The main observation here is that this is the same setup as in [8] except for the fact that samples $\mathcal{M}^{(S)}$ are dependent instead of being independent as assumed in [8], although for large S and given that this is a Gibbs-Sampling, dependence becomes mild. Deriving the posterior distribution of cells probability is the same as deriving an estimation of $\pi(\mathcal{M}_\gamma|y)$ and thus the τ. That is, from $\pi(\tau|\mathcal{M}^{(S)})$ we define $\widehat{\tau}_{\mathfrak{b}} = E_{\pi(\tau|\mathcal{M}^{(S)})}\tau$ and this is one of the proposed estimators that we will compare against $\widehat{\tau}_{\mathfrak{e}}$ and $\widehat{\tau}_{\mathfrak{r}}$ allowing also to properly account for the uncertainty around the obtained value of $\widehat{\tau}$.

The following Section 2 will illustrate the Bayesian model for analyzing $\mathcal{M}^{(S)}$ and its approximation that is used to obtain $\widehat{\tau}_{\mathfrak{b}}$ and a $\widehat{\tau}_f$ (specified below). Further, model implementation, simulation study, and real data application on Travel and Tourism data are considered in Section 3. Conclusion and final remarks are left for Section 4.

2. The Dirichlet Process Mixture Model for Estimating Posterior Model Probabilities

The Dirichlet process prior model was initially considered for sparse contingency tables in [8] and here it is applied to the more specific context of covariate selection and consequent estimation of covariate inclusion probabilities.

Parameters of interest are $\mathfrak{ß} = \left\{\pi_{\gamma_1\gamma_2\cdots\gamma_p}, \gamma_j = 0, 1, j = 1, \ldots, p\right\} \in \Pi$ which is the set of all probabilities tensor on the space Γ of size 2^p. These are the joint probabilities of all covariates and thus of all $\mathcal{M}_\gamma$ on space Γ, namely the 2^p cells probabilities, where $\|\mathfrak{ß}\|_1 = \sum_{\gamma_1=0}^{2}\cdots\sum_{\gamma_p=1}^{d_p}\left|\pi_{\gamma_1\gamma_2\cdots\gamma_p}\right| = 1$ and every $0 \leq \pi_{\gamma_1\gamma_2\cdots\gamma_p} \leq 1$.

In the actual setup, these probabilities are attempted to be estimated with the Gibbs-sampling output, $\mathcal{M}^{(S)}$ regarded here as the data used to estimate $\mathfrak{ß}$. Specifically, a sample $\gamma_j^{(s)} \in \mathcal{M}^{(S)}$ is a dichotomous unordered categorical variable and we denote the two categories of $\gamma_j^{(s)}$, by $c_j = 0, 1$. The key idea in the Dirichlet process mixture model is representing probability $\mathfrak{ß}$ by decomposing it as an addictive mixture of k (possibly infinite) sets of probabilities,

$$\mathfrak{ß} = \sum_{h=1}^{k} v_h \Psi_h, \quad \Psi_h = \psi_h^{(1)} \otimes \psi_h^{(2)} \otimes \cdots \otimes \psi_h^{(p)}$$

where $\mathbf{v} = (v_1 \geq \ldots \geq v_k)'$ is the probability vector of the $h = 1, \ldots, k$ sets of distributions $\Psi = (\Psi_1 \ldots \Psi_k)$, with $\Psi_h \in \Pi_{1\ldots p}$, where $\psi_h^{(j)}$ is the probability for covariate j to be included into the set of predictors given the probability distribution over all Γ labeled by h.

It is important to note that $\psi_h^{(j)}$ is the j covariate inclusion probability and its estimator is also an estimator of τ if the distribution h has high posterior probability v_h given $\mathcal{M}^{(S)}$. In particular, if sets of probabilities are ordered according to their posterior probabilities, and $v_1 \approx 1$ then estimation of $\Psi_1^{(j)}$ could be a good candidate for being a conditional (to $h = 1$) estimator of τ.

For this purpose, we define the estimator

$$\widehat{\tau}_f = \Psi_1,$$

understood for all j, given that the first component of the mixture has the highest probability of v_1.

The likelihood of the two parameters $\mathbf{v}$ and Ψ given $\mathcal{M}^{(S)}$ is

$$\Pr\left(\gamma_1^s = c_1, \ldots, \gamma_p^s = c_p | \mathbf{v}, \Psi\right) = \pi_{c_1 \cdots c_p} = \sum_{h=1}^{k} v_h \prod_{j=1}^{p} \psi_{hc_j}^{(j)}.$$

Introducing the latent class indicator $z_s \in \{1, \ldots, k\}$, the conditional probability to a specific set h is $\Pr\left(\gamma_j^s = c_j | z_s = h\right) = \psi_{hc_j}^{(j)}$ and we have that ψs are the inclusion probabilities of covariates conditional to the latent class indicator.

Thus the marginal distribution of τ is obtained by marginalizing over all latent class indicators, namely $\tau = \psi^{(j)} = \sum_{h=1}^{k} \Pr\left(\gamma_j^s = 1 | z_s = h\right) \Pr\left(z_s = h\right)$, is the parameter of interest which leads to the definition of our estimator

$$\widehat{\tau}_{\mathfrak{b}} = E_{\pi(\tau|\mathcal{M}^{(S)})}\tau = E(\psi^{(j)}|\mathcal{M}^{(S)}) = \sum_{h=1}^{k} \Pr\left(\psi_{hc_j}^{(j)}|\mathcal{M}^{(S)}\right) \Pr\left(v_h|\mathcal{M}^{(S)}\right).$$

The larger the k is, the better is the representation of **ß**, thus we allow for $k = \infty$ by using the following non-parametric prior:

$$\begin{aligned} \text{ß} &= \sum_{h=1}^{\infty} v_h \Psi_h, \quad \Psi_h = \text{-}_h^{(1)} \otimes \cdots \otimes \text{-}_h^{(p)} \\ \text{-}_h^{(j)} &\sim P_{0j}, \quad \text{independently for } j = 1, \ldots, p \text{ and } h = 1, \ldots, \infty \\ \mathbf{v} &\sim Q, \end{aligned} \tag{1}$$

where P_{0j} corresponds to a Dirichlet measure and Q to a Dirichlet process.

For the usual stick-breaking stochastic representation of this model, we have

$$\begin{aligned} \gamma_j^s &\sim \text{Multinomial}\left(\{0,1\}, \psi_{z_i}^{(j)}, \ldots, \psi_{z_i d_j}^{(j)}\right) \\ z_i &\sim \sum_{h=1}^{\infty} V_h \prod_{l<h} (1 - V_l)\, \delta_h, \quad V_h \sim \text{beta}(1, \alpha) \\ \psi_h^{(j)} &\sim \text{Dirichlet}\left(a_{j1}, \ldots, a_{jc_j}\right) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha), \end{aligned}$$

where the Multinomial notation is here over-engineered as $\mathcal{M}^{(S)}$ observations are indicator variables. Parameters $a_{j1} = \ldots = a_{jc_j} = 1$ induce non informative *a priori* information about the probabilities of each covariate being in the model. α is the usual concentration parameter on the space of latent classes of model probability distributions. That is for small values of α, the probability of having many classes of different probability distributions decreases. In what follows we will consider either α fixed or under another layer of uncertainty by setting a gamma prior with shape a_α and scale b_α with $a_\alpha + b_\alpha = 1/2$ and $a_\alpha = 1/4$ as in [8].

With this model at hand, we can obtain justification for the popular empirical estimator of τ. That is, $S \to \infty$ $\widehat{\tau}_{\mathfrak{b}} \approx \widehat{\tau}_{\mathfrak{e}}$ as this is a usual regular Bayesian model in which the prior information is washed out by the sample size S. This is an asymptotic, although not Bayesian, justification of using $\widehat{\tau}_{\mathfrak{e}}$ instead of the proposed $\widehat{\tau}_{\mathfrak{b}}$. In this comparison scenario, $\widehat{\tau}_f$ instead plays the role of a more robust estimation as it is based on that distribution which has the largest probability. Therefore, comparing $\widehat{\tau}_{\mathfrak{b}}$ against $\widehat{\tau}_f$ is the same as comparing means over modes except that the underlying randomness is on probability distributions instead of random variables. Finally, comparing $\widehat{\tau}_{\mathfrak{b}}$ and $\widehat{\tau}_f$ against $\widehat{\tau}_{\mathfrak{e}}$ and $\widehat{\tau}_{\mathfrak{r}}$ is the same as comparing sampling-based estimators against Bayesian ones which incorporate the shrinkage effect of the prior. The claim here is that such an effect can be arbitrarily large for S finite and p arbitrary large.

2.1. Variational Algorithm for Approximate $\widehat{\tau}_b$

The above stochastic representation suggests to obtain $\pi(\mathcal{M}_\gamma|y)$ by using another Gibbs-sampling exposed in [8] and it was used to obtain simulations from the posterior distribution of ψs and vs. However, this algorithm can be very slow for large p and the benefits of the proposed Bayesian approach can be compensated by just calculating $\widehat{\tau}_e$ or $\widehat{\tau}_r$ over larger values of S (if these were possible to be obtained).

To avoid this drawback of the proposed modelling approach we make use of a recent and faster variational algorithm illustrated in [9] and also implemented in the R package `mixdir`. The algorithm relies on using approximated distribution, for the posterior of $\mathbf{v}$ and ψ. Such distributions are derived by applying the mean-field theory to variational inference (see [10]). The approximating q distributions of the variational approach lay down to be a mixture of Dirichlet distributions. For more details, see [9]. This algorithm is much faster than the initial Gibbs-sampling in [8] and thus may compensate for the need of a larger S to obtain good estimations of τ.

3. Implementation and Examples

The `R` implementation of the proposed model is straightforward and it does not even need an ad hoc appendix because major packages are already available. In particular, the implementation requires the following packages for Gibbs-sampling $\mathcal{M}^{(S)}$: `BayesVarSel` for BF based on the conventional prior approach [2] and/or `mombf` for BF based on non-local priors [11]. Finally, package `mixdir` contains the variational Bayes method sketched above in Section 2.1.

The minimal implementation for a response y and a design matrix X with p columns requires two steps:

1. Obtain a sample of $\mathcal{M}^{(S)}$ use:
   - for conventional prior: `gammas<- GibbsBvs(y,X)$modelslogBF[,-p+1]`
   - for non-local prior: `gammas<- modelSelection(y,X)$postSample`
2. Estimate ß by `cellsprob=mixdir(gammas)` and calculate, for generic j covariate, the inclusions probabilities in each of the component of the mixture, `pp=unlist(cellsprob$category_prob[[j]])`. Then
   - $\widehat{\tau}_b$ is `sum(pp[names(pp)==''1'']*cellsprob$lambda)`
   - $\widehat{\tau}_f$ is `pp[names(pp)==''1''][1]`

In the examples and simulation studies illustrated below, we will mainly play with the number of rows S of above-calculated matrix `gammas`.

3.1. Riboflavin Simulation Study

In what follows we will consider the Riboflavin dataset (see [12]) related to the riboflavin production by *Bacillus subtilis*. We have $n = 71$ observations and $p = 4088$ predictors (gene expressions) and a one-dimensional response (riboflavin production), y. We assume that $BF_{\gamma 0}$ is obtained from the conventional prior and non-local prior. Priors on models, $\pi(\mathcal{M}_\gamma)$ is the Uniform prior.

We use the Riboflavin dataset only for fixing X_p and simulating 10,000 times the response vector y of size n according to $y = \sum_{i=1}^{3} i \times X_{p_i} + \epsilon$, $\epsilon \sim N(0, 2)$ and p_1, p_2, p_3 are three columns of X_p picked at random for each simulated response (the rest of columns of X_p are supposed to have no effect on y).

The Gibbs algorithm starts at the null model and S has different sizes $S = 100, 500, 1000$. The goal is to compare the proposed estimator against existing ones.

Results regarding the estimation of τ over all simulations with the Conventional and Non-Local priors are shown in Figure 1.

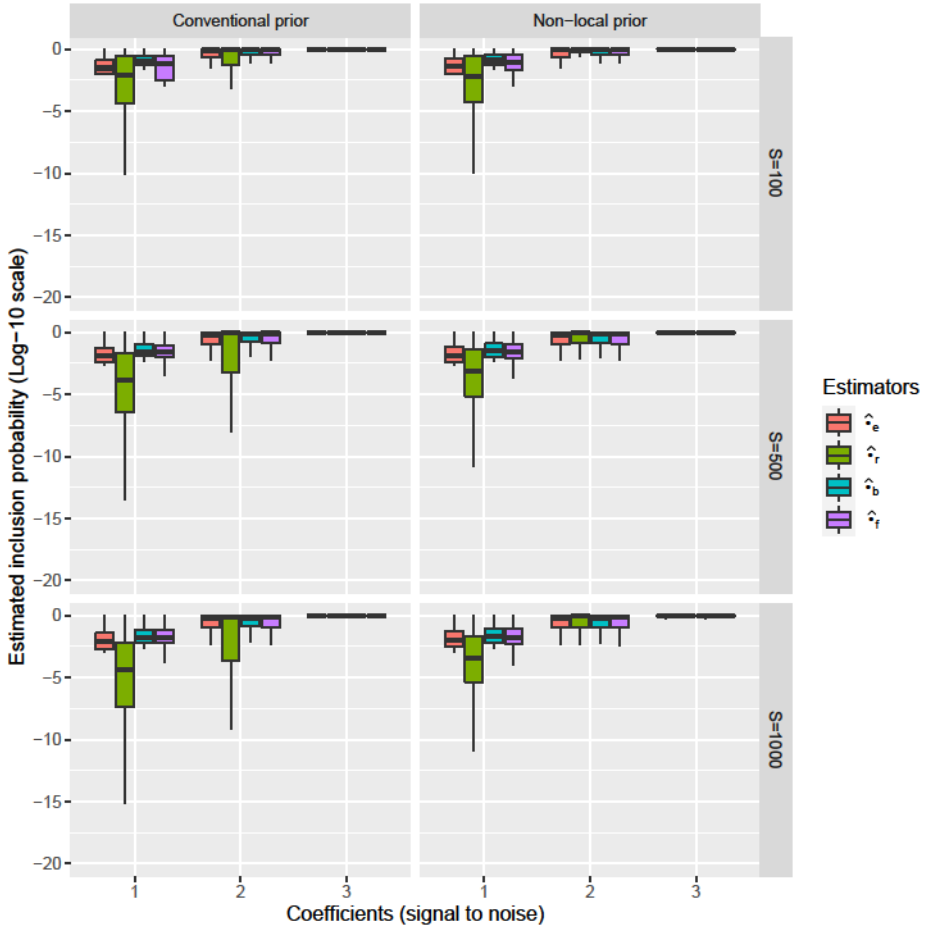

Figure 1. Riboflavin simulation results. Distributions of $\widehat{\tau}_{\mathrm{e}}$, $\widehat{\tau}_{\mathrm{r}}$, $\widehat{\tau}_{\mathrm{b}}$ and $\widehat{\tau}_f$ over the 10,000 datasets for only covariates inside the model, ordered by the magnitude of their corresponding coefficients and the prior used to calculate BFs. Coefficients also represent the signal into the data (i.e., the value of the coefficients) with respect to the residual standard error (the noise).

It is possible to appreciate that when the signal in the data is small (coefficients are 1 or 2) and S is not large enough, the proposed estimators $\widehat{\tau}_{\mathrm{b}}$ and $\widehat{\tau}_f$ perform better than the existing ones, while the $\widehat{\tau}_{\mathrm{r}}$ performs better under the non-local prior approach. This is because, as mentioned above, the non-local prior approach favors the model learning rate which is reflected in the values of the BF used to renormalize $\widehat{\tau}_{\mathrm{e}}$.

However, such improvements depend on the specific simulation analysis, that is the design matrix, the noise in the response (here, two standard deviations) the specific value of the assumed coefficients (i.e., 1, 2 and 3) and also the values of S. In particular, it seems to disappear when the value of the coefficient is high (i.e., 2 or 3). To generalize these results concerning the choice of specific values of coefficients, regression error standard deviation, and S, we analyze the results using probit regressions (one for each prior). In particular, we transform 10,000 estimators $\widehat{\tau}$ on the probit scale and regress it with respect to the signal into the data (the value of the coefficient) and estimator type (four in total) with interactions among them (in total we have 12 probit regression coefficients: main effects plus interactions).

The resulting coefficient regressions that we focus on indicate the improvement to the $\widehat{\tau}_{\mathrm{e}}$ and the signal when the coefficient is 1. Such an improvement is an increment in the probit scale of the probability of having $\widehat{\tau} = 1$ for each combination of coefficient and estimator type. These increments are between 0.62 and 2.14 (all highly significant) and are important if the original point in the probit scale is low, while they are negligible if the point on the probit scale is large. Such an initial point

corresponds to a signal in the data and the value of S. The increment on $\widehat{\tau}$ and the signal in the data is reported in Figure 2.

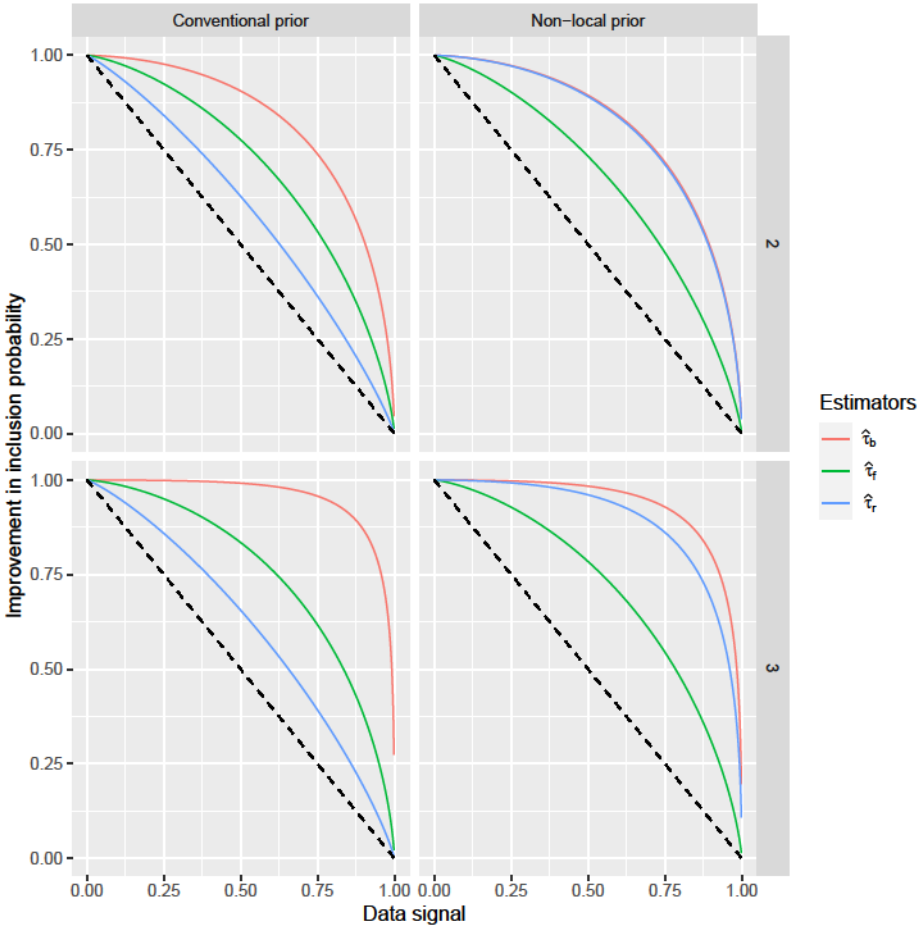

Figure 2. Riboflavin simulation results. Estimate improvements into the inclusion probabilities (vertical axis) with respect to the signal in the data and S (horizontal axis) for $\widehat{\tau}_e$ (dashed line), $\widehat{\tau}_r$, $\widehat{\tau}_b$ and $\widehat{\tau}_f$ estimated over the 10,000 datasets. For only covariates inside the model, ordered by the magnitude of their corresponding coefficients and the prior used to calculate BFs.

From Figure 2 we can see that the proposed estimators $\widehat{\tau}_b$ and $\widehat{\tau}_f$ generally outperform the existing estimator and popular $\widehat{\tau}_e$ as the increment on $\widehat{\tau}$ is larger when the signal in the data is lower. The $\widehat{\tau}_r$ perform better than $\widehat{\tau}_f$ under the non-local prior as the corresponding BFs are very informative for detecting variables.

3.2. Travel and Tourism Data Set

In this application, we want to explore the determinants of Travel and Tourism (T&T) global expenditure using data of the World Economic Forum, based on the latest Travel and Tourism Competitiveness Report (2019). The statistical unit is the country and the response variable is the log of total tourism expenditure, obtained by multiplying the number of arrivals by the reported individual expenditure. Data are obtained at http://www3.weforum.org/docs/WEF_TTCR19_data_for_download.xlsx and further filtered to have a complete dataset https://raw.githubusercontent.com/scabras/varseldmmp/master/tourism-data.csv The actual code to analyze it is here: https://github.com/scabras/varseldmmp/blob/master/s-example-tourism.md.

Tourism worldwide is threatened by the Covid-19 spread and it represents an important source of income for many countries. For instance, in Spain, Mexico, and France, according to The Organization

for Economic Co-operation and Development (OECD) in 2016, tourism represented more than 7% of the Gross Domestic Product. The total number of covariates considered here is $p = 65$ fully observed in $n = 52$ countries.

We apply the analysis using Conventional and non-local priors assuming that the number of latent probability distributions on the model space is $k = 10$ and a fixed concentration parameter $\alpha = 1$ to be as parsimonious as possible on determining the median probability model for $S = 1000$ and a burn-in of 100 Gibbs-sampling steps.

It is important to note that to reach almost a substantial agreement between the actual estimators $\widehat{\tau}_{\mathrm{e}}$, $\widehat{\tau}_{\mathrm{r}}$ on one side and the newly proposed ones $\widehat{\tau}_{\mathrm{b}}$, $\widehat{\tau}_{f}$ it is necessary to have S = 10,000, which corresponds to about seven times more the computational time used to obtain Table 1 which reports the estimated median probability models.

Table 1. Tourism and Travel (T&T) data set. Estimated median probability models according to the different estimators $\widehat{\tau}$ crossed with BF definitions given by conventional and non-local priors. In parenthesis, $\widehat{\tau}$ is reported when the covariate is in the median probability model.

Name of Covariate	**Conventional Prior**				**Non-Local Prior**			
	$\widehat{\tau}_{\mathrm{e}}$	$\widehat{\tau}_{\mathrm{r}}$	$\widehat{\tau}_{\mathrm{b}}$	$\widehat{\tau}_{f}$	$\widehat{\tau}_{\mathrm{e}}$	$\widehat{\tau}_{\mathrm{r}}$	$\widehat{\tau}_{\mathrm{b}}$	$\widehat{\tau}_{f}$
number of operating airlines	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)
timeliness of providing monthly/quarterly t&t data	(0.95)	(1.00)	(0.74)		(0.93)	(1.00)	(0.83)	
individuals using internet, %		(0.91)	(0.57)		(0.76)		(0.72)	
purchasing power parity		(0.96)	(0.51)		(0.64)		(0.53)	
natural tourism digital demand		(0.75)			(0.68)		(0.61)	
active mobile broadband internet subscriptions/100 population		(0.82)						
openness of bilateral air service agreements		(0.96)						
total known species		(0.71)						

The most important factor is precisely the one most affected by Covid-19, which is the number of operating airlines. The other factors are less related to Covid-19 but are still important. One is the organizations of the T&T govern monitoring represented by the timeliness of providing data on T&T. There are also variables related to country features: infrastructures (connection and power) internet search of T&T resources and the presence of natural resources (total known species). As observed above, non-local priors lead to more complex models than the conventional prior, but looking at the proposed estimators they seem to be more robust with respect to the choice of the prior on model parameters, in fact, $\widehat{\tau}_{\mathrm{b}}$ reports almost the same covariates regardless of the BF definition.

4. Remarks

This work aims to analyze the Gibbs-sampling output of model exploration according to a genuine Bayesian analysis and avoid frequentist approaches that are of a critical application for p large and S finite. Further, this is also the first work on comparing the conventional versus the non-local prior definition of BFs, which represent at the moment the state of the art in covariate selection under the normal linear model. The posterior mean of the τ is estimated with $\widehat{\tau}_{\mathrm{b}}$ results to be a robust estimator of the τ. This is because of the Dirichlet model, which shrinks to zero some of the observed empirical proportions, resulting from $\widehat{\tau}_{\mathrm{e}}$, and increases the others. Moreover, it is also clear that the proposed $\widehat{\tau}_{\mathrm{r}}$, studied only for conventional prior in [7], seems to work very well under the improved learning rate of the non-local prior approach.

As a further line of investigation, but beyond the scope of this paper, we note that $\pi(\mathcal{M}_\gamma)$ (here, the uniform) can be substituted by the Dirichlet process model illustrated above and thus used directly as $\pi(\mathcal{M}_\gamma)$ by incorporating it into the Gibbs-sampling. This is important in order to match ß

in (1) with $\pi(\mathcal{M}_\gamma)$. This is incidentally done in [7], with the uniform $\pi(\mathcal{M}_\gamma)$ when using $\widehat{\tau}_{\mathbf{r}}$. In general, the problem of model exploration would deserve more Bayesian methods than the existing ones.

Funding: This research is supported by Ministerio de Ciencia e Innovación of Spain project PID2019-104790GB-I00.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Scott, J.G.; Berger, J.O. Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem. *Ann. Stat.* **2010**, *38*, 2587–2619. [CrossRef]
2. Bayarri, M.J.; Berger, J.O.; Forte, A.; García-Donato, G. Criteria for Bayesian Model Choice with Application to Variable Selection. *Ann. Stat.* **2012**, *40*, 1550–1577. [CrossRef]
3. Nikooienejad, A.; Johnson, V.E. *BVSNLP: Bayesian Variable Selection in High Dimensional Settings Using Nonlocal Priors*; CRAN Package, 2018. Available online: https://cran.r-project.org/web/packages/BVSNLP/index.html (accessed on 27 August 2020).
4. Barbieri, M.M.; Berger, J.O. Optimal Predictive Model Selection. *Ann. Stat.* **2004**, *32*, 870–897. [CrossRef]
5. Barbieri, M.; Berger, J.O.; George, E.I.; Rockova, V. The Median Probability Model and Correlated Variables. Technical Report. *arXiv* **2018**, arXiv:1807.08336v2
6. George, E.I.; McCulloch, R.E. Approaches for Bayesian variable selection. *Stat. Sin.* **1997**, *7*, 339–373.
7. Garcia-Donato, G.; Martinez-Beneito, M.A. On Sampling strategies in Bayesian variable selection problems with large model spaces. *J. Am. Stat. Assoc.* **2013**, *108*, 340–352. [CrossRef]
8. Dunson, D.B.; Xing, C. Nonparametric Bayes modeling of multivariate categorical data. *J. Am. Stat. Assoc.* **2009**, *104*, 1042–1051. [CrossRef] [PubMed]
9. Ahlmann-Eltze, C.; Yau, C. MixDir: Scalable Bayesian Clustering for High-Dimensional Categorical Data. In Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), Turin, Italy, 1–3 October 2018; pp. 526–539.
10. Jaakkola, T.S.; Jordan, M.I. Improving the Mean Field Approximation via the Use of Mixture Distributions. In *Learning in Graphical Models*; Springer: Dordrecht, The Netherlands, 1998; pp. 163–173 [CrossRef]
11. Johnson, V.E.; Rossell, D. On the use of non-local prior densities in Bayesian hypothesis tests. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2010**, *72*, 143–170. [CrossRef]
12. Bühlmann, P.; Kalisch, M.; Meier, L. High-Dimensional Statistics with a View Toward Applications in Biology. *Annu. Rev. Stat. Its Appl.* **2014**, *1*, 255–278. [CrossRef]