Contents

1

Bayes factors based on g-priors for variable selection

Gonzalo García-Donato

Universidad de Castilla-La Mancha (Spain)

Mark F. J. Steel

University of Warwick (U.K.)

CONTENTS

Variable selection can be naturally seen as a model selection problem where the entertained models differ in which subset of variables explains the outcome of interest. In this setting, posterior probabilities of the models are a simple combination of Bayes factors, a well-known inferential tool that is key in the formal Bayesian approach to testing and model choice. This approach to variable selection automatically provides sparse answers that, quite importantly, are accompanied with probabilistic assessments regarding the confidence we have in them. This methodology is, however, not exempt from difficulties including prior elicitation and numerical challenges related with its practical implementation. Particularly in the context of linear and generalized linear models, the so-called g-priors have attracted the interest of many researchers due to their appealing properties formally described in [2]. In this chapter we review the main aspects concerned with the implementation of the Bayesian approach to variable selection based on Bayes factors in linear and generalized linear models, using g-priors. The material presented here has a clear focus on applicability and emphasis is placed on providing: i) practical guides for implementation, including documentation for the use of `R` packages (particularly `BayesVarSel` and `glmBfp`) and ii) the analysis of real examples which illustrate the enormous potential of this approach to variable selection.

1.1 Bayes factors

Variable selection with Bayes factors is a methodology based on the *significance tests* by Sir Harold Jeffreys. These were introduced in a series of papers published during the first decades of the 20th century and culminated in his famous book "Theory of Probability" [19], the first edition of which dates back to 1939. The reader is referred to [12] for an interesting account of the history of Bayes factors, and to [33] for a modern revision of Jeffreys' influential book.

Jeffreys' significance tests provide a solution to testing precise null hypotheses (ie. that a certain parameter is zero) through the evidence that data give to each of the tested hypotheses. A key consideration is that hypotheses define statistical models and hence such evidence can be measured utilizing the relative support that each model receives from the data. The number that contains that relative evidence is what we call today the Bayes factor.

In the simple case with only two hypotheses, the data $\boldsymbol{y}$ follows a certain distribution under the null hypothesis $M_0 : \boldsymbol{y} \sim f_0(\boldsymbol{y} \mid \boldsymbol{\theta}_0)$, while under the alternative $M_1 : \boldsymbol{y} \sim f_1(\boldsymbol{y} \mid \boldsymbol{\theta}_1)$.

To decide which of these models/hypotheses provides a more appropriate representation for the underlying data generating process of $\boldsymbol{y}$ we obtain the Bayes factor of M_1 to M_0 as:

$$B_1 = \frac{m_1(\boldsymbol{y})}{m_0(\boldsymbol{y})}, \qquad m_\gamma(\boldsymbol{y}) = \int f_\gamma(\boldsymbol{y} \mid \boldsymbol{\theta}_\gamma)\, p_\gamma(\boldsymbol{\theta}_\gamma)\, d\boldsymbol{\theta}_\gamma,\ \gamma = 0, 1. \tag{1.1}$$

Above, m_1 and m_0 are the prior predictive marginals (often called marginal likelihoods) and $p_1(\cdot)$ and $p_0(\cdot)$ are the priors for the parameters within each model.

The Bayes factor B_1 is a measure of evidence in favor of M_1 and against M_0 provided by the sample under the chosen prior distributions. The larger B_1, the stronger is the evidence supporting M_1. Several authors have provided rules to interpret B_1 [19, 21] but it is common to use them through their relation with the posterior probabilities of the models being compared. It can easily be seen that posterior odds ratios between models are equal to the prior odds multiplied by the appropriate Bayes factor:

$$\frac{p(M_1 \mid \boldsymbol{y})}{p(M_0 \mid \boldsymbol{y})} = \frac{p(M_1)}{p(M_0)} B_1, \tag{1.2}$$

where $p(M_\gamma)$ is the prior probability assigned to M_γ. Thus, conditionally on either M_0 or M_1 being the true model, the posterior probability of M_1 is

$$p(M_1 \mid \boldsymbol{y}) = \frac{p(M_1) B_1}{p(M_1) B_1 + p(M_0)}.$$

In order to make an explicit model selection we have to choose a threshold for the posterior probability for M_1. If $p(M_1 \mid \boldsymbol{y})$ is larger than this threshold we choose M_1 and otherwise we choose M_0. Such a threshold could, of course, be guided by a utility or loss function in a decision-theoretic setting (see e.g. [6]).

An important characteristic of this approach to model selection is that both models are given full consideration (different from methodologies where the selection is based on consideration of only the largest model). The advantages of this approach are nicely reviewed in [5] and here we want to emphasize two that we find particularly relevant. Firstly, the approach is automatically parsimonious (choosing the simplest model for a similar fit). This is essentially because, in logarithmic scale, Bayes factors can be approximated by a

goodness-of-fit term minus a penalty for complexity that provides the mentioned automatic 'protection' to simpler models (for more details and related references on this issue see eg. [20]). Secondly, the method comes accompanied with a measure of uncertainty regarding the model selection exercise since it is based on the full posterior probability distribution over all considered models. This gives an accurate idea of the remaining uncertainty regarding which model to use. In this chapter we show, through real applications, the potential of the approach based on Bayes factors to variable selection highlighting the richness of the obtained inference.

The probability distribution over models provided by the formal Bayesian approach sketched above also allows us to formally include the uncertainty regarding models in our inference and decision-making by averaging over models with the posterior model distribution. This so-called Bayesian model averaging is the natural Bayesian response to uncertainty and was already described in [23] and used in e.g. [32] and [13]. This is a natural step to fully incorporate the model uncertainty, and is available in the packages used here. Key posterior quantities mentioned in the chapter, such as the posterior probability of inclusion of a regressor are derived by averaging over models.

In the situation where more than two models are entertained, the index γ takes values in a set $\mathcal{M}$ (called model space) and the posterior probability of any of the competing models is

$$p(M_\gamma \mid \boldsymbol{y}) \propto p(M_\gamma) B_\gamma,$$

where B_γ is the ratio of the marginal likelihood $m_\gamma(\boldsymbol{y})$ to the marginal likelihood of a fixed model (say, without loss of generality M_0). A multiple model selection problem naturally arises in *variable selection*. In this situation, the proposed models share a common functional form (e.g. a normal linear regression model) but differ in which explanatory variables, from a given set, are included to explain the response. The focus of this chapter will be on variable selection in the context of normal linear models (Section 1.2) and generalized linear models (Section 1.3).

Although very sound and cogent, the implementation of this methodology has two main challenges that we next describe. The first difficulty is a conceptual issue, namely that the prior used is going to have an important effect on the results. In contrast to the situation where we formulate a prior on the parameters of a single uncontested model, we do not have the luxury of priors that are "non-informative" in the sense that their effect is easily swamped by the data as we collect more observations. In addition, the prior needs to be proper on model-specific parameters. Indeed, any arbitrary constant in $p_\gamma(\boldsymbol{\theta}_\gamma)$ will affect the marginal likelihood $m_\gamma(\boldsymbol{y})$ in (1.1). Thus, if this constant emanating from an improper prior does not multiply the marginal likelihoods of all possible models, it clearly follows from (1.2) that posterior model probabilities are not determined. In this chapter, we pay special attention to the family of priors named g-priors. Strongly inspired by the work of Jeffreys to implement his significance tests, g-priors were introduced by [39, 38] and they have been the topic of renewed research interest over the last fifteen years or so. These types of priors, which some authors have also called *conventional* [5, 1], are introduced in Section 1.2.1.

A second challenge for the practical implementation is computational. In some cases (in particular, the linear Gaussian model with the class of g-priors mentioned in Subsection 1.2.1) the integral in (1.1) can be solved analytically, but in many other cases it does not admit an explicit solution and we need to resort to a simulated or approximated answer. In addition, the number of models in the model space can, for many applications be very large indeed, thus precluding an exhaustive enumeration of all possible models. As an example, the genetic example in Subsection 1.2.1 has a model space with 2^{4088} models, which is far larger than what can be dealt with exhaustively (typically, we can deal with model spaces

up to size 2^{30} or so if we use complete enumeration). We discuss these numerical issues in Section 1.2.2. Fortunately, these methods are easily accesible to practitioners with specific `R` [30] packages. Here, we illustrate the use of the freely available packages `BayesVarSel` and `glmBfp`. All our examples are run on a Macbook pro laptop with 2.6 GHz Intel Core i5 processor without parallel computation, clearly indicating that the analysis of practically relevant problems is readily accessible.

1.2 Variable selection in the Gaussian linear model

Consider a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ with components Y_i being independent, normally distributed with unknown variance σ^2. In the regression setup, the mean μ_i of Y_i is assumed to be a linear combination of a subset of p possible explanatory variables $\{X_1, \ldots, X_p\}$, but it is uncertain which is the relevant subset. This situation implicitly defines 2^p entertained regression models which can be expressed by making use of a vector parameter $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)^T$ where $\gamma_i \in \{0, 1\}$ and $\gamma_i = 1$ indicates that x_i is included in the model. Hence the model space is $\mathcal{M} = \{0, 1\}^p$, where each $\boldsymbol{\gamma} \in \mathcal{M}$ assumes that

$$\mu_i = \alpha + \sum_{j=1}^{p} \gamma_j \, x_{ij} \beta_j, \ \forall i = 1, 2, \ldots, n$$

where x_{ij} denotes the ith observation of variable x_j, i.e. the (i, j)th element of the full $(n \times p)$ covariate matrix $\boldsymbol{X}$. Using the notation in the introduction, now $\boldsymbol{\theta}_\gamma = (\alpha, \boldsymbol{\beta}_\gamma, \sigma)$ and the entertained models

$$M_\gamma : \boldsymbol{Y} \sim f_\gamma(\boldsymbol{y} \mid \alpha, \boldsymbol{\beta}_\gamma, \sigma) = \mathcal{N}_n(\boldsymbol{y} \mid \alpha \mathbf{1}_n + \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 I_n)$$

where $\boldsymbol{X}_\gamma$ is a $(n \times p_\gamma)$-dimensional ($p_\gamma = \sum_{j=1}^{p} \gamma_j$) data matrix formed using the included variables in M_γ (abusing notation, $\boldsymbol{\beta}_\gamma$ is empty when $\boldsymbol{\gamma}$ is the null vector) and $\mathbf{1}_n$ represents a n-dimensional unitary column vector . The covariates are standardized by subtracting their means, which makes them orthogonal to the intercept and renders the interpretation of the intercept common to all models.

Assuming that one of the models in $\mathcal{M}$ is the true model, the posterior probability of any model γ^* is

$$p(M_{\gamma^*} \mid \boldsymbol{y}) = \frac{B_{\gamma^*}(\boldsymbol{y}) p(M_{\gamma^*})}{\sum_\gamma B_\gamma(\boldsymbol{y}) p(M_\gamma)}, \tag{1.3}$$

where $p(M_\gamma)$ is the prior probability of M_γ and B_γ is the Bayes factor of M_γ with respect to a fixed model, say M_0 (without any loss of generality) and hence $B_\gamma = m_\gamma / m_0$ and $B_0 = 1$.

1.2.1 Objective prior specifications

Priors for the within model parameters: the g-priors

The prior on the model parameters assigns posterior point mass at zero for those regression coefficients that are not included in M_γ, which automatically induces sparsity. Without loss of generality, the prior distributions p_γ can be expressed as

$$p_\gamma(\boldsymbol{\beta}_\gamma, \alpha, \sigma) = p_\gamma(\boldsymbol{\beta}_\gamma \mid \alpha, \sigma) p_\gamma(\alpha, \sigma).$$

The common parameters is assumed to be equal for all models and for $p(\alpha, \sigma)$ a very commonly used objective prior is assumed

$$p_\gamma(\alpha, \sigma) = p(\alpha, \sigma) \propto \sigma^{-1}. \tag{1.4}$$

In the g-priors, the model-specific parameters have the following distribution specified conditionally on a hyper-parameter $g > 0$ (which is the reason for the name of these priors):

$$p_\gamma(\boldsymbol{\beta}_\gamma \mid \alpha, \sigma, g) = \mathcal{N}_{p_\gamma}(\boldsymbol{\beta}_\gamma \mid \mathbf{0}_{p_\gamma}, g\sigma^2(\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma)^{-1}). \tag{1.5}$$

This prior structure already appeared in [13, 5] and is now the most commonly used prior in the context of the normal linear model.

Without g, the prior covariance matrix above coincides with the expected Fisher information matrix corresponding to $\boldsymbol{\beta}_\gamma$ obtained from the model M_γ. The parameter g has the role of scaling the resulting matrix, for example such that the prior reflects a similar amount of information as one observation (this corresponds to a fixed value $g = n$ and leads to log Bayes factors that behave asymptotically like the BIC, see [13]). Several authors have argued in favor of treating g as an unknown hyper-parameter for which a hyper-prior needs to be assigned. There are theoretical reasons for the introduction of this extra layer of variability that relate to information consistency, which implies that the posterior probability tends to one for a model with arbitrarily large sampling evidence in its favour. In addition, from a practical perspective, treating g as random provides a prior for $\boldsymbol{\beta}_\gamma$ with flatter tails, hence accommodating regressors with a moderate impact on the response. In the section devoted to the sensitivity analysis, we will see a manifestation of this effect in practice. In Table 1.1 we have collected the most popular proposals for g. There are subtle conceptual differences that have lead the different authors to propose these specific proposals and the reader is referred to the original reference for more details. Hence, the g-priors have, manifestly, been proposed based on constructive arguments. Nevertheless, much later [3] showed that these priors can also be derived using a mathematical formal rule based on the "distance" between competing models.

The ensuing methodology based on g-priors is endorsed by many appealing theoretical properties. A number of these have a frequentist flavor like consistency (ability to select the true model when the sample size grows to infinity or when the evidence in favour of a model becomes overwhelming) but others are specifically related with the desiderata for objective priors for testing and model selection. Within the latter category, we emphasize that g-priors are predictive matching (reporting inconclusive evidence when the sample size is extremely small). Finally, these priors produce Bayes factors that are invariant to affine transformations of the covariates. The reader is referred to [2] for a comprehensive discussion of these properties.

Priors over the model space $\mathcal{M}$

For priors over the model space $\mathcal{M}$, a very popular starting point is

$$p(M_\gamma \mid \theta) = \theta^{p_\gamma}(1-\theta)^{p-p_\gamma}, \tag{1.6}$$

where p_γ is the number of covariates in M_γ, and the hyperparameter $\theta \in (0, 1)$ has the interpretation of the common prior probability that a given variable is included (independently of all others).

For the specific assignment in (1.6), some of the most popular default choices for θ are

- Fixed $\theta = 1/2$, which assigns equal prior probability to each model, i.e $p(M_\gamma) = 1/2^p$;
- Random $\theta \sim \mathcal{U}(0, 1)$, giving equal probability to each possible number of covariates or model size.

Proposal	Reference	Name
Fixed g		
$g = n$	[38, 21]	Unit Information prior
$g = p^2$	[15]	Risk inflation criterion prior
$g = \max\{n, p^2\}$	[13]	Benchmark prior
$g = \log(n)$	[13]	Hannan-Quinn
Random g		
$g \sim IGa(1/2, n/2)$	[19, 39, 40]	Cauchy prior
$g \mid a \sim p(g) \propto (1+g)^{-a/2}$	[27]	hyper-g
$g \mid a \sim p(g) \propto (1+g/n)^{-a/2}$	[27]	hyper-g/n
$g \sim p(g) \propto (1+g)^{-3/2}$, $g > \frac{1+n}{p_\gamma+1} - 1$	[2]	Robust prior

TABLE 1.1
Specific proposals for the hyperparameter g in the literature.

Of course many other choices for θ – both fixed and random– have been considered in the literature. In general, fixed values of θ have been shown to perform poorly in controlling for multiplicity (the occurrence of spurious explanatory variables as a consequence of performing a large number of tests) and can lead to rather informative priors. This issue can be avoided by using random distributions for θ as, for instance, the second proposal above that has been studied in [35]. Additionally, [25] consider the use of $\theta \sim \mathcal{B}eta(1, b)$ which results in a binomial-beta prior for the number of covariates in the model or the model size, W:

$$p(W = w \mid b) = \frac{b}{\Gamma(b+p+1)} \binom{p}{w} \Gamma(1+w)\Gamma(b+p-w), \;\; w = 0, 1, \ldots, p.$$

Notice that for $b = 1$ this reduces to the uniform prior on θ and also on W. As [25] highlight, this setting is useful to incorporate prior information about the mean model size, say $w^\star$. This would translate into $b = (p - w^\star)/w^\star$.

In variable selection, applications with a large number of explanatory variables p are becoming very common. In these situations, depending on the context and the prior information, it is typically a good idea to use a prior which implies a multiplicity correction or a prior which induces sparsity along the lines suggested in [8]. Additionally, in such contexts, we usually have to face situations where $p > n$ (or even $p >> n$) and the set $\mathcal{M}$ contains models with $p_\gamma + 1 > n$ that are hence rank deficient (in the Gaussian setting these models are not estimable). Normally, these models are given zero prior probability (they are discarded) and the prior assignment in (1.6) applies only to those models for which $p_\gamma < n$ with a proportionality sign. A different treatment for these singular models in the normal case is given in [4]. These authors argue that while full rank models may contain decisive information concerning which covariates are related with the response, the rank deficient ones are not informative but will add uncertainty reflecting the fact that p is large compared to n. In this regard, [4] observe that rank deficient models are "copies" (reparameterizations) of the saturated model with $p_\gamma + 1 = n$ with the same marginal likelihood (cf. (1.7) with $SSE_\gamma = 0$ and $p_\gamma + 1 = n$). This justifies the use of unitary Bayes factors for all rank deficient models and thus avoids the need to assign zero prior probability to these models. In practical terms, both approaches are expected to provide similar results, unless n is very small. We have observed this agreement in the second of our applications which concerns a high dimensional study with $n = 71$ and $p = 4088$.

1.2.2 Numerical issues

As already mentioned in Subsection 1.1 there are two main computational challenges in solving a model uncertainty problem through Bayes factors. Firstly, the integral in (1.1) and, secondly, the sum in the denominator of (1.3) which involves many terms if p is moderate or large.

Fortunately, in normal models, g-priors combine easily with the likelihood, and conditionally on g lead to closed forms for $m_\gamma(\boldsymbol{y})$. Hence, at most, a univariate integral needs to be computed when g is taken to be random. Thus Bayes factors have a very manageable expression:

$$B_\gamma(\boldsymbol{y}) = \int \left(1 + g\, Q_\gamma\right)^{-(n-1)/2} \left(1+g\right)^{(n-p_\gamma-1)/2} p(g)\, dg. \tag{1.7}$$

where Q_γ, is the ratio sum of squared errors of model M_γ to the null model M_0. Interestingly, there have been recent proposals for prior distributions, which despite assuming a hyper prior on g induce closed form marginals using special mathematical functions. This characteristic is shared by the robust prior of [2], the prior of [28] and the hyper-g in [27].

The second problem, related with the magnitude of the number of models in $\mathcal{M}$ (i.e. 2^p), could be a much more difficult one. If p is small (say, p in the twenties at most) exhaustive enumeration is possible but if p gets larger, exact approaches quickly become infeasible. Interesting exceptions include the recent work by [9, 37] who have developed, for certain particular problems (where $n = p$), exact algorithms that may handle problems with even very large p. For the general variable selection case, however, it is hard to imagine an exact solution and we will have to rely on some sort of approximation to the posterior distribution. This question has been studied in [17]

who considered a simple Gibbs algorithm that was suggested by [18]. This algorithm begins taking an initial model $\boldsymbol{\gamma}_{(0)} = (\gamma_{1(0)}, \gamma_{2(0)}, \ldots, \gamma_{p(0)})$ with Bayes factor $B_{\gamma_{(0)}}$ then repeating, for $i = 1, \ldots, N$ (N is the number of iterations) the following $p + 1$ steps:

- Step j : $1 \leq j \leq p$. Propose the model $\boldsymbol{\gamma}_* = (\gamma_{1(i-1)}, \ldots, 1 - \gamma_{j(i-1)}, \ldots, \gamma_{p(i-1)})$ and then compute B_{γ_*} and the acceptance ratio $r = B_{\gamma_*} p(M_{\gamma_*}) / B_{\gamma_{(i-1)}} p(M_{\gamma_{(i-1)}})$. With probability $\min\{r, 1\}$ re-define $\boldsymbol{\gamma}_{(i-1)} = \boldsymbol{\gamma}_*$.
- Final step. Define and save $\boldsymbol{\gamma}_{(i)} = \boldsymbol{\gamma}_{(i-1)}$.

The result is $\{\boldsymbol{\gamma}_{(1)}, \boldsymbol{\gamma}_{(2)}, \ldots, \boldsymbol{\gamma}_{(N)}\}$, a sample from the posterior distribution (1.3) which is the only ingredient needed to obtain summaries solely based on the frequency of visits. For instance, the vector of posterior inclusion probabilities is obtained as the sample mean and so on.

Despite the apparent simplicity of the resulting approach, [17] show that it is potentially more precise than heuristic searching methods looking for 'good' models with estimates based on renormalization (i.e with weights defined by the analytic expression of posterior probabilities, cf. (1.3)). They show that these last methods could be biased by the searching procedure.

1.2.3 `BayesVarSel` and applications

There are several `R` packages that make it straightforward to implement variable selection based on g-priors. These have been considered in some detail in [14] who concluded that the results are comparable across the packages, although there are still important differences in cover and focus.

The results in this section are obtained with the `R` package `BayesVarSel` ([16]). This package was first released in December 2012 and has been periodically maintained and

updated with new functionalities since then. The package was conceived as a suite of tools to solve the variable selection problem for Gaussian linear models based on g-priors. It comes armed with many possibilities to summarize the posterior distributions and is very flexible regarding the choices of g (in particular, it incorporates the proposals in Table 1.1) and $p(M_\gamma)$. In our applications, we use the default choice in the package that corresponds to using $\theta \sim \mathcal{U}(0,1)$ for $p(M_\gamma \mid \theta)$ and the Robust prior [2] for $p(g)$. This configuration of prior inputs is the one that we ultimately recommend in general variable selection procedures. The code for running the example is provided as supplementary material.

Example of a moderate p (enumeration is feasible)

OBICE ([41]) was a study conducted in Spain during the years 2007-2008 to determine the association between diet, physical activity and obesity in children under 15 years of age. This study has a case-control design and the collected data come from a questionnaire completed by pediatricians. The survey collected a lot of information, some of which is redundant, and here we are considering $p = 15$ variables that provide a complete description of the aspects considered in the study (see Table 1.2). The model space contains $2^{15} = 32,768$ models, a size that allows us to compute posterior probabilities exactly (through exhaustive enumeration) in a few seconds. To avoid using imputation methods we include here only the children without any missing value leading to a sample size of $n = 996$ (84% of the number of recruited children). The response variable, y_i, is the Body Mass Index (BMI) and the age of the child is a variable that is always included (since it is known to influence BMI).

The possibility of reporting the degree of uncertainty regarding the variable selection problem in several informative ways is an important advantage of the methodology based on Bayes factors illustrated here. For instance, the model that in this study is most probable a posteriori (indicated as HPM in Table 1.2) contains (apart from the fixed one) 9 variables and has a posterior probability of 0.06. The model that follows in probability is the full model with a probability of 0.04. Interestingly, the smaller dimensionality of the HPM indicates that the information in some of the explanatory variables is really contained in others (e.g. the explanatory power of eating fruit and vegetables seem to be contained in the habit of having five meals per day). The individual importance of the variables can be assessed using the posterior inclusion probabilities. These are the aggregated probabilities of all models that contain a certain variable and are presented in Table 1.2. As expected, the variables included in the HPM are assigned large posterior inclusion probabilities (at least "strong evidence" according to the classification in [16]). None of the others are clearly ruled out, so the study doesn't have enough information to clarify whether these have an important role in the explanation of the body mass index. For instance, the sex of the child has an inclusion probability of 0.54 (quite similar to its prior probability) so there is no conclusive evidence whether this variable has any impact on obesity.

Additionally, we can explore the joint effect of variables in relation to their role in explaining the response (for a detailed study on this and related concepts the reader is referred to [24]). The information for such effect is contained in the probability that a certain variable is included, given that another is not, leading to a $p \times p$ matrix represented in Figure 1.1. To ease the interpretation, the row on top of the plot represents the marginal inclusion probabilities and the interrelations worth mentioning are when this probability differs substantially from the conditional probabilities in the table. In the great majority of cases there is barely any change (meaning that not including any other variable has no effect) but nevertheless several interesting facts arise. First is that we clearly see that either not considering the weight or height at birth diminishes the inclusion probability of the other (which makes sense since the response variable depends on both weight and height). More interesting is what happens with the dietary habit of having afternoon snacks. The

TABLE 1.2
Explanatory variables considered from the OBICE study. The dependent variable is Body Mass Index and the table contains posterior inclusion probabilities and an indicator of which variables appear in the highest posterior probability model (HPM).

Variable	Inc. Prob	HPM
Weight at birth	0.99	Y
Height at birth	0.99	Y
Sex	0.54	
The father is obese	1.00	Y
The mother is obese	1.00	Y
The child (regularly):		
...has 5 daily meals	1.00	Y
...eats vegetables	0.37	
...eats fruit	0.33	
...consumes afternoon snacks	0.83	Y
...was breastfed	0.42	
...practices sports	0.88	Y
Daily hours the child:		
...watches TV	1.00	Y
...plays with electronic devices	0.39	
...sleeps	0.42	
Daily candy consumption	0.99	Y

importance of this variable increases substantially if the information contained in "having or not 5 meals per day" is not considered, allowing to conclude that this last variable has a similar role as afternoon snacks. In other words, the two variables are substitutes (even though both appear in the HPM).

Example of a high dimensional setting

In [7], the relation of the production of riboflavin in *Bacillus subtilis* to the expression level of $p = 4088$ genes is studied. The dataset is distributed with the package `hdi` [11], and consists of $n = 71$ samples.

In [7], the authors are primarily interested in comparing the results among different frequentist statistical methodologies for high dimensional variable selection based on controlling false positive statements (type I error) and p-values. In this regard, the authors obtain results that vary "to a certain extent" over the different methods. The LASSO method selects 30 genes and the other methods either select no gene; select only the gene called `YXLD_at` or select three genes (`LYSC_at`, `YOAB_at` and `YXLD_at`). These disparate results provide an idea of the inherent difficulties in these high dimensional problems.

The model space for this problem contains (many) rank deficient models that were assigned zero prior probability. In this case, the posterior probability of rank deficient models is negligible so it does not make any difference whether these models are a priori ruled out or not. With respect to the computation, we used Gibbs sampling with $50,000$ iterations which took slightly more than 2 hours.

The genes with an estimated inclusion probability larger than 0.1 are collected in Table 1.3. The influence of the gene `YOAB_at` on the production of riboflavin is clear and is strongly endorsed by the data, leading to an inclusion probability of 0.97. A main difference

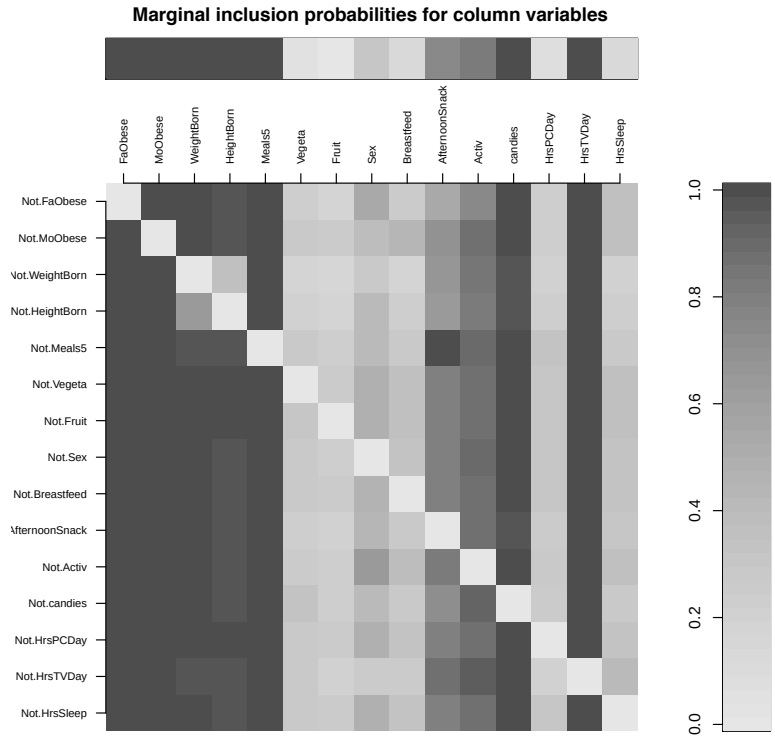

FIGURE 1.1
From the OBICE study, matrix of probabilities of inclusion probability of the column variable given that the variable in the row not included.

with the results in the previous example is that several covariates in the HPM (also shown in Table 1.3) have very small posterior inclusion probabilities and hence the interpretation of the results is more subtle. The HPM has an estimated posterior probability of 0.01 and the model that follows, with a probability of 0.002, contains four genes of which only `YOAB_at` is in the HPM. The third best model, which has a similar posterior probability, also proposes four genes, two of which are not included in the HPM. These results suggest a situation with multiple modes and hence several joint configurations of genes could provide a sensible explanation for the riboflavin detection. On the other hand, the results clearly indicate the great majority of genes are unimportant, with 97% of genes having an inclusion probability below 0.005.

Our interest in this experiment is mainly for illustrative purposes and with the above comments we wanted to highlight the richness of the posterior distribution to provide insight in the nature of the influence of the explanatory variables that goes far beyond a single model selected.

A final note is about the confidence in the numerical method used. As seems customary in MCMC methods, we ran in parallel (thus not requiring extra computational time) two other chains with randomly chosen initial values. The results were very similar to those described suggesting an efficient exploration of this large model space providing a reliable approximation to the posterior distribution.

TABLE 1.3
Posterior inclusion probabilities of genes with a value larger than 0.1 and an indicator of which genes appear in the highest posterior probability model (HPM) in riboflavin dataset.

Gene	Inc. Prob	HPM
YOAB_at	0.97	Y
YXLE_at	0.43	Y
ARGF_at	0.41	
YXLD_at	0.40	
CARB_at	0.21	
YFII_at	0.18	Y
YISU_at	0.17	
ARGB_at	0.14	Y
YHDZ_at	0.11	
YHEA_at	0.03	Y
YLXQ_at	0.08	Y

1.2.4 Sensitivity to prior inputs

In the context of the two previous applications, we now conduct a sensitivity study to asses the impact of the particular choice of $p(g)$ within the g-prior family and that of $p(M_\gamma)$ in (1.6). Recall that in our applications we used the Robust prior and $\theta \sim \mathcal{U}(0,1)$. Here we also computed the posterior distribution using the Cauchy, hyper-g/n and unit information priors for g (cf. Table 1.1) and using the fixed assignment $\theta = 1/2$ on the model space.

Results for the OBICE study, in the form of posterior inclusion probabilities, are collected in Table 1.2.4. These are quite insensitive to the choice of $p(g)$ and $p(M_\gamma)$ and the main conclusions about the importance of entertained covariates remain unchanged. The largest differences are observed between the unit information prior (with a fixed $g = n$) and the rest (with random g, with $p(g)$ a function of n). The extra layer assumed in $p(g)$ provides flatter tails to the prior on $\boldsymbol{\beta}_\gamma$ hence producing methods that are more liberal (more easily allowing for the presence of signals). This essentially explains the increment in the evidence reported towards declaring influential covariates, with the unit information prior being the most conservative, followed by Cauchy, then robust and finally hyper-g/n. We can also see this effect through the comparison of the posterior model size, $W \mid \boldsymbol{y}$, (summarized in this same table with its mean and variance) which clearly shows this same ordering.

In this dataset, which has a moderate number of potential regressors, $p = 15$, the choice of $\theta \sim \mathcal{U}(0,1)$ or $\theta = 1/2$ has barely any impact on the results although we see $\theta = 1/2$ behaving slightly more conservatively. This is a direct consequence of the intrinsic tendency of $\theta \sim \mathcal{U}(0,1)$ to favor models of dimensions that are shared by fewer models (so downweighting models with dimensions around $p/2$, of which there are many) and it so happens that in this problem many of the interesting models have a dimension which is larger than $p/2$ (in the opposite case, we would observe that $\theta \sim \mathcal{U}(0,1)$ favors simpler models).

The results in high dimensional application in the riboflavin dataset were barely sensitive to the prior for the regression parameter. In particular, with the Cauchy, hyper-g/n and Unit Information priors we obtained posterior inclusion probabilities that were very similar to those shown in Table 1.3. Furthermore, the HPM found with these priors coincide.

$\theta \sim \mathcal{U}(0,1)$				
Variable	Robust	Cauchy	hyper-g/n	Unit Inf.
Weight at birth	0.99	0.97	0.99	0.81
Height at birth	0.99	0.96	0.99	0.77
Sex	0.54	0.38	0.62	0.16
The father is obese	1.00	1.00	1.00	1.00
The mother is obese	1.00	1.00	1.00	1.00
...has 5 daily meals	1.00	1.00	1.00	0.99
...eats vegetables	0.37	0.21	0.45	0.06
...eats fruit	0.33	0.18	0.41	0.05
...afternoon snacks.	0.83	0.70	0.86	0.40
...was breastfed	0.42	0.25	0.50	0.07
...practices sports	0.88	0.82	0.90	0.65
...watches TV	1.00	1.00	1.00	1.00
...plays electronic	0.39	0.22	0.47	0.06
...sleeps	0.42	0.25	0.51	0.07
Daily candy consumption	0.99	0.97	0.99	0.92
$E(W \mid \boldsymbol{y})$	13.1	11.9	13.7	10
$\sqrt{V(W \mid \boldsymbol{y})}$	1.8	1.6	1.9	1.4
$\theta = 1/2$				
Weight at birth	0.98	0.96	0.98	0.82
Height at birth	0.97	0.94	0.97	0.78
Sex	0.32	0.25	0.35	0.13
The father is obese	1.00	1.00	1.00	1.00
The mother is obese	1.00	1.00	1.00	1.00
...has 5 daily meals	1.00	0.99	1.00	0.99
...eats vegetables	0.15	0.10	0.17	0.04
...eats fruit	0.12	0.09	0.14	0.03
...afternoon snacks.	0.67	0.58	0.69	0.36
...was breastfed	0.19	0.13	0.21	0.06
...practices sports	0.80	0.76	0.81	0.64
...watches TV	1.00	1.00	1.00	1.00
...plays electronic	0.16	0.11	0.18	0.04
...sleeps	0.19	0.13	0.22	0.06
Daily candy consumption	0.97	0.96	0.97	0.92
$E(W \mid \boldsymbol{y})$	11.5	11	11.7	9.9
$\sqrt{V(W \mid \boldsymbol{y})}$	1.2	1.1	1.2	1.1

TABLE 1.4
Posterior inclusion probabilities and summaries of the posterior distribution of the model size, W, for different prior inputs.

Nevertheless, results dramatically change if instead of $\theta \sim \mathcal{U}(0,1)$ we use the prior with a fixed probability $\theta = 1/2$. In this case, if singular models are considered (with a unitary Bayes factor) then the posterior distribution tends to essentially mimic the prior distribution with all genes having posterior inclusion probabilities very close to 0.5. Results are not more satisfactory if, still using $\theta = 1/2$, singular models are given zero probability or if, for instance, only models with a number of regressors up to certain fixed value are given non null prior probability. Any of these assignments leads to a posterior distribution that strongly concentrates on the largest possible dimension (showing a clear dependence on the prior) and without identifying any sensible model (all posterior inclusion probabilities again being approximately 0.5). In this problem, with such a large p, multiplicity correction is of crucial importance and none of these prior assignments, based on fixed $\theta = 1/2$, provides any such control, letting the posterior distribution concentrate where there are more models leading to useless results. A different path would be a a prior inducing strong sparsity, strongly favoring models with few regressors through for instance the proposal in [25] with a mean model size $w^\star << p$. Although the original motivation for using such prior is sparsity, it seems to work well in practice although without explicitly addressing the issue of multiplicity. Furthermore, and unlike for $\theta \sim \mathcal{U}(0,1)$, the prior input $w^\star$ has to be specified and results depend critically on its assumed value.

1.3 Non-Gaussian variable selection

Consider a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ with components Y_i being independent and with a distribution in the exponential family ([29]):

$$Y_i \sim f(y_i \mid \theta_i, \phi) = \exp\Big\{ \frac{y_i\theta_i - b(\theta_i)}{\phi/\omega_i} + c(y_i, \phi/\omega_i) \Big\},$$

where $b(\cdot)$ and $c(\cdot)$ are functions that specify a distribution for the random variables. The mean of Y_i is $\mu_i = b'(\theta_i)$. This defines a wide family of models denoted by generalized linear models (GLMs). In our covariate selection context, each $\boldsymbol{\gamma} \in \mathcal{M}$ assumes that

$$g(\mu_i) = \alpha + \sum_{j=1}^{p} \gamma_j \, x_{ij}\beta_j, \ \forall i = 1, 2, \ldots, n$$

where $g(\cdot)$ is the link function. In the special case of a linear Gaussian model, the link function is the identity function.

The implementation of g-priors in GLMs is less advanced and in particular there is no consensus about which proposal for the prior covariance matrix in (1.8) (see the Appendix) best generalizes the arguments in the linear model case. There are various such matrices that can be inspired by the information in $\boldsymbol{y}$ about $\boldsymbol{\beta}_\gamma$ and the expected Fisher information matrix (which would be the obvious candidate) cannot be used directly since it depends on the parameters $\boldsymbol{\beta}_\gamma$ themselves. We refer the reader to [34] and [26] (and references therein) for a detailed exposition of the different possibilities. The proposal in [34] is a natural extension of g-priors to non-Gaussian models. The authors propose using the expected information matrix for $\boldsymbol{\beta}_\gamma$ evaluated at $(\alpha, \boldsymbol{\beta}_\gamma) = (0, \mathbf{0})$ (imposing a similar prior scheme as in the normal linear model in which $\boldsymbol{\beta}_\gamma$ does not depend on α).

Another source of concern is the form assumed for the prior for the common parameters. The resulting joint prior is improper and hence it is not guaranteed that the prior predictive

marginal exists (or equivalently that the posterior is proper) for all models in $\mathcal{M}$. As far as we know, there is no general result that ensures that this condition holds and, given the substantial mathematical differences within the GLM class, conditions under which posterior propriety holds must be checked on a case-by-case basis. In our applications the response y_i is Bernoulli, with the intercept α as the only common parameter. As a theoretical contribution of this chapter, the Appendix shows that for this likelihood and under very mild conditions, the use of a constant prior in (1.4) for the common parameters and in (1.8) for $\boldsymbol{\beta}_\gamma$ with fixed g ensures the existence of the posterior for the usual link functions.

The discussion about the prior over the model space, $p(M_\gamma)$, in Section 1.2.1 remains valid in GLM variable selection as it is fully independent on the statistical models entertained. Finally, the expression for the Bayes factor $B_\gamma(\boldsymbol{y})$ does not have a closed form and it has to be computed with numerical methods, usually based on Laplace integration. This idea was already proposed in [31].

1.3.1 `glmBfp` and applications

The package `glmBfp` is distributed as accompanying software for the proposals in [34]. Its main command is `glmBayesMfp` which provides a user-friendly interface since its usage is similar to the base command `glm` to fit GLMs. Unfortunately, `glmBfp` seems to be in an early phase of development (current version is 0.0.60) and its ability to explore the results is limited. We expect that more functionalities will be incorporated in the near future. Concerning the prior inputs, in our examples we use the unit information prior for g (see Table 1.1). For $p(M_\gamma)$ we use $\theta \sim \mathcal{U}(0,1)$ as in [35], assuming that the prior probability for rank deficient models is zero. The code is provided as supplementary material.

Variable selection in logit models with moderate p

As a first example we will consider again the OBICE study ([41]) about child obesity but now the dependent variable is the indicator of whether the child was classified as obese ($y_i = 1$) or not ($y_i = 0$) recorded by the pediatrician. In this situation, y_i follows a Bernoulli distribution (a member of the exponential family) with the probability of success as the mean μ_i. It is well known that in this model $\phi = 1$ and $\omega_i = 1$ and we will employ the logit link function $g(\mu_i) = \log\{\mu_i/(1-\mu_i)\}$ which corresponds to the canonical link function.

The potential explanatory variables are the same as in Subsection 1.2.3 plus age (which now is not fixed). This makes a total of $p = 16$ variables and exhaustive enumeration is still feasible (taking less than 5 minutes to run). Summaries of the posterior distribution in the form of inclusion probabilities and the model which has the highest posterior probability are displayed in Table 1.5.

The results are along the lines of those obtained in Table 1.2 but with interesting differences. In general, the posterior distribution points to simpler models with fewer explanatory variables influencing the response. In this approach, where the obesity condition is to be explained (and not the BMI as before), the weight and height at birth lose their explanatory capacity while the family genetics remain key variables. Among the habits, afternoon snacks and the practice of sports are no longer important determinants for being an obese child, a role that is now assumed by consuming five daily meals, the intensity of watching TV and daily candy intake.

Variable selection in probit models with large p

In our last applied example, we analyze the arthritis data in [36] which concern $n = 31$ patients with rheumatoid arthritis and osteoarthritis and $p = 755$ gene expression measurements. This dataset has been used to study the influence of the prior assignments in [22] in

TABLE 1.5
The dependent variable is the indicator of obesity. The table contains posterior inclusion probabilities and indicates which variables belong to the highest posterior probability model (HPM).

Variable	Inc. Prob	HPM
Weight at birth	0.11	
Height at birth	0.06	
Sex	0.10	
The father is obese	1.00	Y
The mother is obese	1.00	Y
Age	0.12	
The child (regularly):		
...has 5 daily meals	0.87	Y
...eats vegetables	0.03	
...eats fruit	0.07	
...consumes afternoon snacks	0.16	
...was breastfed	0.03	
...practices sports	0.18	
Daily hours the child:		
...watches TV	0.86	Y
...plays with electronic devices	0.03	
...sleeps	0.03	
Daily candy consumption	1.00	Y

a context similar to ours. The response variable is a binary variable (the indicator of having the disease) and the link function is the probit function.

We draw 10^6 simulations from the posterior distribution taking approximately 47 minutes. We also run two other independent chains as a check on the reliability of the results. We find that the great majority of genes do not have any impact on the classification of the disease and only variables $V290$ and $V258$ have some role as determinants: their inclusion probabilities are 0.31 and 0.29, while all the others are smaller than 0.1. It is also relevant that the highest posterior probability model (with a probability of 0.13) is the one with only $V290$ followed very closely by the one with only $V258$ (probability of 0.12). These results are in agreement with [22] who also found these genes to be the most relevant.

1.4 Conclusion

Variable selection in regression models is a pervasive problem that occurs in a very wide variety of applied fields. This chapter focuses on principled Bayesian methods based on Bayes factors in the context of g-prior structures. Through empirical examples, we illustrate the ease of implementation of these methods using freely available `R` packages for both normal linear models and generalized linear models. We analyse applications with the number of possible covariates ranging from 15 to 4088 and show that reliable inference can be obtained quite rapidly with standard computing equipment. A rich tapestry of possible questions can then be answered and the results are easily interpretable. We also highlight that prior assumptions often have an important effect on the results, and we recommend robustifying

the prior structures through priors on hyperparameters. In the Appendix we prove posterior propriety for commonly used generalized linear models.

Acknowledgments

The work of the first author has been partially funded by grant PID2019-104790GB-I00 from the Spanish Ministerio de Ciencia e Innovación and by grant SBPLY/17/180501/000491 from the Consejería de Educación, Cultura y Deportes de la Junta de Comunidades de Castilla-La Mancha.

Appendix

Theorem 1 *Consider the problem of variable selection within the generalized linear model where Y_i follows a Bernoulli distribution. Suppose that i) the link function $g(\cdot)$ is either the probit, the logit or the log-log function; ii) not all observed y_i are equal and iii) the matrix of covariates in the full model $\boldsymbol{X}$ is of full rank. If the prior assumed for M_γ is*

$$p_\gamma(\boldsymbol{\beta}_\gamma, \alpha) \propto N_{p_\gamma}(\boldsymbol{\beta}_\gamma \mid \mathbf{0}_{p_\gamma}, g\Sigma_\gamma), \tag{1.8}$$

where g is a fixed scalar and Σ_γ is a positive definite matrix, then the marginal $m_\gamma(\boldsymbol{y})$ exists for every model in $\mathcal{M}$.

Proof

We have to show that the integral

$$\int\int f(\boldsymbol{y} \mid \alpha, \boldsymbol{\beta}) N_p(\boldsymbol{\beta} \mid \mathbf{0}_p, g\Sigma)\, d\alpha\, d\boldsymbol{\beta}$$

is finite (for simplicity the subscript γ has been removed). We partially base our proof on [10]: given the identity in their (4.4) the above integral can be expressed as:

$$\int\int\int 1\{\alpha\boldsymbol{\iota}^\star + \boldsymbol{X}^\star\boldsymbol{\beta} \leq \boldsymbol{u}\}\, N_p(\boldsymbol{\beta} \mid \mathbf{0}_p, g\Sigma)\, d\alpha\, d\boldsymbol{\beta}\, d\boldsymbol{F}(\boldsymbol{u}),$$

where $\boldsymbol{F}(\boldsymbol{u}) = (g^{-1}(u_1), \ldots, g^{-1}(u_n))^T$ (i.e. component by component, the inverse of the link function); $\boldsymbol{\iota}^\star = (z_1, \ldots, z_n)^T$ where $z_i = 1$ if $y_i = 0$ and $z_i = -1$ if $y_i = 1$; $\boldsymbol{X}^\star$ is the matrix with rows $z_i(x_{i1}, \ldots, x_{ip})$ (i.e. $\boldsymbol{X}$ with row i multiplied by $z_i, i = 1, \ldots, n$). Obviously, the above integral equals:

$$\int\int\int 1\{\alpha\boldsymbol{\iota}^\star \leq \boldsymbol{u} - X^\star\boldsymbol{\beta}\}\, N_p(\boldsymbol{\beta} \mid \mathbf{0}_p, g\Sigma)\, d\alpha\, d\boldsymbol{\beta}\, d\boldsymbol{F}(\boldsymbol{u}),$$

and now we apply Lemma 4.1 in [10] to bound the integral over α, resulting in the following upper bound (up to a constant)

$$\int\int ||\boldsymbol{u} - X^\star\boldsymbol{\beta}||\, N_p(\boldsymbol{\beta} \mid \mathbf{0}_p, g\Sigma) d\boldsymbol{\beta}\, d\boldsymbol{F}(\boldsymbol{u}), \tag{1.9}$$

where $||\cdot||$ represents the Euclidean norm and the conditions in Lemma 4.1 apply given our conditions ii) and iii). Finally, we use the triangle inequality to bound (1.9) by the sum

$$\int ||\boldsymbol{u}||\, d\boldsymbol{F}(\boldsymbol{u}) + \int ||X^{\star}\boldsymbol{\beta}||\, N_p(\boldsymbol{\beta} \mid \mathbf{0}_p, g\Sigma) d\boldsymbol{\beta},$$

and both integrals exist because the first moment of $d\boldsymbol{F}(\cdot)$ (for the cases assumed in i)) and that of $N_p(\boldsymbol{\beta} \mid \mathbf{0}_p, g\Sigma)$ exist.

Bibliography

[1] M. J. Bayarri and G. García-Donato. Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, 94(1):135–152, 2007.

[2] M.J. Bayarri, J.O. Berger, A. Forte, and G. García-Donato. Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40:1550–1577, 2012.

[3] M.J. Bayarri and G. García-Donato. Generalization of Jeffreys divergence-based priors for Bayesian hypothesis testing. *Journal of the Royal Statistical Society: Series B*, 70(5):981–1003, 2008.

[4] J. O. Berger, G. García-Donato, M.A. Martinez-Beneito, and V. Peña. Bayesian variable selection in high dimensional problems without assumptions on prior model probabilities. arXiv:1607.02993, July 2016.

[5] J.O. Berger and L.R. Pericchi. *Objective Bayesian Methods for Model Selection: Introduction and Comparison*, volume 38 of *Lecture Notes–Monograph Series*, pages 135–207. Institute of Mathematical Statistics, Beachwood, OH, 2001.

[6] J.M. Bernardo and A.F.M. Smith. *Bayesian Theory*. Chichester: Wiley, 1994.

[7] P. Bühlman, M. Kalisch, and L. Meier. High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and its Applications*, 1:255–278, 2014.

[8] I. Castillo, J. Schmidt-Hieber, and A. van der Vaart. Bayesian linear regression with sparse priors. *Annals of Statistics*, 43:1986–2018, 2015.

[9] I. Castillo and A. van der Vaart. Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *Annals of Statistics*, 40(4):2069–2101, 2012.

[10] M.H. Chen and Q. Shao. Propriety of posterior distribution for dichotomous quantal responses. *Proceedings of the American mathematical society*, 129(1):293–302, 200.

[11] R. Dezeure, P. Bühlman, L. Meier, and N. Meinshausen. High-dimensional inference: confidence intervals, p-values and R-software hdi. *Statistical Science*, 30:533–558, 2015.

[12] A. Etz and E.-J. Wagenmakers. J.B.S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, 32(2):313–329, 2017.

[13] C. Fernández, E. Ley, and M.F. Steel. Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100:381–427, 2001.

[14] A. Forte, G. García-Donato, and M.F. Steel. Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *International Statistical Review*, 86(2):237–258, 2018.

[15] D.P. Foster and E. I. George. The Risk Inflation Criterion for Multiple Regression. *The Annals of Statistics*, 22:381–427, 1994.

[16] G. García-Donato and A. Forte. Bayesian Testing, Variable Selection and Model Averaging in Linear Models using R with BayesVarSel. *The R Journal*, 10(1):155–174, 2018.

[17] G. Garcia-Donato and M.A. Martinez-Beneito. On Sampling strategies in Bayesian variable selection problems with large model spaces. *Journal of the American Statistical Association*, 108(501):340–352, 2013.

[18] E. George and R. McCulloch. Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373, 1997.

[19] H. Jeffreys. *Theory of Probability*. Oxford University Press, 3rd edition, 1961.

[20] R.E. Kass and A.E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[21] R.E. Kass and L. Wasserman. A reference Bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934, 1995.

[22] D. Lamnisos, J.E. Griffin, and M.F.J. Steel. Cross-validation prior choice in Bayesian probit regression with many covariates. *Statistics and Computing*, 22(2):359–373, 2012.

[23] E.E. Leamer. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: Wiley, 1978.

[24] E. Ley and M.F. Steel. Jointness in Bayesian variable selection with applications to growth regression. *Journal of Macroeconomics*, 29(3):476 – 493, 2007. Special Issue on the Empirics of Growth Nonlinearities.

[25] E. Ley and M.F. Steel. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674, 2009.

[26] Y. Li and M. Clyde. Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, 113:1828–1845, 2018.

[27] F. Liang, R. Paulo, G. Molina, M.A. Clyde, and J.O. Berger. Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423, 2008.

[28] Y. Maruyama and E.I. George. Fully Bayes factors with a generalized g-prior. *The Annals of Statistics*, 39(5):2740–2765, 2011.

[29] P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, 1989.

[30] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

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

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

[33] C.P. Robert, N. Chopin, and J. Rousseau. Harold Jeffreys' theory of probability revisited. *Statistical Science*, 24(2):141–172, 2009.

[34] D. Sabanes and L. Held. Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6:387–410, 2011.

[35] 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, 2010.

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

[37] T. van Erven and B. Szabo. Fast exact Bayesian inference for sparse signals in the normal sequence model. *Bayesian Analysis*, 16:forthcoming, 2021.

[38] A. Zellner. On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In A. Zellner, editor, *Bayesian Inference and Decision techniques: Essays in Honor of Bruno de Finetti*, pages 389–399. Edward Elgar Publishing Limited, 1986.

[39] A. Zellner and A. Siow. Posterior odds ratio for selected regression hypotheses. In J. M. Bernardo, M.H. DeGroot, D.V. Lindley, and Adrian F. M. Smith, editors, *Bayesian Statistics 1*, pages 585–603. Valencia: University Press, 1980.

[40] A. Zellner and A. Siow. *Basic Issues in Econometrics.* Chicago: University of Chicago Press, 1984.

[41] O. Zurriaga, J. Perez-Panades, J. Izquiero, M. Gil, Y. Anes, C. Quiñones, M. Margolles, A. Lopez-Maside, A. T. Vega-Alonso, and M.T. Miralles. Factors associated with childhood obesity in spain. the OBICE study: a case–control study based on sentinel networks. *Public Health Nutrition*, 14(6):1105–113, 2011.