PNAS RESEARCH ARTICLE | STATISTICS OPEN ACCESS

Comparing methods for statistical inference with model uncertainty

Anupreet Porwal[a] and Adrian E. Raftery[a,b,1]

Contributed by Adrian Raftery; received November 15, 2021; accepted February 18, 2022; reviewed by Mark Steel and Jonathan Wright

Probability models are used for many statistical tasks, notably parameter estimation, interval estimation, inference about model parameters, point prediction, and interval prediction. Thus, choosing a statistical model and accounting for uncertainty about this choice are important parts of the scientific process. Here we focus on one such choice, that of variables to include in a linear regression model. Many methods have been proposed, including Bayesian and penalized likelihood methods, and it is unclear which one to use. We compared 21 of the most popular methods by carrying out an extensive set of simulation studies based closely on real datasets that span a range of situations encountered in practical data analysis. Three adaptive Bayesian model averaging (BMA) methods performed best across all statistical tasks. These used adaptive versions of Zellner's *g*-prior for the parameters, where the prior variance parameter *g* is a function of sample size or is estimated from the data. We found that for BMA methods implemented with Markov chain Monte Carlo, 10,000 iterations were enough. Computationally, we found two of the three best methods (BMA with $g = \sqrt{n}$ and empirical Bayes-local) to be competitive with the least absolute shrinkage and selection operator (LASSO), which is often preferred as a variable selection technique because of its computational efficiency. BMA performed better than Bayesian model selection (in which just one model is selected).

Bayesian model averaging | interval estimation | LASSO | model selection | parameter estimation

Significance

Choosing a statistical model and accounting for uncertainty about this choice are important parts of the scientific process and are required for common statistical tasks such as parameter estimation, interval estimation, statistical inference, point prediction, and interval prediction. A canonical example is the choice of variables in a linear regression model. Many ways of doing this have been proposed, including Bayesian and penalized regression methods, and it is not clear which are best. We compare 21 popular methods via an extensive simulation study based on a wide range of real datasets. We found that three adaptive Bayesian model averaging methods performed best across all the statistical tasks and that two of these were also among the most computationally efficient.

Statistical analysis is often carried out using probability models for the data at hand. In this context, five of the most important statistical tasks are parameter estimation, interval estimation, inference about model parameters, point prediction, and producing prediction intervals.

These tasks often have to be carried out in the context of model uncertainty, where several different statistical models are plausible. One canonical example is variable selection in linear regression, where a set of candidate variables is considered, and all possible subsets of these candidate variables define possible models. Consider the linear regression model:

$$\boldsymbol{Y} = \alpha \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}),$$

where $\boldsymbol{Y} \in \mathcal{R}^n$ is a response variable of interest, and $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathcal{R}^{n \times p}$ is a set of p possible covariates, α is the scalar intercept, and $\boldsymbol{\beta}$ is the $p \times 1$ vector of regression coefficients. For concreteness, we focus on this example here. Other examples include the choice of functional forms of the variables and the choice of error distribution, for instance, to account for potential outliers.

Many methods have been proposed for statistical analysis using linear regression models in the presence of model uncertainty. When the model is known in advance and only its parameters have to be estimated, there is consensus on how to do statistical analysis using it, using either a frequentist or Bayesian approach. When the model is to be determined as part of the analysis, however, things are less clear, and the large number of competing approaches can leave it unclear how to proceed. Here we compare 21 of the most prominent methods.

Historically, one approach has been to determine the variables in a model subjectively using subject matter expertise, but this often leaves open questions, and a data-based approach is desired for at least some of the variables. Another approach is to always include all the candidate variables, but this can lead to poor statistical performance when there are many such variables. Many of the early statistical approaches were stepwise methods, in which variables were sequentially added or removed on the basis of significant tests, but these have not been found to have good theoretical or empirical properties (1, 2).

In the past 30 y, many more satisfactory methods have been proposed. Most of these are either Bayesian techniques or penalized likelihood-based approaches.

Many of the Bayesian techniques are some form of Bayesian model averaging (BMA) (3–6); several reviews of the BMA literature are available (7–13). The basic idea of BMA is that the predictive distribution of a quantity of interest (either a parameter or an observable

Author affiliations: [a]Department of Statistics, University of Washington, Seattle, WA 98195; and [b]Department of Sociology, University of Washington, Seattle, WA 98195

Author contributions: A.P. and A.E.R. designed research, performed research, analyzed data, and wrote the paper.

Reviewers: M.S., University of Warwick; J.W., Johns Hopkins University.

The authors declare no competing interest.

[1] To whom correspondence may be addressed. Email: raftery@uw.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2120737119/-/DCSupplemental.

Published April 11, 2022.

future quantity) is a weighted average of its predictive distributions under the different candidate models, where the weights are equal to the models' posterior probabilities given the data at hand.

BMA has some good theoretical properties (14). BMA point estimators and predictions minimize mean squared error; BMA estimation and prediction intervals are calibrated, and BMA predictive distributions have optimal performance in the log score sense (6). These properties hold on average over the prior distribution, extending similar results for Bayesian estimation (15), but the results are somewhat robust to this assumption (16). Used in this way, as a distribution of parameter values over which performance is averaged, the prior distribution has been referred to as the world distribution (17), the practical distribution (14), or the effect-size distribution (18), and analysis using this concept has been called empirical frequentist[*].

The implementation of BMA involves several choices by the user, including the prior distribution of the model parameters under each model and the prior model probabilities. Also, the number of candidate models can be too large for them all to be feasibly evaluated. For example, the number of possible subsets of p regression variables is 2^p; for p much beyond 25 or 30 this can be computationally prohibitive. Thus, the choice of analytic or computational approximations must also be made. Together these choices lead to many possible implementations of BMA.

For the parameter prior distribution in linear regression, several default choices have been proposed. Among the first was the Zellner–Siow Cauchy prior, with a standard Jeffreys prior for the intercept and error variance (17, 19). We treat this as a reference method and call it the Jeffreys–Zellner–Siow (JZS) prior.

Another early prior was the Zellner g-prior (20). Consider a binary vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ that indicates which explanatory variables are part of model $\mathcal{M}_{\boldsymbol{\gamma}}$, so that $\gamma_j = 1$ if the variable X_j is present in $\mathcal{M}_{\boldsymbol{\gamma}}$ and 0 if not. We use Zellner's g-prior in the form

$$\begin{aligned}\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \alpha, \sigma^2, g) &\sim \mathcal{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | 0, g\sigma^2 (X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}})^{-1}),\\ \pi_{\boldsymbol{\gamma}}(\alpha, \sigma) &\propto \sigma^{-1},\end{aligned}$$

where $\mathcal{N}$ denotes the multivariate normal distribution, and $X_{\boldsymbol{\gamma}}$ is the $n \times p_{\boldsymbol{\gamma}}$ matrix consisting of the covariates X_j for which $\gamma_j = 1$ (9). The prior variance of the regression parameters is controlled by the user-specified value g, and the effective prior sample size is n/g, where n is the sample size.

Various choices of g have been proposed (13). Zellner proposed using $g = n$, corresponding to a prior sample size of 1; this has been called the unit information prior (UIP) (21). Another choice is $g = 1$, corresponding to a prior sample size of n (22), one justification being that studies have sample sizes designed to have the power to detect effects of known sizes, so that the prior and sampling variances are similar. An intermediate choice is $g = \sqrt{n}$ (9), with a prior sample size of $\sqrt{n}$; this has been found to work well in high-dimensional settings (23). The benchmark prior where $g = \max\{n, p^2\}$ has also been recommended (9); it combines the consistency properties of the UIP with the good small sample performance of the risk inflation factor (RIC) (24).

The UIP can also be approximated by the Bayesian information criterion (BIC) (25, 26). The Akaike information criterion (AIC) can be used as the basis for an approximation to the posterior model probabilities under a prior that is similar to Zellner's g-prior with $g = 1$, i.e., with an equivalent prior sample size of n (27, 28).

An alternative is not to use a specified g but instead to estimate g from the data. This can be done in an empirical Bayes way, either for each model separately (29) or globally (30, 31). It can also be done in a more fully Bayesian way, by specifying a prior on g, such as the hyper-g approach (32).

A different type of prior used in BMA is the nonlocal prior (NLP) (33, 34), which removes mass close to zero. The horseshoe (35) is a Bayesian method but not a BMA method, with a prior that favors sparsity. The spike and slab method approximates the zero values of lower-dimensional models with continuous distributions around zero (5, 36).

In the frequentist setting, penalized likelihood approaches convert the variable selection problem into an optimization problem. The function to be optimized usually involves the squared error loss function with a penalty term $h_\lambda(\boldsymbol{\beta})$ on the coefficients $\boldsymbol{\beta}$, in which case

$$\begin{aligned}\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathcal{R}^p}{\arg\min} (\boldsymbol{Y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta})^T \\ (\boldsymbol{Y} - \alpha \mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}) + h_\lambda(\boldsymbol{\beta}).\end{aligned} \quad \textbf{[1]}$$

The estimates from these techniques can also be viewed as maximum a posteriori (MAP) estimates under a prior of the form $p(\boldsymbol{\beta}) \propto \exp\{-h_\lambda(\boldsymbol{\beta})\}$. The least absolute shrinkage and selection operator (LASSO) (37) was the first and remains perhaps the most widely used technique in this class, where the penalty takes the form $h_\lambda(\boldsymbol{\beta}) = \lambda \sum_{j=1}^{p} |\beta_j|$ and constrains the l_1 norm of the parameter vector. The popularity of the LASSO is due to factors that include the computational efficiency of the least angle regression and coordinate ascent algorithms that can be used to estimate it (38, 39); its ability to provide a sparse estimate of $\boldsymbol{\beta}$; and the oracle property, namely, that the LASSO will asymptotically find a superset of the correct predictors (40).

However, LASSO also suffers from several known issues. The oracle property ensures only that the true predictors will be asymptotically part of the selected model but not the converse, so that there can be many false positive selections, even asymptotically. LASSO also tends to overshrink the true signals in the observed data and hence produce biased estimates (41). It can also be unstable in the presence of highly correlated covariates. As pointed out by Holmes (ref. 42, p. 280), "In the presence of strong correlations between predictors with differing effect sizes, frequentist sparsity approaches, including the lasso, will tend to select a single variable within a group of collinear predictors, discarding the others in the pursuit of sparsity. However, the choice of the particular predictor might be highly variable and by selecting one we may ignore weak (but important) predictors which are highly correlated with stronger predictors."

The LASSO has a constant rate of penalization for all coefficients which can cause excessive shrinking of nonzero components. Some of the other popular penalty methods vary in the shape or rate of penalty applied to the coefficients. The smoothly clipped absolute deviation (SCAD) (43) and minimax concave penalty (MCP) (44) methods involve a nonconvex penalty which is constant for smaller coefficients and decreases to 0 for larger coefficients. The elastic net (45) involves a convex combination of ridge and LASSO penalties, encouraging grouping effects among strongly correlated variables, and thus addresses one concern mentioned above for the LASSO. Like the LASSO, these methods threshold some coefficients to zero, leading to simultaneous variable selection and estimation.

A common issue with penalized likelihood approaches is the lack of uncertainty quantification since variable selection is an outcome of the constrained optimization problem and not a

[*] J. O. Berger, World Meeting of the International Society for Bayesian Analysis, June 28–July 2, 2021, online.

probabilistic statement of inclusion (46–48). As a result, the zeros induced may not be the same zeros that one would get from a full variable selection approach (49). They also do not provide a way to account for model uncertainty. Expectation–maximization variable selection (EMVS) (50) and the spike-and-slab LASSO (SS LASSO) (51) are two methods that synthesize ideas from BMA and penalized likelihood. In principle, they could quantify uncertainty, but that has not yet been implemented in the associated software.

It is not clear which of the many proposed methods to use. Among penalized likelihood methods, LASSO probably remains the most used, perhaps because it was the first one proposed, there is a well-defined software package to implement it (the glmnet R package), and it is fast (52). Among Bayesian methods there is less clarity, and the relative performance of Bayesian and penalized likelihood methods is also not clear.

To clarify this, we carried out an extensive set of simulation studies based closely on real datasets that span a range of situations encountered in practical data analysis. This is in contrast with many simulation studies in the statistical literature whose design is determined by the investigators without direct reference to data. The simulation design, the metrics, and the underlying datasets are described in *Materials and Methods*. Fig. 1 shows the sample size and the number of candidate variables for the different datasets. These include classic statistical situations where the sample size is much larger than the number of variables, high-dimensional situations where the number of variables exceeds the sample size, and intermediate situations where the two are of the same order of magnitude.

Results

The results are shown in Fig. 2. Performance metrics are shown for all 21 methods for each of point estimation, interval estimation, inference, prediction, and interval prediction. All metrics are relative to the score for the JZS method, taken as the reference, and averaged across datasets. Detailed results of performance metrics for the simulation studies based on each of the 14 datasets can be found in *SI Appendix*. The score column shows the average of the five metrics for each method. For seven of the methods, interval estimation and interval prediction metrics were not available as the methods did not provide uncertainty assessments, and so we calculated the "PartScore," which is the average of the three remaining metrics. In all cases, a lower score is better.

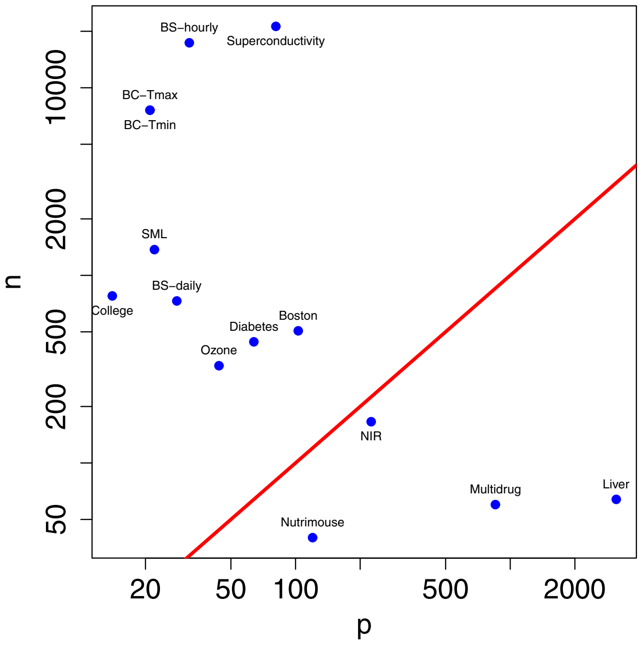


Fig. 1. Sample size n versus the number of candidate variables p for the 14 datasets on which our simulation studies are based. The $n = p$ line is shown in red.

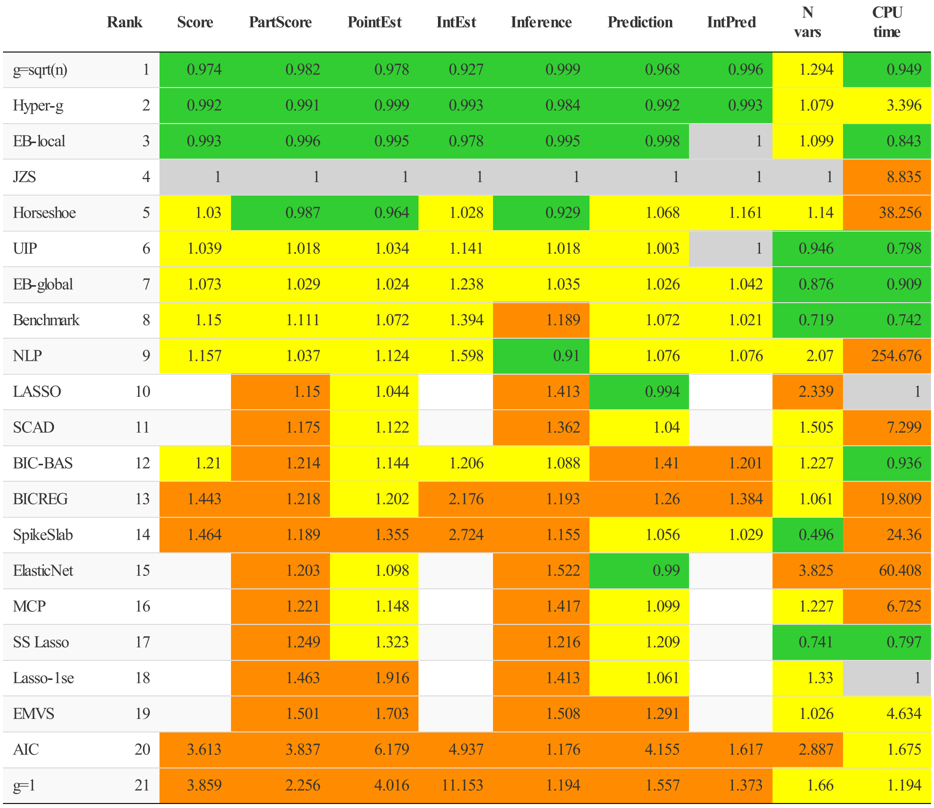
We first ranked the methods according to Score. We then ranked the methods for which Score was not available according to PartScore, ranking each one as highly as possible without changing the Score order. Results are colored green if the method performed better than the reference JZS method, while they are colored red if there was a substantial gap between them and the best methods. Yellow indicated that the method did not perform as well as the reference method but was not substantially worse than competing methods either. We also showed the average number of variables used and the central processing unit (CPU) time. For CPU time, LASSO was taken as the reference as it has generally been viewed as a computationally efficient method.

Overall, the ranking of the methods was similar from the different metrics. Strikingly, the venerable JZS method, now in its fifth decade, performed well and was competitive with all other methods, except that it required more CPU time than many. The top scoring methods were three adaptive g-prior methods: $g = \sqrt{n}$, the hyper-g method, and the local empirical Bayes method, which were the only methods to consistently outperform the reference method. Other Bayesian methods with nonadaptive priors rounded out the top eight spots. Interestingly, $g = 1$ and AIC were the worst performing methods. An advantage of the Bayesian methods is that they organically yield uncertainty statements, unlike the penalized likelihood methods.

LASSO was the top penalized likelihood method, doing particularly well for point prediction, as did the Elastic Net—comparable to the top Bayesian methods for this task, although not for the other tasks. However, they both selected far more variables on average than the Bayesian methods—twice as many or more in most cases without any noticeable increase in predictive performance. Plots of prediction accuracy, given by R^2, versus average model size, denoted by $\hat{p}$, for all datasets are available in *SI Appendix*.

A surprise was that two of the top three methods were efficient computationally even though they were Bayesian, comparable to LASSO despite the reputation of Bayesian methods for being slow. This is partly because we used a default of 10,000 Markov chain Monte Carlo (MCMC) iterations, which is far fewer than the default in the BAS R package used to implement these methods (53). This clearly gave adequate performance. Performance might be improved slightly with more iterations but at the cost of computational efficiency. The hyper-g method is substantially slower, which seems to be due to its greater complexity, but this may be a worthwhile tradeoff given its good performance. Several of the other methods were extremely slow. One needs to be cautious in interpreting the CPU time results as they reflect the coding efficiency of the implementations as well as the intrinsic computational efficiency of the methods. For most methods we used the developers' packages with default settings, and these could clearly often be sped up.

One question is whether inferences are sensitive to the choice of model selection/model averaging method. To provide a partial answer, we compared the results for our 14 datasets for the top three methods identified by our study. Scatterplots of parameter estimates and posterior inclusion probabilities are shown in *SI Appendix* for all 14 datasets. We found that the (model-averaged) parameter estimates were very similar between the three methods for the 10 tall datasets (with $p < n$) and less similar but still highly correlated for the four wide datasets (with $p > n$). The



	Rank	Score	PartScore	PointEst	IntEst	Inference	Prediction	IntPred	N vars	CPU time
g=sqrt(n)	1	0.974	0.982	0.978	0.927	0.999	0.968	0.996	1.294	0.949
Hyper-g	2	0.992	0.991	0.999	0.993	0.984	0.992	0.993	1.079	3.396
EB-local	3	0.993	0.996	0.995	0.978	0.995	0.998	1	1.099	0.843
JZS	4	1	1	1	1	1	1	1	1	8.835
Horseshoe	5	1.03	0.987	0.964	1.028	0.929	1.068	1.161	1.14	38.256
UIP	6	1.039	1.018	1.034	1.141	1.018	1.003	1	0.946	0.798
EB-global	7	1.073	1.029	1.024	1.238	1.035	1.026	1.042	0.876	0.909
Benchmark	8	1.15	1.111	1.072	1.394	1.189	1.072	1.021	0.719	0.742
NLP	9	1.157	1.037	1.124	1.598	0.91	1.076	1.076	2.07	254.676
LASSO	10		1.15	1.044		1.413	0.994		2.339	1
SCAD	11		1.175	1.122		1.362	1.04		1.505	7.299
BIC-BAS	12	1.21	1.214	1.144	1.206	1.088	1.41	1.201	1.227	0.936
BICREG	13	1.443	1.218	1.202	2.176	1.193	1.26	1.384	1.061	19.809
SpikeSlab	14	1.464	1.189	1.355	2.724	1.155	1.056	1.029	0.496	24.36
ElasticNet	15		1.203	1.098		1.522	0.99		3.825	60.408
MCP	16		1.221	1.148		1.417	1.099		1.227	6.725
SS Lasso	17		1.249	1.323		1.216	1.209		0.741	0.797
Lasso-1se	18		1.463	1.916		1.413	1.061		1.33	1
EMVS	19		1.501	1.703		1.508	1.291		1.026	4.634
AIC	20	3.613	3.837	6.179	4.937	1.176	4.155	1.617	2.887	1.675
g=1	21	3.859	2.256	4.016	11.153	1.194	1.557	1.373	1.66	1.194

Fig. 2. Performance of 21 methods for inference in linear regression under model uncertainty: "PointEst" is the RMSE for point estimation, "IntEst" is the MIS for interval estimation, "Inference" is 1 – the AUPRC, "Prediction" is the RMSE for point prediction, and "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 JZS method. See *Results* and *Materials and Methods* for more information about the ranking and coloring and the definitions of the methods and metrics. Note that BICREG denotes the BICREG-SIS method, in which sure independence screening is used first to reduce the number of variables to 30.

posterior inclusion probabilities were similar between methods for the tall datasets but less so for the wide datasets. The $g = \sqrt{n}$ method tended to favor models with slightly more variables than the hyper-g and Empirical Bayes (EB)–local methods.

Comparison of BMA with Bayesian Model Selection. An alternative to BMA is Bayesian model selection (BMS), in which just one model is selected. When several candidate models are available, a researcher can choose to select one model or perform model averaging. BMS refers to selection of one model from a list of candidate models based on the data (7, 10). One choice for BMS is to select the model with the highest posterior probability in model search, also known as the MAP model. We compared the performance of BMA and BMS for the top three methods identified in the previous section: $g = \sqrt{n}$, hyper-g, and EB-local.

We used the same performance metrics as before. As before, all metrics are relative to BMA under the JZS prior, except for computation time, for which LASSO was used as the reference. The results are shown in Table 1. The BMS versions of the top three methods performed worse than the corresponding BMA versions in terms of all the metrics.

Discussion

Several previous comparisons of existing methods have been carried out. They have tended to be based on a narrower range of methods than we consider here, to be based on simulation experiments whose connection to empirical data is less clear, and to base comparisons on a subset of the statistical tasks of interest.

Fernández et al. (9) did a simulation study based on a nonempirical design (54) and compared methods based on their ability to recover the true underlying model as the MAP model and assess predictive performance using log-predictive scores. Hence, their comparisons were based on only two statistical tasks, namely inference and point prediction. They considered only BMA methods. They found a UIP-based method with $g = n$ to work best when $n < p^2$ and an RIC-based method (24) with $g = p^2$ to work best otherwise, but they pointed out that the RIC-based method is not model-selection consistent. We have included the resulting combined method in our study under the name "benchmark prior." The only other method in their study that is also in ours is the $g = \sqrt{n}$ method, which they found to be outperformed by BIC, in contrast with our findings here.

Table 1. Comparison of BMA and BMS for top three methods

Method	Type	Score	PointEst	IntEst	Inference	Prediction	IntPred	N vars	CPU time
g = sqrt(n)	BMA	0.974	0.978	0.927	0.999	0.968	0.996	1.294	0.949
	BMS	1.596	1.098	1.816	2.906	1.100	1.060	1.009	1.222
Hyper-g	BMA	0.992	0.999	0.993	0.984	0.992	0.993	1.079	3.396
	BMS	1.731	1.123	2.242	3.061	1.114	1.117	0.837	3.339
EB-local	BMA	0.993	0.995	0.978	0.995	0.998	1	1.099	0.843
	BMS	1.742	1.127	2.228	3.060	1.142	1.155	0.861	1.096
JZS	BMA	1	1	1	1	1	1	1	8.835

Eicher et al. (55) considered the same BMA methods as in ref. 9, considered prediction for a well-known economic growth dataset and for several simulations with the same nonempirically based design, and again found BIC and UIP to do best. Our results here are based on a wider and more empirically based set of simulations, which may help explain the different results. Liang et al. (32) also carried out a nonempirically based simulation study using the design of Cui and George (56) and found the hyper-g prior to be competitive with other BMA methods in terms of parameter estimation, including several that we have considered here (but they did not include the $g = \sqrt{n}$ method).

Celeux et al. (57) carried out another nonempirical simulation study to assess quality of inference and assessed point prediction using two small real datasets; they assessed 15 methods, of which 7 were in common with ours. They focused on the situation where *p* is close to *n*. Like us, they found Bayesian methods to outperform non-Bayesian ones.

Deckers et al. (58) compared a subset of the Bayesian techniques discussed in our study, specifically the UIP and RIC, or benchmark prior, and LASSO with multiple testing procedures (MTPs) controlling false discovery rate. They focused their comparison on the model inference performance of these procedures using size-adjusted power, i.e., comparison of power (number of correctly selected variables) in situations where procedures infer similar size models. In their comparison over a nonempirical simulation study, they found that BMA was slightly more powerful given the size than the MTPs and LASSO. Their comparison did not focus on other statistical inference tasks of prediction and estimation.

Bhadra et al. (59) compared variants of the horseshoe, LASSO, and SCAD in terms of their performance in variable selection, using the nonempirically based simulation design of Zhao and Yu (60). They found the horseshoe to do best, then SCAD, both substantially dominating LASSO. This agrees with our ranking in terms of inference from Fig. 2, but we found LASSO to overtake SCAD when other statistical tasks were also taken into account. Forte et al. (13) compared different BMA software packages in terms of computational performance and found the BAS package (53) to dominate others in terms of speed, as we also found. However, they also warned against the use of the MCMC + BAS method within BAS, which they reported does not provide reliable estimates of the inclusion probabilities, and instead recommended the method MCMC. They also commented on the very high memory demands of BAS. Here we used their recommended method MCMC and found it to work well.

One can also evaluate methods in terms of theoretical properties. One is model-selection consistency (9), which says that if the true model is among the candidate models considered, the method will select it with probability approaching 1 as the sample size increases indefinitely. All three of our top-ranked methods satisfy this unless the true model is the null model with no predictors (9, 32). However, LASSO does not have this property (60).

A second property is whether the method is subject to Bartlett's paradox (61), according to which if the data are held fixed and the prior variance increases without bound, then BMA will select the null model with probability tending to 1, regardless of the data. None of our top three methods is subject to this as they do not allow the prior variance to increase without bound.

A third consideration is whether the method is subject to the so-called "information paradox" (32). This arises when, for fixed *n* and *p*, the data provide maximal support for a larger model, for example, when $R^2 \longrightarrow 1$. One could argue that in this case, the Bayes factor for this model against any submodel should tend to infinity with the sample size. However, *g*-priors with fixed *g* do not have this property, and indeed in that case the Bayes factor has a finite (although usually very high) upper bound. It has been argued that this is undesirable, making them subject to the information paradox. The hyper-g and EB-local methods are not subject to this, but the $g = \sqrt{n}$ prior is, which could be argued to be a disadvantage of the latter.

However, one might question the relevance of the information paradox to the choice of method (62). If $R^2 = 1$ when *n* is small, this will often be because of the inherent discreteness of most data, which are rarely measured or recorded with full precision but rather to within a certain measurement tolerance (for example, a certain number of significant digits). In that case, the fact that the Bayes factor for an additional variable is bounded above could be viewed as an advantage. The linear regression model models the observed response variable as a continuous variable, thus measured with infinite precision. This is actually an approximation, which is usually inconsequential, but is relevant for assessing the relevance of the information paradox. If the discreteness of observed data were accounted for in the model, the information paradox would never arise.

For example, the famous data on heights of fathers and sons in England (63, 64) are reported to the nearest inch. If one took a sample of size 3 from these data, say (father, son) = (62.5, 64.5), (67.5, 69.5), (70.5, 72.5), and regressed son's height on father's height, one would find that $R^2 = 1$ and the standard F statistic is infinite. In this case, one would not want the Bayes factor for the effect of father's height to be infinite, but it is infinite for the hyper-g and EB-local priors, while for the $g = \sqrt{n}$ prior it is 1.65. The latter represents positive but weak evidence for an effect, which seems more reasonable than an infinite Bayes factor corresponding to absolute certainty based on three data points.

Beyond that, the upper bound on the Bayes factor is typically very high for even moderate *n*. For example, for *n* as low as 20, it is over 4 million. So even if the existence of an upper bound on the Bayes factor were to be viewed as undesirable, it would have no practical effect. Overall, this suggests that the information paradox may not be a disadvantage for the $g = \sqrt{n}$ prior and others that it affects and may even be a positive feature.

We have focused here on the choice of prior distribution for model parameters. BMA also requires a prior on the models themselves, and we have used default choices: either a uniform prior over all models or a uniform prior on model size. It would be worth carrying out a similar analysis to the present one to compare different possible model priors.

Given the good performance of the $g = \sqrt{n}$ prior of ref. 9, it is of interest how it relates to the popular BIC criterion, which corresponds approximately to $g = n$ and performed less well in our experiments. Let us consider just two models: the null model and a regression model of interest, with *d* variables. Then if *B* is the Bayes factor for the regression model against the null model, the exact result is $-2 \log B = (n-1) \log\{1 + \sqrt{n}(1 - R^2)\} - (n - 1 - d) \log(1 + \sqrt{n})$. The BIC approximation is $-2 \log B \approx n \log(1 - R^2) + d \log(n)$. A similar approximation with the $g = \sqrt{n}$ prior is $-2 \log B \approx n \log(1 - R^2) + d(\log(n)/2) + \sqrt{n}(1 - R^2)R^2$. The last term does not involve the number of parameters directly, and so the complexity penalty in the Bayes factor with the $g = \sqrt{n}$ prior is effectively half that in the BIC.

We have focused on one specific type of model uncertainty in one statistical setting, namely, uncertainty about which variables to include in a linear regression model. This has been much studied and arises frequently in science, as well as being a canonical example for other statistical models. However, there are many other statistical settings in which the same issue arises, and it would be of interest to carry out similar comparative studies. In

Table 2. Variable selection methods compared in this study

Method	Authors	Implementation (R package–version)	Function
$g = \sqrt{n}$	Fernández et al. (9)	BAS-V1.5.5 (53)	bas.lm(..., prior=''g-prior'', alpha = sqrt(n))
Hyper-g	Liang et al. (32)	BAS-V1.5.5 (53)	bas.lm(..., prior=''hyper-g'')
EB-local	Hansen and Yu (29)	BAS-V1.5.5 (53)	bas.lm(..., prior=''EB-local'')
JZS	Zellner and Siow (19)	BAS-V1.5.5 (53)	bas.lm(..., prior=''JZS'')
Horseshoe	Carvalho et al. (35)	horseshoe-V0.2.0 (65)	horseshoe()
UIP	Kass and Wasserman (21)	BAS-V1.5.5 (53)	bas.lm(..., prior=''g-prior'', alpha = n)
EB-global	Clyde and George (30) and George and Foster (31)	BAS-V1.5.5 (53)	bas.lm(..., prior=''EB-global'')
Benchmark	Fernández et al. (9)	BAS-V1.5.5 (53)	bas.lm(..., prior=''g-prior'', alpha = max(n,p^2))
NLP	Rossell and Telesca (34) and Johnson and Rossell (33)	mombf-V2.2.9 (66)	modelSelection()
LASSO*	Tibshirani (37)	glmnet-V3.0.2 (52)	cv.glmnet()
SCAD	Fan and Li (43)	ncvreg-V3.11.2 (67)	cv.ncvreg(..., penalty=''SCAD'')
BIC-BAS	George and Foster (31)	BAS-V1.5.5 (53)	bas.lm(..., prior=''BIC'')
BICREG-SIS	Raftery (26) and Fan and Lv (68)	BMA-V3.18.12 (69)	bicreg()
Spike slab	George and McCulloch (36)	BoomSpikeSlab-V1.2.3 (70)	lm.spike()
Elastic net	Zou and Hastie (45)	glmnet-V3.0.2 (52)	cv.glmnet(, alpha)
MCP	Zhang et al. (44)	ncvreg-V3.11.2 (67)	cv.ncvreg(..., penalty=''MCP'')
SS lasso	Ročková and George (51)	SSLASSO-V1.2.2 (51)	SSLASSO()
EMVS	Ročková and George (50)	EMVS-V1.1 (71)	EMVS()
AIC	George and Foster (31)	BAS-V1.5.5 (53)	bas.lm(..., prior=''AIC'')
$g = 1$	van Zwet (22)	BAS-V1.5.5 (53)	bas.lm(..., prior=''g-prior'', alpha = 1)

*LASSO-1se has the same reference as LASSO.

linear regression itself, there are the choices of error distribution and functional form of the variables. The same issues arise in generalized linear models such as logistic regression and Poisson regression, in addition to the choice of link function and mean-variance relationship. Similar model choice issues arise with Bayesian hierarchical models and many other model classes. We expect that our main conclusion, that BMA with an adaptive parameter prior performs well, would carry over to other settings.

Materials and Methods

Statistical Methods for Comparison. The 21 methods we compare are listed in Table 2, along with references, the R package used, and the function call used. All the g-prior methods implemented using the BAS package, and the NLP methods implemented using the mombf package, use the beta-binomial (1, 1) prior as the default model space prior. For high-dimensional datasets with $p > n$, a truncated beta-binomial (1, 1) prior is used as the model space prior; this assigns probability zero to any model with size greater than $n - 2$. The BICREG-SIS method assumes a uniform prior over the model space. For methods implemented using the BAS package, a combination of the Metropolis–Hastings algorithm, as in the MCMC model composition algorithm (54), with a random swap between a currently included and a currently excluded variable, is used for model space exploration.

Datasets. We carried out 14 simulation studies, each one based on a different publicly available real dataset, from a variety of fields including social sciences, healthcare, genome sciences, physical sciences, chemistry, and engineering (Table 3). We selected several of our datasets by filtering all the datasets in the University of California, Irvine, machine learning repository as follows. We filtered datasets with default task as regression, attribute type as numerical, data type as multivariate/univariate, and number of attributes between 10 and 100. We further restricted our attention to datasets with $p > 20$ and $n < 25,000$. This reduced our list of UCI datasets to four: the bias correction, bike sharing, SML, and superconductivity datasets. Note that the bias correction and bike sharing datasets each have two versions based on choice of outcome and frequency.

We also included several datasets that have been used as examples in the literature. We included the college dataset (78) as an example dataset where full enumeration of models is feasible. We included the diabetes (38)

Table 3. Datasets used in the study

Dataset name	Sample size (N)	Covariates (p)	Source
College	777	14	ISLR (72)
Bias Correction-Tmax	7,590	21	UCI ML repository
Bias Correction-Tmin	7,590	21	UCI ML repository
SML2010	1,373	22	UCI ML repository
Bike sharing-daily	731	28	UCI ML repository
Bike sharing-hourly	17,379	32	UCI ML repository
Superconductivity	21,263	81	UCI ML repository
Diabetes	442	64	spikeslab (73)
Ozone	330	44	gss (74)
Boston housing	506	103	mlbench (75)
NIR	166	225	chemometrics (76)
Nutrimouse	40	120	mixOmics (77)
Multidrug	60	853	mixOmics (77)
Liver toxicity	64	3,116	mixOmics (77)

and ozone (1, 32) datasets and the Boston housing dataset with squares and interaction terms between its covariates. Finally, we included four high-dimensional datasets from chemometrics and genomics from the mixOmics (77) and chemometrics R packages (76). For all the datasets, the continuous predictors were standardized to have mean zero and variance 1, and the response variable was centered to have mean zero. The 14 datasets used in the simulation study are listed in Table 3. Details of dataset preprocessing are given in *SI Appendix*.

Determining the Generating Model for the Simulation Study. For our simulation study, we require a data generating model based on each of our real datasets. For datasets for which $p < 30$, we performed all subsets regression using the leaps package in R (79) and selected the largest model with all variables significant at 0.05 level. For datasets with $p > 30$, all subsets regression can be computationally intensive, and so we performed iterative sure independence screening (ISIS) (68) to reduce the number of variables. If the filtered list contained more than 30 variables, we further selected the top 30 variables with the highest R^2 values under univariate regression. We then applied all subsets regression to the filtered list of covariates with the above criteria to find the data generating model for our simulation study.

Consider the Boston housing dataset ($n = 506, p = 103$) as an example. This includes 14 geographic housing variables, plus interactions and squares for each continuous variable, leading to 103 possible predictors. All subsets regression is not computationally feasible, so instead we used ISIS to get a filtered list of 81 variables. We then performed univariate regressions for each of the filtered variables to select the top 30 variables with the highest R^2 values. Finally, we performed all subsets regression using the screened variables to get our data generating model with 23 variables and an R^2 of 0.86.

Simulation Design. For each dataset, we chose a data generating model as described above to closely approximate the data. Using this model, we used the parametric bootstrap to generate 100 bootstrapped datasets with the same design matrix $\boldsymbol{X}$ but different simulated response vectors. We compared the performance of the different techniques for parameter estimation, interval estimation, and variable selection on these datasets for our simulation study using the metrics described below.

To evaluate the predictive performance of the methods, we divided each of the simulated datasets 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. We calculated point predictions for each of the methods and posterior predictive intervals for Bayesian techniques that allow for uncertainty quantification.

We used the following metrics to compare the methods.

PointEst. For point estimation, we calculated the root mean squared error (RMSE) of the parameter estimates as follows:

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

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 for the Bayesian techniques and the estimated optimal coefficients for penalized likelihood based approaches.

IntEst. The interval score (IS) (80) provides a balance between the narrowness of the intervals and the accuracy of the coverage. It is a sum of two components: the first rewards narrow intervals, and the second rewards accurate coverage. For a variable z, the IS is given by

$$\begin{aligned} MIS_\alpha(l, u, z) = (u - l) &+ \frac{2}{\alpha}(l - z)\mathbb{1}\{z < l\} \\ &+ \frac{2}{\alpha}(z - u)\mathbb{1}\{u < z\}, \end{aligned} \qquad \textbf{[3]}$$

where l and u denote the upper and lower bounds of the $(1 - \alpha) \times 100\%$ posterior intervals of z. In order to assess the quality of the interval estimation, we considered the mean interval score (MIS) for the coefficients and calculated the average MIS across coefficients for each of the datasets. We used $\alpha = 0.05$.

Inference. To compare the performance of the techniques for identifying the appropriate variables, we calculated the area under the precision recall curve (AUPRC) for each of the techniques. This gives an overall assessment of model selection quality and does not require a threshold to be chosen for the posterior inclusion probability of a covariate.

For penalized likelihood based approaches, the AUPRC was obtained by varying the cross-validation parameter λ from close to 0 (no penalization) to λ_{max}, defined as the smallest value of λ for which none of the variables is included in the model (81). For the horseshoe, the AUPRC was obtained by varying the credible set levels leading to different number of variables being selected by the method. We report Inference with (1 – AUPRC) as our metric, and a lower value is better.

Prediction. In order to assess the accuracy of point prediction, we calculated R^2_{test} as follows:

$$R^2_{test} = 1 - \frac{\sum_{i \in test}(y_i - \hat{y}_i)^2}{\sum_{i \in test}(y_i - \bar{y}_{train})^2}, \qquad \textbf{[4]}$$

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. Note that this quantity can be less than zero, if the predictions perform worse than the baseline $\bar{y}_{train}$.

IntPred. To assess the quality of the prediction intervals, we calculated the interval score using Eq. **3** for each of the test set observations. Here l and u represent the lower and upper bounds of the $(1 - \alpha) \times 100\%$ posterior predictive interval for the test observation. We calculated the MIS, averaging IS over test set observations for each of the train–test splits. A lower MIS corresponds to a better interval forecast.

N vars. To report sparsity levels, we recorded the average model size for the BMA techniques and the number of nonzero estimated coefficients for the penalized likelihood based approaches. For the horseshoe, we calculated a 95% credible interval and checked whether 0 was included in it to arrive at the model size. We denote the average model size by $\hat{p}$.

CPU time. We recorded the average computation time (in seconds) taken by each technique to fit the model for one bootstrapped dataset.

Data Availability. Previously published data were used for this work (https://archive.ics.uci.edu/ml/index.php) (76, 77).

ACKNOWLEDGMENTS. This research was supported by NIH grant R01 HD-070936.

1. A. Miller, *Subset Selection in Regression* (CRC Press, 2002).
2. D. A. Freedman, A note on screening regression equations. *Am. Stat.* **37**, 152–155 (1983).
3. E. E. Leamer, *Specification Searches: Ad hoc Inference with Nonexperimental Data* (Wiley, 1978), vol. 53.
4. A. E. Raftery, "Approximate Bayes factors for generalized linear models" (Tech. Rep. 121, Department of Statistics, University of Washington, 1988).
5. E. I. George, R. E. McCulloch, Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **88**, 881–889 (1993).
6. D. Madigan, A. E. Raftery, Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Am. Stat. Assoc.* **89**, 1535–1546 (1994).
7. R. E. Kass, A. E. Raftery, Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795 (1995).
8. J. A. Hoeting, D. Madigan, A. E. Raftery, C. T. Volinsky, Bayesian model averaging: A tutorial. *Stat. Sci.* **14**, 382–417 (1999).
9. C. Fernández, E. Ley, M. F. J. Steel, Benchmark priors for Bayesian model averaging. *J. Econom.* **100**, 381–427 (2001).
10. L. Wasserman, Bayesian model selection and model averaging. *J. Math. Psychol.* **44**, 92–107 (2000).
11. M. Clyde, E. I. George, Model uncertainty. *Stat. Sci.* **19**, 81–94 (2004).
12. T. M. Fragoso, W. Bertoli, F. Louzada, Bayesian model averaging: A systematic review and conceptual classification. *Int. Stat. Rev.* **86**, 1–28 (2018).
13. A. Forte, G. Garcia-Donato, M. F. J. Steel, Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *Int. Stat. Rev.* **86**, 237–258 (2018).
14. A. E. Raftery, Y. Zheng, Discussion: Performance of Bayesian model averaging. *J. Am. Stat. Assoc.* **98**, 931–938 (2003).
15. D. B. Rubin, N. Schenker, Efficiently simulating the coverage properties of interval estimates. *J. R. Stat. Soc. Ser. C Appl. Stat.* **35**, 159–167 (1986).
16. P. A. Mattei, A parsimonious tour of Bayesian model uncertainty. arXiv [Preprint] (2020). https://arxiv.org/abs/1902.05539 (Accessed 15 August 2021).
17. H. Jeffreys, *Theory of Probability* (Oxford University Press, Oxford, United Kingdom, ed. 3, 1961).
18. J. H. Park *et al.*, Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).

19. A. Zellner, A. Siow, Posterior odds ratios for selected regression hypotheses. *Trab. Estad. Invest. Oper.* **31**, 585–603 (1980).
20. A. Zellner, On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference Decis. Tech.* **6**, 233–243 (1986).
21. R. E. Kass, L. Wasserman, A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* **90**, 928–934 (1995).
22. E. van Zwet, A default prior for regression coefficients. *Stat. Methods Med. Res.* **28**, 3799–3807 (2019).
23. W. C. Young, A. E. Raftery, K. Y. Yeung, Fast Bayesian inference for gene regulatory networks using ScanBMA. *BMC Syst. Biol.* **8**, 47 (2014).
24. D. P. Foster, E. I. George, The risk inflation criterion for multiple regression. *Ann. Stat.* **22**, 1947–1975 (1994).
25. G. Schwarz, Estimating the dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
26. A. E. Raftery, Bayesian model selection in social research. *Sociol. Methodol.* **25**, 111–163 (1995).
27. H. Akaike, Information measures and model selection. *Bull. Int. Stat. Inst* **44**, 277–291 (1983).
28. K. P. Burnham, D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretical Approach* (Springer-Verlag, New York, ed. 2, 2002).
29. M. H. Hansen, B. Yu, Minimum description length model selection criteria for generalized linear models. *Lect. Notes Monogr. Ser.* **40**, 145–163 (2003).
30. M. Clyde, E. I. George, Flexible empirical Bayes estimation for wavelets. *J. R. Stat. Soc. Series B Stat. Methodol.* **62**, 681–698 (2000).
31. E. I. George, D. P. Foster, Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731–747 (2000).
32. F. Liang, R. Paulo, G. Molina, M. A. Clyde, J. O. Berger, Mixtures of g priors for Bayesian variable selection. *J. Am. Stat. Assoc.* **103**, 410–423 (2008).
33. V. E. Johnson, D. Rossell, Bayesian model selection in high-dimensional settings. *J. Am. Stat. Assoc.* **107**, 649–660 (2012).
34. D. Rossell, D. Telesca, Nonlocal priors for high-dimensional estimation. *J. Am. Stat. Assoc.* **112**, 254–265 (2017).
35. C. M. Carvalho, N. G. Polson, J. G. Scott, The horseshoe estimator for sparse signals. *Biometrika* **97**, 465–480 (2010).
36. E. I. George, R. E. McCulloch, Approaches for Bayesian variable selection. *Stat. Sin.* **7**, 339–373 (1997).
37. R. Tibshirani, Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.* **58**, 267–288 (1996).
38. B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression. *Ann. Stat.* **32**, 407–499 (2004).
39. J. Friedman *et al.*, Pathwise coordinate optimization. *Ann. Appl. Stat.* **1**, 302–332 (2007).
40. R. Tibshirani, Regression shrinkage and selection via the lasso: A retrospective. *J. R. Stat. Soc. Ser. B* **73**, 273–282 (2011).
41. G. Casella, M. Ghosh, J. Gill, M. Kyung, Penalized regression, standard errors, and Bayesian lassos. *Bayesian Anal.* **5**, 369–411 (2010).
42. C. C. Holmes, Discussion of 'Regression shrinkage and selection via the lasso: A retrospective'. *J. R. Stat. Soc. Ser. B* **73**, 279–280 (2011).
43. J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**, 1348–1360 (2001).
44. C. H. Zhang *et al.*, Nearly unbiased variable selection under minimax concave penalty. *Ann. Stat.* **38**, 894–942 (2010).
45. H. Zou, T. Hastie, Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **67**, 301–320 (2005).
46. T. Park, G. Casella, The Bayesian lasso. *J. Am. Stat. Assoc.* **103**, 681–686 (2008).
47. D. Fouskakis, I. Ntzoufras, D. Draper, Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Anal.* **10**, 75–107 (2015).
48. A. J. Womack, L. León-Novelo, G. Casella, Inference from intrinsic Bayes' procedures under model selection and uncertainty. *J. Am. Stat. Assoc.* **109**, 1040–1053 (2014).
49. C. Hans, Bayesian lasso regression. *Biometrika* **96**, 835–845 (2009).
50. V. Ročková, E. I. George, EMVS: The EM approach to Bayesian variable selection. *J. Am. Stat. Assoc.* **109**, 828–846 (2014).
51. V. Ročková, E. I. George, The spike-and-slab lasso. *J. Am. Stat. Assoc.* **113**, 431–444 (2018).
52. J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
53. M. Clyde, BAS: Bayesian variable selection and model averaging using Bayesian adaptive sampling. R package version 1.5.5. https://cran.r-project.org/web/packages/BAS/BAS.pdf. 15 August 2021.
54. A. E. Raftery, D. Madigan, J. A. Hoeting, Bayesian model averaging for linear regression models. *J. Am. Stat. Assoc.* **92**, 179–191 (1997).
55. T. S. Eicher, C. Papageorgiou, A. E. Raftery, Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *J. Appl. Econ.* **26**, 30–55 (2011).
56. W. Cui, E. I. George, Empirical Bayes vs. fully Bayes variable selection. *J. Stat. Plan. Inference* **138**, 888–900 (2008).
57. G. Celeux, M. El Anbari, J. M. Marin, C. P. Robert, Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Anal.* **7**, 477–502 (2012).
58. T. Deckers, C. Hanck, Variable selection in cross-section regressions: Comparisons and extensions. *Oxford Bull. Econ. Stat.* **76**, 841–873 (2014).
59. A. Bhadra, J. Datta, N. G. Polson, B. Willard, Lasso meets horseshoe: A survey. *Stat. Sci.* **34**, 405–427 (2019).
60. P. Zhao, B. Yu, On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–2563 (2006).
61. M. S. Bartlett, A comment on D. V. Lindley's statistical paradox. *Biometrika* **44**, 533–534 (1957).
62. A Zellner, Comments on 'Mixtures of g-priors for Bayesian variable selection (2008).' http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.523.2147&rep=rep1&type=pdf. Accessed 28 March 2022.
63. K. Pearson, A. Lee, On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika* **2**, 357–462 (1903).
64. M. Friendly, HistData: Data sets from the history of statistics and data visualization. R package version 0.8-7. https://CRAN.R-project.org/package=HistData. Accessed 15 August 2021.
65. S. van der Pas, J. Scott, A. Chakraborty, A. Bhattacharya, horseshoe: Implementation of the Horseshoe Prior. R package version 0.2.0. https://CRAN.R-project.org/package=horseshoe. Accessed 15 August 2021.
66. D. Rossell, J. D. Cook, D. Telesca, P. Roebuck, O. Abril, mombf: Bayesian Model Selection and Averaging for Non-Local and Local Priors. R package version 2.2.9. https://CRAN.R-project.org/package=mombf. Accessed 15 August 2021.
67. P. Breheny, J. Huang, Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Ann. Appl. Stat.* **5**, 232–253 (2011).
68. J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B* **70**, 849–911 (2008).
69. A. E. Raftery, J. Hoeting, C. Volinsky, I. Painter, K. Y. Yeung, BMA: Bayesian model averaging. R package version 3.18.12. https://CRAN.R-project.org/package=BMA. Accessed 15 August 2020.
70. S. L. Scott, BoomSpikeSlab: MCMC for spike and slab regression. R package version 1.2.3. https://CRAN.R-project.org/package=BoomSpikeSlab. Accessed 15 August 2020.
71. V. Rockova, G. Moran, EMVS: The expectation-maximization approach to Bayesian variable selection. R package version 1.1. https://CRAN.R-project.org/package=EMVS. Accessed 15 August 2021.
72. G. James, D. Witten, T. Hastie, R. Tibshirani, ISLR: Data for an Introduction to Statistical Learning with Applications in R. R package version 1.2. https://CRAN.R-project.org/package=ISLR. Accessed 15 August 2021.
73. H. Ishwaran, J. Rao, U. Kogalur, spikeslab: Prediction and variable selection using spike and slab regression (manual). R package version 1.1.5. https://CRAN.R-project.org/package=spikeslab. Accessed 15 August 2021.
74. C. Gu, Smoothing spline ANOVA models: R package gss. *J. Stat. Softw.* **58**, 1–25 (2014).
75. D. Newman, S. Hettich, C. Blake, C. Merz, UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/index.php. Accessed 15 August 2021.
76. P. Filzmoser, K. Varmuza, Chemometrics: Multivariate statistical analysis in chemometrics. R package version 1.4.2. https://CRAN.R-project.org/package=Chemometrics. Accessed 15 August 2021.
77. F. Rohart, B. Gautier, A. Singh, K. A. Lê Cao, mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
78. G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning* (Springer, 2013), vol. 112.
79. T. Lumley, leaps: Regression subset selection. R package version 3.1. https://CRAN.R-project.org/package=leaps. Accessed 15 August 2021.
80. T. Gneiting, A. E. Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
81. M. Vignes *et al.*, Gene regulatory network reconstruction using Bayesian networks, the Dantzig Selector, the Lasso and their meta-analysis. *PLoS One* **6**, e29165 (2011).