SCALABLE MODEL SELECTION WITH MIXTURES OF G-PRIORS IN LARGE DATA SETTINGS

BY JACOB FONTANA[1,a], BRUNO SANSÓ[2,b]

[1]*Department of Statistics, University of California Santa Cruz,* [a]*jafontan@ucsc.edu*

[2]*Department of Statistics, University of California Santa Cruz,* [b]*bsanso@ucsc.edu*

We consider the variable selection problem for linear models using a mixture of g-priors. While a commonly studied aspect of these models is their posterior consistency properties in the case where the true model is a subset of the possible set of covariates ($\mathcal{M}$-closed), we examine the $\mathcal{M}$-open case, where the data generating process is not an element of the model space. Two unique problems present themselves in this setting when examining large data sets (on the order of 10^6 observations): Shrinkage deficient estimation SDE, and model superinduction (MS). SDE refers to the phenomenon where the posterior shrinkage decays at a linear rate with the sample size. MS refers to the tendency of model selection procedures to select larger models as the sample size grows. We prove that when comparing nested models using Bayes factors, for a sufficiently large sample size the larger model will be selected. We consider many cases where this behavior results in overparametrized models which induce severe computational difficulties. We show that this phenomena is inescapable (affecting even oracle estimators), so we instead seek to minimize the severity of its effects on large sample sizes while preserving posterior consistency. To that end, we propose a beta-prime hyper-prior on g, with hyper-parameters chosen to result in a sub-linear decay of the posterior shrinkage. We also propose a model space prior which biases the posterior odds ratio towards smaller models asymptotically. These two priors introduce two new hyper-parameters, for which we propose default values. We demonstrate the aforementioned phenomena, and the efficacy of our proposed solutions, via several synthetic data examples, as well as a case study using albedo data from GOES satellites.

1. Introduction. Consider a class of candidate models $M_\gamma : Y = X\alpha + K_\gamma\beta_\gamma + \epsilon$. Here X denotes a set of parameters common to all models and K_γ denotes the matrix of covariates specific to M_γ, such that the columns of K_γ are orthogonal to those of X ($K_\gamma^T X = 0$). The model that contains only the covariates in X is referred to as the null model and denoted as $M_\emptyset$. There are two tasks of interest here. The first, is to estimate the values of β_γ, given a model M_γ. There are numerous methods, in both the Bayesian and frequentist literature, to address this problem. The second problem, is to choose a model that is "optimal" in some sense from a larger space of models $\mathcal{M} = \{M_\emptyset, M_1, \dots\}$. There are two common settings for this task. In the first, the $\mathcal{M}$-closed view (Bernardo and Smith, 2009), we assume that the data generating process (denoted M_*) is in the space of models considered: $M_* \in \mathcal{M}$. In the second, the $\mathcal{M}$-open view, we assume that $M_* \notin \mathcal{M}$.

In both of these settings, we adopt the *optimal prediction view* of Vehtari and Ojanen (2012). According to this view, our task is to choose a model which minimizes the expected loss taken with respect to the predictive distribution of the data conditioned on the data generating model. Common choices of loss functions are the log-utility function or the squared error loss (in the $\mathcal{M}$-closed case, 0-1 losses can also be considered). However, in practice, since

MSC2020 subject classifications: Primary 62C10, 62F15; secondary 62F05, 62J99.

Keywords and phrases: Model selection, Bayesian, g-prior, Linear Models, M-open model comparison, Bayesian decision theory.

we do not have knowledge of the underlying data generating process, and model spaces can be quite large, we can never know if a given model is optimal. Instead, we use a *model selection procedure*, common examples of which include step-wise searches, stochastic searches (Hans, Dobra and West, 2007), cross-validation, etc.

We consider a Bayesian approach to the variable selection problem. A popular choice of non-informative prior for the linear coefficients, in Bayesian regression settings, is the g-prior. Although g-priors will be the focus of this paper, our results are likely extensible to other choices of priors for variable selection, as well as classical approaches to the variable selection problem. First introduced by Zellner (1986), g-priors assume that

$$\pi(\alpha) \propto 1 \quad \pi(\sigma^2) \propto \frac{1}{\sigma^2} \quad \beta \sim \mathcal{N}\left(0, g\sigma^2 (K_\gamma^T K_\gamma)^{-1}\right).$$

These priors have two desirable properties. First, they produce a posterior mean that provides shrinkage with respect to the ordinary least squares estimator $\hat{\beta}_{OLS}$. In fact, $\mathbb{E}(\beta|y) = g/(g+1)\hat{\beta}_{OLS}$. Second, they result in easily computable marginal distributions, which allows for fast step-wise model selection procedures. Values of g that are independent from the sample size, have been shown to be undesirable, as for any model M_γ, $BF_{\gamma:\emptyset} \to 0$ as $g \to \infty$, where $BF_{\gamma:\gamma'}$ denotes the Bayes factor to compare M_γ to $M_{\gamma'}$. This phenomenon is referred to as the Barlett-Jeffrey Paradox (Liang et al., 2008). Early literature thus considered values for g dependent on the sample or model size, such as the Unit Information Prior ($g = n$) (Kass and Wasserman, 1995), the Risk Inflation Criteria ($g = p^2$) (Foster and George, 1994), and the Benchmark prior ($g = \max\{n, p^2\}$) (Fernández, Ley and Steel, 2001). Unfortunately these choices result in the Lindley Information Paradox, a phenomenon where sequences of samples that indicate progressively stronger support for a model ($R^2 \to 1$), result in the Bayes factor converging to a constant ($BF_{\gamma,\emptyset} \to (g+1)^{(n-p-q)/2}$), preserving a non-zero posterior probability of the null model (Liang et al., 2008). To resolve this paradox, hyper-prior distributions need to be placed on g (or equivalently a function of g, such as $g/(g+1)$). Cui and George (2008) consider placing a $\text{Be}(1, a/2-1)$ prior on $g/(g+1)$, while Carvalho, Polson and Scott (2010) considers a $\text{Be}(1/2, 1/2)$ prior. These priors, where the distribution does not incorporate information on the sample size, result in inconsistency in the case when the null model is the true model. To resolve this problem, Liang et al. (2008) consider placing a $\text{Be}(1, a/2-1)$ prior on $g/(g+n)$, and recasts early work by Zellner and Siow (1980) (which examined Cauchy priors on β) as a g-prior, with $g \sim \text{IG}(1/2, n/2)$. Maruyama and George (2011) proposes letting $g \sim \text{BetaPrime}((n-p-q)/2 - a, a)$, which has the added benefit of closed form marginal distributions, while Bayarri et al. (2012), considers a prior of the form $\pi(g) = a_r[p_r(b_r+n)]^{a_r}(g+b_r)^{-(a_r+1)}1_{\{g \geq p_r(b_r+n)-b_r\}}$, the so-called "robust" prior. Recent attention has turned to casting Power Expected Posterior Priors as g-priors, where $\pi(g) \propto g^{-(n-p-q)}(g-n)^{(n-p-q)/2-1}1_{\{g \geq n\}}$ (Consonni et al., 2018).

While the properties of these priors have been thoroughly explored in the $\mathcal{M}$-closed case, little attention has been paid to the $\mathcal{M}$-open setting. Mukhopadhyay, Samanta and Chakrabarti (2015) considered this setting when the model space grows sub-linearly. They let $g = kn^c$, where $c > 0$, and show that the model selected by maximizing the marginal probability is asymptotically loss efficient: $\lim_{n\to\infty} L_n(\hat{M}_\gamma)/\min_{M_\gamma} L_n(M_\gamma) \xrightarrow{p} 1$. Their later work (Mukhopadhyay and Samanta, 2017), extends this result to a specific case of inverse gamma hyper-prior on g, and further shows that many of the popular choices of g-priors are not asymptotically loss efficient in the case where the model space grows (sub-linearly) with n. However, these results do not extend to more general state spaces, such as when the model space grows linearly, or super-linearly. Moreover, we will show that the proposed solutions still present numerous practical difficulties

In this paper we demonstrate three important results. First, we examine what we call *shrinkage deficient estimation*, where the posterior shrinkage factor $\mathbb{E}(1/(g+1)|y) \in$

$\mathcal{O}(1/n)$. In practice, for any large data set, this results in posteriors that essentially reproduce OLS regression estimates. Second, we show that when comparing any two nested finite models, as the sample size gets large, the Bayes factor will always favor the larger model (what we call *model superinduction*). This presents obvious issues for any step-wise model selection procedure. Lastly, we show that a Bayes factor based step-wise procedure is pairwise consistent, even in the $\mathcal{M}$-open setting. Thus, far from being an issue that only plagues g-priors, *model superinduction* is a problem that manifests itself for any model selection procedure in the $\mathcal{M}$-open setting. Despite the inevitable nature of this problem, we propose a class of hyper-priors for g that minimize the effects of these problems for a given sample size, while retaining many of the desirable properties of g-priors. We show the efficacy of this method with both simulated and real data examples.

The rest of the paper will proceed as follows. In Section 2, we will introduce, in more detail, the theory of model-selection. In Section 3, we will review the literature of g-priors, the common choices in the literature, and their properties. In Section 4, we will rigorously introduce the problem of model superinduction and shrinkage deficient estimation. In Section 4, we will introduce our solution to these problems, and, in Section 6, demonstrate its effectiveness on simulated and real data-sets.

2. Model selection and set-up. We consider models of the form:

$$M_\gamma: \qquad Y = X\alpha + K_\gamma \beta_\gamma + \epsilon \qquad \text{where} \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$

We let α be a set of q parameters common to all models, and let X denote the corresponding matrix of covariates $\{X_1, \ldots, X_q\}$. Similarly, we let β_γ be the set of p parameters specific to model M_γ, and K_γ be the corresponding matrix of covariates $\{K_1, \ldots, K_p\}$. In the simplest case, X is just a vector of ones, but we allow for more general sets of common covariates. We further assume that X and K_γ are orthogonal. More generally, we can consider a set of m (possibly infinite) covariates $\mathcal{K}_m = \{K_1, K_2, \ldots, K_m\}$. We then consider some set of models $\{M_i\}$, such that each model represents some subset of the covariates. In the $\mathcal{M}$-closed setting (Bernardo and Smith, 2009), we assume that our data is generated from model M_γ. In the $\mathcal{M}$-open setting, our data is generated according to $Y \sim \mathcal{N}\left(X\alpha + \mu, \sigma^2 I\right)$, where $\mu \notin \textit{Span}\{\mathcal{K}_m\}$.

Following the set-up in Bernardo and Smith (2009) and Vehtari and Ojanen (2012), we adopt the optimal prediction view. In this setting, the observations are generated by some model M_*, which may, or may not, be included in the space of considered models $\mathcal{M}$. The goal is to maximize the expected utility, with respect to the posterior predictive distribution

$$\bar{U}(a) = \int_{\tilde{y}} U(a, \tilde{y}) p(\tilde{y}|y, M_*).$$

Common choices of utility functions are the log-utility $U(a, \tilde{y}) = \log(\tilde{y})$ and the squared error utility $U(a, \tilde{y}) = (a - \tilde{y})^2$. In the former case, the goal is to choose an optimal distribution $a_k(\tilde{y})$, while the latter is to choose the optimal point estimator.

If we consider an oracle model selection procedure, where the true model is known, then the maximizers for the above utility functions are $\log(p(\tilde{y}|y, M_*))$ and $\mathbb{E}(\tilde{y}|M_*, y)$ respectively. In the $\mathcal{M}$-closed setting, $M_* \in \mathcal{M}$, so these maxima are obtainable. In the $\mathcal{M}$-open setting, they are not, so instead we choose the models that minimize respectively

$$\int_{\tilde{y}} \log\left(\frac{p(\tilde{y}|M_k, y)}{p(\tilde{y}|M_*, y)}\right) p(\tilde{y}|M_*, y) \qquad \text{or} \qquad \left[\mathbb{E}(\tilde{y}|y, M_K) - \mathbb{E}(\tilde{y}|, y, M_*)\right]^2.$$

The former is equivalent to minimizing the KL-divergence between the predictive distributions of the data generating process and the best candidate model. The latter minimizes the squared error loss.

The above formulation of the selection problem addressed oracle selection, where the true data generating process is known. This is not the case in any actual modeling scenario, so other approaches will have to suffice. In practice, we use a model selection procedure which searches the space of possible models, and we evaluate the respective loss functions over a validation set (cross-validation) or use information criterion to score each model. A procedure is said to be *consistent* with respect to a loss function L, if $\lim_{n\to\infty} \mathbb{P}\left(\hat{M}_n = \text{argmin}_{M_\gamma} L_n(M_\gamma)\right) = 1$ (Shao, 1997). Consistency is well defined when the true model is nested in the space of models considered, or, more generally, when the set of possible models is finite. When the set of possible models is infinite, or grows with n, there will not always be a unique limiting minimum for the loss function.

A slightly weaker condition, but one that may be more practical in many settings is *pairwise consistency*. In this setting, we consider two finite models, M_1 and M_2. A model selection procedure is pairwise consistent if $\lim_{n\to\infty} \mathbb{P}\left(L_n(\hat{M}_{M_1,M_2}) = \min_{M_1,M_2} L_n(M_\gamma)\right)$. In other words, when comparing two models, asymptotically, our model selection procedure will select the same model as an oracle method.

3. *g*-priors. As previously discussed, one choice of prior on β, proposed in Zellner (1986), is to let $\pi(\beta|\alpha,\sigma^2) = \pi(\beta|\sigma^2) = \mathcal{N}\left(\beta_0, g\sigma^2(K^TK)^{-1}\right)$. As commented in the introduction, these priors result in a posterior mean of $\mathbb{E}(\beta|y) = g/(g+1)\hat{\beta}_{OLS}$, where $g/(1+g)$ is the "shrinkage factor". Furthermore, if we let $RSS_\emptyset = (Y - X\hat{\alpha})^T(Y - X\hat{\alpha})$, and $RSS_\gamma = (Y - X\hat{\alpha} - K_\gamma\hat{\beta})^T(Y - X\hat{\alpha} - K_\gamma\hat{\beta})$ (with $\hat{\alpha}$ and $\hat{\beta}$ being the least squares estimates of α and β respectively), we can then define $S = RSS_\emptyset/RSS_\gamma$. This allows us to generalize the coefficient of determination, R^2, to cases where the null model may not simply be the intercept (in the case that it is, then $S = (1-R^2)^{-1}$).

If we consider a model where there are q common parameters and p model specific parameters, such that $p + q < n$, then we have that

$$BF_{\gamma:\emptyset} = (1+g)^{\frac{n-p-q}{2}}\left(\frac{g}{S}+1\right)^{-\frac{n-q}{2}}.$$

3.1. *Hyper-priors on g.* As commented in the introduction, all fixed choices of g (exempting some empirical Bayes methods discussed in Liang et al. (2008)), fail to resolve the Lindley information paradox. Alternatively, g can be considered random. Several of these hyper-priors on g are discussed below.

Beta priors. A common choice of prior is to let $g/(g+1) \sim \text{Be}(a,b)$. The most popular choice in this class is the hyper-g prior proposed by Cui and George (2008), and discussed extensively by Liang et al. (2008). With this choice, we let $a = 1$ and $b = a_g/2 - 1$, where $2 < a_g < 4$, with choices near 2 beginning to exhibit information paradox type behavior. This prior leads to a quasi-closed form marginal likelihood in the form of a Gaussian hypergeometric distribution, and satisfies nearly all desirable criteria, except for null model consistency. Another popular choice is the horseshoe prior (Carvalho, Polson and Scott, 2010), which sets $a = b = 1/2$. Other choices are discussed in Ley and Steel (2012).

Several authors have proposed ideas to address the null model consistency problems. But there has been little discussion of an underlying theory explaining under what conditions these consistency issues arise. Liang et al. (2008) notes that priors that don't depend on n tend to result in inconsistency, while priors that do, tend to resolve them. We formalize this as follows

THEOREM 1. *Let $M_\emptyset$ be the data-generating model. Then, $BF_{\gamma,\emptyset} \xrightarrow{\mathcal{P}} 0$ if, and only if,* $\lim_{n\to\infty} \mathbb{E}_{\pi(g)}\left(1/(g+1)\right) \to 0$.

PROOF. Let $S = RSS_{\emptyset}/RSS_{\gamma}$, then the Bayes factor $BF_{\gamma:\emptyset}$ comparing M_{γ} to M_0 is:

$$BF_{\gamma:\emptyset} = S^{\frac{n-q}{2}} \int_0^{\infty} (1+g)^{-\frac{p}{2}} \left(\frac{g+1}{g+S}\right)^{\frac{n-q}{2}} \pi(g) dg,$$

then $\mathbb{E}(1/(g+1))^{p/2} \leq BF_{\gamma:\emptyset} \leq S^{(n-q)/2}\mathbb{E}(1/(g+1))^{p/2}$. Assume that $\mathbb{E}(g+1)^{-1} \to 0$, then, from Lemma 2 in the appendix, $\mathbb{E}(g+1)^{-p/2} \to 0$. That, in turn, implies that $(g+1)^{-p/2} \xrightarrow{p} 0$. Then, since $S^{(n-q)/2} \xrightarrow{d} \exp\{\chi^2_p/2)\}$ (See Fernández, Ley and Steel (2001)), we have that $S^{(n-q)/2}\mathbb{E}(g+1)^{-p/2} \xrightarrow{p} 0$. Thus, $BF_{\gamma:\emptyset} \xrightarrow{p} 0$. For the reverse direction, we have that $BF_{\gamma:\emptyset} \xrightarrow{p} 0$. Then, since $0 \leq \mathbb{E}(g+1)^{-p/2} \leq BF_{\gamma:\emptyset}$, we have that $\mathbb{E}(g+1)^{-p/2} \to 0$. Then, from Lemma 2, we have that $\mathbb{E}(g+1)^{-1} \to 0$. □

In Bayarri et al. (2012), a weaker version of this result is given. Namely that: $BF_{\gamma,\emptyset} \xrightarrow{\mathcal{P}} 0$ if $\lim_{n\to\infty} \mathbb{E}(g+1)^{-p/2} \to 0$. While this result establishes a class of null-consistent priors, it does not fully enumerate the class, whereas our result shows that the class only contains priors such that $\mathbb{E}(g+1)^{-1} \to 0$. Notable priors that are not in this class include the hyper-g prior and horse-shoe g priors, that have non-zero expectation of $1/(g+1)$. Later development of hyper-priors on g have been focused on ways to introduce sample size dependencies into the hyper-parameters of the prior distribution, most of these resulting in $\mathbb{E}(1/(g+1)) \in \mathcal{O}(1/n)$.

Hyper g/n prior. This prior, proposed by Liang et al. (2008), places a $\text{Be}(1, a/2-1)$ prior on $g/(g+n)$, which gives a prior mean on $1/(g+1)$ that is decreasing in n. There exists a similar extension to the Horseshoe prior (Ley and Steel, 2012), where $g/(g+n) \sim \text{Be}(1/2, 1/2)$.

Maruyama and George prior. Maruyama and George (2011) propose a hyper-prior such that $g \sim \text{BetaPrime}((n-p-q)/2-a, a)$, where $0 < a < (n-p-q)/2$, with a recommended value of $1/4$. This prior has an added benefit of having truly closed form marginal distributions, and a conjugate posterior (in the scaled beta-prime family) for $g|y \sim \text{BetaPrime}((n-p-q)/2-a, a+p/2, S)$.

Other adaptive beta priors. Some general families are obtained by letting the hyperparameters of the hyper-g prior depend on n, e.g. letting $a_g = 2(1+1/n)$. More generally, Zeugner and Feldkircher (2009) propose letting $a_g = 2(1+1/\max\{n, p^2\})$. Ley and Steel (2012) similarly propose letting $g/(1+g) \sim \text{Be}(c\max\{n, p^2\}, c)$ for some $c > 0$.

Zellner-Siow prior. This prior, based on the early work by Zellner and Siow (1980), and formalized by Liang et al. (2008), places a Cauchy prior on $\pi(\beta|\sigma^2)$. Since the Cauchy admits representation as a mixture of Normals, this prior is equivalent to letting $g \sim \text{IG}(1/2, n/2))$. This prior does not admit closed form marginal distributions, and is often computed via a Laplace approximation, which degrades as p gets large.

Truncated priors. Ideas include truncating g in such a way that $\mathbb{E}(g)$ will tend to infinity. Bottolo and Richardson (2008) propose truncating the prior space such that $g < \max\{n, p^2\}$ and then imposing a uniform prior on $\log(g+1)$. One important class of truncated priors are truncated gamma priors (Li and Clyde, 2018) on $1/(g+1)$: $u \sim TG_{(0,v)}(a,b) = b^a/\gamma(a, bv)\, u^{a-1} \exp\{-bu\} 1_{\{u<v\}}$. These priors are conjugate with an approximation of the likelihood, but are intractable with the proper likelihood function. The Robust prior (under default parameters) proposed by Bayarri et al. (2012) falls into this family ($v = (p+1)/(n+1), a = 1/2, b = 0$). Furthermore, the ZS-adaptive prior proposed by Held, Bové and Gravestock (2015) (which mimics the behavior of the Zellner-Siow prior by matching the mode) is a member of this family, with parameters ($v = 1, a = 1/2, b = (n+3)/2$).

Robust prior. This prior was developed in Bayarri et al. (2012) to satisfy their model selection criteria. They let $\pi(g) = a_r[p_r(b_r+n)]^{a_r}(g+b_r)^{-(a_r+1)} 1_{\{g \geq p_r(b_r+n)-b_r\}}$, where

Prior	L	b	a	ϕ	Null Consistent
Hyper-g	0	1	$\frac{a_g}{2} - 1$	1	No
Horseshoe	0	$\frac{1}{2}$	$\frac{1}{2}$	1	No
F-Z	0	1	$1/\max\{p^2, n\}$	1	Yes
Benchmark	0	$c\max\{n, p^2\}$	c	1	Yes
M and G	0	$\frac{n-p-q}{2} - a$	a	1	Yes
hyper-g/n	0	1	$\frac{a}{2} - 1$	n	Yes
Horseshoe g/n	0	$\frac{1}{2}$	$\frac{1}{2}$	n	Yes
Robust	$p_r(b_r + n) - b_r$	1	a_r	$p_r(b_r + n)$	Yes
PEP	n	$\frac{n-p-q}{2}$	$\frac{n-p-q}{2}$	n	Yes

TABLE 1

Hyper-priors on g as Pearson Type-VI distributions. F-Z denotes the prior in Zeugner and Feldkircher (2009). M and G denotes the hyper-prior in Maruyama and George (2011).

$p_r \geq b/(b+n)$ and $a, b > 0$. They recommend taking $p_r = 1/(p+q)$ and $b_r = 1$. This prior both truncates the prior support of g from below, and also applies a scaling to g that grows in n.

Power expected posterior prior. The PEP prior was developed in a different context. The Expected Posterior Prior lets the prior on β be set to the expected posterior distribution with respect to an imaginary training sample. The PEP prior modifies this approach such that the likelihood of the marginal imaginary data is taken to the power of $1/n$. However, as noted by Consonni et al. (2018), the PEP is equivalent to a g-prior, with a hyper-prior on g of the form $\pi(g) \propto g^{-(n-p-q)}(g-n)^{(n-p-q)/2-1}$, where $g \geq n$.

Many of the above mentioned priors can be represented as a Pearson Type-VI prior:

$$\pi(g) = \frac{\left(\frac{(g-L)}{\phi}\right)^{b-1}\left(1 + \frac{(g-L)}{\phi}\right)^{-(a+b)}}{\phi \mathbf{B}(a,b)} 1_{\{g \geq L\}} \qquad \phi = \frac{L+1}{\theta} \geq 0.$$

Table 1 describes each of the above priors in terms of the Pearson Type VI distribution. We notice that not every prior proposed in the literature falls under this categorization. Most notably, the inverse Gamma prior proposed by Zellner and Siow (1980) is not a member of the Pearson Type-VI distribution. Nor is the prior proposed by Bottolo and Richardson (2008), or the truncated gamma priors proposed by Li and Clyde (2018). In fact, Li and Clyde (2018) proposed a broader class of priors, the compound confluent hyper-geometric distributions, of which the Pearson is just a special case.

3.2. *An approximation for posterior means and Bayes factors.* For further analysis, we are interested in the properties of the posterior means and Bayes factors for large, but finite, n. However, all prior choices for g, with the exception of the one proposed by Maruyama and George (2011), do not result in closed form expressions and are often represented as a ratio of Hyper-geometric functions (see for example Liang et al. (2008) or Bayarri et al. (2012)). Numerical libraries, such as the Cephes routine, are available for the ${}_2F_1$ function, but their accuracy tends to deteriorate rapidly as the number of parameters grow large. Alternatively, Laplace approximations perform well in cases where the shape parameters in the Pearson Type-VI distribution are fixed in n. However, while these are often easily computable, they have complicated analytic forms, obscuring the underlying behavior of the quantities of interest.

We develop asymptotic approximations for special cases of the Pearson-Type VI prior. In particular, we consider approximations for the Beta priors of $1/(1+g)$, where a and

Prior	$\mathbb{E}\left(\frac{1}{g+1}\vert y\right)$	$BF_{\gamma_1:\gamma_2}$
Hyper-g	$\frac{p+2a_g-2}{n-p-q-a_g+2}\left[\frac{1}{S_\gamma-1}\right]$	$(1-\Delta)^{\frac{n-p-q-a_g}{2}+1}\left[\frac{\Gamma(\frac{p+a_g}{2}-1)\Gamma(\frac{n-p-q-a_g+2}{2})}{\Gamma(\frac{p+a_g+1}{2}-1)\Gamma(\frac{n-p-q-a_g+1}{2})}\right]U^{\frac{p+a_g}{2}-1}(S_{\gamma_2}-1)^{\frac{1}{2}}$
Horseshoe	$\frac{p+1}{n-p-q-1}\left[\frac{1}{S_\gamma-1}\right]$	$(1-\Delta)^{\frac{n-p-q-1}{2}}\left[\frac{\Gamma(\frac{p+1}{2})\Gamma(\frac{n-p-q-1}{2})}{\Gamma(\frac{p+2}{2})\Gamma(\frac{n-p-q-2}{2})}\right]U^{\frac{p+1}{2}}(S_{\gamma_2}-1)^{\frac{1}{2}}$
F-Z	$\frac{p+2/n}{n-p-q-2/n}\left[\frac{1}{S_\gamma-1}\right]$	$(1-\Delta)^{\frac{n-p-q-1/n}{2}}\left[\frac{\Gamma(\frac{np+1}{2n})\Gamma(\frac{n-p-q-1/n}{2})}{\Gamma(\frac{n(p+1)+1}{2n})\Gamma(\frac{n-p-q-1-1/n}{2})}\right]U^{\frac{np+1}{2n}}(S_{\gamma_2}-1)^{\frac{1}{2}}$

TABLE 2
Approximations of posterior means and Bayes Factors under different beta priors.
$U=(RSS_\emptyset - RSS_{\gamma_2})/(RSS_\emptyset - RSS_{\gamma_1})$

b do not grow in n, which are equivalent to letting $L=0$ and $\phi=1$ in the Pearson-Type VI distribution. These hyper-priors include the Hyper-g, Horseshoe-g, F-Z, Benchmark, and Maruyama and George priors.

PROPOSITION 1. *Let $\frac{1}{g+1}\sim Be(a,b)$, with $a,b>0$ and bounded in n, and let the data generating process $M_*\neq M_\emptyset$, then*

1. $BF_{\gamma:\emptyset}\approx S_\gamma^{\frac{n-q}{2}}(S_\gamma-1)^{-(\frac{p}{2}+a)}\frac{\mathbf{B}(\frac{n-p-q}{2}-a,\frac{p}{2}+a)}{\mathbf{B}(a,b)}$ *where* $S_\gamma=\frac{RSS_\emptyset}{RSS_\gamma}$ *and RSS denotes the residual sums of squares.*
2. $\mathbb{E}(\frac{1}{1+g}\vert y)\approx\frac{p+2a}{n-p-q-2a}\left[\frac{1}{S_\gamma-1}\right]$

The proof of this proposition can be found in the appendix. Importantly, if we consider a model M_{γ_1} and augment it with one additional parameter to obtain model M_{γ_2}, then the Bayes Factor comparing the two is approximately:

$$BF_{\gamma_1:\gamma_2}\approx(1-\Delta)^{\frac{n-p-q}{2}-a_1}\left[\frac{\Gamma(a_1+\frac{p}{2})\Gamma(a_1+b_1)\Gamma(a_2)\Gamma(b_2)\Gamma(\frac{n-p-q}{2}-a_1)}{\Gamma(a_2+\frac{p+1}{2})\Gamma(a_2+b_2)\Gamma(a_1)\Gamma(b_1)\Gamma(\frac{n-p-q-1}{2}-a_2)}\right]\left[\frac{RSS_\emptyset-RSS_{\gamma_2}}{RSS_\emptyset-RSS_{\gamma_1}}\right]^{\frac{p}{2}+a_1}(S_{\gamma_2}-1)^{\frac{1}{2}+a_2-a_1},$$

where $1-\Delta=RSS_{\gamma_2}/RSS_{\gamma_1}$. Thus, Δ can be thought of as the relative improvement in the RSS from adding a new parameter. Table 3.2 displays the approximate means and Bayes factors for the hyper-g, horseshoe-g, and F-Z priors. In all the cases considered, we see that $\mathbb{E}(1/(g+1)\vert y)$ is $\mathcal{O}(1/n)$ and the Bayes factor of the smaller model to the larger is $\mathcal{O}((1-\Delta)^{(n-p-q)/2-a}n^{1/2})$. Thus, it is important to note that, despite the wide variety of prior choices for g, all of then produce posterior means and Bayes factors that have essentially identical behavior when n is sufficiently large.

4. Two New Problems. Two problematic behaviors of posterior distributions based on g-priors for large sample sizes, that have not received much attention in the literature are: *Shrinkage deficient estimation* and *model superinduction*. The former refers to the phenomenon where $\mathbb{E}(1/(g+1)\vert y)\in\mathcal{O}(1/n)$, implying that the posterior expectation of the shrinkage factor decays at least linearly in n. This implies that for large sample sizes, the posterior expectation of β, will have no shrinkage and be essentially equal to $\hat{\beta}_{LS}$. The latter refers to the phenomenon, where, in the $\mathcal{M}$-open setting, the Bayes factor comparing any two nested models, $M_\gamma\subset M_{\gamma'}$, is such that $\lim_{n\to\infty}BF_{M_\gamma,M_{\gamma'}}=0$, implying that, for large

sample sizes, the more complex model will be preferred. This creates both theoretical parsimony problems, as well as practical computation issues, for any step-wise model selection procedure.

4.1. *SDE.* In order for g-priors to be consistent,$\mathbb{E}(g/(1+g)|y, M_\gamma) \to 1$ as $n \to \infty$. Thus, as $n \to \infty$, all g-priors produce no-shrinkage, and result in OLS estimates: $\mathbb{E}(\beta|y) \to \hat{\beta}_{LS}$. The rate at which this convergence occurs is linear in n, or faster. To demonstrate this result, we first consider the following assumptions:

ASSUMPTIONS 1. Let $y \sim \mathcal{N}(X\alpha + \mu_n, \sigma^2 I_n)$, such that $X^T \mu_n = 0$.
A1. Let $\mu_n^T \mu_n / n \to C_\mu$, where $C_\mu \in (0, \infty)$.
A2. Consider a class of priors such that $\mathbb{E}\left((g+1)^{-1}|M_\gamma\right) \leq B_\gamma n^{-1}$ for all M_γ, where $B_\gamma \in (0, \infty)$.
A3. Furthermore, let $\mathbb{E}\left(n/(g+1)|M_\gamma\right) \xrightarrow{\mathcal{L}_1} R_\gamma$ for all M_γ, where $R_\gamma \in (0, \infty)$.

THEOREM 2 (Shrinkage Deficient Estimation). *Under the Hyper-g, Horseshoe-g, Maruyama and George g-priors, and all priors satisfying A.2 and A.3, we have that* $\mathbb{E}(1/(g+1)|y) \in \mathcal{O}(\frac{1}{n})$.

PROOF. For the Beta priors, the approximations developed in Proposition 1 suffice to show the result. For the general case where the prior mean of $1/(g+1)$ decays linearly, we have that

$$\mathbb{E}\left(\frac{1}{g+1}|y\right) = \frac{\int_g (1+g)^{-\frac{p+2}{2}} \left(\frac{g+1}{g+S_{\gamma,n}}\right)^{\frac{n-q}{2}} \pi_n(g|M_\gamma)\mathrm{d}g}{\int_g (1+g)^{-\frac{p}{2}} \left(\frac{g+1}{g+S_{\gamma,n}}\right)^{\frac{n-q}{2}} \pi_n(g|M_\gamma)\mathrm{d}g}.$$

Suppose towards a contradiction that $\lim_{n\to\infty} n\mathbb{E}\left((g+1)^{-1}|y\right) \geq B_\gamma$, for all B_γ. Then we have that

$$\lim_{n\to\infty} \int_g \left((1+g)^{-1} - \frac{B_\gamma}{n}\right)(1+g)^{-\frac{p}{2}} \left(\frac{g+1}{g+S_{\gamma,n}}\right)^{\frac{n-q}{2}} \pi_n(g|M_\gamma)\mathrm{d}g \geq 0.$$

The left hand side is

$$\leq \lim_{n\to\infty} \int_g \left(\left(\frac{n}{g+1}\right) - B_\gamma\right) \pi_n(g|M_\gamma)\mathrm{d}g = \mathbb{E}\left(\left(\frac{n}{g+1}\right)|M_\gamma\right) - B_\gamma = K_\gamma - B_\gamma$$

There exists a $B_\gamma > K_\gamma$, so that the quantity is less than zero, creating a contradiction. Therefore, under the assumptions in A.1 and A.2, $\mathbb{E}\left((g+1)^{-1}|y\right) \in \mathcal{O}\left(1/n\right)$. □

While the consistency properties are desirable, in practice, SDE means that for any large data set, a g-prior is essentially reproducing OLS estimates. We seek to slow the rate at which this convergence occurs, such that we retain some level of posterior shrinkage for large, but finite, n.

4.2. *Model Superinduction.* Model superinduction refers to a phenomenon where a model selection procedure will overwhelmingly favor larger models as n increases. We first establish the following result, which illustrates the phenomenon. Consider a class of priors $\mathcal{G}$, on g, which includes

- The Maruyama and George Prior
- Any prior such that $\frac{1}{g+1} \sim \text{Be}(a, b)$ (for fixed a and b)
- Any prior such that $\mathbb{E}\left(\frac{n^k}{g+1}\right) \to C$, and $\mathbb{E}\left(\frac{g+1}{n^k}\right) \to C^*$ such that $k \geq 1$, $C \in (0, \infty)$, and $C^* \in (0, \infty)$.

THEOREM 3 (Model Superinduction). *For the class of priors $\mathcal{G}$, consider data generated as $y \sim \mathcal{N}(X\alpha + \mu, \sigma^2)$. Consider linear models M_γ and $M_{\gamma'}$ with p and m parameters, respectively, and such that $M_\gamma \subset M_{\gamma'}$. Then $\lim_{n\to\infty} BF_{\gamma,\gamma'} \to 0$, unless*

1. *M_γ is the true model*
2. *There is perfect collinearity between the terms in $M_{\gamma'} \setminus M_\gamma$ and M_γ in the limit*
3. *$\mu \notin ker(\{M_{\gamma'} \setminus M_\gamma\})$*

PROOF. To establish the result for the fixed beta priors, we then note that under the approximations in the previous section, we have that

$$BF_{M_\gamma:M_{\gamma'}} \approx \left[\left(\frac{S_\gamma}{S_{\gamma'}}\right)^{\frac{n-q}{2}}\right]\left[\frac{\Gamma(\frac{n-p-q}{2}-a)}{\Gamma(\frac{n-m-q}{2}-a)}\right]\left[\left(\frac{S_{\gamma'}-1}{S_\gamma-1}\right)^{\frac{p}{2}+a}(S_{\gamma'}-1)^{\frac{m-p}{2}}\frac{\Gamma(\frac{p}{2}+a)}{\Gamma(\frac{m}{2}+a)}\right].$$

From Lemma 8, we see that the first term goes to zero in probability at an exponential rate. The second term goes to infinity, but at a polynomial rate (in $m-p$), and the third term converges to a constant (from Lemma 4). Thus $\lim_{n\to\infty} BF_{M_\gamma,M_{\gamma'}} \to 0$.

To establish the result in the case where $\mathbb{E}\left(n^k/(g_n+1)\right) \to C$, we note that

$$BF_{M_\gamma:M_{\gamma'}} = \left(\frac{S_{\gamma,n}}{S_{\gamma',n}}\right)^{\frac{n}{2}} \frac{\int_g (1+g)^{\frac{-p}{2}}(\frac{g+1}{g+S_{\gamma,n}})^{\frac{n-q}{2}}\pi_n(g|M_\gamma)\mathrm{d}g}{\int_g (1+g)^{\frac{-m}{2}}(\frac{g+1}{g+S_{\gamma',n}})^{\frac{n-q}{2}}\pi_n(g|M_{\gamma'})\mathrm{d}g}.$$

From Corollary 3, we know that the ratio of integrals goes to infinity at a polynomial rate. Since the first term diverges at an exponential rate, we know that $BF_{M_\gamma:M_{\gamma'}} \to 0$ as $n \to \infty$, for all M_γ. □

It should be noted that the second exception here is rather trivial, as perfect collinearity is rarely an issue (especially in the limit). The third condition is similarly trivial, as if it implies perfect prediction of μ (in the limit), with $X\hat{\alpha}_{LS} + K\hat{\beta}_{LS} + (I - P_{M_{\gamma'}\setminus M_\gamma})y$. An immediate corollary (from Mukhopadhyay, Samanta and Chakrabarti (2015)) is

COROLLARY 1. *Let $0 \leq s < 1$. If $\max_{\{M_{\gamma_1}, M_{\gamma_2}\}} \frac{p(M_{\gamma_1})}{p(M_{\gamma_2})} \leq \exp\{n^{(1-s)}\}$, then $p(M_\gamma|y) \to 0$ as $n \to \infty$.*

As we discuss below, these results have strong implications for model comparison in $\mathcal{M}$-open settings.

4.2.1. *$\mathcal{M}$-Closed Case.* In the $\mathcal{M}$-closed case, the true data generating model is contained within the set of linear models we are considering. In this case, the problem of model superinduction is fairly mild, as the posterior probability of the true model will always tend to one. Theorem 3 indicates, though, that for any two nested models that aren't the true model, the larger one will always be preferred. This can create issues for model search algorithms based on iteratively adding and dropping parameters until convergence. While such a procedure will always converge to the true model, it may tend to explore regions of heavily parameterized models before selecting the true model.The resulting computational issues are especially salient when dealing with large model spaces.

4.2.2. *$\mathcal{M}$-Open Case.* In the $\mathcal{M}$-open case the problem of model superinduction can be quite severe, as for any large data set, a step-wise model selection procedure will continue to add parameters indefinitely. Moreover, in the case where we have a finite set of possible covariates, we have the following corollary to Theorem 3:

COROLLARY 2. *Consider a model selection problem with a finite set of m possible covariates and let M_F denote the model that contains all of them. In the $\mathcal{M}$-open case, under the conditions established in Theorem 3 and Corollary 1,* $\lim_{n\to\infty} \mathbb{P}\left(M_F|y\right) \to 1$.

This result follows immediately from Corollary 1 and Theorem 3 by considering the Bayes factor of any model to the full model. This creates two major issues. First, as a general philosophical problem, we prefer parsimonious models. Very minor improvements in model fit should not result in extremely strong evidence in favor of the more heavily parameterized model. Second, and more practically, this behavior forces step-wise selection algorithms to explore areas of the model space consisting of very large models. This adds considerable computational burdens, making it impractical to actually evaluate these models.

One question one may ask, is to what extent Bayes factor based model selection even makes sense in the $\mathcal{M}$-open case. Indeed, since the true model is outside the space of considered model, and thus is allocated 0 prior probability, this procedure seems to start out with a prior mis-specification issue. Unfortunately, the Model superinduction problems is a deeper issue than that, as seen with the following two results. First, we establish a useful Lemma:

LEMMA 1. *Consider a model selection problem between two nested models, M_{γ_1} with p covariates and M_{γ_2} with 1 additional covariate. Then, in the $\mathcal{M}$-open case, under the conditions in Theorem 3,* $\operatorname{argmin}_{M_\gamma} (\mu_n - P_{M_\gamma}\mu_n)^T(\mu_n - P_{M_\gamma}\mu_n) = M_{\gamma_2}$.

PROOF. Consider the partition of the projection matrix $P_{M_{\gamma_2}}$ into $P_{M_{\gamma_1}} + P_{M_{\gamma_2}\setminus M_{\gamma_1}}$. We note that $P_{M_{\gamma_2}\setminus M_{\gamma_1}}$ is the projection matrix of the components of M_{γ_2} orthogonalized to the components of M_{γ_1} (ie. $P\{(I - P_{M_{\gamma_1}})K_{M_{\gamma_2}\setminus M_{\gamma_1}}\}$. We can then write:

$$\lim_{n\to\infty} (\mu_n - P_{M_{\gamma_2}}\mu_n)^T(\mu_n - P_{M_{\gamma_2}}\mu_n) = \lim_{n\to\infty} (\mu_n^T\mu_n - \mu_n^T P_{M_{\gamma_1}}\mu_n - \mu_n^T P_{M_{\gamma_2}\setminus M_{\gamma_1}}\mu_n).$$

The last term will always be non-zero under the final two conditions established in Theorem 3. □

We develop two final results. The first one extends the problem of model superinduction to oracle estimators in the $\mathcal{M}$-open setting. The second one, establishes the pairwise consistency properties of Bayes factor based model selection in the $\mathcal{M}$-open case.

THEOREM 4. *Consider data generated from a model M_*, and an "oracle" model selection procedure, using either the predictive squared error loss function*

$$M_k = \underset{M_\gamma \in \mathcal{M}}{\operatorname{argmin}} \lim_{n\to\infty} \int_{\tilde{y}} (\mathbb{E}(\tilde{y}|y_n, M_\gamma) - \tilde{y})^2 p(\tilde{y}|y_n, M_*)$$

or the predictive log loss

$$M_k = \underset{M_\gamma \in \mathcal{M}}{\operatorname{argmin}} \lim_{n\to\infty} \int_{\tilde{y}} \frac{\log\left(p(\tilde{y}|M_*, y_n)\right)}{\log\left(p(\tilde{y}|M_\gamma, y_n)\right)} p(\tilde{y}|y_n, M_*).$$

In choosing between models M_{γ_1} with p covariates, and M_{γ_2} with one additional covariate, the oracle will always favor the larger model, M_{γ_2}.

PROOF. The proofs follow from a simple rearrangement of the loss functions, that shows that minimization, in both cases, is equivalent to the minimization done in Lemma 1. □

THEOREM 5. *When considering two models, M_γ and $M_{\gamma'}$, in the limit as $n \to \infty$, the oracle model selection procedure (under both the predictive squared error loss and predictive log loss functions) will select M_γ if, and only if, for any priors in the class $\mathcal{G}$, $\lim_{n\to\infty} BF_{\gamma,\gamma'} \to \infty$.*

PROOF. Under both the squared error loss function, and the log loss function, the Oracle estimator will only choose M_γ if $\lim_{n\to\infty} \mu_n^T P_\gamma \mu_n > \mu_n^T P_{\gamma'} \mu_n$. Equivalently, if we define $\alpha_\gamma = \mu_n^T P_\gamma \mu_n / \mu_n^T \mu_n$, and $\alpha_{\gamma'} = \mu_n^T P_{\gamma'} \mu_n / \mu_n^T \mu_n$, then $\alpha_\gamma > \alpha_{\gamma'}$. Following the same proof method from Lemma 4 (see Supplement (Fontana and Sansó, 2023)), we have that

$$\frac{S_{\gamma,n}}{S_{\gamma',n}} \xrightarrow{p} \frac{1 + \frac{(1-\alpha_{\gamma'})C_\mu}{\sigma^2}}{1 + \frac{(1-\alpha_\gamma)C_\mu}{\sigma^2}} = \frac{\sigma^2 + (1-\alpha_{\gamma'})C_\mu}{\sigma^2 + (1-\alpha_\gamma)C_\mu}.$$

If $\alpha_\gamma > \alpha_{\gamma'}$, then the numerator is larger than the denominator. We recall that the Bayes Factor has the form

$$BF_{M_\gamma : M_{\gamma'}} = \left(\frac{S_{\gamma,n}}{S_{\gamma',n}}\right)^{\frac{n}{2}} \frac{\int_g (1+g)^{\frac{-p}{2}} \left(\frac{g+1}{g+S_{\gamma,n}}\right)^{\frac{n-q}{2}} \pi_n(g|M_\gamma)\mathrm{d}g}{\int_g (1+g)^{\frac{-m}{2}} \left(\frac{g+1}{g+S_{\gamma',n}}\right)^{\frac{n-q}{2}} \pi_n(g|M_{\gamma'})\mathrm{d}g}.$$

From Corollary 3, we know that the ratio of integrals is bounded polynomially. Then, since $S_{\gamma,n}/S_{\gamma',n} > 1$, $BF_{M_\gamma, M_{\gamma'}} \xrightarrow{p} \infty$. Conversely, if $BF_{M_\gamma, M_{\gamma'}} \xrightarrow{p} \infty$, and assuming that that $S_{\gamma,n}/S_{\gamma',n} \not\to 1$, then we know that $\alpha_\gamma > \alpha_{\gamma'}$, which implies that $\lim_{n\to\infty} \mu_n^T P_\gamma \mu_n > \mu_n^T P_{\gamma'} \mu_n$. Thus, the Oracle estimator will also select model M_γ. □

From the previous results we conclude that, in the $\mathcal{M}$-open case, as any additional parameter will produce a *slightly* better fit in the limit, *any* consistent model selection procedure will ultimately suffer from model superinduction. Indeed, to the extent that we would like to minimize the square error loss, or the log loss, these results are not altogether problematic, as the selected models are marginally better fitting in an asymptotic sense. However, the rapidity at which this problem manifest in the finite n case *is* problematic. As we show in the proofs in the appendix, the growth of the Bayes factors is exponential in n. Not only do we have strong philosophical arguments to be skeptical of additional complexity in exchange for insignificant improvements in model fit, but when utilizing the Bayes factor for stochastic search algorithms, this behavior creates additional computational burdens.

Unfortunately, the behavior described here can not be entirely resolved, as the problem manifests even with oracle estimators. Indeed, any true solution, either in the prior on g or the model space prior, will result in posterior inconsistency issues in the $\mathcal{M}$-open case, as well as the finite model case. Thus, we must content ourselves with an approach that minimizes the severity of the behavior for large sample sizes used in practice (ie. $n = 10^6$).

5. A new hyper-prior. As we have discussed above, the problem of model superinduction is ultimately unavoidable, as it is even exhibited by oracle estimators. To minimize the impacts of this phenomenon for a given sample size we propose a modification of the Maruyama and George Hyper-prior, for which a closed form Bayes factor is available. Under such prior, we have that $g|y, M_\gamma \sim \text{BetaPrime}\left((n-p-q)/2-a, a+p/2, S_\gamma\right)$ and $BF_{\gamma:\emptyset} = S_\gamma^{(n-p-q)/2-a}\Gamma(a+p/2)\Gamma((n-p-q)/2)/(\Gamma(a)\Gamma((n-q)/2))$.

Thus, the severity of the model superinduction effect is determined by the leading term $S^{n-p-q/2-a}$, that depends on the hyper-parameter a. Choosing a large value for a minimizes the severity of the superinduction problem, yet a is subject to some constraints. First, $(n-p-q)/2-a \geq 0$, so $a \leq (n-p-q)/2$. If a grows linearly with n, then in the typical $\mathcal{M}$-closed setting, we will lose posterior model consistency. With this in mind, we let $a = (n^r-p-q)/2$, where $r \in (0,1)$. Then

$$g|y, M_\gamma \sim \text{BetaPrime}\left(\frac{n-n^r}{2}, \frac{n^r-q}{2}, S_\gamma\right) \quad \text{and} \quad BF_{\gamma:\emptyset} = S_\gamma^{\frac{n-n^r}{2}} \frac{\Gamma(\frac{n^r-q}{2})\Gamma(\frac{n-p-q}{2})}{\Gamma(\frac{n^r-p-q}{2})\Gamma(\frac{n-q}{2})},$$

and for any pair-wise comparison, the Bayes factor will decay (or explode) at the rate of $(S_\gamma/S_{\gamma'})^{(n-n^r)/2}$. Thus, large values of r will slow the effects of super-induction. Yet, convergence is still achieved as the n term in the exponential dominates in the limit. The posterior mean of the shrinkage factor is not available in closed form. We can instead obtain the bounds

$$\mathbb{E}\left(\frac{1}{g+S_\gamma}\right) \leq \mathbb{E}\left(\frac{1}{g+1}\right) \leq \mathbb{E}\left(\frac{1}{g}\right) \implies \frac{1}{S_\gamma}\frac{n^r-q}{n-q} \leq \mathbb{E}\left(\frac{1}{g+1}\right) \leq \frac{1}{S_\gamma}\frac{n^r-q}{n-n^r-2}.$$

Thus, the posterior shrinkage factor decays at a rate of $\mathcal{O}\left(n^{r-1}\right)$, which is slower than the linear decay rates with typical hyper-prior choices minimizing the problem of shrinkage deficient estimation discussed above. In addition this model tackles the conditional Lindley paradox, by removing the dependence on the model size p, in the Bayes factor.

One important feature of objective priors is student tails, which enables robustness to the mis-specification of the prior mean (Bayarri et al., 2012). Our prior produces student tails for $\mathbb{P}\left(\beta|\sigma^2\right)$, with $(n^r-p-q)/2$ degrees of freedom. Notably, this results in less heavy tailed t-distributions than those resulting from most default parameter choices. The proposed choice of a will also result in an effective truncation of the model space as $p \leq n^r - q$. Finally, we notice that the modified hyper-g prior, although fairly good at correcting for shrinkage deficient estimation, can only apply an additive penalty to the rate of the exponential growth or decay of the Bayes factor. As n grows large, even extreme choices of r, very near 1, may not be enough to compensate for the effect of model super-induction.

5.1. *Model Space Priors.* In order to effectively control model super-induction we consider the prior on the space of models $\{M_\gamma\}$. One common approach in the literature is to place an inclusion probability π, on each covariate. Thus, $\Pr[M_\gamma|\pi] = \pi^p(1-\pi)^{m-p}$. Here, m is used to denote the total number of possible covariates. As noted by Scott and Berger (2010), a fixed choice of π results in multiplicity issues. Instead, they propose the use of a $Be(a_\pi, b_\pi)$ prior on π. This results in

$$\Pr[M_\gamma] = \frac{\mathbf{B}(a_\pi+k, b_\pi+m-k)}{\mathbf{B}(a_\pi, b_\pi)} \qquad \text{and} \qquad \frac{\Pr[M_\gamma]}{\Pr[M_{\gamma^+}]} = \frac{b+m-k-1}{a+k}.$$

Here M_{γ^+} denotes the model with one additional covariate. In order to choose the values of a_π and b_π we seek a strategy that is adaptive in n. Kirsner and Sansó (2020), let $a_\pi = 1$ and $b_\pi = 10^{\frac{3n}{10^4}}$. This choice addressed the immediate computational concerns in their application, but resulted in a penalty that is too extreme, leading to overly smoothed spatial fields.

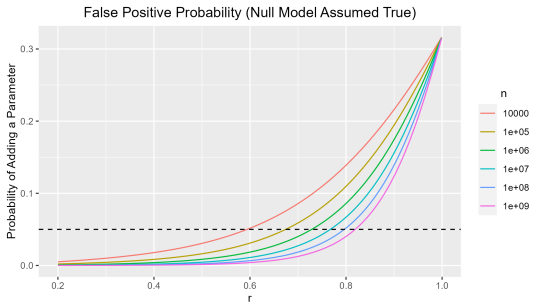

FIG 1. *Probability of adding a spurious parameter for different values of r and n (assuming a flat prior on the model space)*

Moreover, as the above theoretical results indicate, sub-exponential priors odds are necessary for consistency. In view of this, we let $a_\pi = 1$ and $b_\pi = Cn^{\log(n)}$ which is sub-exponential. Indeed, $b_\pi = n^{n^\epsilon}$, for $\epsilon < 1$ is another natural choice. We found that polynomials such as n^c, are ill-suited, as small values of c do little to counter the exponential growth of the Bayes Factor, and large values of c lead to overly extreme penalties when n is moderate ($\approx 10^4$). The value $Cn^{\log(n)}$ reflects the need for a prior odds ratio that is super-polynomial, but sub-exponential. The hyper-parameter C needs to be chosen such that the behavior for moderate n is reasonable.

5.2. *Choosing r and C.* Our proposed hyper-g prior and model space prior, require the specification of r and C. A choice needs to be made for a default value of these parameters. There are a few key points to consider that provide guidance. First, as r decreases, so does the size of the model space. Second, as r increases, so does the amount of posterior shrinkage. Third, as r increases, the Bayes factor's growth is slowed, but the Bayes factor is not sensitive to the choice of r with large n. Fourth, the larger C is, the more prior weight will be placed on smaller models, for all values of n. This can lead to undesirable behavior, where we over-penalize model space exploration for moderate values of n.

A possible strategy to fix the values of r is to consider the case where the null model is the true model, and examine the behavior of the Bayes factor under different values of n and r. The Bayes factor comparing the null model to a model with one parameter is a random variable distributed as a transformation of an F distribution. We can examine the false positive rate (FPR), defined as the $\Pr[BF_{\gamma:\emptyset} \geq 1]$. Figure 5.2 shows the FPR as a function of r, with a dotted line representing a 0.05 FPR for reference. We notice that the FPR is increasing in r and decreasing in n. Thus, we set a threshold for sufficiently small n and find the largest r that meets that threshold. As an example, for 0.05 and $n = 10^4$ we have $r = 0.593$. A lower bound for r can be obtained by selecting a minimum model space size, for example, $p_{\text{max}} \geq \sqrt{n}$. To set value of C we notice that the RFP is also a function of C. So we can obtain an RFP surface varying r and C jointly, and obtain its minimum. This is illustrated in Figure 5.2

Another criteria worth considering is the probability of correctly rejecting the null in favor of some other model. This, in a sense, is an analog to the true negative rate, or power of the test. Unfortunately, we are operating in settings where there isn't a true model, so these probabilities cannot be calculated in general. However, since our goal is ultimately to ensure that our penalties are not so extreme as to prevent the addition of important predictors, we can simply assume that some model in our space is the data generating process (in particular, we consider a model with 1 parameter). Since the Bayes Factor itself is a transformation of a

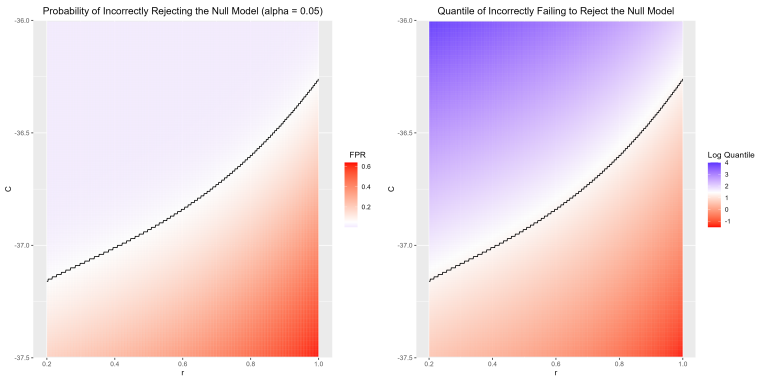

FIG 2. *False Positive Rate and Quantile of Power Function for various values of r and C. Here $n = 10,000$, and the black line represents the decision boundary of maximizing power such that the $FPR < 0.05$*

doubly non-central F-distribution, we have that

$$\mathbb{P}\left(BF_{\gamma:\emptyset} \geq 1\right) = \mathbb{P}\left(F_{1,n-2}(\lambda_1, \lambda_2) \geq (n-2)\left(\frac{\Gamma(\frac{n^r-p-q}{2})\Gamma(\frac{n-q}{2})}{\Gamma(\frac{n^r-q}{2})\Gamma(\frac{n-p-q}{2})}\right)^{\frac{2}{n-n^r}} \mathbb{P}\left(M_\gamma\right)^{-1}\right).$$

To maximize the power, we need only find the values of r and C that maximize the right side of the above expression (since the CDF is increasing and the F random variable does not depend on r or C).

For model selection procedures that operate in a series of add-drop moves, optimizing the FPR and TNR provides the calibration. In addition to setting the upper bound for the FPR, both the FPR and power expressions depend on n, for which we will need to select a reference value. If we see values of n larger than our reference value, our procedure will be much more conservative about accepting new parameters. As an example, for the values $n_0 = 10^4$, with $\alpha = 0.05$ we have $r = 0.924$ and $C = 10^{-36.4}$.

6. Data Examples. To illustrate the behavior of our proposed hyper-prior we consider two examples with 10^6 data simulated from the function:

$$y_i = \begin{cases} \sin(2\pi x_i) + 5 & 0 \leq x_i < 2 \\ -|\sin(x_i - 3)|^3 + 5 & 2 \leq x_i < 4 \\ 5|\sin(x_i - 5)| + 5 & 4 \leq x_i < 6 \\ -\sin(2\pi x_i)x_i + 5 & 6 \leq x_i \leq 10 \end{cases}.$$

The inputs x_i were generated uniformly between $(0, 10)$, and $\mathcal{N}(0, 1)$ error was added to each response. For both examples we consider two approaches: One with the standard hyper-g prior on β and the model space prior of Scott and Berger (2010) with a uniform covariate inclusion probability; and another with our modified hyper prior on g, and model space prior. We then examine the performance of our priors for different sample sizes, using MSE and continuous rank probability score (CRPS) as metrics.

6.1. *Kernel-Regression.* We fit a kernel regression using the hockey stick basis functions popularized by multivariate adaptive regression splines (MARS). We randomly select 100 points from the simulated input, denote them as t_j, and compute

$$B_{ij} = \max\{0, s_j(x_i - t_j)\}, \text{ with } s_j = \begin{cases} -1 & \text{with prob } \frac{1}{2} \\ 1 & \text{with prob } \frac{1}{2} \end{cases}.$$

n	$-\log(\mathbb{E}(g\|y)+1)$		Avg Model Size		Run-time (s)	
	M&G g-prior	New g-prior	M&G g-prior	New g-prior	M&G g-prior	New g-prior
10^3	-2.680	-0.780	17.20	14.35	56.03	46.34
5×10^3	-3.108	-0.912	32.58	28.69	176.17	186.75
10^4	-3.369	-0.960	35.64	34.81	272.93	233.92
5×10^4	-4.049	-1.050	38.03	34.39	1216.81	1048.68
10^5	-4.308	-1.092	42.00	42.03	2736.81	2279.47
5×10^5	-4.887	-1.178	55.00	38.00	15266.16	10874.73
10^6	-5.095	-1.212	67.86	41.00	40941.06	25396.30

TABLE 3
Model comparison of step-wise kernel regression using the Maruyama and George g-prior and our novel g-prior

n	MSE		CRPS	
	M&G g-prior	New g-prior	M&G g-prior	New g-prior
10^3	1.4208	1.7355	0.6389	0.8004
5×10^3	0.9452	1.1798	0.4323	0.6147
10^4	0.9190	1.0802	0.4262	0.5520
5×10^4	0.8990	1.0481	0.4068	0.5384
10^5	0.8967	1.0135	0.4081	0.5297
5×10^5	0.8926	0.9568	0.4005	0.4828
10^6	0.8913	0.9465	0.4007	0.4750

TABLE 4
Predictive performance comparison of step-wise kernel regression using the Maruyama and George g-prior and our novel g-prior

We then use a randomized add-drop procedure to explore the model space (which is of size 2^{100}), with the randomized acceptance probability being the marginal probability under both the Maruyama and George prior and our new prior. We repeat this procedure for samples sizes of 1000, 5000, 10000, 50000, 100000, 500000, and 1000000. After 100 iterations, we take the models with the the ten highest posterior probabilities, and examine the average model size (averaged over the normalized probabilities of the top ten models), as well as the average shrinkage factor. The results are presented in Table 6.1. Note that we computed $(\mathbb{E}(g)+1)^{-1}$ instead of $\mathbb{E}(1/(g+1))$. This is because the Maruyama and George prior does not lend itself to closed form expectations on the latter, but does under the former, as it is the estimator that minimizes the posterior scaled-square error loss.

Performance metrics can be found in tables 6.1 and 6.1. We observe that under the hyper-g prior, the expectation is decreasing approximately linearly in n, whereas the expectation under our prior is decaying at a much slower rate. We also see that our proposed priors induces the selection of significantly smaller models, appreciably shortening computing times. Our prior choices do tend to result in slightly larger mean squared errors and CRPS scores, taken with respect to the true mean function over a test set.

Figure 3 shows plots for the predicted mean functions under both priors, and the corresponding 95% probability intervals for the mean function. The fit under the new prior tend to be slightly smoother, at the cost of accuracy in some local regions where the mean function fluctuates rapidly. These inaccuracies decrease as the sample size gets large, and the fit from both approaches begins to look very similar. Notably, for large sample sizes, there are some regions where the default g-prior tends to produce over-fitting, whereas the proposed prior produces a smoother and more accurate fit. This suggests that overall MSE as a benchmark is

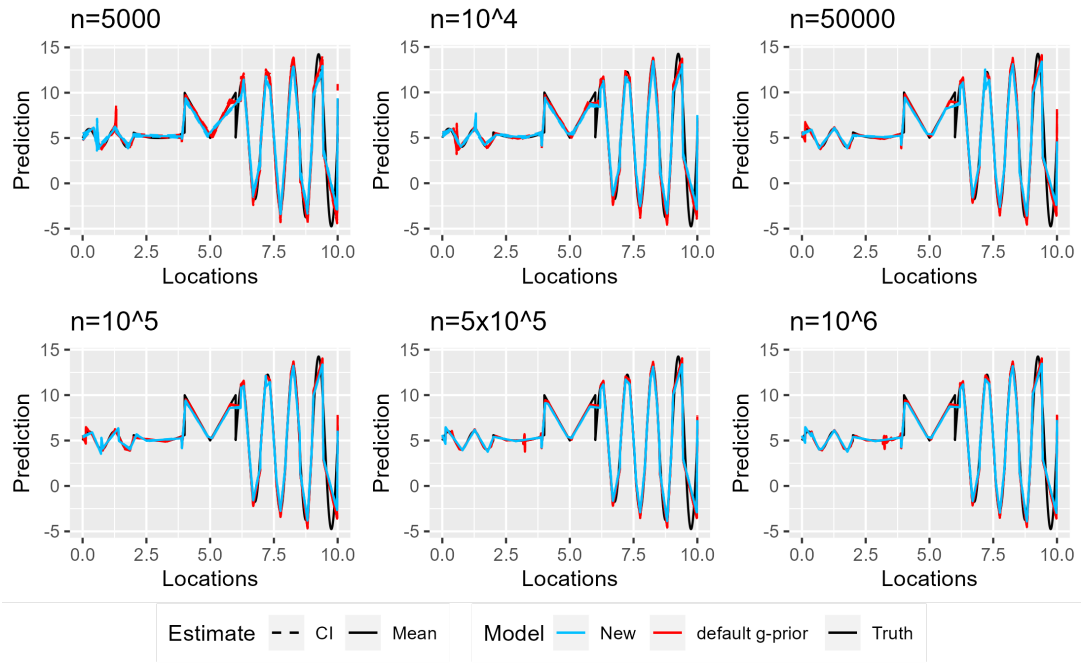

FIG 3. *Predictive fits using Kernel Regression with the standard g-prior and our new prior*

obscuring important information, as the differences in performance are best assessed locally. Thus, in Figure 4, we look at the partial MSE for the locations to the left of a given point ($\sum_{x_i \leq x}(x_i - \mu)^2 / \sum_{x_i < x} 1_{\{x_i \leq x\}}$). The plots reveal a complicated picture of performance. For all models, the total MSE favors the hyper-g prior, but there are many regions where the new prior performs better. These plots indicate very similar performance profiles, with the new prior requiring far less complex models and significant time savings.

6.2. *MSSS.* Multi-resolutional stochastic shotgun search is a method of fitting non-stationary multi-resolutional spatial data sets developed by Kirsner and Sansó (2020). Data is assumed to be generated as $y(s) = X(s)\alpha + w(s) + \epsilon(s) \quad \epsilon(s)_i \sim \mathcal{N}(0, \sigma^2)$. The spatial process is represented by a discrete process convolution: $w(s) = \sum_{r=1}^{\infty} \sum_{j=1}^{J(r)} K(s, s_j^r | \phi_r, \nu) \beta_j^r$. Here, s represents a given spatial point, s_j^r represents the center of the kernel, ϕ_r and ν are kernel parameters. Kernels are typically chosen to be compact, with the support varying based on ϕ_r. For example, a common choice is a Bezier Kernel:

$$K(s, s_j^r | \phi_r, v) = \begin{cases} \left(1 - \left(\frac{||s - s_j^r||}{\phi_r}\right)^2\right)^v & ||s - s_j^r|| \leq \phi_r \\ 0 & \text{Otherwise} \end{cases}.$$

To determine the centers of the kernel, a branching process is defined under the spatial domain. The region is first divided into $J(1)$ square sub-regions, each one of which can be iteratively split into 2^d additional sub-regions. The prior probability of a given tree-structure is given by a branching process, with a probability of splitting chosen to ensure termination of the tree with probability 1. This can be represented as

$$\mathbb{P}\left(\gamma_j^1 = 1\right) = 1 \qquad \mathbb{P}\left(\gamma_j^r = 1 | \gamma_{\lfloor \frac{j-1}{2^d} \rfloor - 1}^{r-1} = 1\right) = \pi \times \gamma_{\lfloor \frac{j-1}{2^d} \rfloor - 1}^{r-1}$$

$$\mathbb{P}\left(\gamma_j^r = 1 | \gamma_{\lfloor \frac{j-1}{2^d} \rfloor - 1}^{r-1} = 0\right) = 0 \qquad p(\pi) = \text{Beta}(a_\pi, b_\pi).$$

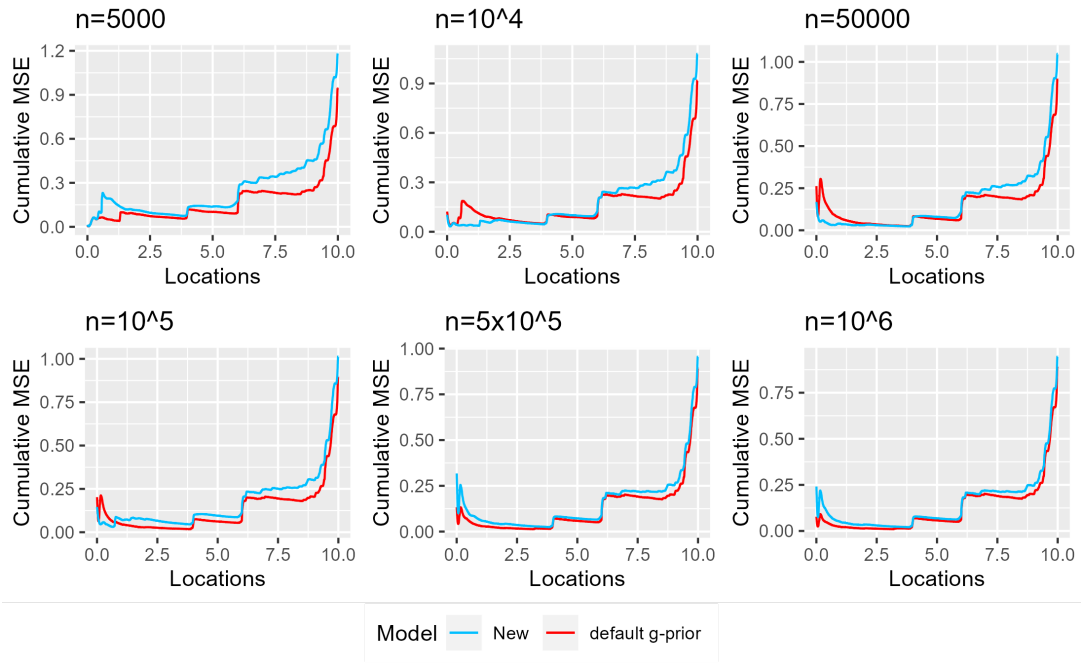

FIG 4. *Comparison of cumulative MSE over the spatial domain using kernel regression with the standard g-prior and our new prior*

We then have that $\beta_j^r = 0$, if $\gamma_j^r = 0$, ensuring sparsity. A g prior is then placed on any non-zero β_j^r, with flat priors on α and σ^2. This gives a quasi-closed form expression for the posterior probability of a given tree structure, which allows for posterior odds based step-wise model selection. The major advantage of this approach, is that for a given tree structure, the posterior probabilities of adjacent tree structures (those with one additional or one less split), can be computed via a rank-1 update, and thus can be parallelized. A major drawback of this procedure, is that for large spatial data sets, the tree structures are often too complex. This results in excessively long computation times and can result in overfitting. This problem was addressed in Kirsner (2020) by letting the beta prior that controls the probability of splitting have hyper-parameters $a_\pi = 1$, $b_\pi = 10^{\frac{3n}{10^4}}$. This creates theoretical consistency issues, as the prior odds-ratio is exponential in n. Practically, this approach resulted in overly smooth model fits that didn't accurately capture the multi-resolutional structure of the data.

We fit the simulated data used in the previous example with a MSSS model. In one case we use the standard hyper-g prior on β, and the Scott and Berger model space prior with $a_\pi = 1$, and $b_\pi = 5$, and in the other we use our modified hyper prior on g, and model space prior. See the supplement (Fontana and Sansó, 2023) for plots of the predictive fits for the different models, and the corresponding 95% predictive intervals. The results indicate that our method preforms as well as the approach proposed in Kirsner and Sansó (2020), although, visually, there are a few regions of sharp discontinues that our approach fails to capture. Yet, the benefits of our approach can clearly be seen in Table 6.2. We see that for large values of n, our method is nearly 5 times faster and results in models roughly half the size of the standard approach. The trees generated by our models are also less deep, and we see only slight decreases in MSE and CRPS performance. For the results in the table, in-sample MSE was calculated using 10000 points common to all the different sample sizes. Out-of-sample MSE was calculated with respect to the true mean function on a regular grid of $10,000$ points from $[0, 10]$. Our proposed prior performs slightly worse on MSE and CRPS metrics, this is

n	Run-time		Avg Size		Avg Depth	
	New	hyper-g	New	hyper-g	New	hyper-g
10^4	21.4	31.3	58.2	64.9	8.92	9.11
5×10^4	75.6	449.8	44.0	89.0	5.00	11.00
10^5	94.6	951.0	40.5	103.8	4.00	12.38
5×10^5	2555.8	14704.1	72.8	133.7	10.00	14.01
10^6	5041.1	24641.7	77.0	150.8	10.00	14.00
n	CRPS		MSE (in-sample)		MSE (out-sample)	
	New	hyper-g	New	hyper-g	New	hyper-g
10^4	0.2696	0.2479	1.0437844	0.9924766	0.0927	0.0400
5×10^4	0.2892	0.2375	1.1005105	0.9856406	0.1303	0.0105
10^5	0.3077	0.2340	1.1422150	0.9836267	0.1742	0.0077
5×10^5	0.2480	0.2349	1.0158044	0.9833313	0.0323	0.0040
10^6	0.2592	0.2505	1.0404386	1.0184548	0.0587	0.0405

TABLE 5

MSSS performance under our new prior (New) and the standard g-prior (hyper-g). MSE (in-sample) refers to the in sample MSE taken over a shared set of 10^4 observations. MSE (out-sample) refers to the the MSE from a predicted observation to the true mean function taken over a regular grid over the space

to be expected, as the models under the hyper-g prior are much more complex. The deficits, especially for large sample sizes, are not significant.

6.3. *Albedo Data.* To demonstrate the efficacy of our method with real data we consider use white-sky albedo data collected by two Geostationary Operational Environmental Satellites (GOES): GOES East, located at 75°W, and GOES West, located at 135°W. The data correspond to July 1st, 2000, and consist of 664,911 observations. White-sky Albedo, refers to BHR_{iso}, or Bi-Hemispherical Reflectance (the ratio between upward and downward radiation fluxes), under isotropic diffuse sky irradiance (Pinty et al., 2005). BHR_{iso} takes values between $(0, 1)$ at a given location s. In the July 1st data, only 3% observations were above 0.5. These values are likely attributable to aerosols or cloud coverage, and thus were removed. Although data is collected across the Western Hemisphere, we limit our analysis to the data collected within the Continental US (CONUS).

The measurements from the two satellites are aggregated to create a single dataset. Previous analyses have suggested that in some areas of the spatial field, there is a bias in the measurements between GOES East and GOES West. To account for this, we include the view zenith angle as a covariate. If we let (ψ_s, λ_s) and (ψ_e, λ_e) be the satellite latitude and longitude, and earth latitude and longitude, respectively, of a measurement, then we have that the view zenith angle, θ, is:

$$\sin(\theta) = \frac{42164 \sin(\beta)}{\sqrt{1.8084 \times 10^9 - 5.3725 \times 10^8 \cos(\beta)}}, \ \cos(\beta) = \cos(\psi_e - \psi_s)\cos(\lambda_e - \lambda_s).$$

The dataset was split, withholding 10% of the data at randomly sampled locations as test data, with the remainder being used as training data. MSSS was then used to fit the training data, with the two approaches described in the previous section. The maximum-iteration stopping condition was set to 1000 iterations, which the standard g-prior eventually reached (notably, as shown in the tables below, this took nearly a week of computing time). Plots of dense predictive surfaces over the CONUS for July 1st can be found in Figure 6.3, and a table of relevant metrics of model performance can be found in Table 6.3.

We see that our model results in a much smoother surface than the hyper-g prior, while still managing to capture the core features of the spatial surface. Although the hyper-g prior

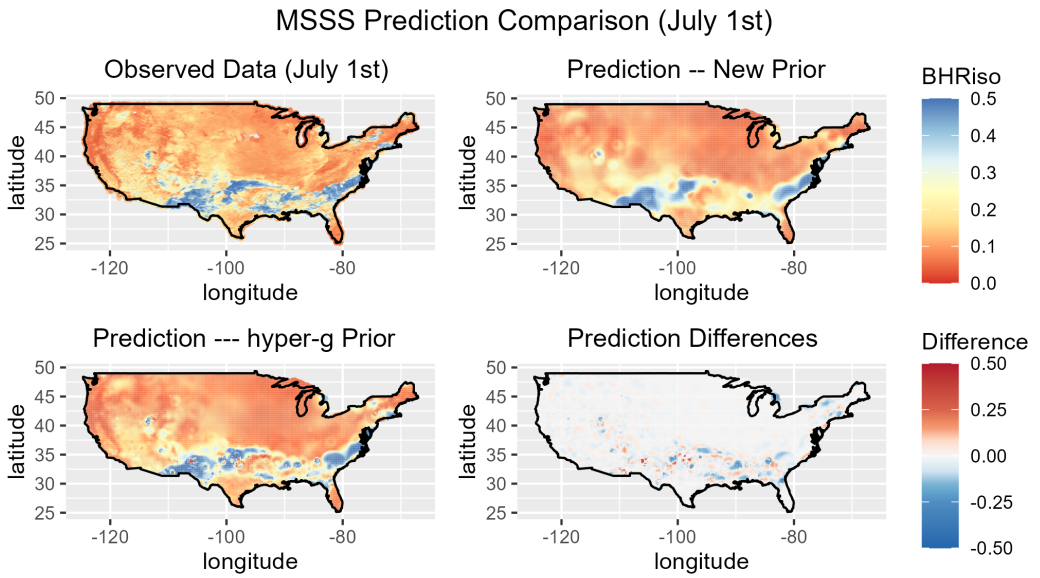

FIG 5. *Dense Predictive Surfaces generated by MSSS under our novel g-prior and the standard hyper-g prior*

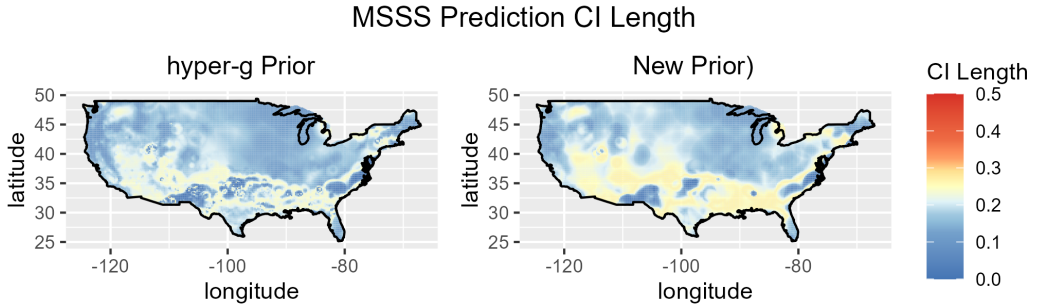

FIG 6. *Length of* 95% *predictive intervals generated by MSSS under our novel g-prior and the standard hyper-g prior*

Prior	Run-Time (hours)	Average Model Size	Average Max Depth	MSE	CRPS
New Prior	28.09	264.872	4	0.01239538	0.07185805
hyper-g	193.31	1156	8	0.005553336	0.03743759

TABLE 6
MSSS performance on GOES Albedo Data under our new prior (New) and the standard g-prior (hyper-g)

produces a more detailed predictive surfaces, and better captures some small features of the surface, it comes at a steep cost. Our model is nearly $1/5$ the complexity of the model under the hyper-g prior, and is fitted in $1/7$ the time. Moreover, is is important to note that MSSS applies a cap of 1000 iterations, at which point the model search is terminated. Under the hyper-g prior, the model search reached this hard cap, and did not terminate under the normal

stopping rules. In addition to a much smaller number of total knots, our model also produces significantly shallower trees. On the test-set, performance was assessed using the MSE and CRPS, and, as expected, we performed slightly worse than the hyper-g prior (results can be found in table 6.3). The length of the predictive CI for each point on the dense surface can be found in Figure 6.3. We see that our model does result in slightly larger intervals over the regions where the surface has the most volatility, and elsewhere performs comparable.

7. Discussion. We have explored two problems afflicting g-prior based model selection in $\mathcal{M}$-open settings. While much attention has been dedicated in the literature to the development of novel hyper-priors, the asymptotic behavior of the resulting posteriors, both on β and the model space, remain very similar for the different choices. For all choices, the posterior mean of β converges to the MLE at a linear rate, and the Bayes factor comparing two nested models, converges exponentially to zero in favor of the larger model. In this paper we have presented an approach to tackle those issues.

The model superinduction result fits firmly in line with the existing literature on $\mathcal{M}$-open model selection. Berk (1966) showed that in the $\mathcal{M}$-open setting, the posterior will eventually concentrate on the set of the parameter space that minimizes the KL-divergence. Theorem 5 can be read as a special case of this result for nested linear models. Similarly, results concerning the exponential rate of convergence of the Bayes Factor are also well known (Tadesse and Vannucci, 2021). While these theoretical properties have been well studied, the practical consequence that arises from them have received less attention. We have shown that in large data settings, model complexity grows in a way that ultimately makes any search of the model space impractical. As our data examples have demonstrated, a standard approach may take days to analyze a data-set of a few hundred thousand observations. Moreover, the additional complexity offers little of value for either inference or prediction.

In this paper we have focused on g-priors, but our results are much more general. Information criteria, such as the BIC, will clearly suffer from the same problem. Assuming normally distributed data, we have that $\text{BIC}_{\gamma_1} - \text{BIC}_{\gamma_2} = n \log\left(RSS_{\gamma_1}/RSS_{\gamma_2}\right) + (p_1 - p_2)\log(n)$. If $M_{\gamma_1} \subset M_{\gamma_2}$, then the first term will tend towards infinity exponentially faster than the second term will tend towards $-\infty$. Thus, since we choose the model with the smallest BIC score, we will always choose the larger model. In contrast, the log-posterior odds ratio, under our priors, will look like

$$\log\left(\frac{\pi(M_{\gamma_1}|y)}{\pi(M_{\gamma_1}|y)}\right) \approx \left(\frac{n-n^r}{2}\right)\log\left(\frac{RSS_{\gamma_2}}{RSS_{\gamma_1}}\right) + \left(\frac{p_{\gamma_2}-p_{\gamma_1}}{2}\right)\log\left(n^{1-r}\right) + (p_{\gamma_2}-p_{\gamma_1})\left[\log(n)\right]^2 + (p_{\gamma_2}-p_{\gamma_1})\log(C) + K$$

where K denotes a constant. It should be noted that the choice of $b = Cn^{\log(n)}$ was merely *a* choice, not the only one. Indeed, as noted in Section 5.1, there is a class of functions $\mathcal{H}$, such that $h(n) \in \mathcal{H}$ is super-polynomial, but sub-exponential. Thus, more generally, we can consider a Superinduction-Resistant-Information Criterion (SRIC):

$$\text{SRIC} = \left(\frac{n-n^r}{2}\right)\log\left(\frac{RSS_{\gamma_2}}{RSS_{\gamma_1}}\right) + (p_{\gamma_2}-p_{\gamma_1})\log(h(n)) + K \quad \text{where } K \text{ is a constant.}$$

One can choose functions in the class $\mathcal{H}$, and tune the constant K to obtain certain desirable properties such as bounded false positive rates. One can then preform step-wise or stochastic searches, optimizing the SRIC, instead of a fully Bayesian approach. While model selection using the SRIC will not be free of superinduction, the effects for moderate to large n will be significantly minimized. Indeed, Theorem 4 indicates that this problem afflicts even oracle estimators, so different choices of priors will only have an effect on the rate of convergence

of the Bayes factor, not on the convergence itself. Future work will examine the rates of convergence of the spike and slab, the non-local priors, and other popular objective prior choices.

We note that, ultimately, as long as a model has posterior consistency, it will converge to the model optimal under the KL-divergence, which inevitably leads to the super-induction problem. Thus, one area for future investigation would be to consider different loss functions that are not asymptotically equivalent to the KL-divergence. Certain predictive scoring methods may show promise here. One example of this idea is presented in Tallman and West (2022), where the posterior model weights are replaced with

$$\widetilde{\pi}(M_j) = C^* \pi(M_j) \alpha_j(y^*, \hat{\mu}_j | y),$$

where C^* is a normalizing constant and $\alpha_j(y^*, \hat{\mu}_j|y)$ is a constrained utility or scoring function. While Tallman and West (2022) discuss this approach in the context of Bayesian model averaging, it should be noted that the problem of superinduction is essentially a problem with the growth rate of the posterior odds ratio. This method would rescale that growth rate, potentially alleviating the aforementioned issues.

APPENDIX

LEMMA 2. $\mathbb{E}\left((1+g)^{-1}\right) \to 0$ *if, and only if,* $\mathbb{E}\left((1+g)^{-p/2}\right) \to 0$.

PROOF. This lemma is a consequence of Jensen's inequality (both the Concave and Convex version) and Monotonicity. □

LEMMA 3. *For* $a^*, b, z > 0$, *and* $b < 1$*:*

$$e^z \int_{u=0}^{1} (1-u)^{b-1} u^{a^*-1} e^{-zu} \mathrm{d}u \leq e^z \int_{u=0}^{1} u^{a^*-1} e^{-zu} \mathrm{d}u + \frac{\mathbf{B}(b, a^*+1)}{\mathbf{B}(1, a^*+1)} e^z \int_{u=0}^{1} u^{a^*} e^{-zu} \mathrm{d}u.$$

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

PROOF OF PROPOSITION 1.1. We let $\pi(\sigma^2) \propto 1/\sigma^2$ and $u = 1/(g+1) \sim Be(a, b)$. For convenience, we also let $z_{\sigma^2} = (RSS_\emptyset - RSS_\gamma)/2\sigma^2$, $C_X = [\det(X^T X)(2\pi)^{(n-q)}]^{-1/2}$, and $Q_Y = RSS_\gamma/(RSS_\emptyset - RSS_\gamma)$. We have that

$$f(y) = \int_{\sigma^2} \frac{C_X}{\mathbf{B}(a,b)} \left(\frac{1}{\sigma^2}\right)^{\frac{n-q}{2}-1} \exp\left\{-\frac{RSS_\gamma}{2\sigma^2}\right\} \int_u (1-u)^{b-1} (u)^{\frac{p}{2}+a-1} \exp\{-z_{\sigma^2} u\} \mathrm{d}u \mathrm{d}\sigma^2$$

$$\tilde{f}(y) = C_X \frac{\Gamma(a+\frac{p}{2})\Gamma(\frac{n-p-q}{2}-a)}{\mathbf{B}(a,b)} (Q_Y)^{\frac{p}{2}+a} \left[\frac{2}{RSS_\gamma}\right]^{\frac{n-q}{2}}.$$

Here $f(y)$ is the true marginal density, and $\tilde{f}(y)$ is an approximating function. We consider two cases. First, let $b \geq 1$. Then, since $(1-u)^{b-1} \leq 1$, we have that

$$\begin{aligned} f(y) &\leq \int_{\sigma^2} \frac{C_X}{\mathbf{B}(a,b)} \left(\frac{1}{\sigma^2}\right)^{\frac{n-q}{2}-1} \exp\left\{-\frac{RSS_\gamma}{2\sigma^2}\right\} \Gamma(a+\frac{p}{2}) \left[\frac{1}{z_{\sigma^2}}\right]^{\frac{p}{2}+a} \\ &= C_X \frac{\Gamma(a+\frac{p}{2})\Gamma(\frac{n-p-q}{2}-a)}{\mathbf{B}(a,b)} (Q_Y)^{\frac{p}{2}+a} \left[\frac{2}{RSS_\gamma}\right]^{\frac{n-q}{2}} = \tilde{f}(y). \end{aligned}$$

For a lower bound on $f(y)$, we note that Bernoulli's inequality gives us $(1-u)^{b-1} \geq (1-u)^b \geq 1-bu$. We further note that for the upper incomplete gamma function, Theorem 2.4 from Borwein and Chan (2009) gives us :

$$\Gamma(\alpha, z) \leq \begin{cases} 2z^{\alpha-1}e^{-z} & , \alpha \leq 1 \\ 2z^{\alpha-1}e^{-z}\sum_{k=0}^{K-1}\left(\frac{\alpha-1}{z}\right)^k \quad \text{where } K = \lceil \alpha \rceil & , \alpha > 1 \end{cases}.$$

Then (here we will abbreviate z_{σ^2} as z) we have that:

$$e^z \int_{u=0}^{1} (1-bu)u^{a+\frac{p}{2}-1}e^{-zu}\mathrm{d}u \geq e^z z^{-(\frac{p}{2}+a)}\left[\Gamma(a+\frac{p}{2}) - \frac{b}{z}\Gamma(a+\frac{p}{2}+1) - \Gamma(a+\frac{p}{2}, z)\right]$$

$$\geq \begin{cases} e^z z^{-(\frac{p}{2}+a)}\Gamma(a+\frac{p}{2})\left[1+\frac{b}{z}\left(\frac{p}{2}+a\right)\right] + \frac{2}{z} & a+\frac{p}{2} \leq 1 \\ e^z z^{-(\frac{p}{2}+a)}\Gamma(a+\frac{p}{2})\left[1+\frac{b}{z}\left(\frac{p}{2}+a\right)\right] + \frac{1}{2a+p-2}\sum_{k=0}^{\lceil a+\frac{p}{2}\rceil-1}\left(\frac{a+\frac{p}{2}-1}{z}\right)^{k+1} & a+\frac{p}{2} \geq 1 \end{cases}$$

Then, after integrating out σ^2, we have that:

$$f(y) \geq \tilde{f}(y)\left[1 - \frac{2b(2a+p)}{n-p-q-2a-2}Q_Y - 2\left(\frac{Q_Y^{-(a+\frac{p}{2}-1)}\Gamma(\frac{n-q}{2}-1)}{\Gamma(\frac{n-p-q}{2}-a)\Gamma(a+\frac{p}{2})}\right)\left(\frac{RSS_\gamma}{RSS_\emptyset}\right)^{\frac{n-q}{2}}\right.$$
$$\left. \times \begin{cases} 1 & a+\frac{p}{2} < 1 \\ \sum_{k=0}^{\lceil a+\frac{p}{2}\rceil-1}(a+\frac{p}{2}-1)^k Q_Y^k\left[\frac{\Gamma(\frac{n-q}{2}-k-1)}{\Gamma(\frac{n-q}{2}-1)}\right] & a+\frac{p}{2} \geq 1 \end{cases}\right].$$

We note that $1/Q_Y$ is bounded as n grows and tends towards a constant with probability 1. Thus, assuming a and b are bounded in n, the second term is $\mathcal{O}(1/n)$. For the third term, we note that the ratios of gamma functions are bounded polynomially in n, while $(RSS_\gamma/RSS_\emptyset)^{(n-q)/2}$ tends to zero exponentially fast with probability 1. Thus, we have that $f(y) \geq \tilde{f}(y)[1-\mathcal{O}(1/n)]$.

For the case where $b \leq 1$, we obtain a similar lower bound by noting that $(1-u)^{b-1} \geq 1$. The second $\mathcal{O}(1/n)$ term is dropped, but the exponentially fast third term is still present. Thus, $f(y) \geq \tilde{f}(y)[1-\mathcal{O}(c_y^{n/2})]$ with probability 1, where $c_y \leq 1$. For the upper bound, we apply Lemma 3, and integrate out σ^2. This gives us:

$$f(y) \leq \tilde{f}(y)\left[1 + \alpha\frac{\mathbf{B}(b,\alpha+1)}{\mathbf{B}(1,\alpha+1)}\left(\frac{1}{Q_Y}\right)\left[\frac{\Gamma(\frac{n-p-q}{2}-a-1)}{\Gamma(\frac{n-p-q}{2}-a)}\right]\right] = \tilde{f}(y)\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right].$$

Thus, with probability 1, $\tilde{f}(y)/f(y) = 1+\mathcal{O}(1/n)$. Using the fact that $f_\emptyset(y) = C_X\Gamma((n-q)/2)(RSS_\emptyset/2)^{-(n-q)/2}\det(X^TX)^{-1/2}$ we obtained the desired approximation.

□

PROOF OF PROPOSITION 1.2. The result is obtained by noting that: $\mathbb{E}((g+1)^{-1}|y) = BF_{\gamma:\emptyset,p+2}/BF_{\gamma:\emptyset,p}$. □

LEMMA 4. *$S_{\gamma,n} \xrightarrow{P} S^*_\gamma \in [1,\infty)$ for all models M_γ, where S^*_γ is a constant*

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

LEMMA 5. *$\frac{n}{g_n+1} \xrightarrow{P} W$, where W is a random variable with $\mathbb{E}(W) \in (0,\infty)$.*

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

LEMMA 6. $\frac{n}{g_n+S_{\gamma,n}} \xrightarrow{P} W$

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

LEMMA 7.

$$\left(\frac{g_n+S_{1,n}}{g_n+S_{2,n}}\right)^{\frac{n}{2}} \xrightarrow{P} \exp\left\{-\frac{(S_2^*-S_1^*)}{2}W\right\}$$

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

LEMMA 8. *If* $S_1^* < S_2^*$, *then* $(S_{1,n}/S_{2,n})^{n/2} \xrightarrow{P} 0$. *Equivalently, if* $S_1^* > S_2^*$, *then* $(S_{1,n}/S_{2,n})^{n/2} \xrightarrow{P} \infty$.

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

LEMMA 9. *We have that the following sequences of random variables are uniformly integrable.*

1. $\left(\frac{n}{g_n+1}\right)^{\frac{1}{2}}$
2. $\left(\frac{g_n+S_{1,n}}{g_n+S_{2,n}}\right)^{\frac{n}{2}}$, *for* $S_1^* \leq S_2^*$
3. $\left(\frac{n}{g_n+1}\right)^{\frac{1}{2}}\left(\frac{g_n+S_{1,n}}{g_n+S_{2,n}}\right)^{\frac{n}{2}}$, *for* $S_1^* \leq S_2^*$

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

LEMMA 10. *We have that the following sequences of random variables converge in mean, and that the means are non-zero and finite.*

1. $(n/(g_n+1))^{1/2} \xrightarrow{\mathcal{L}_1} W^{1/2}$ *and* $\mathbb{E}(W^{1/2}) \in (0,\infty)$
2. $\left((g_n+S_{1,n})/(g_n+S_{2,n})\right)^{n/2} \xrightarrow{\mathcal{L}_1} \exp\{-(S_2^*-S_1^*)W/2\}$
   and $\mathbb{E}\left(\exp\{-(S_2^*-S_1^*)W/2\}\right) \in (0,\infty)$
3. $(n/(g_n+1))^{1/2}\left((g_n+S_{1,n})/(g_n+S_{2,n})\right)^{n/2} \xrightarrow{\mathcal{L}_1} W^{1/2}\exp\{-(S_2^*-S_1^*)W/2\}$
   and $\mathbb{E}\left(W^{1/2}\exp\{-(S_2^*-S_1^*)W/2\}\right) \in (0,\infty)$

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

LEMMA 11. *Under the assumptions above, and with* $p_1 \in \{p_2-1, p_2, p_2+1\}$, *we have that:*

$$\frac{\int_g (1+g)^{\frac{-p_1}{2}}\left(\frac{g+1}{g+S_{1,n}}\right)^{\frac{n-q}{2}}\pi_n(g|M_\gamma)\mathrm{d}g}{\int_g (1+g)^{\frac{-p_2}{2}}\left(\frac{g+1}{g+S_{2,n}}\right)^{\frac{n-q}{2}}\pi_n(g|M_{\gamma'})\mathrm{d}g} \in [B_L n^{-\frac{1}{2}}, B_U n^{\frac{1}{2}}] \quad \text{as } n\to\infty.$$

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

COROLLARY 3. *Under the assumptions above, and with arbitrary positive* p_1 *and* p_2*, we have that:*

$$\frac{\int_g (1+g)^{\frac{-p_1}{2}} \left(\frac{g+1}{g+S_{1,n}}\right)^{\frac{n-q}{2}} \pi_n(g|M_\gamma)\mathrm{d}g}{\int_g (1+g)^{\frac{-p_2}{2}} \left(\frac{g+1}{g+S_{2,n}}\right)^{\frac{n-q}{2}} \pi_n(g|M_{\gamma'})\mathrm{d}g} \in [B_L n^{-\frac{|p_1-p_2|}{2}}, B_U n^{\frac{|p_1-p_2|}{2}}] \quad \textit{as } n \to \infty.$$

PROOF. See the supplementary material (Fontana and Sansó, 2023). □

Acknowledgments. The research was supported in part by the National Science Foundation under awards MMS 2050012, DMS 2153277. The analysis of the albedo data presented in this paper is part of a collaborative effort with Dr Jessica Matthews of NOAA's National Centers for Environmental Information.

SUPPLEMENTARY MATERIAL

Predictive Plots and Proofs of Lemmas

This supplement contains the predictive plots for the MSSS data example in section 6.2, and the proofs of Lemmas 3, 4, 5, 6, 7, 8, 9, 10, 11, and Corollary 3

REFERENCES

BAYARRI, M. J., BERGER, J. O., FORTE, A. and GARCÍA-DONATO, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics* **40** 1550-1577. https://doi.org/10.1214/12-AOS1013

BERK, R. H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *The Annals of Mathematical Statistics* **37** 51-58. https://doi.org/10.1214/aoms/1177699597

BERNARDO, J. M. and SMITH, A. F. (2009). *Bayesian theory* **405**. John Wiley & Sons.

BORWEIN, J. and CHAN, O.-Y. (2009). Uniform bounds for the incomplete complementary gamma function. *Mathematical Inequalities and Applications* **12** 115–121.

BOTTOLO, L. and RICHARDSON, S. (2008). Fully Bayesian variable selection using g-priors Technical Report, Working paper, Imperial College.

CARVALHO, C. M., POLSON, N. G. and SCOTT, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480.

CONSONNI, G., FOUSKAKIS, D., LISEO, B. and NTZOUFRAS, I. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis* **13** 627 – 679. https://doi.org/10.1214/18-BA1103

CUI, W. and GEORGE, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference* **138** 888-900. https://doi.org/10.1016/j.jspi.2007.02.011

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

FONTANA, J. and SANSÓ, B. (2023). Supplement to "Scalable Model Selection with Mixtures of g-Priors in Large Data Settings".

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

HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association* **102** 507-516. https://doi.org/10.1198/016214507000000121

HELD, L., BOVÉ, D. S. and GRAVESTOCK, I. (2015). Approximate Bayesian model selection with the deviance statistic. *Statistical Science* **30** 242 – 257. https://doi.org/10.1214/14-STS510

KASS, R. E. and WASSERMAN, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association* **90** 928–934.

KIRSNER, D. (2020). Nonstationary Models for Large Spatial Datasets Using Multi-resolution Process Convolutions, PhD thesis, University of California, Santa Cruz.

KIRSNER, D. and SANSÓ, B. (2020). Multi-scale shotgun stochastic search for large spatial datasets. *Computational Statistics & Data Analysis* **146** 106931.

LEY, E. and STEEL, M. F. J. (2012). Mixtures of g-priors for Bayesian model averaging with economic applications. *Journal of Econometrics* **171** 251-266. Bayesian Models, Methods and Applications. https://doi.org/10.1016/j.jeconom.2012.06.009

LI, Y. and CLYDE, M. A. (2018). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association* **113** 1828-1845. https://doi.org/10.1080/01621459.2018.1469992

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

MARUYAMA, Y. and GEORGE, E. I. (2011). Fully Bayes factors with a generalized g-prior. *The Annals of Statistics* **39** 2740–2765.

MUKHOPADHYAY, M., SAMANTA, T. and CHAKRABARTI, A. (2015). On consistency and optimality of Bayesian variable selection based on g-prior in normal linear regression models. *Annals of the Institute of Statistical Mathematics* **67** 963–997.

MUKHOPADHYAY, M. and SAMANTA, T. (2017). A mixture of g-priors for variable selection when the number of regressors grows with the sample size. *Test* **26** 377–404.

PINTY, B., LATTANZIO, A., MARTONCHIK, J. V., VERSTRAETE, M. M., GOBRON, N., TABERNER, M., WIDLOWSKI, J.-L., DICKINSON, R. E. and GOVAERTS, Y. (2005). Coupling diffuse sky radiation and surface albedo. *Journal of the Atmospheric Sciences* **62** 2580–2591.

SCOTT, J. G. and BERGER, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics* **38** 2587 – 2619. https://doi.org/10.1214/10-AOS792

SHAO, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7** 221–242.

TADESSE, M. G. and VANNUCCI, M. (2021). *Handbook of Bayesian variable selection*. CRC Press.

TALLMAN, E. and WEST, M. (2022). Bayesian Predictive Decision Synthesis. *arXiv preprint arXiv:2206.03815*.

VEHTARI, A. and OJANEN, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys* **6** 142–228.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.) 15, 233-243. Elsevier Science Publishers, Amsterdam: North-Holland.

ZELLNER, A. and SIOW, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadística e Investigación Operativa* **31** 585-603.

ZEUGNER, S. and FELDKIRCHER, M. (2009). *Benchmark priors revisited: on adaptive shrinkage and the supermodel effect in Bayesian model averaging*. International Monetary Fund.