Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

On Sampling Strategies in Bayesian Variable Selection Problems With Large Model Spaces

G. García-donato & M. A. Martínez-beneito

To cite this article: G. García-donato & M. A. Martínez-beneito (2013) On Sampling Strategies in Bayesian Variable Selection Problems With Large Model Spaces, Journal of the American Statistical Association, 108:501, 340-352, DOI: 10.1080/01621459.2012.742443

To link to this article: https://doi.org/10.1080/01621459.2012.742443

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

On Sampling Strategies in Bayesian Variable Selection Problems With Large Model Spaces

G. García-Donato and M. A. Martínez-Beneito

One important aspect of Bayesian model selection is how to deal with huge model spaces, since the exhaustive enumeration of all the models entertained is not feasible and inferences have to be based on the very small proportion of models visited. This is the case for the variable selection problem with a moderately large number of possible explanatory variables considered in this article. We review some of the strategies proposed in the literature, from a theoretical point of view using arguments of sampling theory and in practical terms using several examples with a known answer. All our results seem to indicate that sampling methods with frequency-based estimators outperform searching methods with renormalized estimators. Supplementary materials for this article are available online.

KEY WORDS: Bayesian model selection; g-priors; Searching strategies

1. INTRODUCTION

This article is rooted in the model selection problem, that is, with uncertainty surrounding the probabilistic model, which, from an initial set $\mathcal{M}$ of candidates, better explains certain data $\boldsymbol{y}$. In particular, we address the variable selection problem where the competing models differ regarding which subset of variables are to be included as explanatory covariates for $\boldsymbol{y}$.

One special characteristic of the variable selection problem is that $\mathcal{M}$, the model space, easily becomes extremely large. For instance, a problem with $p = 40$ potential covariates has $2^{40} \approx 10^{12}$ different models. Merely the binary representation of such a model space would occupy 5 terabytes of memory. We focus on the difficulties that arise as a consequence of a moderately large size of $\mathcal{M}$. This article does not address the question of what happens when p is very large (hundreds or more). We consider the problem from a Bayesian point of view and the context we use for the development of our ideas is the problem of variable selection in Gaussian linear regression models.

The Bayesian approach to the problem is conceptually straightforward. Any feature of interest, say τ, is a deterministic function of the posterior distribution over the model space. Examples of such features are the highest posterior probability model (HPM), the inclusion probabilities of covariates, or the posterior predictions of a new value for the dependent variable. Unfortunately, three major difficulties arise when putting the Bayesian approach into practice: (1) the choice of the prior distributions; (2) the computation of the integrated likelihood (or equivalently the Bayes' factors) for single models in $\mathcal{M}$; and, in large model spaces, (3) the estimation of τ, since its exact value is virtually unknown due to the size of $\mathcal{M}$. The few but comprehensive articles on these aspects of Bayesian model selection are George and McCulloch (1997), Hoeting et al. (1999), Berger and Pericchi (2001), Chib and Jeliazkov (2001), and Clyde and George (2004). To keep the impact of difficulties (1) and (2) above under control, we use Zellner's (1986) g-priors that produce closed-form Bayes factors. The corresponding formulas are also briefly given in Section 2.

This article is basically concerned with (3), which is intimately related to strategies for exploring the model space, that is, visiting a very reduced number of models (hereafter denoted $\mathcal{M}^*$), since covering the whole model space is unfeasible. Our main aim is to shed new light on a topic (almost an implicit debate) that appears in the literature from time to time (see references). The subject concerns the estimation of τ and more precisely whether it should be based on the empirical distribution (observed frequencies in $\mathcal{M}^*$) or on the renormalized analytical expression of Bayes factors (a concise definition given later in this article) of models in $\mathcal{M}^*$. In the first approach, until quite recently the general one, models in $\mathcal{M}^*$ are a Markov chain Monte Carlo (MCMC) sample with stationary distribution being the posterior probabilities over the model space. In the second approach, the emphasis is normally placed on visiting, usually without replacement, models with high posterior probability (Berger and Molina 2005). In the rest of the article, for ease of comprehension, we, respectively, refer to *empirical* and *renormalized* for each of the approaches outlined above. Renormalization was originally introduced to estimate parameters defined over a much smaller subset of $\mathcal{M}$, like Occam's Window in Madigan and Raftery (1994), Clyde, DeSimone, and Parmigiani (1996), and Raftery et al. (1997), although in this article the set of models' target is the whole model space.

The theoretical basis behind these two paradigms to construct estimates in large model spaces is simple but not conclusive enough so as to make one preferable over the other. These bases are revised within a common envelope of sampling theory in Section 3. On the one hand, empirical estimates are shown to be consistent while the renormalized are biased (an alternative study in this direction can also be found in Clyde and Ghosh

G. García-Donato, Departamento de Análisis Económico y Finanzas, Universidad de Castilla La Mancha, Albacete, Spain (E-mail: *Gonzalo.GarciaDonato@uclm.es*). M. A. Martínez-Beneito, Centro Superior de Investigación en Salud Pública (CSISP), Valencia, Spain (E-mail: *miguel.a.martinez@uv.es*); and CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain. We thank Susie Bayarri and Jim Berger for very fruitful discussions about the problem of estimation in large model spaces. We also thank Rafael Espinosa at the Supercomputing Center in the Institute for Research in Information Technology at the Universidad de Castilla-La Mancha for providing us with technical support. We also thank the suggestions of two anonymous reviewers as they have led to a much improved version of the article. This study has been partially funded by a project granted by the Spanish Ministry of Science and Education coded MTM2010-19528.

Journal of the American Statistical Association
March 2013, Vol. 108, No. 501, Review Article
DOI: 10.1080/01621459.2012.742443

2012). On the other hand, it is usually perceived that the frequency of visits in such huge model spaces is a poor basis for estimation since the number of repeated visits (if any) is very small. This appealing and, at first sight, quite convincing reasoning is the starting point for the renormalized approach, where the mathematical expression of Bayes factors serves as the basis for estimation.

For large enough $\mathcal{M}$, either *exact* independent sampling from the posterior distribution or identifying the most probable models is not feasible in practical terms. Hence, different methods to put into practice either the empirical or the renormalized approach have been proposed. For the first approach, the use of MCMC-based methods to obtain a sample of models from the posterior distribution is very widespread (see, e.g., George and McCulloch 1993; Madigan and York 1995; George and McCulloch 1997; Kuo and Mallick 1998; Dellaportas, Forster, and Ntzoufras 2000; Ntzoufras 2002; Nott and Kohn 2005; Casella and Moreno 2006; Ntzoufras 2009). In contrast, the renormalized approach uses specifically designed heuristic methods with the aim of discovering, usually without replacement, the most probable models (see, e.g., Berger and Molina 2005; Hans, Dobra, and West 2007; Carvalho and Scott 2009; Scott and Carvalho 2009; Clyde, Ghosh, and Littman 2011). In Section 4, a number of such methods are compared with the exact values of the parameters of interest in moderate to large datasets.

This article is the result of our work in trying to obtain convincing evidence in favor of one of these approaches. In Section 6, we have brought together the main conclusions extracted from that work.

2. BAYESIAN VARIABLE SELECTION

We consider a model selection problem in the context of normal regression models. More specifically, we let $\boldsymbol{X} = \{x_{ij}\}$ be an $N \times p$ full rank matrix and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ be a p-dimensional vector of binary variables. Denote $k_\gamma = \sum \gamma_i$ and for each $\boldsymbol{\gamma}$, let $\boldsymbol{X}_\gamma$ be the $N \times k_\gamma$ design matrix corresponding to the columns with ones in $\boldsymbol{\gamma}$.

The variable selection problem has 2^p competing models, each proposed as a plausible explanation of an N-dimensional vector $\boldsymbol{Y}$. More concisely

$$M_\gamma : \boldsymbol{Y} \sim N_N(\alpha\mathbf{1} + \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, \sigma^2\boldsymbol{I}), \quad \boldsymbol{\gamma} \in \{0,1\}^p. \tag{1}$$

In this problem, the model space $\mathcal{M}$ can be represented by $\{0,1\}^p$. The simplest model among the proposed ones is $M_0 : \boldsymbol{Y} \sim N_N(\alpha\mathbf{1}, \sigma^2\boldsymbol{I})$.

2.1 Posterior Distribution Over the Model Space

Without loss of generality, posterior probabilities of the models can be expressed as

$$\Pr(M_\gamma \mid \boldsymbol{y}) = C\, B_{\gamma 0} \Pr(M_\gamma), \tag{2}$$

where $B_{\gamma 0}$ is the Bayes factor of M_γ to M_0, $\Pr(M_\gamma)$ is the prior probability of M_γ, and $C^{-1} = \sum_\gamma B_{\gamma 0} \Pr(M_\gamma)$ is the constant of proportionality.

Bayes factors are the ratio of the marginal prior predictive distributions evaluated at $\boldsymbol{y}$, that is, $B_{\gamma 0} = m_\gamma(\boldsymbol{y})/m_0(\boldsymbol{y})$, where

$$m_\gamma(\boldsymbol{y}) = \int N_N(\boldsymbol{y} \mid \alpha\mathbf{1} + \boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma, \sigma^2\boldsymbol{I})\, \pi_\gamma(\alpha, \boldsymbol{\beta}_\gamma, \sigma)\, d\alpha\, d\boldsymbol{\beta}_\gamma\, d\sigma.$$

The function π_γ is the prior distribution for the parameters under model M_γ, which can be expressed, without any loss of generality, as the product $\pi_\gamma(\alpha, \sigma)\pi_\gamma(\boldsymbol{\beta}_\gamma \mid \alpha, \sigma)$. It is well known that the conditional $\pi_\gamma(\boldsymbol{\beta}_\gamma \mid \alpha, \sigma)$ can be neither improper nor *vague* (with a very large variance), since otherwise the resulting Bayes factors would be essentially arbitrary (see Berger and Pericchi 2001). Our default choice for this prior is the g-prior proposed by Zellner (1986), which, jointly with the usual noninformative prior for the common parameters α, σ, leads to the popular proposal:

$$\pi_\gamma(\alpha, \boldsymbol{\beta}_\gamma, \sigma \mid g) = \sigma^{-1} N_{k_\gamma}(\boldsymbol{\beta}_\gamma \mid \mathbf{0}, g\sigma^2 (\boldsymbol{X}_\gamma^t(\boldsymbol{I} - N^{-1}\mathbf{1}\mathbf{1}^t)\boldsymbol{X}_\gamma)^{-1}),$$

for $\boldsymbol{\gamma} \neq \mathbf{0}$ and $\pi_0(\alpha, \sigma) = \sigma^{-1}$ for M_0.

The g-priors seem to be greatly inspired by Jeffreys' (1961) ideas and by the corresponding extension to regression problems in Zellner and Siow (1980, 1984). The assignment of the constant g has been analyzed by several authors (see Liang et al. 2008, and references therein). This parameter g must increase with N to avoid an asymptotically degenerate prior. The default assignment $g = N$ gives rise to a "unit information" prior in the sense that the covariance matrix is corrected by the sample size (see Raftery 1998).

The g-prior provides closed-form expressions for the Bayes factors. In fact, it can easily be shown that

$$B_{\gamma 0}(g) = \left(1 + g\frac{\mathrm{SSE}_\gamma}{\mathrm{SSE}_0}\right)^{-(N-1)/2} (1+g)^{(N-k_\gamma-1)/2}, \tag{3}$$

where SSE_γ is the sum of the squared errors for M_γ. Therefore, if all models $\boldsymbol{\gamma}$ can be visited, their posterior probabilities can be calculated without great computational effort.

An alternative to the choice of the constant g is to assume, hierarchically, a proper prior on g, say $\pi(g \mid \boldsymbol{\gamma})$. In general, the resulting prior for $\boldsymbol{\beta}_\gamma$ has heavy tails, this being an appealing characteristic of a model selection prior, which is related to properties like information consistency (Bayarri and García-Donato 2008). Examples of such priors are the multivariate Cauchy for $\boldsymbol{\beta}_\gamma$ proposed by Jeffreys–Zellner–Siow (Jeffreys 1961; Zellner and Siow 1980, 1984), which corresponds to $\pi(g \mid \boldsymbol{\gamma}) = \mathrm{Gamma}^{-1}(1/2, N/2)$, the hyper-$g$-priors of Liang et al. (2008), the conventional robust prior in Forte (2011), and the extension of the g-prior in Maruyama and George (2011). Interestingly, the proposals in Forte (2011) and Maruyama and George (2011) also lead to closed-form Bayes' factors.

2.2 Summaries of the Posterior Distribution

As in any other Bayesian procedure, the posterior distribution over the model space (2) contains all the information concerning the problem under study. Nevertheless, and depending on the specific context of the study, researchers mainly focus on certain summaries τ of the posterior distribution. Obviously, these quantities are deterministic functions of the posterior distribution. Examples of such summaries are given below.

One relevant question in model selection is which model is the most probable in the light of the data (the HPM), that is,

$$\tau = \mathrm{HPM} = \arg\max_{\gamma \in \mathcal{M}} \Pr(M_\gamma \mid \boldsymbol{y}). \tag{4}$$

Notice that in the determination of the HPM, there is no explicit contribution of the rest of the models in $\mathcal{M}$. This is a very special aspect of the HPM, which is not shared by many other quantities of interest. In fact, most such quantities have the following expression in terms of an expectation over the posterior distribution

$$\tau(a) = E_{Pr(\cdot|\mathbf{y})}(a(M_\gamma)) = \sum_{\gamma\in\mathcal{M}} a(M_\gamma)\Pr(M_\gamma \mid \mathbf{y}), \quad (5)$$

where $a(M_\gamma)$ is a known function of M_γ. Note that $\tau(a)$ is the population total of the variable $a(M_\gamma)\Pr(M_\gamma \mid \mathbf{y})$. For example, the posterior probability of a single model $M_{\gamma*}$ can be expressed as (5) with $a(M_\gamma) = 1$ if $M_\gamma = M_{\gamma*}$, and 0 otherwise. Other important parameters of interest with the above representation are as follows:

2.2.1 Inclusion Probabilities and Median Probability Model. For a given explanatory variable x_l, the inclusion probability is defined as

$$q_l = \sum_{\gamma:\,\gamma_l=1} \Pr(M_\gamma \mid \mathbf{y}).$$

These probabilities have interesting theoretical properties as shown in Barbieri and Berger (2004) and are useful summaries of the posterior distribution. In particular, they can be helpful when the number of models is very large and the posterior probabilities of single models are essentially useless. Apart from their intrinsic interest, inclusion probabilities are the basis of the median probability model in Barbieri and Berger (2004). This model, hereafter called MPM, is defined as the one having variables with $q_l > 0.5$ and the theory in Barbieri and Berger (2004) suggests that the MPM model has optimal properties, being even better for prediction purposes than the HPM (a surprising fact). The probability, q_l can be expressed as (5) with $a(M_\gamma) = \gamma_l$.

2.2.2 Dimension of the "True" Model. The probability of the "true" model having exactly k^* explanatory covariates is

$$d(k^*) = \sum_{\gamma:\,k_\gamma=k^*} \Pr(M_\gamma \mid \mathbf{y}).$$

This corresponds to the expression in (5) with $a(M_\gamma) = 1$ if $k_\gamma = k^*$, and 0 otherwise.

2.2.3 Model Averaged Estimates and Predictions. Suppose that Δ is a quantity of interest, then its posterior distribution $\Pr(\Delta \mid \mathbf{y})$ corresponds to (5) with $a(M_\gamma) = \Pr(\Delta \mid \mathbf{y}, M_\gamma)$.

What arises is, of course, the methodology called model averaging, which is just the Bayesian way of accounting for the uncertainty regarding which the true model is (see, e.g., Hoeting et al. 1999).

Special mention should be made of the case where Δ is a future observable y^{new}, given certain values of the explanatory covariates $\boldsymbol{x}^{\text{new}}$. In this case, the distribution $\Pr(y^{\text{new}} \mid \mathbf{y})$ is the posterior predictive distribution and summaries of this distribution are special cases of (5), like the posterior predictive expectation $E(y^{\text{new}} \mid \mathbf{y}, \boldsymbol{x}^{\text{new}})$ with $a(M_\gamma) = E(y^{\text{new}} \mid M_\gamma, \mathbf{y}, \boldsymbol{x}^{\text{new}})$. It can be easily seen that, with the g-prior

$$E(y^{\text{new}} \mid M_\gamma, \mathbf{y}, \boldsymbol{x}^{\text{new}}) = \hat{\alpha} + \hat{\boldsymbol{\beta}}_\gamma^t \left(\frac{1}{N(g+1)}\boldsymbol{X}_\gamma^t\mathbf{1} + \boldsymbol{x}_\gamma^{\text{new}}\right),$$

where $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}_\gamma$ are the corresponding maximum likelihood estimators under model M_γ. This expression will be used to calculate posterior predictive expectations in Section 4.

3. SAMPLING-BASED STRATEGIES IN LARGE MODEL SPACES

The problem of inferring about τ in large model spaces can be seen as a sampling problem derived from the unfeasibility of covering the whole model space. Therefore, τ needs to be estimated, based on a sample of say n models $\mathcal{M}^* = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$, where n is very small compared with 2^p, the size of $\mathcal{M}$. From now on n will be the number of models in $\mathcal{M}^*$ that may contain replicates. It will be evident in the sequel that, under certain strategies, it is preferable not to repeat sampled models.

The goodness of an estimator depends heavily on the sampling mechanism followed to obtain $\mathcal{M}^*$. For instance, the obvious estimator of the HPM (see (4)) is simply

$$\arg\max_{\gamma\in\mathcal{M}^*}\Pr(M_\gamma \mid \mathbf{y}) = \arg\max_{\gamma\in\mathcal{M}^*}\Pr(M_\gamma)\, B_{\gamma 0}.$$

This equality holds because posterior probabilities are proportional to the Bayes factors times the prior probabilities. In this case, the estimator does not depend on the proportionality constant C and the success of the estimation relies *strictly* on the ability of the sampling method to discover good models. Unfortunately, such a dependency becomes crucial in the estimation of many other summaries and in particular of quantities expressible as $\tau(a)$ in (5), for which estimation based only on a sample of good models is not necessarily an optimal strategy. Note in passing that in this situation, the estimator of $\tau(a)$ will generally depend on C, the unknown proportionality constant, through the dependence of this parameter on $\Pr(M_\gamma|\mathbf{y})$.

There have been few attempts in the literature to approximate C (see George and McCulloch 1997) to estimate $\tau(a)$. Nevertheless, the most popular strategies followed for inferring about $\tau(a)$ do not rely on an explicit estimation of C, but are based on estimators of $\tau(a)$ whose mathematical expression does not depend on this quantity. These estimators, plus a specific sampling plan to obtain $\mathcal{M}^*$, give rise to what we called in Section 1 renormalized and empirical approaches. Both can be properly understood from a sampling theory viewpoint as we describe in the next two sections. In sampling language (see, e.g., Thompson 2002), the renormalized approach is strongly related with the *ratio* estimator and the empirical approach with the Hansen–Hurwitz estimator.

3.1 Ratio Estimator

In sampling theory, ratio estimators are proposed to efficiently estimate population totals by taking advantage of some available auxiliary information (Thompson 2002). Translated to our problem and making use of $\Pr(M_\gamma|\mathbf{y})$ as the auxiliary variable, the ratio estimator is

$$\hat{\tau}_R(a) = \frac{\sum_{\gamma\in\mathcal{M}^*} a(M_\gamma)\Pr(M_\gamma|\mathbf{y})}{\sum_{\gamma\in\mathcal{M}^*}\Pr(M_\gamma|\mathbf{y})} = \frac{\sum_{\gamma\in\mathcal{M}^*} a(M_\gamma)\Pr(M_\gamma)B_{\gamma 0}}{\sum_{\gamma\in\mathcal{M}^*}\Pr(M_\gamma)B_{\gamma 0}},$$

which, remarkably, does not depend on C. This functional form, in combination with strategies for sampling good models, defines what we called the renormalized approach. The term "renormalized" comes from the fact that the denominator in $\hat{\tau}$ acts as a renormalizer over the sample models.

Ratio estimators are generally biased, although they tend to be very precise in the sense of a small mean squared error, particularly when the auxiliary variable and the variable of interest ($a(M_\gamma)\Pr(M_\gamma|\boldsymbol{y})$ here) are approximately proportional (Thompson 2002). If the proportionality assumption held reasonably (in our problem this would imply $\Pr(M_\gamma|\boldsymbol{y}) \approx \kappa\, a(M_\gamma)\Pr(M_\gamma|\boldsymbol{y})$ for a certain constant κ, independent of γ), Theorem 4 in Royall (1970) would prove that the sample $\mathcal{M}^*$ that leads to the most accurate estimate $\hat{\tau}_R(a)$ consists precisely of those models having the largest posterior probability. This would give full theoretical support to the renormalized approach, that is, sampling the best n models without replacement. Also relevant is that the violations on the proportionality assumption (something that it is expected to happen in the model selection context) provide an explanation of the biased results we encountered in practice (fully described in Section 4).

Clyde and Ghosh (2012) gave a formal expression for the bias in $\hat{\tau}_R(a)$ as a function of the probabilities that models are included in $\mathcal{M}^*$. Nevertheless, renormalized methods are normally put into practice in nonprobabilistic samplings (see, e.g., Berger and Molina 2005; Carvalho and Scott 2009; Scott and Carvalho 2009; Clyde, Ghosh, and Littman 2011). For them it is, thus, more useful to have a quantification of estimation error, which is provided in the following result.

Result 1. Let $\bar{\mathcal{M}}^*$ denote the complement set of $\mathcal{M}^*$. If $\mathcal{M}^*$ does not contain repetitions, then

$$\hat{\tau}_R(a) - \tau(a) = \frac{\sum_{\gamma\in\mathcal{M}^*}\sum_{\delta\in\bar{\mathcal{M}}^*}\Pr(M_\gamma \mid \boldsymbol{y})\Pr(M_\delta \mid \boldsymbol{y})\big(a(M_\gamma) - a(M_\delta)\big)}{\sum_{\gamma\in\mathcal{M}^*}\Pr(M_\gamma \mid \boldsymbol{y})}.$$

Proof. See Appendix B. □

According to the above result, the error sign in the estimation depends only on $a(M_\gamma) - a(M_\delta)$, where $M_\gamma \in \mathcal{M}^*$ and $M_\delta \in \bar{\mathcal{M}}^*$. Notice that if $a(\cdot)$ and $\mathcal{M}^*$ are related in such a way that $a(M)$ has a tendency to give larger results (or vice versa) when M is in $\mathcal{M}^*$ than when it is not, then we can expect systematic errors in the estimation of $\tau(a)$. We will see a clear manifestation of the above result in the section devoted to the examples.

3.2 Hansen–Hurwitz Estimators

In sampling with replacement of a finite population, Hansen and Hurwitz (1943) proposed a type of unbiased estimators under unequal probability sampling. In our problem, if models in $\mathcal{M}^*$ have been independently sampled with replacement and the probability that $M_\gamma \in \mathcal{M}^*$ is proportional to a certain probability Π_γ, then the Hansen–Hurwitz estimator of $\tau(a)$ is

$$\frac{1}{n}\sum_{\gamma\in\mathcal{M}^*}\frac{a(M_\gamma)\Pr(M_\gamma \mid \boldsymbol{y})}{\Pi_\gamma}.$$

If models in $\mathcal{M}^*$ are sampled from the posterior distribution over the model space, then $\Pi_\gamma = \Pr(M_\gamma \mid \boldsymbol{y})$ (a so-called probability proportional to size sampling or PPS, see Lohr 1999) and the above estimator becomes simply

$$\hat{\tau}_p(a) = n^{-1}\sum_{\gamma\in\mathcal{M}^*} a(M_\gamma),$$

regardless of the value of the proportionality constant C. That is, the former estimator unseemingly integrates information about the probability of the models sampled. The resulting strategy of PPS sampling plus the Hansen–Hurwitz estimator $\hat{\tau}_p(a)$ provides a justification for the empirical approach from within the sampling theory point of view. Note that in the case of estimating the probability of a single model M_{γ^*} ($a(M_\gamma) = 1$ if and only if $\gamma = \gamma^*$), $\hat{\tau}_p(a)$ will be just the relative frequency of that model.

From a practical point of view, the main problem with the empirical approach is how to carry out a PPS sampling of models. Generating independent draws from that sampling scheme in huge model spaces is usually a nontrivial task. Therefore, the main solution used in practice to solve this problem is to generate dependent samples by means of MCMC techniques. A great majority of these proposals are to a certain extent based on the seminal work by George and McCulloch (1993), greatly improved and extended by George and McCulloch (1997). A number of interesting contributions in this area are Kuo and Mallick (1998), Dellaportas, Forster, and Ntzoufras (2000), Nott and Kohn (2005), Ntzoufras (2002), Ntzoufras (2009), and Casella and Moreno (2006).

As we show in the following result, $\hat{\tau}_p(a)$ estimators obtained from an MCMC sample of models from the posterior distribution are, under very mild conditions, consistent.

Proposition 1. Let $\{\boldsymbol{\gamma}^{(i)}\}$ be an irreducible Markov chain over the state space $\mathcal{M}$ with stationary distribution $\Pr(M_\gamma \mid \boldsymbol{y})$, and let $a()$ be any function defined on $\mathcal{M}$. Also define,

$$\hat{\tau}_p^{(n)}(a) = n^{-1}\sum_{i=1}^{n} a(M_{\gamma^{(i)}}).$$

Then $\sqrt{n}(\hat{\tau}_p^{(n)}(a) - \tau(a))$ converges weakly to a normal distribution of zero mean and variance $\sigma^2(a)$ for any initial distribution.

Further, if $a()$ is such that $\sum_{k=1}^{\infty}\mathrm{cov}_k(a) < \infty$, where $\mathrm{cov}_k(a)$ is the lag-k autocovariance of $a()$, then

$$\sigma^2(a) = \mathrm{cov}_0(a) + 2\sum_{k=1}^{\infty}\mathrm{cov}_k(a).$$

Proof. See Appendix B. □

The sufficient conditions in Proposition 1 are not strong at all, but a key fact in the proof is that the state space is finite. That is, if the regression parameters $\boldsymbol{\beta}, \sigma$ could not be integrated out, the previous result would not necessarily be true. In our opinion, this is an important issue to guarantee a good performance of the MCMC. In particular, the algorithm used in the examples, fully detailed in Appendix A, completely satisfies those conditions. The former result proves the consistency (and therefore, the asymptotic unbiasedness) of $\hat{\tau}_p^{(n)}(a)$ as an estimator of $\tau(a)$. These are appealing properties of $\hat{\tau}_p^{(n)}(a)$ that the ratio estimator did not fulfill.

Geyer (1992) proposed several ways of consistently estimating the variance in the proposition above. These ways are mainly based on the restriction of the sum in the above expression to a finite number of summands, leaving out the covariance terms mainly responding to random noise according to different criteria (Geyer 1992). This way, although MCMC methods generate dependent samples of models $\{\boldsymbol{\gamma}^{(i)}\}$, it is still possible to derive a measure of uncertainty on the estimation of $\tau(a)$ by taking into account the dependence on the sample drawn. The following expression will be used repeatedly in the examples to quantify the uncertainty on the different quantities of interest studied:

$$\hat{V}(\hat{\tau}_p(a)) = \widehat{V}(a(M_\gamma)) + 2\sum_{k=1}^{m} \widehat{\mathrm{cov}}_k(a(M_\gamma)). \tag{6}$$

The $\widehat{V}$ and $\widehat{\mathrm{cov}}_k$ above are the sample variance and the lag-k autocovariance of the sequence $a(M_{\boldsymbol{\gamma}^{(i)}})$, and m is an upper bound determined following the *initial positive sequence estimator* in Geyer (1992).

Another interesting result is the following. Independently of the approach adopted to sample the models in $\mathcal{M}^*$ under the renormalized approach, the errors in the estimation of $\tau(a)$ with the ratio and the Hansen–Hurwitz estimator are related through the following simple identity:

Result 2.

$$\hat{\tau}_R(a) - \tau(a) = \frac{\widehat{\mathrm{cov}}(a(M_\gamma), \Pr(M_\gamma \mid \mathbf{y}))}{n^{-1}\sum_{\gamma\in M^*} \Pr(M_\gamma \mid \mathbf{y})} + \hat{\tau}_p(a) - \tau(a),$$

where $\widehat{\mathrm{cov}}(,)$ stands for observed covariance.

Proof. See Appendix B. □

The error in the ratio estimator can be expressed as the sum of two terms: the sample covariance between $a(M_\gamma)$ and the posterior probabilities of models $P(M_\gamma \mid \mathbf{y})$, and the error in the Hansen–Hurwitz estimator. The first of these terms will be substantially different from 0 in the case that some values of $a(M_\gamma)$ are more frequent for those models with a higher probability. The estimation of the inclusion probability of a variable with a high (respectively, low) probability of being present in the model is a clear example of this, as $a(M_\gamma)$ will be mainly 1 (respectively, 0) for those models with a high (respectively, low) posterior probability. This seems to be quite a general situation, therefore, this term is expected to have a nonnegligible effect on the error of the ratio estimator.

For the ratio estimator, this second term may have 0 as its expected value, for example, under simple random sampling, but it is not guaranteed to be 0 in general and it may be heavily influenced by the sampling mechanism. As a consequence, when only high probability models are visited, it could have a considerable impact on the estimation error.

4. A COMPARISON IN PRACTICE

In the previous section, we have seen that, in theory, estimations under the empirical approach are consistent, while the renormalized approach produces very precise estimations, especially if the hypothesis of proportionality between $a(M_\gamma)$ and $a(M_\gamma)\Pr(M_\gamma|y)$ holds approximately. Nevertheless, it is usually perceived that the large size of the model space precludes the appealing theoretical properties of empirical estimates being translated into good results in practice.

If such an argument is seen as plausible, the only way it seems convincing to us to elucidate which approach is better is through the application of the methodologies in specific datasets. In this section, we present, in the form of empirical studies, the results of applying the renormalized and empirical approaches in moderately large datasets. The results presented here clearly support empirical estimates and highlight that they are barely affected by the size of $\mathcal{M}$. On the contrary, renormalized estimates do not scale properly with the size of $\mathcal{M}$.

In our analysis, we work on three different aspects of the problem, giving rise to the three examples introduced. The datasets analyzed were taken from the literature (references provided in the sections) trying to guarantee, as much as possible, the fairness of comparisons.

In Section 4.1, we consider a clean comparison between paradigms. That is, we analyze the question of which strategy (empirical or renormalized) gives more accurate results in the ideal scenario where both the set with the n most probable models and the set with an exact independent and identically distributed (iid) sample of size n from the posterior distribution are available. Interest in this question predates comparisons between specific methods and tries to answer the relevant question of whether inferences in large model spaces should focus on optimal ways of sampling from the posterior or on finding the most probable models. That is, our intention with this example is to compare both empirical and renormalized paradigms instead of specific methods, which put these paradigms into practice.

A second question is to know what the behavior of specific methods is to put into practice their corresponding paradigms. This is analyzed in Section 4.2 where we face the more realistic question of comparing methods. In particular, we analyze the Gibbs sampling in George and McCulloch (1997) as representative of the empirical approach; the Bayesian adaptive sampling of Clyde, Ghosh, and Littman (2011) and the stochastic search of Berger and Molina (2005) as representatives of the renormalized approach.

In the above two examples, we put considerable emphasis on computing exactly the bias of the estimates in problems with large model spaces. It is generally understood that problems with p larger than 25–30 are intractable (Clyde, Ghosh, and Littman 2011), and we are considering a step beyond this limiting size (the dataset in Section 4.1 has $p = 30$ and the one in Section 4.2 has $p = 35$). Of course, computing the bias implies knowing the exact values of the parameter of interest and hence, an involved computational challenge. To tackle this problem, we wrote a specific optimized code in ansi-C, making use of the gsl library (Galassi et al. 2009), which was parallelized in a cloud with 150 cores. It took approximately 12 min to run the problem in Section 4.1 (which has $p = 30$) and 20 hr for the problem in Section 4.2 (with $p = 35$). The source code is available upon request.

Third, in Section 4.3, we analyze a larger problem (with $p = 65$), where the exact values of the parameters of interest are unknown. Because of the impossibility of assessing the bias exactly, we study the results in terms of behavior under repeated runs of the methods analyzed.

With respect to the parameters of interest, we focus on τ being the HPM, the inclusion probabilities q_l, the MPM and expected

posterior predictions of an outcome of the dependent variable:

$$\mu_j = E\big(y^{\text{new}} \mid \boldsymbol{y}, \boldsymbol{x}^{\text{new}} = j\text{th row of } \boldsymbol{X}\big),$$

for certain j. Finally, for all the examples we use the g-prior with $g = N$ and a constant prior for the prior probabilities of models.

4.1 An Exact Comparison Between Paradigms

We use a dataset generated following the simulation scheme in Kuo and Mallick (1998), where the interest was also to compare methods in large model spaces. In particular, these data have a design matrix $\boldsymbol{X}$ with $p = 30$ explanatory variables $\boldsymbol{x}_j$, each with $N = 60$ observations, where $\boldsymbol{x}_j = \boldsymbol{x}_j^* + \boldsymbol{z}$ and $\boldsymbol{x}_1^*, \ldots, \boldsymbol{x}_{30}^*, \boldsymbol{z}$ independently follow an $N_N(\boldsymbol{0}, \boldsymbol{I})$. As Kuo and Mallick (1998) noted, this scheme induces a pairwise correlation of about 0.5. The dependent variable is generated as $\boldsymbol{Y} = \boldsymbol{1} + [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{30}]\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_N(\boldsymbol{0}, 4\boldsymbol{I})$ and

$$\boldsymbol{\beta}^t = (0, \ldots^{10)}, 0, 1, \ldots^{10)}, 1, 2, \ldots^{10)}, 2).$$

This and subsequent datasets are provided as annex material to the article. The code used for this and the rest of the examples is available on request from the authors.

The exact inclusion probabilities q_l, for $l = 1, \ldots, 30$ are shown in Figure 1(a). As expected, in general, these are small for $l = 1, \ldots, 10$, higher for $l = 11, \ldots, 20$, and attain their maximum levels for $l = 21, \ldots, 30$. In particular, the inclusion probabilities of covariates $\{21, 22, 27, 28, 29, 30\}$ are very close to one. The expected value of the posterior predictive distribution associated with the first three design points are $\mu_1 = 44.3$, $\mu_2 = -13.6$, and $\mu_3 = -11.2$.

To compare paradigms, we calculated the renormalized estimators corresponding to the most n probable models and the empirical estimates of an iid sample of n models from the posterior distribution (using a time-consuming importance sampling, with the importance function being the posterior probability of the HPM model). The observed biases $\hat{\tau} - \tau$ for $n = 1000$, $n = 5000$, and $n = 10{,}000$ in the predictions are represented in Figure 1(b), while those for the inclusion probabilities are in Figures 1(c) and 1(d).

The results clearly demonstrate that the renormalized strategy produces biased results, especially in the estimation of the inclusion probabilities. Furthermore, there is a strong tendency in the inclusion probabilities: the error is negative when q_l is small and positive otherwise (cf. Figures 1(a) and 1(c)). This is a clear manifestation of Result 1 since the models sampled (they

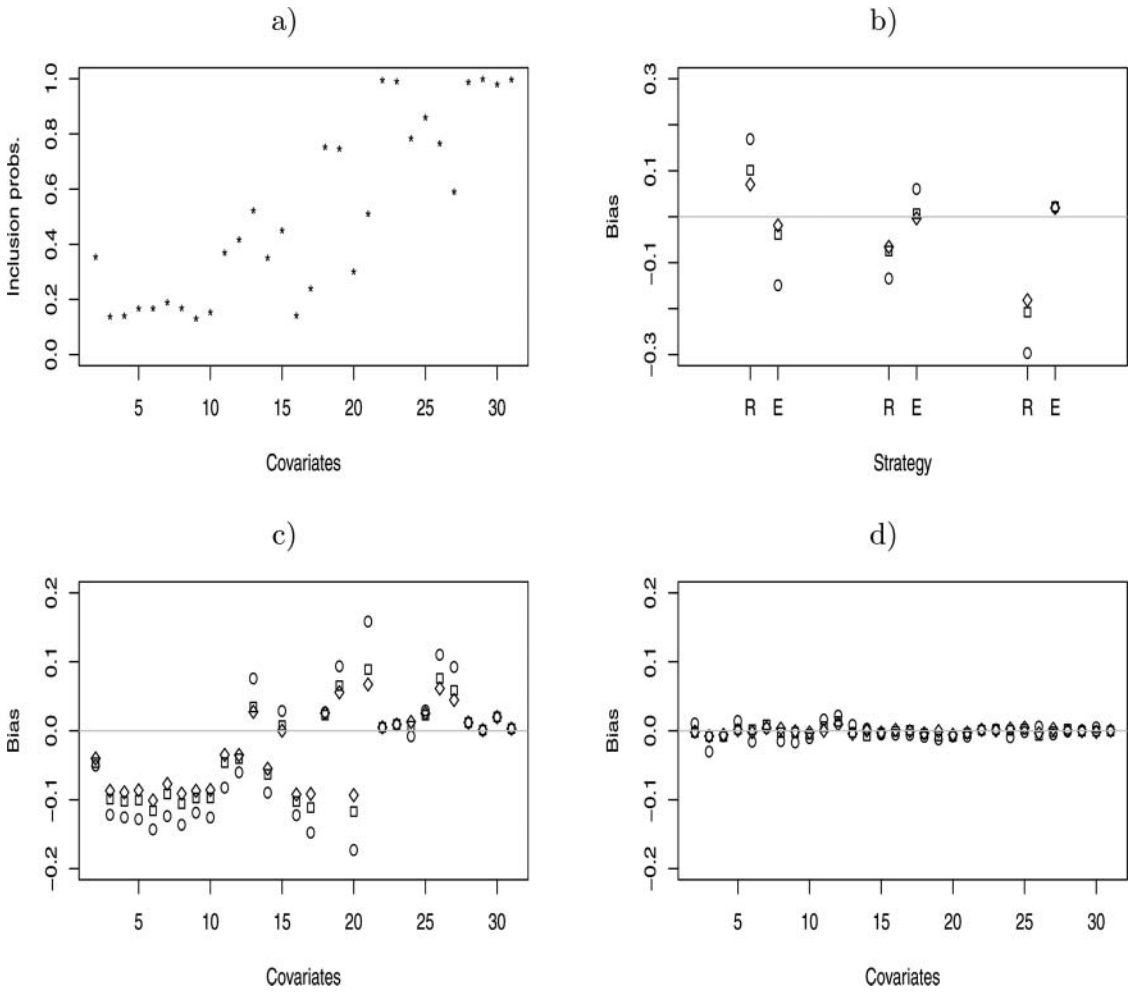

Figure 1. Simulated dataset with $p = 30$ and $n = 60$. (a) Exact inclusion probabilities q_l; (b) error in the predictions; (c) error in estimating q_l within the renormalized paradigm; and (d) same for the empirical paradigm. Keys: circles stand for $n = 1000$; squares for $n = 5000$, and diamonds for $n = 10,000$; R for renormalized and E for empirical.

are good models) very frequently include those covariables with high q_l (and do not include the rest) and hence, the difference $a(M_\gamma) - a(M_\delta)$ (see Result 1) is expected to be 1 when q_l is large and -1 when q_l is small. We will see this peculiar effect of renormalized approaches again in the next sections. In contrast, empirical estimates are much more precise, without any observed systematic bias (Figures 1(b) and 1(d)). Particularly revealing is the effect of n on the results, being very small for the empirical approach (almost negligible in estimating q_l) but with a large impact on the renormalized approach. In the context of huge model spaces where the number of visited models is by definition extremely small, the empirical paradigm seems to be a more convenient procedure.

Of course, none of the paradigms compared are feasible in practice for problems with a large number of variables. Nevertheless, the results above clearly point in the direction that, regardless of the specific method used to implement each of these paradigms, the empirical approach shows clear advantages over the renormalized one. For the examples analyzed in the following sections, we see that standard sampling strategies can lead to very good results.

4.2 Ozone35 Dataset

Mainly for comparative purposes but also to report the exact results on a moderately large problem (something that has not been done before to the best of our knowledge), here we present the exact solution for a problem with $p = 35$ covariates, and hence, with $34,359,738,368 \approx 3 \times 10^{10}$ different models.

The data we analyzed were previously used by Casella and Moreno (2006) and Berger and Molina (2005) and concern $N = 178$ measures of ozone concentration in the atmosphere. Details on the data can be found in Casella and Moreno (2006). A full description of the variables in the original ozone dataset appears as annex material to the article. Of the 10 main effects originally considered, we only make use of those with an atmospheric meaning x_4–x_{10}, as was done by Liang et al. (2008). We then have seven main effects, which, jointly with the quadratic terms and second-order interactions, produce the above-mentioned $p = 35$ possible regressors. We call this dataset Ozone35, for which we now present the exact results.

The sum of all Bayes factors is $1.13\,10^{50}$. The highest probability model, HPM, has $\{1, x_{10}, x_4x_6, x_6x_8, x_7^2, x_7x_{10}\}$ with a posterior probability of 0.0009 (a Bayes factor in its favor and compared with M_0 of $1.02\,10^{47}$). The first 1000 most probable models accumulated a total probability of 0.07 and a sum of Bayes factors (expressed in decimal logarithm) of 48.92 (this value is used later).

The expected posterior predictive value μ_1 and inclusion probabilities of each variable are shown in Table 1. The median inclusion probability model, MPM, is $\{1, x_6^2, x_6x_7, x_6x_8, x_7x_{10}\}$, which has a posterior probability which is 23 times lower than the probability of the highest posterior probability. Moreover, there are 851 models that are more probable than the median inclusion probability model.

We then run the following methods 10 times, each with $n =$ 10,000 iterations.

`Freq` Gibbs sampling with the algorithm in Appendix A with the initial model $M^{(0)} = M_0$ (we did not observe any differences starting with the full model or with a randomly chosen model). For a fair comparison among the methods compared, we did not exclude any model sampled and did not use any burn-in period.

`BAS` Bayesian adaptive sampling of Clyde, Ghosh, and Littman (2011) through the corresponding R-package `BAS`. As recommended (personal communication, Clyde 2010), we used `method` = "`MCMC` + `BAS`," which uses an MCMC method to initialize the search (this is a clear improvement over other options like `eplogp`, which uses a rough approximation of inclusion probabilities with p-values to initialize the search). We tuned the parameter `update` = `500` so that sampling probabilities were updated every 500 iterations.

`SSBM` The stochastic search in Berger and Molina (2005). This method was originally proposed for a particular prior but the searching algorithm can be easily adapted to accommodate the g-prior.

Using the labels introduced in Section 1, `Freq` is a particular method within the empirical approach, while `BAS` and `SSBM` are methods of the renormalized approach. The sampling methods underlying `BAS` and `SSBM` only contain unique models. Results are summarized in Table 1.

For the first of the 10 runs of each method, Table 1 shows estimates of the expected posterior prediction, inclusion probabilities, the MPM, and the HPM. With this same run, we estimated the standard deviation of the estimators with `Freq` using (6). In addition, with the 10 runs we computed the observed standard deviation among runs as this provides a measure of variability in `BAS` and `SSBM` (for which an expression like the one in (6) does not exist). In Figure 2, we have represented the estimations of μ_1 obtained in the 10 runs.

The main conclusions that we have extracted from the former simulations can be summarized as follows:

4.2.1 Prediction. The estimation of μ_1 is extremely accurate in the `Freq` approach with very low uncertainty. On the contrary, the estimation given by `BAS` and `SSBM` is far from the exact value. In `BAS` the variation among the different runs is small, confirming the existence of a systematic bias.

4.2.2 MPM and Inclusion Probabilities. Of the 10 experiments conducted, `Freq` correctly identified the MPM 10 times, while with `BAS` and `SSBM` the estimated MPM and the real MPM did not succeed in any of the 10 runs. `Freq` also provides very accurate estimations of the (exact) inclusion probabilities with a small variability, which confirms the high efficiency of such estimators. The observed variability with `BAS` and `SSBM` is large, so in general we expect large differences in repetitions of these methods in a similar manner to that observed in this experiment.

One great advantage of empirical methods over renormalized is that the first come with a measure of precision in the estimates (6). In this experiment, we see that these estimates and the observed standard deviation are quite close to each other, suggesting that (6) is quite a good estimator.

The most worrisome aspect observed of `BAS` and `SSBM` is that they are clearly biased: for certain covariates we have to move the point estimation more than 10 times the standard deviation to cover the exact value of the inclusion probabilities (see, e.g., x_6x_8 in `BAS` and x_5x_{10} in `SSBM`). The nature and origin of this bias

Table 1. Expected posterior predictive value and inclusion probabilities (exact τ and estimates $\hat{\tau}$ in one run of 10,000 iterations) for the Ozone35 dataset ($q_{i,j}$ stands for the inclusion probability of $x_i x_j$). Also, $\hat{V}(\hat{\tau})$ is the estimated variance (6) using this same run and $S(\hat{\tau})$ is the deviation of the estimators observed in 10 independent runs. Symbols (†) are for those variables in the estimated MPM and asterisks (∗) are for those variables in the estimated HPM

	Method	μ_1	q_4	q_5	q_6	q_7	q_8	q_9	q_{10}	$q_{4,4}$
τ	exact	10.224	0.164	0.096	0.297	0.195	0.200	0.291	0.368*	0.164
$\hat{\tau}$	Freq	10.221	0.157	0.099	0.300	0.191	0.200	0.292	0.368*	0.162
$[\hat{V}(\hat{\tau})]^{1/2}$	Freq	0.017	(0.004)	(0.003)	(0.007)	(0.005)	(0.005)	(0.007)	(0.007)	(0.004)
$S(\hat{\tau})$	Freq	0.021	(0.005)	(0.002)	(0.007)	(0.008)	(0.007)	(0.004)	(0.005)	(0.004)
$\hat{\tau}$	BAS	9.532	0.022	0.01	0.231	0.032	0.025	0.092	0.508*†	0.023
$S(\hat{\tau})$	BAS	0.109	(0.007)	(0.003)	(0.046)	(0.017)	(0.018)	(0.027)	(0.078)	(0.006)
$\hat{\tau}$	SSBM	11.063	0.105	0.03	0.04	0.053	0.073	0.297	0.131	0.125
$S(\hat{\tau})$	SSBM	0.515	(0.034)	(0.006)	(0.154)	(0.021)	(0.046)	(0.086)	(0.277)	(0.037)
		$q_{4,5}$	$q_{4,6}$	$q_{4,7}$	$q_{4,8}$	$q_{4,9}$	$q_{4,10}$	$q_{5,5}$	$q_{5,6}$	$q_{5,7}$
τ	exact	0.095	0.325*	0.252	0.208	0.301	0.361	0.124	0.107	0.094
$\hat{\tau}$	Freq	0.094	0.320*	0.244	0.210	0.303	0.360	0.127	0.104	0.095
$[\hat{V}(\hat{\tau})]^{1/2}$	Freq	(0.003)	(0.008)	(0.007)	(0.005)	(0.008)	(0.007)	(0.003)	(0.003)	(0.003)
$S(\hat{\tau})$	Freq	(0.002)	(0.01)	(0.006)	(0.005)	(0.008)	(0.006)	(0.002)	(0.003)	(0.003)
$\hat{\tau}$	BAS	0.019	0.373*	0.164	0.061	0.078	0.416	0.019	0.013	0.012
$S(\hat{\tau})$	BAS	(0.003)	(0.09)	(0.043)	(0.024)	(0.049)	(0.061)	(0.005)	(0.004)	(0.003)
$\hat{\tau}$	SSBM	0.037	0.03	0.082	0.092	0.348*	0.132	0.047	0.049	0.035
$S(\hat{\tau})$	SSBM	(0.007)	(0.295)	(0.285)	(0.16)	(0.098)	(0.33)	(0.008)	(0.011)	(0.008)
		$q_{5,8}$	$q_{5,9}$	$q_{5,10}$	$q_{6,6}$	$q_{6,7}$	$q_{6,8}$	$q_{6,9}$	$q_{6,10}$	$q_{7,7}$
τ	exact	0.098	0.088	0.124	0.532†	0.636†	0.560*†	0.126	0.115	0.450*
$\hat{\tau}$	Freq	0.098	0.087	0.124	0.524†	0.634†	0.564*†	0.127	0.113	0.465*
$[\hat{V}(\hat{\tau})]^{1/2}$	Freq	(0.003)	(0.003)	(0.003)	(0.008)	(0.011)	(0.008)	(0.003)	(0.003)	(0.009)
$S(\hat{\tau})$	Freq	(0.002)	(0.003)	(0.004)	(0.008)	(0.012)	(0.007)	(0.003)	(0.001)	(0.009)
$\hat{\tau}$	BAS	0.009	0.014	0.014	0.282	0.493	0.929*†	0.025	0.019	0.793*†
$S(\hat{\tau})$	BAS	(0.003)	(0.003)	(0.004)	(0.078)	(0.117)	(0.034)	(0.004)	(0.007)	(0.066)
$\hat{\tau}$	SSBM	0.027	0.017	0.023	0.98*†	1*†	0.077	0.078	0.031	0.112
$S(\hat{\tau})$	SSBM	(0.005)	(0.01)	(0.009)	(0.301)	(0.37)	(0.339)	(0.017)	(0.043)	(0.342)
		$q_{7,8}$	$q_{7,9}$	$q_{7,10}$	$q_{8,8}$	$q_{8,9}$	$q_{8,10}$	$q_{9,9}$	$q_{9,10}$	$q_{10,10}$
τ	exact	0.349	0.431	0.743*†	0.142	0.263	0.236	0.434	0.103	0.117
$\hat{\tau}$	Freq	0.346	0.430	0.756*†	0.140	0.264	0.231	0.440	0.103	0.116
$[\hat{V}(\hat{\tau})]^{1/2}$	Freq	(0.008)	(0.009)	(0.006)	(0.004)	(0.007)	(0.004)	(0.004)	(0.003)	(0.003)
$S(\hat{\tau})$	Freq	(0.008)	(0.01)	(0.006)	(0.004)	(0.008)	(0.004)	(0.005)	(0.003)	(0.002)
$\hat{\tau}$	BAS	0.091	0.124	0.965*†	0.017	0.045	0.127	0.393	0.022	0.018
$S(\hat{\tau})$	BAS	(0.047)	(0.073)	(0.026)	(0.013)	(0.032)	(0.028)	(0.032)	(0.005)	(0.007)
$\hat{\tau}$	SSBM	0.975*†	0.663*†	0.879*†	0.026	0.57*†	0.597*†	0.244	0.059	0.064
$S(\hat{\tau})$	SSBM	(0.425)	(0.193)	(0.209)	(0.01)	(0.164)	(0.178)	(0.084)	(0.015)	(0.015)

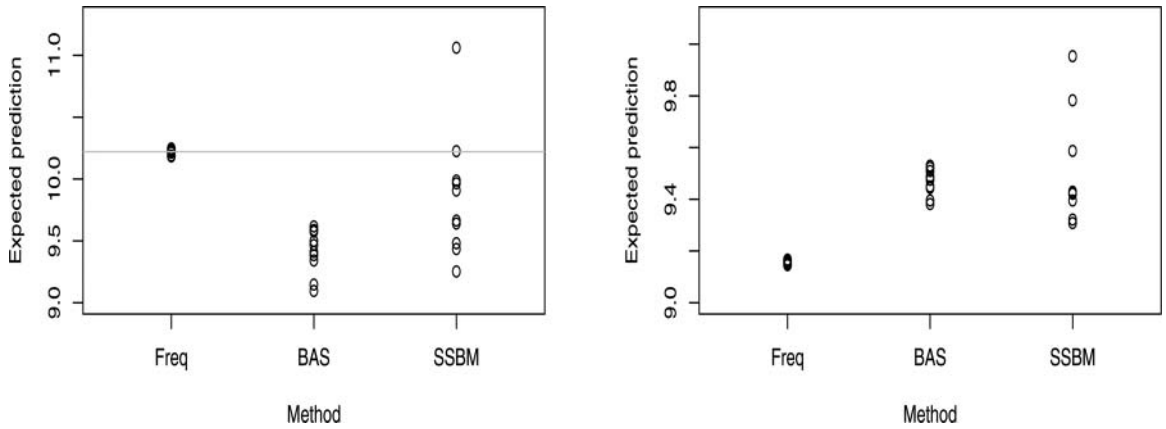

Figure 2. Estimation of the expected posterior prediction μ_1. Left: Ozone35 dataset with $n = 10{,}000$ iterations of each method (horizontal line for the exact value); right: Ozone65 dataset with $n = 100{,}000$.

has an easy interpretation after a careful reading of the table. In BAS, six inclusion probabilities are overestimated, of which five are in the estimated HPM; the rest are underestimated. A similar pattern is found in SSBM. This means that inclusion probabilities within these methods are very influenced by the estimated HPM, leading to a bias in the direction of the HPM model. Notice also that there is a tendency to overestimate the larger inclusion probabilities (the covariables in the MPM are strongly overestimated either by BAS or SSBM) in clear agreement with Result 1.

4.2.3 HPM and Probability Mass Discovered. One interesting question is which method is visiting better models. BAS and SSBM are, in some sense, specifically designed with this aim while this characteristic is only presumed in MCMC methods (since more probable models should be more visited). In our experiment, BAS correctly identified the HPM nine times while SSBM did so five times. The exact HPM was among the models visited in Freq in the 10 runs, showing that Freq is visiting good models.

Finally, we calculated the mean and standard deviation (over the 10 runs) of the sum of the Bayes factors of the 1000 (in decimal logarithm) most probable different models explored (to be compared with the exact value given above of 48.92). The results were 48.77(0.01), 48.64(0.05), and 48.50(0.20), for Freq, BAS, and SSBM, respectively. In this respect, the three methods analyzed behave similarly, although Freq gives more stable answers.

4.3 Ozone65 Dataset

We now consider the full ozone dataset with the 10 main effects, the quadratic terms and the second-order interactions. The same problem was considered by Berger and Molina (2005) and, as before, we keep the same notation for the covariates as they. This problem has $p = 65$, and hence, $2^{65} \approx 3.7 \times 10^{19}$ models in $\mathcal{M}$. In what follows we call this dataset Ozone65. The size of $\mathcal{M}$ precludes having the exact answer to the problem. To have an approximate idea of the infeasibility, notice that with the C code that we used for Ozone35, it would take more than 350 years to compute the answer using a cloud with 10^6 processors.

For this dataset, we repeated the comparison in Section 4 and performed 10 different runs of Freq, BAS, and SSBM, each now with $n = 100{,}000$ iterations. Table 2 shows the statistics of all the variables included in the estimated HPM or MPM in any of the runs for the methods being compared. In Figure 2, we have represented the estimation of μ_1 among the 10 runs. In essence, these results are in clear agreement with our findings in the previous analyses, and confirm the conclusions drawn there. Notice that in this example the exact answer is unknown, but results with Freq provide a very reasonable and consistent picture of the solution.

4.3.1 Prediction. The estimation of μ_1 is quite concentrated around 9.16 with an observed standard deviation among the 10 runs of 0.006 (estimated as $\hat{V}(\hat{\tau})^{1/2} = 0.005$ with the first run). BAS estimations are also relatively centered on 9.47(0.05), while estimations with SSBM are more disperse: 9.51(0.21).

Table 2. For the Ozone65 dataset, the number of times each covariate is included in the estimated HPM and the estimated MPM in 10 independent runs of $n = 100{,}000$ iterations of Freq, BAS, and SSBM methods. Asterisks identify the best model encountered in the full experiment. Also, $\hat{q}_l$'s are the estimation of the inclusion probabilities from the first run of Freq and $\hat{V}(\hat{q}_l)$ are the corresponding estimated variances (6) using this same run

	HPD			MPM			
Method	Freq	BAS	SSBM	Freq	BAS	SSBM	$\hat{q}_l\,(\hat{V}(\hat{q}_l)^{1/2})$
x_1*	7	7	8	10	8	8	0.575(0.004)
x_6*	8	9	4	–	10	2	0.427(0.002)
x_7	–	–	3	–	–	4	0.290(0.002)
x_8	1	–	1	–	–	1	0.297(0.002)
x_{10}*	3	5	4	–	2	2	0.292(0.002)
x_1x_1*	10	10	10	10	10	10	1.000(<0.001)
x_1x_4	3	3	2	1	3	2	0.497(0.002)
x_2x_8*	9	8	4	–	–	–	0.184(0.001)
x_4x_4	–	–	1	–	–	1	0.266(0.004)
x_4x_6*	8	9	1	–	10	2	0.418(0.002)
x_4x_7	2	1	5	–	–	5	0.337(0.003)
x_4x_8	–	2	3	–	1	3	0.303(0.002)
x_4x_{10}	6	3	4	–	2	4	0.309(0.002)
x_5x_5*	10	9	9	–	–	–	0.261(0.001)
x_5x_7	–	1	–	–	–	–	0.156(0.001)
x_6x_6	–	–	–	–	–	2	0.218(0.002)
x_6x_7	1	2	7	10	3	7	0.614(0.003)
x_6x_8	–	–	–	–	–	6	0.328(0.001)
x_6x_{10}	1	1	2	–	–	2	0.237(0.002)
x_7x_7*	7	7	2	–	6	2	0.372(0.003)
x_7x_8	1	2	2	–	2	2	0.479(0.003)
x_7x_{10}*	9	9	8	10	9	8	0.623(0.003)
x_9x_9*	10	10	10	10	10	10	0.966(<0.001)
Number of different variables							
	17	18	20	6	13	20	

4.3.2 MPM and Inclusion Probabilities. In Freq, except for one run, there is unanimity in the estimation of the MPM. Furthermore, and quite appealing, is that we can give an explanation of the disagreement in terms of estimation errors. The discordant run differs from all the others in that it includes x_1x_4. This variable has an estimated inclusion probability of 0.497 with an estimated error of 0.002.

In BAS and SSBM, results vary considerably over the different runs. In BAS (SSBM), 8(10) different models were estimated as the MPM and hence, this method has incorrectly identified the true MPM at least 8(9) times. Of greater concern is that these bad results do not seem to be due to variability. We can find a more likely explanation in Table 2, where we can clearly see that on many occasions the estimated MPM mimics the estimated HPM. For instance, x_6 and x_4x_6 (which are in the estimated HPM) are always in the MPM estimated by BAS. This also seems to be the case for x_4x_7 and x_6x_7 in SSBM. We interpret these results as a manifestation of the bias produced with methods conceived to look for good models.

4.3.3 HPM and Probability Mass Discovered. In this aspect, the three methods behave quite similarly, perhaps BAS and Freq performing slightly better than SSBM. The best model

found in the whole experiment, the 10 runs of the three different methods, had a Bayes factor (in its favor and against the null) of 50.87 (expressed in decimal logarithm). This model, identified in Table 2 with asterisks, was visited in 4 of the 10 runs by BAS, in 3 runs by Freq, and in 1 run by SSBM. The means (standard deviations) over the 10 runs of the sum of the Bayes factors of the 1000 (in decimal logarithm) most probable different models explored were 52.77(0.02) in Freq, 52.78(0.15) in BAS, and 52.78(0.16) in SSBM. These results confirm the popular hypothesis that good models also show up when sampling from the posterior distribution.

5. DISCUSSION

A number of interesting questions have been raised throughout the reviewing process of this article. We have opted to include these along with our responses since we think these could be of great interest to the reader. Alternatively, the resulting material can be viewed as complementary material to this article.

5.1 Comparability of the Methods

The computational burden of the different methods has not been considered explicitly and the results (see the previous sections) are based on the same number of iterations. This is not totally fair since, for example, the BAS approach needs one marginal likelihood evaluation per iteration while the Gibbs sampler in Appendix A needs p of such evaluations.

To address this question, instead of comparing methods based on equal running time (a comparison that depends heavily on how the algorithms have been coded and implemented) we have used an alternative approach. We compared BAS and Freq for an equal number of marginal evaluations (i.e., estimations based on n iterations for Freq and on np for BAS) in the Ozone35 dataset. The results (partially shown in Figure 3) clearly point to the superiority of Freq since it converges on the true value much faster than BAS.

A similar comparison with other methods (each with their own particularities) is not that straightforward and results will depend heavily on the implementation of the algorithm. Our experience is that SSBM is quite demanding (although a more thorough implementation of the method like the one in Scott and Carvalho (2009) for graphical models can alleviate the problem substantially), much more than the Freq algorithm based on Gibbs sampling with very simple full conditionals. Furthermore, and independently of this, the results in Figure 3 suggest that Freq produces reliable results with much fewer iterations than those reported in Table 1 (with $n = 10{,}000$). Recall that the results in this table clearly demonstrated that SSBM was still far from convergence with $n = 10{,}000$.

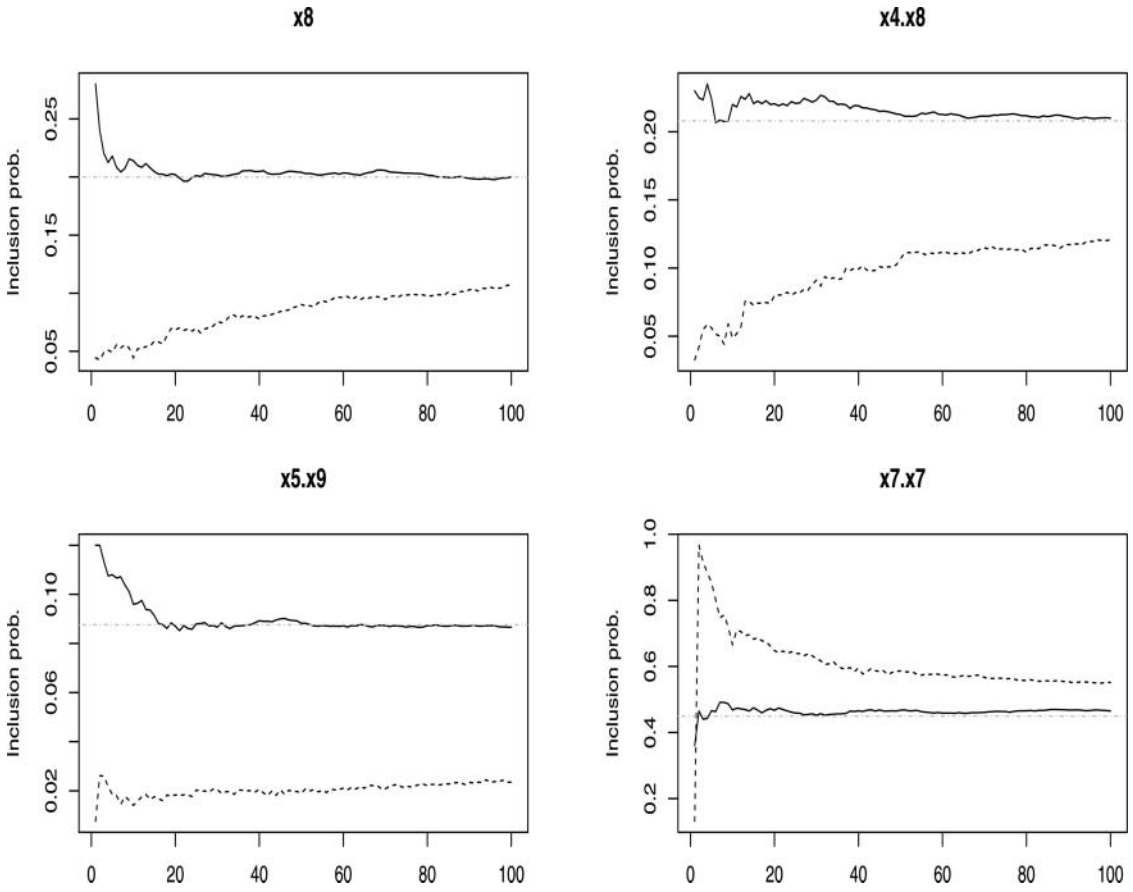

Figure 3. Estimates of the inclusion probabilities as a function of the number of iterations for four randomly selected covariates in the Ozone35 dataset. Each (x, y) point in the graph is the estimation (y) obtained based on $100x$ iterations in method Freq (solid line), and on $3500x$ iterations in method BAS (dashed line). Horizontal lines are at exact values of inclusion probabilities.

5.2 Renormalization and MCMC methods.

How do MCMC methods behave with renormalized estimators? Answering this question would help to distinguish between the effect of renormalization from the effect of the searching algorithm.

To quantify this effect, we have computed the renormalized estimators for the Ozone35 dataset for the models visited by the Gibbs sampling approach. The results obtained (see Table 1 of the supplementary material) are, for the most part, close to the other renormalized method, clearly outperformed by the `Freq` strategy. In our opinion, these results reinforce the idea that renormalization usually produces biased results.

5.3 Generalizability of the Results

The examples considered are somewhat difficult because the designs used (with squared terms and interactions) induce very high multicollinearity. It may be possible that MCMC works well here, this being one reason why the frequency-based approaches do better than the renormalized approaches. It would be illuminating to see what happens in other examples, like in the simplest case with orthogonal design matrices (under orthogonality, `BAS` has additional properties and is expected to provide more reliable results). The results would help to separate the effect of large model spaces with multicollinearity and will speak about the generalizability of the results to other examples.

We have addressed this question by considering two new datasets of a very different nature to the ozone dataset. The first is based on the data in Ley and Steel (2007) concerning potential regressors to model the average per capita gross domestic product (GDP) over the period 1960–1992 for a sample of $N = 72$ countries. Of the 41 initial covariates, we randomly selected $p = 30$ (detailed in the supplementary material) to make it feasible to obtain exact results. These data only contain the main effects (not interactions), so we found them very suitable for the question raised. The exact values of the parameters and estimates based on `Freq`, `BAS`, and `SSBM` are provided as supplementary material. The conclusions are along the lines drawn for the Ozone35 dataset, once more supporting the idea of the superiority of the `Freq` approach.

In second place, we have compared `BAS` and `Freq` in an orthogonalized transformation (via the Cholesky decomposition) of the above-mentioned dataset. The results (again given as annex material) show a modest improvement of `BAS`, although it still highly biased compared with `Freq`. This is in agreement with the results already reproduced by Heaton and Scott (2010) for orthogonal designs.

5.4 Comparison With Related Work

A couple of contemporary articles have worked around the idea of the properties of renormalized approaches in the context of large model spaces. How does this article compare with those articles?

The preliminary results on which this article is based were first presented in a poster in 2010 at the Ninth Valencia International Meeting on Bayesian Statistics. Since then, to the best of our knowledge, and apart from the present article, two other interesting studies have emerged focusing on the performance of renormalized and empirical approaches: Clyde and Ghosh (2012) and Heaton and Scott (2010). In essence, all three agree in that renormalized approaches can be biased; Clyde and Ghosh (2012) substantiated the bias in the estimation approximately and concluded that the only case when renormalization will not be biased is that of equal probability sampling. In this article, we have added new and complementary insights into the existence and nature of such bias. Furthermore, we have established both theoretical and solid practical evidence pointing out that MCMC methods with estimates based on the frequency of visits (which we have called the empirical approach) can be extremely precise. This is perhaps the most interesting and surprising message that emerges from our work.

6. MAIN CONCLUSIONS

In large model spaces, estimations obtained with the empirical and renormalized approaches usually differ dramatically. We have provided several examples where such differences are exhibited. In those examples, the empirical approach shows quite accurate estimations that largely outperform those obtained with renormalized methods.

Our results do not have to be understood as a guarantee that *any* MCMC method plus an estimation based on the frequency of visits will produce reliable results. The main conclusion is that they have the potential to provide accurate results, but of course this will depend on the characteristics of the MCMC. In our case, we think that a key reason why the Gibbs sampling used (see Appendix A) worked so well is because all the parameters (except $\boldsymbol{\gamma}$) have been integrated out, leading to a finite sample space.

To give broad generality to the results, we have provided a theoretical justification embedding both approaches in a common framework of sampling theory. A revealing fact that arises is that the success of renormalized results relies strongly on the proportionality assumption between $a(M_\gamma)P(M_\gamma \mid \boldsymbol{y})$ and $P(M_\gamma \mid \boldsymbol{y})$, an assumption that is rarely appropriate and definitely wrong for quantities like the posterior inclusion probabilities. In our opinion, the inadequacy of this hypothesis is the main reason underlying the poor behavior of renormalized methods in our examples. Moreover, we have also quantified the resulting bias in a simple equation (Result 1), giving a full explanation for the systematic errors observed in the estimation of inclusion probabilities using renormalized strategies. This is quite a characteristic phenomenon for renormalized methods already highlighted by Clyde and Ghosh (2012) and Heaton and Scott (2010) that, to our knowledge, remained unexplained. A second advantage of connecting large variable selection problems with sampling theory, already shown in the article, is the possibility of quantifying the uncertainty of estimates if sampling probabilities of models are known. This fact enables us to know the precision of the estimates derived from the empirical approach, as illustrated in our examples.

The empirical estimates are, under very mild conditions, consistent and hence (for a large enough sample size) unbiased (that, of course, being a very well-known property). Moreover, some implementations of the empirical approach, like the one used in this study, are known to have reasonable convergence properties. Surprisingly, there is not normally so much confidence put in the frequency of visits as the basis for the estimations.

Furthermore, it is usually perceived that the large size of the model space precludes the appealing theoretical properties of empirical estimates being translated into good results in practice. Our argument is quite the opposite and is based on a simple but convincing idea: in a *random* sampling of a population with approximately infinite elements, the sample size needed to reach certain precision tends to a finite number. Therefore, results from the empirical approach scale up satisfactorily as a function of p. Still, it could be argued that MCMC-based methods run the risk of getting caught on local modes of the posterior distribution. No one can deny that possibility, which may occur in virtually any Bayesian analysis solved with simulation-based techniques with a dependent sampling. Nevertheless, there are many simple strategies to prevent this from happening (or at least to warn us that it could be happening) like simulating several chains if multiple isolated modes are expected. More importantly, it is difficult to figure out why heuristic methods designed to find good models (those used under the renormalized approach) are by definition free of this same difficulty. Nonetheless, these difficulties are intrinsic to the specific methods putting the approaches into practice, and are not related to the paradigms themselves, which are the main focus of study in this work. Connections between variable selection in large model spaces and sampling theory have been revealed as an interesting tool, shedding some light on the qualitative properties of proposed model selection methods. Further exploration of this connection could provide some interesting clues about how to sample models or to modify ratio estimators to improve their performance. This is, in our opinion, a promising new line of research for variable selection problems in large model spaces whose exploration has already been started in Clyde and Ghosh (2012). Finally, notice that this article has not dealt with huge model spaces, such as $p = 100$ or $p = 1000$, though we would expect the results in that case to be similar to those in our article as the theoretical results stated here also apply for them. Nevertheless, a thorough exploration of this particular issue would surely be of sufficient interest to be carried out in future works.

APPENDIX A: GIBBS SAMPLING ALGORITHM

Once the parameters $\boldsymbol{\beta}_\gamma, \sigma$ have been analytically integrated out (see (3)), the only unknown parameters in the problem are the components in $\boldsymbol{\gamma}$. Those have full conditional distributions (specified below), which are straightforward to sample:

$$\gamma_i \mid \gamma_1, \ldots, \gamma_{i-1}, \gamma_{i+1}, \ldots, \gamma_p, \boldsymbol{y} \sim \text{Bernoulli}(p_i), \tag{A.1}$$

where

$$p_i = \frac{B_{a0}(g)\Pr(M_a)}{B_{a0}(g)\Pr(M_a) + B_{b0}(g)\Pr(M_b)},$$
$$\boldsymbol{a} = (\gamma_1, \ldots, \gamma_{i-1}, 1, \gamma_{i+1}, \ldots, \gamma_p),$$

and

$$\boldsymbol{b} = (\gamma_1, \ldots, \gamma_{i-1}, 0, \gamma_{i+1}, \ldots, \gamma_p).$$

APPENDIX B: PROOFS

In the proofs of Result 1 and Result 2, for simplicity on notation, denote $\tau = \tau(a)$, $\hat{\tau}_R = \hat{\tau}_R(a)$, $P_\gamma = Pr(M_\gamma \mid \boldsymbol{y})$, and $a_\gamma = a(M_\gamma)$.

Proof of Result 1. Clearly,

$$(\hat{\tau}_R - \tau) \sum_{\gamma \in \mathcal{M}^*} P_\gamma = \sum_{\gamma \in \mathcal{M}^*} a_\gamma P_\gamma - \sum_{\gamma \in \mathcal{M}} a_\gamma P_\gamma \sum_{\gamma \in \mathcal{M}^*} P_\gamma.$$

Now, denote D the right term on the identity above. Then

$$\begin{aligned} D &= \sum_{\gamma \in \mathcal{M}^*} a_\gamma P_\gamma \Big(1 - \sum_{\gamma \in \mathcal{M}^*} P_\gamma\Big) - \sum_{\delta \in \bar{\mathcal{M}}^*} a_\delta P_\delta \sum_{\gamma \in \mathcal{M}^*} P_\gamma \\ &= \sum_{\gamma \in \bar{\mathcal{M}}^*} \sum_{\delta \in \bar{\mathcal{M}}^*} a_\gamma P_\gamma P_\delta - \sum_{\delta \in \bar{\mathcal{M}}^*} \sum_{\gamma \in \bar{\mathcal{M}}^*} a_\delta P_\delta P_\gamma \\ &= \sum_{\gamma \in \bar{\mathcal{M}}^*} \sum_{\delta \in \bar{\mathcal{M}}^*} P_\gamma P_\delta (a_\gamma - a_\delta), \end{aligned}$$

which proves the result.

Proof of Result 2. Additionally denote $\hat{\tau}_p = \hat{\tau}_p(a)$ and $\bar{P} = \sum_{\gamma \in \bar{\mathcal{M}}^*} P_\gamma / n$, then

$$\hat{\tau}_R = \hat{\tau}_p + \frac{\sum_{\gamma \in \bar{\mathcal{M}}^*} (a_\gamma - \hat{\tau}_p) P_\gamma}{\sum_{\gamma \in \bar{\mathcal{M}}^*} P_\gamma} = \hat{\tau}_p + \frac{\sum_{\gamma \in \bar{\mathcal{M}}^*} (a_\gamma - \hat{\tau}_p)(P_\gamma - \bar{P})}{\sum_{\gamma \in \bar{\mathcal{M}}^*} P_\gamma},$$

from which the identity on this result follows easily.

Proof of Proposition 1. Since $\{\boldsymbol{\gamma}^{(i)}\}$ is irreducible and $\mathcal{M}$ (the state space) is finite, then the transition probabilities between any two states can be bounded below by a positive constant. In consequence, a minorization condition holds and the state space is *small* (Tierney 1994). Hence, by Proposition 2 in Tierney (1994) the chain $\{\boldsymbol{\gamma}^{(i)}\}$ is uniformly ergodic. Furthermore, by Theorem 5 (in Tierney 1994) together with the hypothesis of stationarity on the posterior distribution and since $a()$ is L^2 summable with respect to the measure $\Pr(M_\gamma \mid \boldsymbol{y})$, that is,

$$\sum_{\gamma \in \mathcal{M}} a^2(M_\gamma) \Pr(M_\gamma \mid \boldsymbol{y}) < \infty,$$

there is a real number $\sigma(a)$ such that $\sqrt{n}(\hat{\tau}_p^{(n)} - \tau(a))$ converges weakly to a normal distribution with a zero mean and variance $\sigma^2(a)$ for any initial distribution.

Finally, the lag-k autocovariance of $a()$ exists (since $a()$ is L^2) and according to Chan and Geyer (1994)

$$\sigma^2(a) = \text{cov}_0(a) + 2 \sum_{k=1}^{\infty} \text{cov}_k(a),$$

provided that the sum on the right-hand side converges.

SUPPLEMENTARY MATERIAL

The supplementary material contains the description of the variables in the Ozone35 dataset and results to which we have referred to in Section 5.

[Received March 2012. Revised September 2012.]

REFERENCES

Barbieri, M. M., and Berger, J. O. (2004), "Optimal Predictive Model Selection," *The Annals of Statistics*, 32, 870–897. [342]

Bayarri, M. J., and García-Donato, G. (2008), "Generalization of Jeffreys Divergence-Based Priors for Bayesian Hypothesis Testing," *Journal of the Royal Statistical Society,* Series B, 70, 981–1003. [341]

Berger, J. O., and Molina, G. (2005), "Posterior Model Probabilities via Path-Based Pairwise Priors," *Statistica Neerlandica*, 59, 3–15. [340,341,343,344,346,348]

Berger, J. O., and Pericchi, L. R. (2001), "Objective Bayesian Methods for Model Selection: Introduction and Comparison," *Lecture Notes-Monograph Series*, 38, 135–207. [340,341]

Carvalho, C., and Scott, J. (2009), "Objective Bayesian Model Selection in Gaussian Graphical Models," *Biometrika*, 96, 497–512. [341,343]

Casella, G., and Moreno, E. (2006), "Objective Bayesian Variable Selection," *Journal of the American Statistical Association*, 101, 157–167. [341,343,346]

Chan, K. S., and Geyer, C. T. (1994), Discussion to "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1747–1758. [351]

Chib, S., and Jeliazkov, I. (2001), "Marginal Likelihood From the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281. [340]

Clyde, M. A., DeSimone, H., and Parmigiani, G. (1996), "Prediction via Orthogonalized Model Mixing," *Journal of the American Statistical Association*, 91, 1197–1208. [340]

Clyde, M. A., and George, E. I. (2004), "Model Uncertainty," *Statistical Science*, 19, 81–94. [340]

Clyde, M. A., and Ghosh, J. (2012), "Finite Population Estimators in Stochastic Search Variable Selection," Technical report 2010-11, Department of Statistics of Duke University. Available at *http://stat.duke.edu/research/papers/2010-11*. [341,343,350]

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011), "Bayesian Adaptive Sampling for Variable Selection and Model Averaging," *Journal of Computational and Graphical Statistics*, 20, 80–101. [341,343,344,346]

Dellaportas, P., Forster, J., and Ntzoufras, I. (2000), "Bayesian Variable Selection Using the Gibbs Sampler," in *Generalized Linear Models: A Bayesian Perspective*, eds. D. Dey, S. K. Ghosh, and B. K. Mallick, New York-Basel: Marcel Dekker, Inc. [341,343]

Forte, A. (2011), "Objective Bayes Criteria for Variable Selection," Ph.D. dissertation, Department of Statistics at the University of Valencia. [341]

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., and Rossi, F. (2009), *GNU Scientific Library Reference Manual* (v1.12, 3rd ed.), United Kingdom: Network Theory Ltd. [344]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [341,343]

George, E. I., and McCulloch, R. E. (1997), "Approaches for Bayesian Variable Selection," *Statistica Sinica*, 7, 339–373. [340,341,342,343,344]

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo," *Statistical Science*, 7, 473–511. [344]

Hans, C., Dobra, A., and West, M . (2007), "Shotgun Stochastic Search for Regression With Many Candidate Predictors," *Journal of the American Statistical Association*, 102, 507–516. [341]

Hansen, M. H., and Hurwitz, W. N. (1943), "The Theory of Sampling From Finite Populations," *Annals of Mathematical Statistics*, 14, 333–362. [343]

Heaton, M. J., and Scott, J. G. (2010), "Bayesian Computation and the Linear Model," in *Frontiers of Statistical Decision Making and Bayesian Analysis*, eds. M.-H. Chen, P. Mller, D. Sun, K. Ye, and D. K. Dey, New York: Springer, pp. 527–544. [350]

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial," *Statistical Science*, 14, 382–401. [340,342]

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford: Oxford University Press. [341]

Kuo, L., and Mallick, B. (1998), "Variable Selection for Regression Models," *Sankhya: The Indian Journal of Statistics*, 60, 65–81. [341,343,345]

Ley, E., and Steel, M. F. J. (2007), "Jointness in Bayesian Variable Selection with Applications to Growth Regression," *Journal of Macroeconomics*, 29, 476–493. [350]

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008), "Mixtures of g Priors for Bayesian Variable Selection," *Journal of the American Statistical Association*, 103, 410–423. [341,346]

Lohr, S. L. (1999), *Sampling: Design and Analysis*, Pacific Grove, CA: Duxbury Press. [343]

Madigan, D., and Raftery, A. E. (1994), "Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window," *Journal of the American Statistical Association*, 89, 1535–1546. [340]

Madigan, D., and York, J. (1995), "Bayesian Graphical Models for Discrete Data," *International Statistical Review*, 63, 215–232. [341]

Maruyama, Y., and George, E. I. (2011), "Fully Bayes Factors With a Generalized g-prior," *The Annals of Statistics*, 39, 2740–2765. [341]

Nott, D., and Kohn, R. (2005), "Adaptive Sampling for Bayesian Variable Selection," *Biometrika*, 92, 747–762. [341,343]

Ntzoufras, I. (2002), "Gibbs Variable Selection Using BUGS," *Journal of Statistical Software*, 7, 1–19. [341,343]

——— (2009), *Bayesian Modeling Using WinBUGS: Wiley Series in Computational Statistics*, New York: Wiley. [341,343]

Raftery, A. E. (1998), "Bayes Factors and BIC: Comment on Weakliem," Technical Report 347, Department of Statistics, University of Washington. [341]

Raftery, A. E., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191. [340]

Royall, R. M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," *Biometrika*, 57, 377–387. [343]

Scott, J., and Carvalho, C. (2009), "Feature-Inclusion Stochastic Search for Gaussian Graphical Models," *Journal of Computational and Graphical Statistics*, 17, 790–808. [341,343,349]

Thompson, S. K. (2002), *Sampling* (2nd ed.), New York: Wiley. [342,343]

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," (with discussion and a rejoinder by the author), *The Annals of Statistics*, 22, 1701–1762. [351]

Zellner, A. (ed.) (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis With g-prior Distributions," in *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, North Holland: Edward Elgar Publishing Limited, pp. 389–399. [341]

Zellner, A., and Siow, A. (1980), "Posterior Odds Ratio for Selected Regression Hypotheses," in *Bayesian Statistics 1*, eds. J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and Adrian F. M. Smith, Valencia: University Press, pp. 585–603. [341]

Zellner, A., and Siow, A. (1984), *Basic Issues in Econometrics*, Chicago: University of Chicago Press. [341]