Effect of Model Space Priors on Statistical Inference with Model Uncertainty

Anupreet PORWAL* and Adrian E. RAFTERY

Abstract

Bayesian model averaging (BMA) provides a coherent way to account for model uncertainty in statistical inference tasks. BMA requires specification of model space priors and parameter space priors. In this article we focus on comparing different model space priors in the presence of model uncertainty. We consider eight reference model space priors used in the literature and three adaptive parameter priors recommended by Porwal and Raftery [37]. We assess the performance of these combinations of prior specifications for variable selection in linear regression models for the statistical tasks of parameter estimation, interval estimation, inference, point and interval prediction. We carry out an extensive simulation study based on 14 real datasets representing a range of situations encountered in practice. We found that beta-binomial model space priors specified in terms of the prior probability of model size performed best on average across various statistical tasks and datasets, outperforming priors that were uniform across models. Recently proposed complexity priors performed relatively poorly.

KEYWORDS AND PHRASES: Bayesian model averaging, Zellner's g-prior, Model space prior, Beta-Binomial prior, Complexity prior, Model selection, Prediction.

1. INTRODUCTION

Analysis of data in the presence of model uncertainty is a critical problem in statistical modeling applications. Accounting for model uncertainty, rather than selecting a single statistical model, improves predictive performance and robustness in estimation and inference of model parameters [40, 37].

One common instance of model uncertainty is that of variable selection in linear regression model. Given an n-dimensional continuous response variable, $\boldsymbol{Y}$, and a set of p potential predictor variables $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathcal{R}^{n \times p}$, the aim is to do statistical analysis of the data when it is not known in advance which of the 2^p possible models is most appropriate. Consider a binary indexing vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ for the model space which indicates which explanatory variables are part of a model $\mathcal{M}_{\boldsymbol{\gamma}}$. Under $\mathcal{M}_{\boldsymbol{\gamma}}$, the linear regression model can then be expressed as:

$$\mathcal{M}_{\boldsymbol{\gamma}} : \boldsymbol{Y} = \boldsymbol{1}_n \alpha + X_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 I_n)$, and $X_{\boldsymbol{\gamma}}$ is a $n \times p_{\boldsymbol{\gamma}}$ matrix where each column is centered around its mean and $p_{\boldsymbol{\gamma}}$ denotes the number of explanatory variables in the model $\mathcal{M}_{\boldsymbol{\gamma}}$.

The Bayesian framework provides a straightforward way to account for model uncertainty by treating the model as a parameter itself, using Bayesian model averaging (BMA) [29, 38, 20, 34]. BMA requires the specification of a model

space prior and a parameter space prior. However, subjective elicitation of these priors is often not feasible, particularly when p is large, motivating the use of default reference priors.

Several default parameter prior choices have been proposed in the last thirty years (see Porwal and Raftery [37] for an overview) and several other comparisons of these methods have been carried out [4, 8, 11, 13, 15, 32]. However, similar comparisons of default model space priors remain limited. In this article, our focus is on understanding the effect of model space priors on the statistical tasks of parameter estimation, interval estimation, statistical inference, point and interval prediction.

We compare combinations of three default parameter priors with eight choices of model space priors that have been advocated in the literature. These model space priors correspond to different flavors of Bayesian inference with: (i) fixed hyper-parameter choices, (ii) with Bayesian treatment of hyper-parameters, and (iii) estimation of hyper-parameters in an empirical Bayes (EB) manner. The comparison is carried out over an extensive simulation study closely based on 14 real datasets that span a wide range of practical data analysis situations.

The article is organized as follows. Section 2 provides a brief review of BMA and discusses in detail the parameter and model prior choices considered in this article. We discuss the simulation design, metrics and datasets used for comparison, and the results in Section 3, followed by discussion and concluding remarks in Section 4.

---

*Corresponding author.

2. BAYESIAN MODEL AVERAGING: A REVIEW

Bayesian model averaging [24, 28] provides a formal way to account for model uncertainty in statistical inference. Several reviews of the BMA literature are available [28, 25, 13, 50, 7, 15]. The basic idea of BMA is that it uses prior probabilities for the models considered, and Bayes' theorem to deal with model uncertainty by calculating their posterior probabilities.

Assuming that there is one true model among the set of 2^p candidate models, the posterior probability of a model $\mathcal{M}_{\gamma}$ is

$$P(\mathcal{M}_{\gamma}|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\mathcal{M}_{\gamma})P(\mathcal{M}_{\gamma})}{\sum_{\gamma'} P(\boldsymbol{Y}|\mathcal{M}_{\gamma'})P(\mathcal{M}_{\gamma'})},$$

where $P(\mathcal{M}_{\gamma})$ is the prior model probability of $\mathcal{M}_{\gamma}$, and $P(\boldsymbol{Y}|\mathcal{M}_{\gamma})$ is the marginal likelihood of the model after integrating out parameters with respect to the prior π_{γ}, namely:

$$P(\boldsymbol{Y}|\mathcal{M}_{\gamma}) = \int \mathcal{N}(\boldsymbol{Y}|\mathbf{1}_n\alpha + X_{\gamma}\boldsymbol{\beta}_{\gamma}, \sigma^2 I_n)$$
$$\pi_{\gamma}(\boldsymbol{\beta}_{\gamma}, \alpha, \sigma^2)d\boldsymbol{\beta}_{\gamma}d\alpha d\sigma^2.$$

Under BMA inference, we can express the predictive distribution of a quantity of interest, Δ, such as a parameter or an observable future quantity, as a weighted average of its predictive distributions under the different candidate models:

$$P(\Delta|\boldsymbol{Y}) = \sum_{\gamma} P(\Delta|\mathcal{M}_{\gamma})P(\mathcal{M}_{\gamma}|\boldsymbol{Y}),$$

where the posterior model probabilities $P(\mathcal{M}_{\gamma}|\boldsymbol{Y})$ serve as averaging weights. In the case where Δ is a regression coefficient, the resulting posterior distribution, $P(\Delta|\boldsymbol{Y})$, is a mixture of a point mass at 0 and a continuous density.

BMA has several desirable theoretical properties [39]. When choosing between two models, one of which is nested within the other, choosing the one with the higher posterior probability minimizes the total error rate (sum of Type I and Type II error probabilities); BMA point estimators and predictions minimize mean squared error (MSE); BMA estimation and prediction intervals are calibrated; and BMA predictive distributions have optimal performance in the log score sense.

The next subsection discusses the choice of parameter and model space priors that need to be specified by the user when implementing BMA.

2.1 Choice of Parameter Priors

Despite the wide adoption of Bayesian methods in linear models, prior elicitation for linear models is still an open problem. The parameter prior distribution π_{γ} can be expressed as

$$\pi_{\gamma}(\boldsymbol{\beta}_{\gamma}, \alpha, \sigma^2) = \pi_{\gamma}(\boldsymbol{\beta}_{\gamma}|\alpha, \sigma^2)\pi_{\gamma}(\alpha, \sigma^2).$$

A standard Jeffreys' prior is often used for the intercept and error variance, which are often common to all models considered, i.e. $\pi_{\gamma}(\alpha, \sigma^2) \propto \sigma^{-2}$. One of the most popular priors used for the model parameters $\boldsymbol{\beta}_{\gamma}$ is Zellner's g-prior [53], namely

$$\boldsymbol{\beta}_{\gamma}|\sigma^2, \mathcal{M}_{\gamma} \sim \mathcal{N}(\mathbf{0}, g\sigma^2(X_{\gamma}^T X_{\gamma})^{-1}).$$

This is popular because of its computational efficiency in evaluating marginal likelihoods and performing model search. It is also attractive because of its intuitive interpretation arising from analysis of a conceptual sample generated using the same design matrix $\boldsymbol{X}_{\gamma}$ as in the data at hand. It is a special case of spike-and-lab family with the slab density given by the Normal density above and the spike being a point mass at zero. Also, g-priors are appealing in variable selection problems since they require the user to specify only value (or hyper-prior) for the scalar hyper-parameter g. This controls the prior variance of the parameters; the effective prior sample size is n/g. Several choices of g have been proposed [6, 13, 19, 23, 28, 32, 48, 54].

Based on an extensive simulation study, Porwal and Raftery [37] found that in comparing parameter priors for BMA, three adaptive g-priors performed the best among many popular choices across the statistical tasks of parameter estimation, interval estimation, model inference, point prediction and interval prediction. In what follows, we shall focus only on these three parameter prior choices, namely:

- **$g = \sqrt{n}$**: First proposed by [13], it corresponds to a prior sample size equal to $\sqrt{n}$ and has been found to work well in high dimensional settings [52]. The complexity penalty for a model using this specification is effectively half that in the BIC [37].
- **EB-local**: An alternative to fixing g to a pre-specified value is to instead estimate it from the data in an empirical Bayes (EB) manner. The local EB approach estimates a different g for each model. Let $P(\boldsymbol{Y}|\mathcal{M}_{\gamma}, g)$ denotes the marginal likelihood of the data under a g-prior. Then

$$\hat{g}_{\gamma} = \arg\max_{g \geq 0} P(\boldsymbol{Y}|\mathcal{M}_{\gamma}, g).$$

For a linear model, Hansen and Yu [23] showed that it reduces to $\hat{g}_{\gamma} = \max\{F_{\gamma} - 1, 0\}$ where F_{γ} is the F statistic for testing $\boldsymbol{\beta}_{\gamma} = \mathbf{0}$.
- **Hyper-g**: A natural Bayesian way to account for uncertainty about the scale parameter g is to specify a hyper-prior for g and perform fully Bayesian inference. Liang et al [32] proposed the hyper-g prior

$$\pi(g) = \frac{a-2}{2}(1+g)^{-a/2},$$

which is proper for $a > 2$. Liang et al [32] recommended $a = 3$ as a default choice for the hyper-g prior. One advantage of

using a hyper-g prior is that the posterior distribution of g given the model $\mathcal{M}_{\boldsymbol{\gamma}}$ is available in closed form, simplifying Bayesian inference.

In terms of theoretical properties, all three priors are model-selection consistent [32], except when the true model is the null model. This means that if the true model, denoted by $\mathcal{M}_{\boldsymbol{\gamma}_T}$ belongs to the model space, then the posterior probability of the true model, $P(\mathcal{M}_{\boldsymbol{\gamma}} = \mathcal{M}_{\boldsymbol{\gamma}_T}|\boldsymbol{y}) \to 1$ as the sample size $n \to \infty$. None of the above priors suffers from Bartlett's paradox [1]. BMA with $g = \sqrt{n}$ is subject to the "information paradox", but it has been argued that information consistency is of little practical importance in real data applications [37]. The EB-local and Hyper-g BMA methods are not subject to the information paradox.

2.2 Choice of Model Space Priors

Model space priors require specification of the prior probabilities of all models $\mathcal{M}_{\boldsymbol{\gamma}}$, indexed by the binary variable inclusion vector $\boldsymbol{\gamma}$. A common approach is to consider the inclusion of each variable as an independent and exchangeable Bernoulli random variable with a common prior probability of inclusion θ, i.e.

$$p(\mathcal{M}_{\boldsymbol{\gamma}}|\theta) = \prod_{i=1}^{p} \theta^{\gamma_i}(1-\theta)^{1-\gamma_i} = \theta^{p_{\boldsymbol{\gamma}}}(1-\theta)^{p-p_{\boldsymbol{\gamma}}}, \quad (2.1)$$

where θ is the prior expected fraction of the β_j's that are not zero and $p_{\boldsymbol{\gamma}} = \sum_{i=1}^{p} \gamma_i$ is the total number of covariates included in the model $\mathcal{M}_{\boldsymbol{\gamma}}$.

In the absence of prior information, a common choice is to set $\theta = 0.5$. This induces a uniform prior over all models with $p(\mathcal{M}_{\boldsymbol{\gamma}}) = 2^{-p}$ for all $\boldsymbol{\gamma} \in \{0,1\}^p$, where p is the total number of covariates considered. The expected prior model size under the uniform model prior is $p/2$. However, choosing $\theta = 0.5$ does not provide any multiplicity control [19].

For a fixed value of θ, the above prior induces a binomial prior for model size $S = \sum_{i=1}^{p} \gamma_i$ i.e. $S \sim Bin(p, \theta)$, with prior mean $p\theta$ and prior variance $p\theta(1-\theta)$. Another way to specify θ is by using the researcher's prior belief about expected model size $E[S]$. Sala-i-Martin et al [46] (hereafter SDM) recommended a prior expected model size of 7 based on their growth regression analysis. Similar priors for expected model size have also been proposed elsewhere [45, 30]. Hence, we can define an SDM version of the above prior with $\theta_{SDM} = 7/p$.

Any fixed choice of θ can lead to rather informative priors on model size $p_{\boldsymbol{\gamma}}$. One way to address this issue is by estimating θ from the data using an empirical Bayes (EB) approach [19]. The EB approach involves maximizing the marginal likelihood of the data given θ:

$$\hat{\theta}_{EB} = \arg\max_{\theta \in [0,1]} \sum_{\boldsymbol{\gamma}} P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}})P(\mathcal{M}_{\boldsymbol{\gamma}}|\theta). \quad (2.2)$$

However, maximization of (2.2) can be computationally challenging, especially when p is large since the sum is over

all models. Moreover, when p is large, marginal likelihood evaluation for all models is not feasible and the sum is approximated based on the models explored by MCMC. To optimize the marginal likelihood in (2.2), we implement Algorithm 1, iterating between fitting the BMA approach to find likely models given θ and solving (2.2) using the fitted models to find a new θ:

Algorithm 1 EB optimisation algorithm for $\hat{\theta}_{EB}$.

$probs \leftarrow 0.5$
Fit mod$\leftarrow$ `bas.lm(..., model.prior=bernoulli(probs))`
Initialize $\theta^{(0)} \leftarrow \frac{\hat{p}}{p}$ $\triangleright \hat{p}$ is avg. posterior model size
$i \leftarrow 1$
repeat
    $probs \leftarrow \theta^{(i-1)}$
    mod$\leftarrow$ `bas.lm(..., model.prior=bernoulli(probs))`
    Optimise (2.2) over models explored in `mod` to find $\theta^{(i)}$
    $i \leftarrow i + 1$
until convergence

An alternative way to reduce the sensitivity of the posterior distribution to prior assumptions is to use hierarchical modeling and specify a weak hyper-prior for θ. One choice of such a hyper-prior is a Beta distribution, $\theta \sim \text{Beta}(a, b)$. Marginalizing out θ in (2.1), gives

$$\begin{aligned} P(\mathcal{M}_{\boldsymbol{\gamma}}|a, b) &= \int_0^1 p(\mathcal{M}_{\boldsymbol{\gamma}}|\theta)p(\theta)d\theta \\ &= \frac{B(p_{\boldsymbol{\gamma}} + a, p - p_{\boldsymbol{\gamma}} + b)}{B(a, b)}, \end{aligned} \quad (2.3)$$

where $B(a', b') = \frac{\Gamma(a')\Gamma(b')}{\Gamma(a'+b')}$ is the Beta function. It thus induces a Beta-Binomial(a, b) prior on the model size S with probability mass function

$$P_S(s) = \binom{p}{s}\frac{B(s + a, p - s + b)}{B(a, b)}.$$

Under a uniform prior on θ, i.e. when $a = b = 1$, (2.3) simplifies to $p(\mathcal{M}_{\boldsymbol{\gamma}}) = \frac{1}{p+1}\binom{p}{p_{\boldsymbol{\gamma}}}^{-1}$. This is a combination of a uniform prior over model size with a uniform prior over the models of same size given the model size.

Under a Beta-Binomial (BB) prior, the prior expected model size is $E[S] = \frac{a}{a+b}p$. Similarly to a Bernoulli prior, we can elicit the prior in terms of the prior expected model size $E[S]$. To facilitate prior elicitation, we fix $a = 1$. We can then define an SDM version of the BB prior (BB-SDM) with an expected prior model size, such as $E[S] = 7$, by setting $b_{SDM} = \frac{p}{E[S]} - 1$. Note that SDM themselves [46] did not use a Beta-Binomial prior on models, but only a Bernoulli prior.

Alternatively, we can use an EB approach to learn b from the data. This can be done by maximizing the marginal

likelihood given b, namely

$$\hat{b}_{EB} = \underset{b \in (0, \infty)}{\arg\max} \sum_{\boldsymbol{\gamma}} P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}) P(\mathcal{M}_{\boldsymbol{\gamma}}|a=1, b). \qquad (2.4)$$

We find the optimal value, $\hat{b}_{EB}$, using Algorithm 2.

Algorithm 2 EB optimisation algorithm for $\hat{b}_{EB}$.

$b \leftarrow 1$
Fit `mod`← `bas.lm(..., model.prior=beta.binomial(1,b))`
Initialize $b^{(0)} \leftarrow \frac{p}{\hat{p}} - 1$ $\triangleright$ $\hat{p}$ is avg. posterior model size
$i \leftarrow 1$
repeat
    $b \leftarrow b^{(i-1)}$
    `mod`← `bas.lm(..., model.prior=beta.binomial(1,b))`
    Optimise (2.4) over models explored in `mod` to find $b^{(i)}$
    $i \leftarrow i + 1$
until convergence

For Zellner's g-prior, we require the model size to be no larger than the number of regression coefficients that can be identified from the data, so that $p_{\boldsymbol{\gamma}} < n - 2$. Thus for higher dimensional datasets ($p > n$), we require that $P(\mathcal{M}_{\boldsymbol{\gamma}}) = 0$ for all models with model size greater than $n - 2$. Hence, we use truncated versions of the priors defined in (2.1) and (2.3), namely

$$p(\mathcal{M}_{\boldsymbol{\gamma}}|\theta) \propto \theta^{p_{\boldsymbol{\gamma}}}(1-\theta)^{p-p_{\boldsymbol{\gamma}}} \mathbb{1}\{p_{\boldsymbol{\gamma}} < n-2\}, \qquad (2.5)$$

$$P(\mathcal{M}_{\boldsymbol{\gamma}}|a, b) \propto \frac{B(p_{\boldsymbol{\gamma}}+a, p-p_{\boldsymbol{\gamma}}+b)}{B(a, b)} \mathbb{1}\{p_{\boldsymbol{\gamma}} < n-2\}. \qquad (2.6)$$

Castillo et al [3] introduced complexity priors, also known in the literature as diffusing [35] or power priors [5]. Here the marginal probability of inclusion of any variable decays at the rate $p^{-\kappa}$ for some $\kappa > 0$, where p is the total number of possible covariates. This specifies a vanishing prior probability of large models and leads to a faster rate of rejection of spurious parameters, at the cost of slower rates of detection of active parameters [44]. Similar priors have also been used elsewhere [51, 35, 42].

The complexity prior is defined as

$$p(\mathcal{M}_{\boldsymbol{\gamma}}) \propto p^{-\kappa|\boldsymbol{\gamma}|} \mathbb{1}\{|\boldsymbol{\gamma}| \leq s_0\},$$

where s_0 is a pre-specified integer specifying the maximum number of important covariates and $|\boldsymbol{\gamma}|$ is the model size. In the absence of external information, we set $s_0 = \min\{n-2, p\}$. This prior is implemented in the BAS package as `tr.power.prior(kappa,trunc)`. We implement Complexity priors with $\kappa = 1$ [43, 44], and with $\kappa = 2$ which is the default choice in the BAS package.

2.3 Model Space Priors – A Graphical Illustration

To illustrate the effect of different model space priors, we use two datasets from our analysis: Boston Housing ($n = 506, p = 103$) and Nutrimouse ($n = 40, p = 120$) (Figure 1). The solid lines show the independent Bernoulli prior from (2.1), while the dashed lines represent the Beta-Binomial prior in (2.3) and the dash-dotted lines illustrate
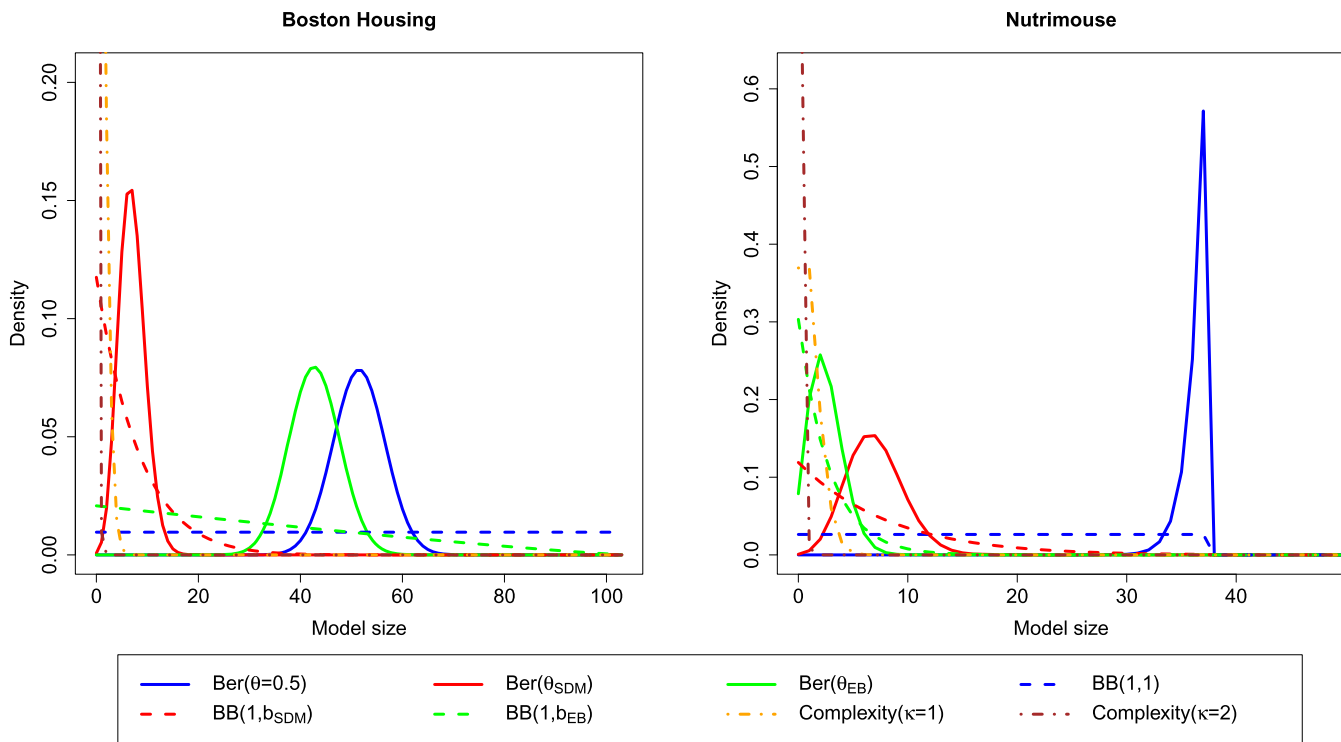

Figure 1: Prior model size distribution for the Boston Housing and Nutrimouse datasets.

Complexity priors. For the Nutrimouse dataset $(p > n)$, we use the truncated versions as discussed above. The colors group different flavors of methods: (i) Uniform versions with $\theta = 0.5$ or $b = 1$ (blue), (ii) SDM versions with expected prior model size 7 (red), (iii) EB versions with θ or b learned from the data (green), (iv) Complexity prior with $\kappa = 1$ (orange), and (v) Complexity prior with $\kappa = 2$ (brown).

The Bernoulli model space priors are very concentrated around their mean, $p\theta$. The complexity priors are concentrated around smaller model sizes with a mode at 0. The BB priors, on the other hand, are more diffuse, implying more prior uncertainty about model size. For the Nutrimouse dataset, all the model space priors assign zero probability to any model with size greater than $(n - 2) = 38$. Among the Bernoulli versions, $\theta = 0.5$ implies a prior mode around $\min\{p/2, n-2\}$ while θ_{SDM} has a prior model size of 7 (the same as the prior mean). The prior model size distribution induced by the $\text{Ber}(\theta_{EB})$ prior adapts based on the data, with a prior mode between $\theta = 0.5$ and θ_{SDM} for the Boston Housing dataset, while having the lowest prior mode among the Bernoulli priors considered for the Nutrimouse dataset. The $\text{BB}(1, 1)$ prior corresponds to a uniform prior over model size. The $\text{BB}(1, \theta_{SDM})$ and $\text{BB}(1, \theta_{EB})$ priors both induce a model size distribution with prior mode at zero.

3. NUMERICAL COMPARISON

We investigate the performance of different model space priors and parameter prior combinations using an extensive simulation study based closely on real datasets. We evaluate the effect of prior choices for the statistical tasks of parameter point and interval estimation, inference, point and interval prediction, and computation time.

All the parameter and model space prior combinations were implemented using the `BAS` R package [5] with skeleton code shown in Table 1. A combination of the MC^3 Metropolis-Hastings algorithm for sampling from the posterior distribution of models [40], along with a random swap between a currently included and a currently excluded variable is used for model space exploration. This is implemented by setting the option `method="MCMC"` in the `bas.lm()` function. We used a default of 10,000 MCMC iterations for all methods.

For the EB methods, we used Algorithm 1 (or 2) to find the optimal θ_{EB} (or b_{EB}) before fitting a BAS model with the estimated hyperparameter value. For

higher dimensional datasets $(p > n)$, a truncated version of the Beta-Binomial prior (2.6) was implemented by setting the option `model.prior="tr.beta.binomial(alpha,beta,trunc=n-2)"` in BAS. Similarly, a truncated version of complexity prior is implemented in the BAS package. A truncated version of the Bernoulli prior (2.5) is not available in BAS. We implemented it by (i) implementing `bas.lm()` with `tr.beta.binomial(1,1,trunc=n-2)`, and then (ii) using importance sampling to calculate updated posterior model probabilities with weights proportional to the ratio of the prior model space densities in (2.5) and (2.6).

3.1 Datasets

We based our analysis on 14 publicly available datasets, of which six are available from UCI machine learning repository and the others are examples available in the literature. The sample size and number of candidate variables along with the data sources for the different datasets are listed in Table 2. These include the classical statistical setting with $n > p$, high dimensional datasets with $p > n$, and intermediate settings where $n \approx p$. For each dataset, continuous covariates are standardized to have zero mean and variance 1 and the response variable is centered to have zero mean. Details of the choice of datasets and additional pre-processing can be found in [37].

Table 2. Datasets used in the study.

Dataset Name	Sample size (N)	Covariates (p)	Source
College	777	14	ISLR [27]
Bias Correction-Tmax	7590	21	UCI ML repository
Bias Correction-Tmin	7590	21	UCI ML repository
SML2010	1373	22	UCI ML repository
Bike sharing-daily	731	28	UCI ML repository
Bike sharing-hourly	17379	32	UCI ML repository
Superconductivity	21263	81	UCI ML repository
Diabetes	442	64	spikeslab [26]
Ozone	330	44	gss [22]
Boston housing	506	103	mlbench [36]
NIR	166	225	chemometrics [14]
Nutrimouse	40	120	mixOmics [41]
Multidrug	60	853	mixOmics [41]
Liver toxicity	64	3116	mixOmics [41]

3.2 Simulation Design

For each dataset, we selected a data generating model that closely approximates the real dataset. We carry out all-subsets regression for datasets with $p < 30$ using the `leaps` package [33] in R. We then selected the largest model with

Table 1. Summary of prior moments of model size S under different model space priors and BAS code to implement them.

Model prior	E[S]	Var(S)	BAS code
Bernoulli(θ)	$p\theta$	$p\theta(1 - \theta)$	`bas.lm(..., model.prior=bernoulli(probs))`
Beta-Binomial(a, b)	$\frac{pa}{a+b}$	$\frac{pab(a+b+p)}{(a+b)^2(a+b+1)}$	`bas.lm(..., model.prior=beta.binomial(alpha,beta))`
Complexity(κ)	–	–	`bas.lm(..., model.prior=tr.power.prior(kappa,trunc))`

all variables significant at the 0.05 level. For datasets with $p > 30$, all subsets regression is not computationally feasible. For these datasets, we obtained a filtered list of variables using iterative sure independence screening [12]. If the filtered list contained more than 30 variables, we selected the top 30 variables with the highest R^2 values under univariate regression. All-subsets regression was then applied to the filtered list of covariates to obtain the data generating model for our study. A summary of the data generating model used for each dataset can be found in the supplementary materials.

We used the data generating model and parametric bootstrapping to generate 100 bootstrapped datasets with the same design matrix $\boldsymbol{X}$ and different simulated response vectors $\boldsymbol{Y}$. Each of the resulting simulated datasets had the same design matrix and error distribution as the real dataset on it was based.

We compared the performance of different parameter and model space prior combinations on these simulated datasets using the following metrics:

- **PointEst**: We use root mean squared error (RMSE) as a metric to evaluate the parameter point estimation performance of different combinations. RMSE is evaluated as

$$RMSE = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(\beta_{i,DG} - \hat{\beta}_i)^2}, \qquad (3.1)$$

  where $\beta_{i,DG}, i = 1, \ldots, p$ denote the coefficients in the data generating model, and $\hat{\beta}_i, i = 1, \ldots, p$ denote the posterior means of the coefficients.
- **IntEst**: The interval score (IS) [21] evaluates the performance of interval estimators in terms of both their coverage and width. It is the sum of two terms, the first of which rewards narrow intervals while the second rewards accurate coverage. For a variable z, the IS is

$$IS_\alpha(l, u, z) = (u - l) + \frac{2}{\alpha}(l - z)\mathbb{1}\{z < l\}$$
$$+ \frac{2}{\alpha}(z - u)\mathbb{1}\{u < z\}, \qquad (3.2)$$

  where l and u denote the upper and lower bounds of the $(1-\alpha)\times100\%$ posterior interval of z. The Mean Interval Score (MIS) is the average of the IS values for the quantities being estimated. In order to assess the quality of the interval estimation, we compute the Mean Interval score (MIS) for the coefficients and calculate the average MIS across coefficients for each of the datasets. We use $\alpha = 0.05$.
- **Inference**: We calculate the area under the precision recall curve (AUPRC) using the posterior inclusion probabilities of the covariates to evaluate the model selection performance of different combinations of priors. We assess the quality of the resulting inference using

$(1-\text{AUPRC})$ as our metric, where a lower value is better.

We also compared methods based on their out-of-sample predictive performance. We divided each dataset into 100 random 75%–25% train-test splits. We trained the methods on the training data and used the test data to assess the predictive performance using the metrics described below:

- **Prediction**: We calculate R_{test}^2 to evaluate accuracy of point predictions as follows:

$$R_{test}^2 = 1 - \frac{\sum_{i\in test}(y_i - \hat{y}_i)^2}{\sum_{i\in test}(y_i - \bar{y}_{train})^2}, \qquad (3.3)$$

  where $\{y_i : i \in test\}$ denotes the response variable of the test set, $\hat{y}_i$ denotes the corresponding predictions, and $\bar{y}_{train}$ denotes the mean of the response variable in the training set.
- **IntPred**: To assess the quality of the prediction intervals, we calculate the interval score using (3.2) for each of the test set observations. Here, l and u represent the lower and upper bounds of the $(1-\alpha)\times100\%$ posterior predictive interval for the test observation. We calculate the mean interval score (MIS), averaging IS over test set observations for each of the train-test splits. A lower MIS corresponds to a better interval forecast.

We also recorded the average size of the sampled models for each dataset and the average CPU time (in seconds) to carry out BMA for one bootstrapped dataset.

3.3 Results

The results are shown in Table 3. We used the combination of the g-prior with $g = \sqrt{n}$ as the parameter prior and the Beta-Binomial$(1, 1)$ model space prior as the reference. Note that the g-prior with $g = \sqrt{n}$ was found to the best parameter prior by [37]. Metrics for all other combinations were calculated relative to the reference metric, and averaged across datasets. Detailed results of performance metrics for the simulation studies based on each of the 14 datasets can be found in the Supplementary materials. The "Score" column contains the average of the scores for PointEst, IntEst, Inference, Prediction and IntPred under each method. We used the Score column to rank the methods.

For each metric, we color the methods based on their performance relative to the reference metric. A method is colored green if it performed similarly or better than the reference method, yellow if it performed somewhat worse, and orange if it performed substantially worse.

For all choices of parameter prior, Beta-Binomial$(1, 1)$ was the top scoring model space prior. The three Beta-Binomial versions with $g = \sqrt{n}$ were the top three methods across statistical tasks. The Complexity priors with $\kappa = 1$ and $\kappa = 2$ were the worst performing model space priors. The uniform prior denoted by Ber$(\theta = 0.5)$ also performed

Table 3. Performance of different parameter prior and model space prior combinations for inference in linear regression under model uncertainty: "PointEst" is the RMSE for point estimation, "IntEst" is the Mean Interval Score (MIS) for interval estimation, "Inference" is the 1- area under the precision-recall curve (AUPRC), "Prediction" is the RMSE for point prediction, while "IntPred" is the MIS for interval prediction. "N vars" is the average number of variables used for the task. All metrics are standardized to equal 1 for the $g = \sqrt{n}$ with $BB(1,1)$ prior on model space. For each column, lower value is better.

Rank	Parameter prior	$P(\mathcal{M}_\gamma)$	Score	PointEst	IntEst	Inference	Prediction	IntPred	N vars	CPU time
1	$g = \sqrt{n}$	$BB(1,1)$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	$g = \sqrt{n}$	$BB(1,b_{SDM})$	1.005	1.005	1.005	1.013	1.001	1.001	0.918	1.009
3	$g = \sqrt{n}$	$BB(1,b_{EB})$	1.012	1.004	1.025	1.026	1.004	1.001	0.960	3.136
4	Hyper-g	$BB(1,1)$	1.015	1.024	1.062	0.988	1.006	0.994	0.837	3.271
5	EB-local	$BB(1,1)$	1.024	1.037	1.071	1.001	1.012	0.999	0.853	0.803
6	EB-local	$BB(1,b_{EB})$	1.025	1.039	1.091	0.992	1.007	0.994	0.811	2.441
7	Hyper-g	$BB(1,b_{EB})$	1.025	1.026	1.091	0.996	1.015	0.996	0.793	9.849
8	EB-local	$BB(1,b_{SDM})$	1.042	1.053	1.124	1.006	1.023	1.002	0.805	0.812
9	Hyper-g	$BB(1,b_{SDM})$	1.042	1.047	1.139	1.009	1.017	0.996	0.786	3.287
10	$g = \sqrt{n}$	$Ber(\theta_{SDM})$	1.052	1.062	1.150	1.036	1.000	1.011	0.837	0.849
11	$g = \sqrt{n}$	$Ber(\theta_{EB})$	1.085	1.091	1.275	1.053	1.004	1.005	1.068	2.595
12	EB-local	$Ber(\theta_{EB})$	1.104	1.115	1.360	1.011	1.026	1.009	0.988	2.163
13	Hyper-g	$Ber(\theta_{EB})$	1.105	1.115	1.358	1.021	1.025	1.010	0.964	9.483
14	Hyper-g	$Ber(\theta_{SDM})$	1.130	1.121	1.472	1.030	1.020	1.007	0.761	3.179
15	EB-local	$Ber(\theta_{SDM})$	1.143	1.132	1.500	1.034	1.037	1.014	0.772	0.730
16	$g = \sqrt{n}$	Uniform	1.200	1.184	1.400	1.165	1.179	1.070	1.343	0.846
17	Hyper-g	Uniform	1.215	1.186	1.482	1.083	1.198	1.125	1.200	3.172
18	EB-local	Uniform	1.256	1.181	1.508	1.092	1.295	1.204	1.211	0.721
19	$g = \sqrt{n}$	Complexity(1)	1.298	1.222	2.045	1.152	1.047	1.023	0.558	0.670
20	EB-local	Complexity(1)	1.409	1.303	2.515	1.162	1.046	1.020	0.546	0.707
21	Hyper-g	Complexity(1)	1.415	1.305	2.526	1.171	1.054	1.017	0.537	3.492
22	$g = \sqrt{n}$	Complexity(2)	1.787	1.490	3.652	1.444	1.255	1.096	0.406	0.680
23	EB-local	Complexity(2)	1.863	1.544	4.032	1.432	1.227	1.081	0.411	0.732
24	Hyper-g	Complexity(2)	1.872	1.553	4.045	1.434	1.245	1.085	0.407	4.023

less well than the Beta-Binomial priors. This ranking of methods was consistent across different performance metrics.

Most parameter and model prior combinations selected sparser models than the $g = \sqrt{n}$ and $BB(1,1)$ prior combination, with the exception of methods involving the uniform model space prior. The complexity priors selected very sparse models compared to our baseline, which may be explained by the strong sparsity induced by the prior. This may also explain the poor performance of the complexity priors across statistical tasks. Notably, the rankings of the prior combinations are similar for the different tasks. In particular, the rankings for prediction are consistent with those for point estimation and parameter inference, with a correlation of 0.77 between scores for point estimation and point prediction.

We also note that the EB model space priors tended to outperform the corresponding SDM model space priors when combined with the Hyper-g and EB-local parameter priors. However, the results with the EB model space priors took longer to computer on average because of the optimisation procedure. The hyper-g parameter priors are the slowest due to the integral calculations required in the posterior computation. In general, the Beta-Binomial priors performed better than the Bernoulli and complexity priors.

4. DISCUSSION

We have compared BMA techniques with different choices of model space priors and parameter priors using an empirical study based closely on real datasets. We found that the Beta-Binomial(1,1) model space prior performed the best across various statistical tasks and choices of parameter priors. We found that the hierarchical model space priors with a hyper-prior on the prior inclusion probability θ was more diffuse and led to more efficient exploration of the model space. Fixed choices of θ led to worse performance across statistical tasks and were often quite concentrated. Complexity priors that induce high sparsity on model complexity performed worst among all the methods considered.

We are not the first to compare model space priors in the presence of model uncertainty. Past comparisons have either focused on a subset of the model priors discussed here, or evaluated BMA methods for only a subset of the statistical tasks considered here. In several cases, they also tended to use simulation designs that are at best loosely related to empirical data observed in practice.

Ley and Steel [31] evaluated the effect of different model priors on model selection performance using three real economic growth regressions datasets. However, they used only two fixed choices of g-priors: the Unit Information prior (UIP) with $g = n$ [28] and the risk inflation criterion (RIC) with $g = p^2$ [16], motivated by the simulation study of [13].

Porwal and Raftery [37] found both of these parameter prior choices to be outperformed by the parameter priors used in this study. Also, their comparison was based only on tall datasets ($n > p$) and their comparison of methods was limited to the statistical tasks of inference and probabilistic prediction using the log-predictive score. They also did not consider EB versions of the Binomial(p, θ) and Beta-Binomial($1, b$) model space priors and complexity priors. Like Ley and Steel, we found that random θ versions (or Beta-Binomial versions) performed better since the hierarchical prior is less sensitive to the choice of prior model size $E[S]$. Similarly, they found that priors specified by a fixed θ tended to be quite informative, casting doubt on their appropriateness as default reference priors.

Scott and Berger [47] discussed the multiplicity correction effect of a subset of the model space priors discussed here, specifically Ber($\theta = 0.5$), Ber(θ_{EB}) and BB($1, 1$). They used a non-empirical simulation design, and did not compare methods based on the statistical tasks discussed here. Eicher et al [11] compared 12 parameter priors (of which $g = \sqrt{n}$ is common with ours) along with two fixed model priors: Uniform model priors with $\theta = 0.5$ and Ber(θ_{SDM}) with a prior expected model size of 7. The comparison was based on non-empirical simulation studies and one real growth regression dataset using predictive performance and inference measures. They found that the UIP with a uniform model prior performed better than Ber(θ_{SDM}) on the three statistical tasks common with ours. In contrast, we found that Ber(θ_{SDM}) was ranked higher than Uniform model priors for all our three preferred parameter priors across the statistical tasks considered.

We found the complexity priors [3] to perform relatively poorly. At first sight, this seems to be in conflict with the theoretical results of Castillo et al [3], who showed that under certain assumptions the posterior distribution contracts optimally to recover an unknown sparse parameter vector and gives optimal predictions. However, their theoretical results assume that the data are generated from a spike and slab prior with the Laplace distribution as the slab density, and that the error variance σ^2 is known, which rarely holds in practice. Also, Rossell [44] argued that complexity priors can introduce very strong sparsity a priori, and showed empirically that when the true model is not sparse, complexity priors may perform suboptimally for finite n. This is consistent with our results.

We have focused attention on independent model priors, i.e. priors in which the inclusion of each variable is statistically independent of that of the other variables. However, non-independent default priors have been proposed as well. George [17, 18] proposed dilution priors which dilute the prior model probability within subsets of similar models with highly correlated predictors. There is also research designing dependent model priors based on domain knowledge [2, 10]. Dellaportas et al [9] proposed a joint specification of the prior distribution across models so that the sensitivity

of posterior model probabilities to the dispersion of prior distributions for the parameters of individual models (Lindley's paradox) is diminished. Villa and Walker [49] assigned prior mass to models on the basis of their *worth*, based on the KL-divergence between densities under different models. However, all of these dependent model space priors lead to increased computational complexity and have been shown to work only when p is relatively small. They have also not yet been implemented in publicly available software.

SUPPLEMENTARY MATERIAL

The supplementary material contains detailed summary results for each metric and dataset used in the study. It also contains a summary of data-generating models for each of the datasets.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

[1] BARTLETT, M. S. (1957). A Comment on D. V. Lindley's Statistical Paradox. *Biometrika* **44** 533–534. https://doi.org/10.1093/biomet/52.3-4.507. MR0207142

[2] BROCK, W., DURLAUF, S. N. and WEST, K. D. (2003). *Policy evaluation in uncertain economic environments*. National Bureau of Economic Research Cambridge, Mass., USA.

[3] CASTILLO, I., SCHMIDT-HIEBER, J. and VAN DER VAART, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43**(5) 1986–2018. https://doi.org/10.1214/15-AOS1334. MR3375874

[4] CELEUX, G., EL ANBARI, M., MARIN, J. Q. M. and ROBERT, C. P. (2012). Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informative Situation. *Bayesian Analysis* **7** 477–502. https://doi.org/10.1214/12-BA716. MR2934959

[5] CLYDE, M. (2020). BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling. R package version 1.5.5.

[6] CLYDE, M. and GEORGE, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B, Statistical Methodology* **62**(4) 681–698. https://doi.org/10.1111/1467-9868.00257. MR1796285

[7] CLYDE, M. and GEORGE, E. I. (2004). Model uncertainty. *Statistical Science* **19**(1) 81–94. https://doi.org/10.1214/088342304000000035. MR2082148

[8] DECKERS, T. and HANCK, C. (2014). Variable Selection in Cross-Section Regressions: Comparisons and Extensions. *Oxford Bulletin of Economics and Statistics* **76**(6) 841–873.

[9] DELLAPORTAS, P., FORSTER, J. J. and NTZOUFRAS, I. (2012). Joint specification of model space and parameter space prior distributions. *Statistical Science* **27**(2) 232–246. https://doi.org/10.1214/11-STS369. MR2963994

[10] DURLAUF, S. N., KOURTELLOS, A. and TAN, C. M. (2008). Are any growth theories robust? *The Economic Journal* **118**(527) 329–346.

[11] EICHER, T. S., PAPAGEORGIOU, C. and RAFTERY, A. E. (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics* **26**(1) 30–55. https://doi.org/10.1002/jae.1112. MR2759908

[12] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B — Statistical Methodology* **70**(5) 849–911. https://doi.org/10.1111/j.1467-9868.2008.00674.x. MR2530322

[13] FERNÁNDEZ, C., LEY, E. and STEEL, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics* **100**(2) 381–427. https://doi.org/10.1016/S0304-4076(00)00076-2. MR1820410

[14] FILZMOSER, P. and VARMUZA, K. (2017). chemometrics: Multivariate Statistical Analysis in Chemometrics. R package version 1.4.2. https://CRAN.R-project.org/package=chemometrics.

[15] FORTE, A., GARCIA-DONATO, G. and STEEL, M. F. J. (2018). Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *International Statistical Review* **86**(2) 237–258. https://doi.org/10.1111/insr.12249. MR3852410

[16] FOSTER, D. P. and GEORGE, E. I. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics* **22**(4) 1947–1975. https://doi.org/10.1214/aos/1176325766. MR1329177

[17] GEORGE, E. (1999). Discussion of "Model averaging and model search strategies" by M. Clyde. In *Bayesian Statistics 6–Proceedings of the Sixth Valencia International Meeting.* MR1723497

[18] GEORGE, E. I. (2010). Dilution priors: Compensating for model space redundancy. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown* 158–165 Institute of Mathematical Statistics. MR2798517

[19] GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**(4) 731–747. https://doi.org/10.1093/biomet/87.4.731. MR1813972

[20] GEORGE, E. I. and MCCULLOCH, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**(423) 881–889.

[21] GNEITING, T. and RAFTERY, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**(477) 359–378. https://doi.org/10.1198/016214506000001437. MR2345548

[22] GU, C. (2014). Smoothing Spline ANOVA Models: R Package gss. *Journal of Statistical Software* **58**(5) 1–25.

[23] HANSEN, M. H. and YU, B. (2003). Minimum description length model selection criteria for generalized linear models. *Lecture Notes-Monograph Series* **40** 145–163. https://doi.org/10.1214/lnms/1215091140. MR2004337

[24] HOETING, J. A., RAFTERY, A. E. and MADIGAN, D. (2002). Bayesian variable and transformation selection in linear regression. *Journal of Computational and Graphical Statistics* **11**(3) 485–507. https://doi.org/10.1198/106186002501. MR1938444

[25] HOETING, J. A., MADIGAN, D., RAFTERY, A. E. and VOLINSKY, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science* **14** 382–417. https://doi.org/10.1214/ss/1009212519. MR1765176

[26] ISHWARAN, H., RAO, J. S. and KOGALUR, U. B. (2013). spikeslab: Prediction and variable selection using spike and slab regression. R package version 1.1.5. http://cran.r-project.org/web/packages/spikeslab/. https://doi.org/10.1214/21-ejp733. MR4366222

[27] JAMES, G., WITTEN, D., HASTIE, T. and TIBSHIRANI, R. (2017). ISLR: Data for an Introduction to Statistical Learning with Applications in R. R package version 1.2. https://CRAN.R-project.org/package=ISLR. https://doi.org/10.1007/978-1-0716-1418-1. MR4309209

[28] KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**(430) 773–795. https://doi.org/10.1080/01621459.1995.10476572. MR3363402

[29] LEAMER, E. E. (1978) *Specification Searches: Ad hoc Inference with Nonexperimental Data* **53**. Wiley. MR0471118

[30] LEVINE, R. and RENELT, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American economic review* 942–963.

[31] LEY, E. and STEEL, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of applied econometrics* **24**(4) 651–674. https://doi.org/10.1002/jae.1057. MR2675199

[32] 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. https://doi.org/10.1198/016214507000001337. MR2420243

[33] LUMLEY, T. (2020). leaps: Regression Subset Selection. R package version 3.1. https://CRAN.R-project.org/package=leaps.

[34] 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**(428) 1535–1546.

[35] NARISETTY, N. N. and HE, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics* **42**(2) 789–817. https://doi.org/10.1214/14-AOS1207. MR3210987

[36] NEWMAN, D. J., HETTICH, S., BLAKE, C. L. and MERZ, C. J. (1998). *UCI Repository of machine learning databases.* http://www.ics.uci.edu/~mlearn/MLRepository.html.

[37] PORWAL, A. and RAFTERY, A. E. (2022). Comparing methods for statistical inference with model uncertainty. *Proceedings of the National Academy of Sciences* **119**(16) 2120737119.

[38] RAFTERY, A. E. (1988). Approximate Bayes factors for generalized linear models. Technical Report No. 121, Department of Statistics, University of Washington. https://stat.uw.edu/sites/default/files/files/reports/1988/tr121.pdf.

[39] RAFTERY, A. E. and ZHENG, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association* **98**(464) 931–938.

[40] RAFTERY, A. E., MADIGAN, D. and HOETING, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**(437) 179–191. https://doi.org/10.2307/2291462. MR1436107

[41] ROHART, F., GAUTIER, B., SINGH, A. and LE CAO, K. Q. A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Computational Biology* **13**(11) 1005752.

[42] ROSSELL, D. (2021). Concentration of posterior model probabilities and normalized l0 criteria. *Bayesian Analysis* **1**(1) 1–27. https://doi.org/10.1214/21-ba1262. MR4483231

[43] ROSSELL, D. and RUBIO, F. J. (2018). Tractable Bayesian Variable Selection: Beyond Normality. *Journal of the American Statistical Association.* https://doi.org/10.1080/01621459.2017.1371025. MR3902243

[44] ROSSELL, D., ABRIL, O. and BHATTACHARYA, A. (2021). Approximate Laplace approximations for scalable model selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **83**(4) 853–879. MR4320004

[45] SALA-I-MARTIN, X. (1997). *I just ran four million regressions.* National Bureau of Economic Research Cambridge, Mass., USA.

[46] SALA-I-MARTIN, X., DOPPELHOFER, G. and MILLER, R. I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American economic review* **94**(4) 813–835.

[47] SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* 2587–2619. https://doi.org/10.1214/10-AOS792. MR2722450

[48] VAN ZWET, E. (2019). A default prior for regression coefficients. *Statistical Methods in Medical Research* **28**(12) 3799–3807. https://doi.org/10.1177/0962280218817792. MR4003623

[49] VILLA, C. and WALKER, S. (2015). An objective Bayesian crite-

rion to determine model prior probabilities. *Scandinavian Journal of Statistics* **42**(4) 947–966. https://doi.org/10.1111/sjos.12145. MR3426304

[50] WASSERMAN, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology* **44**(1) 92–107. https://doi.org/10.1006/jmps.1999.1278. MR1770003

[51] YANG, Y., WAINWRIGHT, M. J. and JORDAN, M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics* **44**(6) 2497–2532. https://doi.org/10.1214/15-AOS1417. MR3576552

[52] YOUNG, W. C., RAFTERY, A. E. and YEUNG, K. Y. (2014). Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Systems Biology* **8**(1) 47.

[53] ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques* **6**. MR0881437

[54] ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadística y de Investigaciów Operativa* **31**(1) 585–603.

Anupreet Porwal. Department of Statistics, University of Washington, Seattle, WA 98195, USA.
E-mail address: porwaa@uw.edu

Adrian E. Raftery. Department of Statistics, University of Washington, Seattle, WA 98195, USA.
E-mail address: raftery@uw.edu