Bayesian Analysis (2022) **17**, Number 2, pp. 565–591

Concentration of Posterior Model Probabilities and Normalized L_0 Criteria*

David Rossell†

Abstract. We study frequentist properties of Bayesian and L_0 model selection, with a focus on (potentially non-linear) high-dimensional regression. We propose a construction to study how posterior probabilities and normalized L_0 criteria concentrate on the (Kullback-Leibler) optimal model and other subsets of the model space. When such concentration occurs, one also bounds the frequentist probabilities of selecting the correct model, type I and type II errors. These results hold generally, and help validate the use of posterior probabilities and L_0 criteria to control frequentist error probabilities associated to model selection and hypothesis tests. Regarding regression, we help understand the effect of the sparsity imposed by the prior or the L_0 penalty, and of problem characteristics such as the sample size, signal-to-noise, dimension and true sparsity. A particular finding is that one may use less sparse formulations than would be asymptotically optimal, but still attain consistency and often also significantly better finite-sample performance. We also prove new results related to misspecifying the mean or covariance structures, and give tighter rates for certain non-local priors than currently available.

Keywords: model selection, Bayes factors, high-dimensional inference, consistency, uncertainty quantification, L_0 penalty, model misspecification.

Selecting a probability model and quantifying the associated uncertainty are two fundamental tasks in Statistics. In Bayesian model selection (BMS), given models and priors one obtains posterior model probabilities that guide model choice and measure the (Bayesian) certainty on that choice. It is interesting to understand how such posterior probabilities relate to the frequentist probability of selecting the optimal model (defined below). L_0 penalties are also powerful selection criteria, but it is less clear how to portray uncertainty. Suppose one selects the model optimizing the Bayesian information criterion (BIC, Schwarz, 1978), how is one to measure the certainty about that choice? Given the connection between the BIC and Bayes factors, it is tempting to define a pseudo-posterior probability via a normalized L_0 criterion (defined below). Again the question is how do pseudo-probabilities relate to frequentist selection probabilities.

Our goals are two-fold. First, we present a general framework to study the L_1 convergence of posterior model probabilities and normalized L_0 criteria, and show that the rates bound the frequentist probabilities of choosing the wrong model and making type I–II errors. There is previous work studying L_1 convergence (see below), our specific construction however is novel (to our knowledge) and reduces the problem to integrating

*DR was partially funded by the Europa Excelencia grant EUR2020-112096, NIH grant R01 CA158113-01, Ramón y Cajal Fellowship RYC-2015-18544, Plan Estatal PGC2018-101643-B-I00 and Ayudas Fundación BBVA a equipos de investigación científica en Big Data 2017.

†Universitat Pompeu Fabra, Department of Business and Economics, Barcelona (Spain), rosselldavid@gmail.com

 https://doi.org/10.1214/21-BA1262

certain Bayes factor tail probabilities. The result on bounding frequentist error probabilities is also new (although elementary), and validates using posterior probabilities and normalized L_0 criteria to quantify model choice uncertainty from a frequentist standpoint. Our second goal is to apply our framework to Gaussian regression to synthesize and extend current theoretical results. We show that posterior model probabilities in high dimensions depend on the same three elements that drive their behavior in finite dimensions. These are the sparsity of the prior on the models, the dispersion of the prior on the parameters, and whether the latter is a local or a non-local prior (Johnson and Rossell, 2010). We impose mild conditions on these prior elements so that, relative to current consistency results for a specific prior sparsity regime, we portray the impact of enforcing sparsity. As novel aspects, we consider model misspecification within (possibly non-linear) regression and we obtain tighter rates for the non-local product MOM prior (pMOM, Johnson and Rossell, 2012) than currently available. A practical implication of our results is that, by using less sparse priors than those leading to optimal asymptotic rates, one can still get consistency and sometimes significantly better finite n performance. Some of our examples may be striking in that regard.

The introduction is organized as follows. First, we lay out notation needed to define the problem. We then review results for fixed and high dimensions, and finally outline the paper. Let $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ be an observed outcome and n the sample size. One wishes to consider a set of K candidate models $M_1, \ldots, M_K$ for $\mathbf{y}$. Each model is defined by a density $p(\mathbf{y} \mid \boldsymbol{\theta}_k, \phi, M_k)$ for $k = 1, \ldots, K$, where $\boldsymbol{\theta}_k \in \Theta_k$ is a parameter of interest and $\phi \in \Phi$ a (potential) nuisance parameter. The model dimension is given by $p_k = \dim(\Theta_k)$ and $d = \dim(\Phi)$. Densities are in the Radon-Nikodym sense, in particular allowing discrete and continuous $\mathbf{y}$. Without loss of generality let $\Theta_k \subseteq \Theta \subseteq \mathbb{R}^p$ for $k = 1, \ldots, K$, i.e. models are nested within a larger model of dimension $p + d$. Although not denoted explicitly in high-dimensional problems both the number of parameters p and models K may grow with n, and this is precisely our main focus. In BMS each model is equipped with a prior density $p(\boldsymbol{\theta}_k, \phi \mid M_k)$, and one obtains posterior probabilities

$$p(M_k \mid \mathbf{y}) = \left(1 + \sum_{l \neq k} \frac{p(\mathbf{y} \mid M_l)}{p(\mathbf{y} \mid M_k)} \frac{p(M_l)}{p(M_k)}\right)^{-1} = \left(1 + \sum_{l \neq k} B_{lk} \frac{p(M_l)}{p(M_k)}\right)^{-1}, \qquad (1)$$

where $p(\mathbf{y} \mid M_k) = \int p(\mathbf{y} \mid \boldsymbol{\theta}_k, \phi, M_k) dP(\boldsymbol{\theta}_k, \phi \mid M_k)$ is the integrated likelihood under model M_k, $p(M_k)$ its prior probability and $B_{lk} = p(\mathbf{y} \mid M_l)/p(\mathbf{y} \mid M_k)$ the Bayes factor between (M_l, M_k). We focus our discussion on BMS, but one can obtain analogous expressions for normalized L_0 criteria, see Section 4.

To fix ideas, consider a Gaussian regression where $\mathbf{y} \in \mathbb{R}^n$ and the data analyst assumes the model $p(\mathbf{y} \mid \boldsymbol{\theta}, \phi) = N(\mathbf{y}; X\boldsymbol{\theta}, \phi I)$ where X is an $n \times p$ matrix, $\boldsymbol{\theta} \in \mathbb{R}^p$ the regression coefficients, and $\phi > 0$ the error variance. Models M_k are defined by selecting subsets of columns in X, i.e. $p(\mathbf{y} \mid \boldsymbol{\theta}_k, \phi) = N(\mathbf{y}; X_k\boldsymbol{\theta}_k, \phi I)$ where X_k is the $n \times p_k$ matrix containing the columns selected by M_k, and $\boldsymbol{\theta}_k \in \mathbb{R}^{p_k}$. Note that X may contain non-linear effects and interactions such as wavelets, splines, or tensor-products. Here one might set a conjugate Normal-inverse Gamma prior $p(\boldsymbol{\theta}_k, \phi \mid M_k) = N(\boldsymbol{\theta}_k; \mathbf{0}, \tau\phi I)\text{IG}(\phi; a_\phi/2, b_\phi/2)$, for example, where (τ, a_ϕ, b_ϕ) are prior parameters.

We consider the following question. Suppose that $\mathbf{y}$ arises from some data-generating density f^*, which may be outside the considered models (model misspecification). Let M_t be the optimal model in that it is the smallest model minimizing Kullback-Leibler (KL) divergence to f^* (see Section 1). For example, in Gaussian regression M_t is the smallest model minimizing mean squared prediction error under f^*. If the models are well-specified, then M_t is simply the smallest model containing f^*, and is often referred to as *true model*. Ideally one wants to assign large $p(M_t \mid \mathbf{y})$, so that one not only selects the optimal model but is also confident about that choice.

Our goal is to study if $p(M_t \mid \mathbf{y})$ converges to 1 as n grows, and at what rate. This problem has been well-studied in finite dimensions where (p, K) do not grow with n. Consider a model M_k that includes the optimal M_t ($\Theta_t \subset \Theta_k$), i.e. M_k contains spurious parameters. We refer to such M_k as a *spurious model*. For fairly general models and priors, the Bayes factor B_{kt} converges in probability to 0 at a polynomial rate in n and in the prior dispersion (τ in our regression example). See Theorem 1 in Dawid (1999) and the proof of our Theorem 1 for models with concave log-likelihood. This polynomial rate holds when $p(\boldsymbol{\theta}_k \mid \phi, M_k)$ is a local prior, for non-local priors the rates are faster (Johnson and Rossell, 2010). In contrast, if M_k is a non-spurious model ($\Theta_t \notin \Theta_k$, i.e. missing parameters from M_t), then B_{kt} vanishes exponentially in n (more precisely, in a non-centrality parameter that is proportional to n, under suitable assumptions). In summary, to help discard spurious models one may either set large τ (i.e. a diffuse prior on parameters), set sparse model priors $p(M_k)$ that penalize model size, and/or set a non-local prior. A caveat with diffuse and sparse priors is that they penalize complexity purely a priori, which can lead to a drop in statistical power.

Extensions of such precise rates to high dimensions are of fundamental interest yet hard to come by. Most results focus on a prior satisfying relatively rigid sparsity conditions and either only study consistency (with no rates) or focus on asymptotic optimality. We show that in high-dimensional regression the finite-dimensional rates above still hold, up to lower-order terms, for the stronger L_1 convergence. The prior features driving consistency are still the use of diffuse, sparse and non-local priors.

We review selected high-dimensional literature. Johnson and Rossell (2012) proved that $p(M_t \mid \mathbf{y})$ converges to 1 in linear regression with $p \ll n$ under NLPs and uniform $p(M_k)$. Narisetty and He (2014) showed that if $p \ll e^n$ then certain diffuse priors $p(\boldsymbol{\theta}_k \mid M_k)$ also attain consistency. In fact, the RIC of Foster and George (1994) is a related early advocate for diffuse priors, and can also be shown to attain consistency for $p \ll e^n$ (Section 4). Shin et al. (2018) extended Johnson and Rossell (2012) to $p \ll e^n$ under certain diffuse NLPs. Yang and Pati (2017) also used diffuse priors (defined implicitly via a prior anti-concentration condition), in a more general framework that allows for non-parametric models. These results proved consistency but no specific rates were given. Castillo et al. (2015) showed that, by using so-called *Complexity priors* $p(M_k)$ and Laplace priors on parameters, one can consistently select the data-generating model in regression. Chae et al. (2016) proved that the same prior structure attains consistency in regression with non-parametric symmetric errors. Gao et al. (2015) extended these results to general structured linear models under misspecified sub-Gaussian errors, and Rockova and van der Pas (2017) to regression trees. Yang et al. (2016) studied a regression setting where one uses diffuse priors on parameters and a type of Complexity prior

on models. These are significant insights, the focus however is showing that complex models are discarded a posteriori under a given, sufficiently sparse, prior.

In summary, the diffuse priors as in Narisetty and He (2014) and Complexity priors as in Castillo et al. (2015) underlie much of the state-of-the-art literature. These priors excel at discarding spurious models, and do not require one to restrict the maximum model complexity. Our results suggest that they should be used with care, however, and that there can be advantages to setting less sparse priors by placing mild restrictions on the model complexity. As an illustration, we preview Figures 1–2 where three prior formulations were used. Although the Complexity prior attains better asymptotic rates, the combined pMOM and Beta-Binomial priors attained better finite n power/sparsity tradeoffs. Although in this example M_t has small dimension $p_t = 5, 10, 20$ (sparse truth), the losses in power due to setting sparse priors are substantial. It is therefore of interest to study consistency allowing for less sparse priors.

We focus on a fully Bayesian framework where no priors are data-dependent. Extending our framework to empirical Bayes approaches where prior features are learned from data is interesting but requires a delicate treatment beyond our scope to avoid certain posterior degeneracy issues, we refer the reader to Petrone et al. (2014).

The paper is organized as follows. Section 1 sets notation, presents our general framework, and shows that the expectation of posterior probabilities such as $E_{f^*}(p(M_t \mid \mathbf{y}))$ bound relevant frequentist error probabilities. Section 2 discusses the priors that we focus attention on, and important technical conditions related to the model complexity and sparsity embedded in the prior. It also outlines necessary conditions that fairly general priors need to satisfy, if one wishes to attain consistency. Section 3 characterizes the posterior probability of individual models in Gaussian regression, under essentially any prior on the models and the priors on parameters from Section 2. Specifically, we consider Zellner priors (with known and unknown error variance) and more general Normal priors, for which Bayes factors have tractable expressions and hence simplify our exposition. We also include the pMOM prior where such an expression is unavailable, and misspecified (possibly non-linear) models where tail probabilities are harder to bound. We show that failing to include true non-linearities or omitting relevant variables causes an exponential drop in power, whereas misspecifying the error covariance (truly correlated and/or heteroskedastic errors) need not do so but may inflate false positives. Section 4 extends Section 3 to normalized L_0 penalties, including the BIC, EBIC and RIC. Section 5 obtains global rates for $p(M_t \mid \mathbf{y})$ and other interesting model subsets. The results show that it is often possible to discard spurious parameters, even when not using particularly sparse priors in problems of fairly large dimension, e.g. by combining a Beta-Binomial prior on models with non-local priors on parameters. Section 6 offers examples, and Section 7 concludes. A significant number of auxiliary lemmas, technical results and all proofs are in the supplementary material (Rossell, 2021).

1 Approach

We first formalize the notion of optimal model M_t and introduce notation used throughout the paper in Section 1.1. Then Section 1.2 presents a framework to study L_1 con-

vergence of $p(M_t \mid \mathbf{y})$ in fully general settings, discusses its tightness, and shows that the associated rates bound relevant frequentist error probabilities.

1.1 Definitions and notation

We define the optimal model M_t as having smallest dimension p_t among models minimizing Kullback-Leibler (KL) divergence to f^*. For simplicity we assume M_t to be unique, but otherwise one may define M_t to be the union of smallest optimal models.

Definition 1. *Let* $(\boldsymbol{\theta}^*, \phi^*) = \arg\min_{\theta\in\Theta,\phi\in\Phi} KL(f^*, p(\mathbf{y} \mid \boldsymbol{\theta}, \phi))$. *Define* $t = \arg\min_{k\in\mathcal{M}^*} p_k$, *where*

$$\mathcal{M}^* = \{k : \exists(\boldsymbol{\theta}_k^*, \phi_k^*) \in \Theta_k \times \Phi : KL(f^*, p(\mathbf{y} \mid \boldsymbol{\theta}_k^*, \phi_k^*)) = KL(f^*, p(\mathbf{y} \mid \boldsymbol{\theta}^*, \phi^*))\}$$

is the set of all models minimizing KL-divergence to f^*.

We denote by $(\boldsymbol{\theta}^*, \phi^*)$ the global optimal parameter value minimizing KL-divergence to f^*, and by $(\boldsymbol{\theta}_k^*, \phi_k^*)$ that under a model M_k. If f^* lies in the assumed model family (well-specified case), then $(\boldsymbol{\theta}^*, \phi^*)$ is the true parameter value.

We shall study the posterior probability assigned to models other than M_t. To that end, denote the set of l-dimensional models that contain M_t plus some spurious parameters by $S_l = \{k : \Theta_t \subset \Theta_k, p_k = l\}$. We refer to S_l as *spurious models* of dimension l. Similarly, let $S_l^c = \{k : \Theta_t \not\subset \Theta_k, p_k = l\}$ be the size l *non-spurious models*, and let $S = \bigcup_{l=p_t+1}^{\bar{p}} S_l$ and $S^c = \bigcup_{l=0}^{\bar{p}} S_l^c$ the complete set of spurious and non-spurious models. Denote by $|S|$ the cardinality of S.

In our study it is often convenient to express certain conditions and results in terms of their asymptotic order as n grows. To this end, $a_n \ll b_n$ denotes $\lim_{n\to\infty} a_n/b_n = 0$ for two deterministic sequences $a_n, b_n > 0$, and similarly $a_n \preceq b_n$ denotes $\lim_{n\to\infty} a_n/b_n \leq c$ for some constant $c > 0$. Finally, $a_n \asymp b_n$ denotes that both $a_n \preceq b_n$ and $a_n \succeq b_n$.

As we discuss later, although p could potentially grow exponentially with n, for certain prior/L_0 penalty settings to achieve consistency it may be necessary to impose restrictions on the model complexity. We assume that the analyst specifies a maximum model size that we denote by $\bar{p} = \max_k p_k$, and describe rates as a function of $\bar{p}$ For instance, in regression one may have $p \gg n$ but restrict attention to models selecting at most $\bar{p} = \min\{n, p\}$ out of the p variables, as choosing a model with $p_k > n$ parameters may not be desirable. The number of models is then $K = \sum_{j=0}^{\bar{p}} \binom{p}{j}$, which is still $\gg n$.

1.2 L_1 convergence

From (1), posterior consistency requires $\sum_{k\neq t} B_{kt} p(M_k)/p(M_t) \xrightarrow{P} 0$. The difficulty in high dimensions is that the number of models $K-1$ grows with n, hence the sum can only vanish if each term $B_{kt}p(M_k)/p(M_t)$ converges to 0 quickly enough. This intuition is clear, but obtaining probabilistic bounds for this stochastic sum is non-trivial, since the B_{kt}'s may exhibit complex dependencies. To avoid dealing with such high-dimensional

stochastic sums, it is simpler to study deterministic expectations. Specifically, we study when $p(M_t \mid \mathbf{y}) \xrightarrow{L_1} 1$, which by definition of L_1 convergence is equivalent to

$$\lim_{n\to\infty} \sum_{k\neq t} E_{f^*}\left(p(M_k \mid \mathbf{y})\right) = 0, \tag{2}$$

where $E_{f^*}(\cdot)$ is the expectation under f^*. Some remarks are in order. First, L_1 convergence in (2) implies convergence in probability. Let $b_n > 0$ be a sequence such that $\lim_{n\to\infty} b_n = 0$, if $E_{f^*}(1 - p(M_t \mid \mathbf{y})) \preceq b_n$ then $1 - p(M_t \mid \mathbf{y}) = O_p(b_n)$. Naturally (2) may require more stringent conditions than convergence in probability, but in regression we obtain essentially tight rates and the gains in clarity are substantial. Second, one can evaluate the sum on the left-hand side of (2) for fixed n, p and K, i.e. the expression can be used in non-asymptotic regimes.

L_1 convergence guarantees bounding relevant frequentist probabilities, supporting the use of posterior probabilities (or normalized L_0 criteria) to quantify model uncertainty. By Proposition 1, $E_{f^*}(p(M_t \mid \mathbf{y}))$ bounds the frequentist selection probability $P_{f^*}(\hat{k} \neq t)$, where $\hat{k}$ is the highest posterior probability model. Proposition 1 also holds when $\hat{k}$ is the median probability model (Barbieri and Berger, 2004) selecting parameters with marginal posterior inclusion probability $P(\theta_j \neq 0 \mid \mathbf{y}) > 0.5$ (see the proof).

Proposition 1. *Let* $\hat{k} = \arg\max_k p(M_k \mid \mathbf{y})$ *be the posterior mode, then*

$$P_{f^*}(\hat{k} \neq t) \leq 2E_{f^*}(1 - p(M_t \mid \mathbf{y})) = 2\sum_{k\neq t} E_{f^*}(p(M_k \mid \mathbf{y})).$$

Corollaries 1–2 relate type I–II error probabilities to expected posterior model probabilities. Corollary 1 uses the trivial observation that family-wise type I–II error rates are both $\leq P_{f^*}(\hat{k} \neq t)$, and hence bounded by Proposition 1. Suppose instead that $\hat{k}$ is obtained by selecting parameters with $P(\theta_j \neq 0 \mid \mathbf{y}) > t$, for some threshold t. For instance, to control the Bayesian False Discovery rate below a level α one sets a certain $t \leq 1 - \alpha$ (Müller et al., 2004). Then, by Corollary 2 the type I–II errors for individual coefficients are bounded by $E_{f^*}(P(\theta_j \neq 0 \mid \mathbf{y}))$, times a factor depending on t.

Corollary 1. *Let* $S(\hat{k})$ *the set of non-zero parameters in model* $\hat{k} = \arg\max_k p(M_k \mid \mathbf{y})$.

- *The family-wise type I error is* $P_{f^*}\left(\bigcup_{j:\theta_j^*=0}\{j \in S(\hat{k})\}\right) \leq P_{f^*}(\hat{k} \neq t)$.
- *The family-wise type II error is* $P_{f^*}\left(\bigcup_{j:\theta_j^*=1}\{j \notin S(\hat{k})\}\right) \leq P_{f^*}(\hat{k} \neq t)$.

Corollary 2. *Let* $S(\hat{k}) = \{j : P(\theta_j \neq 0 \mid \mathbf{y}) > t\}$ *for a given threshold* t.

- *False positives. Assume that* $\theta_j^* = 0$. *Then* $P_{f^*}(j \in S(\hat{k})) \leq \frac{1}{t} E_{f^*}(P(\theta_j \neq 0 \mid \mathbf{y}))$.
- *Power. Assume that* $\theta_j^* \neq 0$. *Then* $P_{f^*}(j \notin S(\hat{k})) \leq \frac{1}{1-t} E_{f^*}(P(\theta_j = 0 \mid \mathbf{y}))$.

To summarize, if one can bound sums of expectations $E_{f^*}(p(M_k \mid \mathbf{y}))$ across models one can then prove that the posterior probability of M_t converges to 1 via (2), as well as bound the frequentist probability of selecting M_t, and of the selected model including type I–II errors. The question is therefore how to bound the right-hand side in Proposition 1, which we discuss next. Our strategy is to use that

$$1 - p(M_t \mid \mathbf{y}) = \sum_{k \neq t} p(M_k \mid \mathbf{y}) \leq \sum_{k \neq t} \left(1 + B_{tk} \frac{p(M_t)}{p(M_k)}\right)^{-1}.$$

Per Lemma 1 below, the L_1 convergence of the right-hand side can be proven by integrating tail probabilities that, conveniently, only involve pairwise Bayes factors B_{kt}. A natural question is whether said right-hand side provides a sufficiently tight bound. Lemma 2 shows that, indeed, whenever $p(M_t \mid \mathbf{y})$ converges to 1 the right-hand side is asymptotically equivalent to $1 - p(M_t \mid \mathbf{y})$.

Lemma 1.

$$E_{f^*}(p(M_k \mid \mathbf{y})) \leq E_{f^*}\left(\left(1 + B_{kt} \frac{p(M_k)}{p(M_t)}\right)^{-1}\right) = \int_0^1 P_{f^*}\left(B_{kt} > \frac{p(M_k)}{p(M_t)(1/u - 1)}\right) du.$$

Lemma 2. *Suppose that* $p(M_t \mid \mathbf{y}) \xrightarrow{L_1} 1$. *Then*

$$\frac{1 - p(M_t \mid \mathbf{y})}{\sum_{k \neq t}(1 + B_{tk} p(M_t)/p(M_k))^{-1}} \xrightarrow{L_1} 1.$$

Our strategy is based on two steps. First, we use Lemma 1 to bound the posterior probability assigned to an individual model $E_{f^*}(p(M_k \mid \mathbf{y}))$. This is achieved by bounding tail probabilities for B_{kt}, for all $n \geq n_{k0}$ and some fixed n_{k0}. Sections 3–4 use such bounds for (possibly non-linear) Gaussian regression for Bayesian and normalized L_0 methods, respectively. The key is that Bayes factors can be bounded by quadratic forms involving least-squares estimators (or Bayesian analogues), for which we derived tail inequalities (Section S2). To facilitate applying our framework to other models, Section S2 also gives finite-n bounds for $E_{f^*}(p(M_k \mid \mathbf{y}))$ in more general cases where suitably re-scaled $\log(B_{tk})$ have exponential or polynomial tails.

The second step is to bound the right-hand side in Proposition 1 for all $n \geq n_0$ and fixed $n_0 = \max_k n_{k0}$ by adding the model-specific bounds. Note that one can similarly bound the posterior probability of other interesting model subsets, e.g. adding spurious parameters to M_t. Section 5 performs this task for Gaussian regression (and implicitly for other settings where rates for $E_{f^*}(p(M_k \mid \mathbf{y}))$ take a similar form). As a technical remark, one must ensure that such fixed n_0 exists. This need not hold in general, since the number of models $k \neq t$ grows with n, but in our regression examples such n_0 indeed exists. We refer the reader to Section S1 (A4) for further discussion.

2 Conditions for consistency

We outline priors and conditions related to the extent to which they encourage sparsity. The conditions feature non-centrality parameters measuring the signal strength when

comparing M_t versus another model M_m. We generically denote these by λ_{tm}, and define them precisely in each setting below. Section 2.1 lists the priors used in our Gaussian regression examples. Section 2.2 sets conditions on these priors (see Section 4 for L_0 criteria), and discusses connections to necessary conditions for $p(M_t \mid \mathbf{y})$ to converge to 1 in a wide class of models and priors (Section S3.1) and to related literature. The main difference to earlier work is that we do not restrict attention to situations where $p(M_k)$ is a sparse prior or one sets diffuse parameter priors. By restricting the maximum model complexity, our study includes the use of less sparse priors, to provide a wider depiction of when one can hope $p(M_t \mid \mathbf{y})$ to converge to 1.

2.1 Prior distributions for regression

The framework from Section 1 applies to any prior but the required algebra varies, for illustration we focus on several popular priors. First we consider Zellner's prior

$$p(\boldsymbol{\theta}_k \mid M_k, \phi) = N(\boldsymbol{\theta}; \mathbf{0}, \tau n\phi (X_k' X_k)^{-1}), \tag{3}$$

where $\tau > 0$ is a known prior dispersion. For simplicity $X_k' X_k$ is assumed invertible for $p_k \leq \bar{p}$. We then extend results to Normal priors

$$p(\boldsymbol{\theta}_k \mid M_k, \phi) = N(\boldsymbol{\theta}_k; \mathbf{0}, \tau n\phi V_k) \tag{4}$$

with general covariance V_k and to the pMOM prior (Johnson and Rossell, 2012)

$$p(\boldsymbol{\theta}_k \mid \phi, M_k) = \prod_{j \in M_k} \theta_j^2 \mathbf{x}_j' \mathbf{x}_j / (\tau n\phi) N(\theta_j; \mathbf{0}, \tau n\phi / \mathbf{x}_j' \mathbf{x}_j), \tag{5}$$

where $\mathbf{x}_j$ is the j^{th} column in X_k.

The idea is that constant τ leads to roughly constant prior variance, e.g. for the pMOM prior if X_k has zero column means and unit column variances then $n/\mathbf{x}_j' \mathbf{x}_j = 1$. Such constant τ may be desirable from a foundational Bayesian point of view, where the prior does not to depend on n. In fact, the default choice $\tau = 1$ leads to the unit information prior, which in turn leads to the BIC (Schwarz, 1978). An alternative is to set τ growing with n, which leads to diffuse priors. For example, one may set $\tau = p^2/n$ (Foster and George, 1994), $\tau = \max\{1, p^2/n\}$ (Fernández et al., 2001) and $\tau \gg p^2$ (Narisetty and He, 2014). As discussed, diffuse priors are used by many high-dimensional methods to induce sparsity. Regarding the error variance ϕ, whenever we treat it as unknown, we set $p(\phi \mid M_k) = \text{IG}(\phi; a_\phi/2, l_\phi/2)$ for fixed $a_\phi, l_\phi > 0$.

For the prior on the models, in Section 3 we allow for a general prior. For concreteness, when discussing prior sparsity conditions below and when providing global rates in Section 5, we focus on three popular choices. These assume that all models with the same dimension p_k receive equal prior probability, that is

$$p(M_k) = P(p_k = l) / \binom{p}{l}, \tag{6}$$

where $P(p_k = l)$ is the prior on the model size, and $\binom{p}{l}$ the number of models selecting l parameters out of p. First, we consider the uniform prior where $P(p_k = l) = \binom{p}{l}$, so that $p(M_k) = 1/K$ for all $k = 1, \ldots, K$. Second, we consider the Beta-Binomial(1,1) prior where $P(p_k = l) = 1/\bar{p}$ (Scott and Berger, 2010), and finally and a so-called Complexity prior where $P(p_k = l) \propto 1/p^{cl}$ for $c > 0$ (Castillo et al., 2015). Note that the Beta-Binomial corresponds to $c = 0$.

2.2 Conditions on model complexity and prior sparsity

We state two sets of conditions. First, B1–B2 constrain the sizes of the optimal and largest allowed models.

(B1) The maximum model size satisfies $\bar{p} \ll \min\{n, p, n\tau\}$.

(B2) The optimal model size satisfies $p_t \ll \min\{\bar{p}, n\}$.

If one assigns non-vanishing τ, as in all default choices above, B1 simplifies to $\bar{p} \ll \min\{n, p\}$. One can allow for larger $\bar{p} = p$, e.g. under Zellner's prior $B_{mt} = 1$ for $p_m \geq n$ and one can immediately bound $E_{f^*}(p(M_m \mid \mathbf{y}))$, but then one must impose stricter prior sparsity conditions that our C1–C2 below. Setting $\bar{p} \ll n$ seems natural, however, as $p_m \geq n$ results in data interpolation. See Martin et al. (2017) (Section 2.1) and references therein for further arguments for setting $\bar{p} \leq n$.

The second set of Conditions C1–C2 restrict the sparsity induced by the model prior and the prior dispersion τ in (3)–(5), and are related to the non-centrality parameter measuring the signal strength. Specifically, let $H_m = X_m(X_m'X_m)^{-1}X_m'$ be the projection matrix onto the column space of X_m. For any non-spurious model $m \in S^c$, denote by

$$\lambda_{tm} = (X_t\boldsymbol{\theta}_t^*)'(I - H_m)X_t\boldsymbol{\theta}_t^*/\phi^* \tag{7}$$

the non-centrality parameter measuring the difference in mean squared prediction error between M_t and M_m under the KL-optimal $(\boldsymbol{\theta}_t^*, \phi^*)$. Equivalently, λ_{tm} is the difference between the L_2 norm of the optimal predictor $X_t\boldsymbol{\theta}_t^*$ relative to its projection onto X_m. This non-centrality parameter can be lower-bounded by

$$\lambda_{tm} \geq n v_{tm}(\boldsymbol{\theta}_t^*)'\boldsymbol{\theta}_t^*/\phi^*,$$

where v_{tm} is the smallest non-zero eigenvalue of $X_t'(I - H_m)X_t/n$.

Conditions C1–C2 suffice for $p(M_m \mid \mathbf{y}) \xrightarrow{L_1} 0$ in high-dimensional regression. As we shall see in Section 5, uniform versions of C1–C2 also guarantee that $p(M_t \mid \mathbf{y}) \xrightarrow{L_1} 1$. C1–C2 are stated for a generic $p(M_k)$, see Section S3 for concrete expressions for the uniform, Beta-Binomial and Complexity priors in (6).

(C1) Let $m \in S$ be a spurious model. As $n \to \infty$, $(\tau n)^{(p_m - p_t)/2} p(M_t)/p(M_m) \gg 1$.

(C2) Let $m \in S^c$ be a non-spurious model. As $n \to \infty$,

$$\frac{\lambda_{tm}}{2\log(\lambda_{tm})} + \frac{p_m - p_t}{2}\log(\tau n) + \log\left(\frac{p(M_t)}{p(M_m)}\right) - \log(p_m) \gg 1.$$

C1–C2 ensure that $p(M_k)$ and τ do not favor M_m over M_t too strongly a priori and, per Theorem 1, are near-necessary. See Section S11 for an extension of Theorem 1 to high-dimensional models for Zellner's prior. For the pMOM prior one can relax slightly C1 to $(\tau n)^{3(p_m-p_t)/2}p(M_t)/p(M_m) \gg 1$ under certain conditions, see Section 3.4.

We compare our conditions to those in Narisetty and He (2014), Castillo et al. (2015), Yang et al. (2016) and Yang and Pati (2017). We offer a summary, see Section S3 for further details. A main difference is on the prior setup. These authors restricted attention to diffuse priors (large τ) and/or Complexity priors akin to that in (6). Specifically Narisetty and He (2014) and Yang et al. (2016) required $\tau n \gg p^2$, whereas Yang and Pati (2017) set a prior anti-concentration condition that also leads to τ growing with n. Castillo et al. (2015) and Yang et al. (2016) required $p(M_k)$ to be a Complexity prior, and Narisetty and He (2014) also used $p(M_k)$ that converges to a Complexity prior as p grows. Our C1–C2 in principle allow more general τ and $p(M_k)$, such as fixed τ and $p(M_k)$ that do not penalize model size exponentially, e.g. the Beta-Binomial. For such choices the asymptotic rates for $p(M_t \mid \mathbf{y})$ are then usually slower, further one may need to restrict the maximum model complexity $\bar{p}$ (see Section 5). Nevertheless, there can be significant improvements for finite n, as illustrated in Section 6.

Regarding conditions on the data-generating truth, Narisetty and He (2014) require that p_t is fixed, Castillo et al. (2015) that $p_t \leq \sqrt{n/\log p}$, Yang et al. (2016) that $p_t \leq n/\log p$ and Yang and Pati (2017) that $p_t \log(p/p_t) \leq n$. These are related to our B1–B2, which require $p_t \ll n$ and $\bar{p} \ll n$, though these authors did not restrict $\bar{p}$. Rather, they set $\bar{p} = p$ and priors that strongly penalize complexity.

Finally, these authors also set assumptions which, under restricted eigenvalue conditions, are related to beta-min conditions. Specifically, Castillo et al. (2015) essentially required that $\min_j |\theta_j^*|^2/\phi^* > p_t(\log p)/n$, Yang et al. (2016) that $\min_j |\theta_j^*|^2/\phi^* > (c + p_t)(\log p)/n$, where c is the Complexity prior's parameter, and Narisetty and He (2014) and Yang and Pati (2017) that $\min_j |\theta_j^*|^2/\phi^* > (\log p)/n$. Under such eigenvalue conditions, if $p(M_k)$ is the Complexity prior then for our C2 to hold it suffices that

$$\min_j |\theta_j^*|^2/\phi^* \gg [\log(\tau n) + (1 + c)\log p]/n, \tag{8}$$

which is similar to these conditions above. Recall that $c = 0$ corresponds to the Beta-Binomial prior, illustrating that using less sparse $p(M_k)$ lowers the required signal strength. These conditions are mild, e.g. Wainwright (2009) showed that $\min_j |\theta_j^*|^2/\phi^* > [\log(p/p_t)]/n$ is a necessary condition for any method to consistently select M_t.

3 Model-specific rates for regression

In this section we bound $E_{f^*}(p(M_m \mid \mathbf{y}))$ for a single model M_m for Gaussian regression and the priors in Section 2.1. Per Lemma 1 the proof strategy is to bound tail probabilities for pairwise Bayes factors B_{mt}. Sections 3.1–3.4 consider the case where the

model is well-specified, that is they assume a data-generating $f^*(\mathbf{y}) = N(\mathbf{y}; X_t\boldsymbol{\theta}_t^*, \phi^* I)$. Then B_{tm} is bounded by chi-square and F distribution tails for Normal priors, and by a slightly more involved term for pMOM priors. Section 3.5 considers the situation where the mean structure has been misspecified, e.g. $E_{f^*}(y)$ features variables or non-linear terms that were omitted from X. Finally, Section 3.6 considers a misspecified covariance case, i.e. f^* has heteroskedastic and/or correlated errors.

The rates in Sections 3.1–3.3 are similar to standard finite-dimensional rates (Theorem 1 in Dawid (1999), proof of our Theorem 1). Roughly speaking, non-spurious models are discarded at an exponential rate in n (more precisely, in the non-centrality parameter λ_{tm} in (7), proportional to n under restricted eigenvalue conditions). More critically, spurious models are discarded at a rate that is essentially

$$E_{f^*}(p(M_m \mid \mathbf{y})) \preceq \frac{p(M_m)}{p(M_t)(\tau n)^{(p_m - p_t)/2}},$$

up to lower-order terms. This result portrays the effect of the prior dispersion τ and model prior probabilities to encourage sparsity in more general regimes that in current high-dimensional literature (see Section 2). As shown in Section 5, the implication is that one can often allow for fixed τ and/or $p(M_k)$ that are not particularly sparse (e.g. the Beta-Binomial prior in (6)), and still attain consistency. The pMOM rates to discard spurious models are faster, as is standard for non-local priors, but we provide tighter rates than currently available (see Section 3.4).

3.1 Zellner's prior with known variance

Under Zellner's prior and known error variance ϕ^*, simple algebra gives

$$B_{tm} = \exp\left\{-\frac{\tau n}{2\phi^*(1+\tau n)} W_{mt}\right\} (1+\tau n)^{\frac{p_m - p_t}{2}} \tag{9}$$

and hence in Lemma 1

$$P_{f^*}\left(B_{mt} > \frac{p(M_t)}{p(M_m)(1/u-1)}\right) = P_{f^*}\left(\frac{W_{mt}}{\phi^*} > \frac{1+\tau n}{\tau n} 2\log\left[\frac{(1+\tau n)^{\frac{p_m - p_t}{2}} p(M_t)}{p(M_m)(1/u-1)}\right]\right), \tag{10}$$

where $W_{mt} = \hat{\boldsymbol{\theta}}_m' X_m' X_m \hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_t' X_t' X_t \hat{\boldsymbol{\theta}}_t$ is the difference between residual sums of squares under M_t and M_m and $\hat{\boldsymbol{\theta}}_m = (X_m' X_m)^{-1} X_m' \mathbf{y}$ the least-squares estimate.

Proposition 2 gives a simple asymptotic expression for the L_1 rate at which $p(M_m \mid \mathbf{y})$ vanishes. Spurious models are discarded at a rate that depends on $p(M_m)$ and τn. Non-spurious models are discarded near-exponentially in the non-centrality parameter, times a factor driven by $p(M_m)$ and τn. The result portrays the effect of favoring sparse models either via $p(M_k)$ or by setting large τ, namely a faster rate in Part (i) at the cost of a slower rate in Part (ii) for any model of size $p_m < p_t$.

Proposition 2. *Assume that* $f^*(\mathbf{y}) = N(\mathbf{y}; X_t\boldsymbol{\theta}_t^*; \phi^* I)$ *and consider* $m \neq t$.

(i) Let $m \in S$ be a spurious model, and $g = (\tau n)^{(p_m - p_t)/2} p(M_t)/p(M_m)$. Assume Conditions B1 and C1. Then, for all fixed $\alpha < 1$,

$$E_{f^*}(p(M_m \mid \mathbf{y})) \preceq \frac{[\log(g)]^{(p_m-p_t)/2}}{g} \ll \left[\frac{p(M_m)}{p(M_t)(\tau n)^{(p_m-p_t)/2}}\right]^{\alpha}.$$

(ii) Let $m \in S^c$ be a non-spurious model. Assume Conditions B2 and C2. Then

$$E_{f^*}(p(M_m \mid \mathbf{y})) \ll e^{-\lambda_{tm}^{\gamma}/2}\left[\frac{p(M_m)}{p(M_t)(\tau n)^{(p_m-p_t)/2}}\right]^{\gamma},$$

for all fixed $\gamma < 1$, where λ_{tm} is as in (7).

We remark that α, γ are taken arbitrarily close to 1, and are introduced to provide simpler expressions, see the proof for slightly tighter bounds where $\alpha = \gamma = 1$, after adding lower-order terms. Proposition 2 gives upper-bounds. To see that they are reasonably tight, Section S11 shows that for spurious models $E_{f^*}(p(M_m \mid \mathbf{y})) \succeq [p(M_m)/p(M_t)](\tau n)^{-(p_m-p_t)/2}$, which equals the rate in Part (i), up to a log term. A similar argument is made for Part (ii).

3.2 Zellner's prior with unknown variance

Proposition 3 extends Proposition 2 to the case where ϕ^* is unknown, and one sets a prior $\phi \sim \mathrm{IG}(a_\phi/2, l_\phi/2)$. The rates are essentially equivalent, up to lower-order terms.

Let $s_k = \mathbf{y}'\mathbf{y} - \mathbf{y}'X_k(X_k'X_k)^{-1}X_k'\mathbf{y}$ be the residual sum of squares under M_k, then

$$B_{tm} = \left(\frac{\tilde{s}_m}{\tilde{s}_t}\right)^{\frac{a_\phi+n}{2}}(1+\tau n)^{\frac{p_m-p_t}{2}} = \left(1 + \frac{p_m - p_t}{n - p_m}\tilde{F}_{mt}\right)^{-\frac{a_\phi+n}{2}}(1+\tau n)^{\frac{p_m-p_t}{2}}, \quad (11)$$

where $\tilde{s}_m = l_\phi + \mathbf{y}'\mathbf{y} - \frac{\tau n}{\tau n+1}\mathbf{y}'X_m(X_m'X_m)^{-1}X_m'\mathbf{y}$ is a Bayesian analogue of s_m and

$$\tilde{F}_{mt} = \frac{(\tilde{s}_t - \tilde{s}_m)/(p_m - p_t)}{\tilde{s}_m/(n - p_m)} \leq \frac{(s_t - s_m)/(p_m - p_t)}{s_m/(n - p_m)} = F_{mt}. \quad (12)$$

F_{mt} is the F-statistic to test M_t versus M_m, $\tilde{F}_{mt}$ is its Bayesian analogue, and the inequality in (12) follows from trivial algebra.

Proposition 3. *Assume $f^*(\mathbf{y}) = N(\mathbf{y}; X_t\boldsymbol{\theta}_t^*; \phi^* I)$ and Conditions B1–B2, C1–C2.*

(i) Let $m \in S$ be a spurious model and $g = (\tau n)^{(p_m-p_t)/2}p(M_t)/p(M_m)$. If $\log(g) \ll n - p_m$, then

$$E_{f^*}(p(M_m \mid \mathbf{y})) \preceq \left(\frac{1}{g}\right)^{1-4\sqrt{\frac{\log(g)}{n-p_m}}} \ll \frac{[p(M_m)/p(M_t)]^{\alpha}}{(\tau n)^{\alpha\frac{p_m-p_t}{2}}},$$

for any fixed $\alpha < 1$. *If* $\log(g) \gg n - p_m$ *then*

$$E_{f^*}(p(M_m \mid \mathbf{y})) \ll \exp\left\{-\frac{(n-p_t-5)}{2}\log\left(\frac{\log^\gamma(g)}{n-p_t-6}\right)\right\} \ll e^{-\kappa n},$$

for any fixed $\gamma < 1$, $\kappa > 1$.

(ii) Let $m \in S^c$ *be a non-spurious model and* λ_{tm} *as in* (7). *Then*

$$E_{f^*}(p(M_m \mid \mathbf{y})) \ll \max\left\{e^{-\lambda_{tm}^\gamma/2}\left[\frac{p(M_m)}{p(M_t)(\tau n)^{(p_m-p_t)/2}}\right]^\gamma, e^{-\kappa n}\right\},$$

for any fixed $\gamma < 1$, $\kappa > 0$.

3.3 Normal prior with general covariance

We extend Proposition 3 to more general Normal priors $p(\boldsymbol{\theta}_k \mid \phi, M_k) = N(\boldsymbol{\theta}_k; \mathbf{0}, \tau n \phi V_k)$. The rates are essentially equivalent, subject to mild eigenvalue conditions. Let $\rho_{k1} \geq \ldots \geq \rho_{kp_k} > 0$ be the p_k non-zero eigenvalues of $V_k X_k' X_k$, F_{mt} the F-test statistic in (12), and $\tilde{F}_{mt}$ be as in (12) after replacing $\tilde{s}_k = l_\phi + \mathbf{y}'\mathbf{y} - \mathbf{y}' X_k (X_k' X_k + (\tau n)^{-1} V_k^{-1})^{-1} X_k' \mathbf{y}$. Then simple algebra gives the following expression for Bayes factors

$$B_{tm} = \left(1 + \frac{p_m - p_t}{n - p_m}\tilde{F}_{mt}\right)^{-\frac{a_\phi + n}{2}} \frac{\prod_{j=1}^{p_m}(\tau n \rho_{mj} + 1)^{\frac{1}{2}}}{\prod_{j=1}^{p_t}(\tau n \rho_{tj} + 1)^{\frac{1}{2}}}. \tag{13}$$

Proposition 4 assumes two further technical conditions D1–D2, beyond those in Proposition 3. Both can be relaxed, but they simplify exposition. D1 allows interpreting τ as driving the prior variance in a similar fashion than for Zellner's prior. D2 ensures that the Bayesian-flavoured F-statistic $\tilde{F}_{mt}$ is close to the classical F_{mt}, and is a mild requirement since typically $\tau n \succeq n \succeq \lambda_{t0}$.

(D1) For some constant $c_{mt} > 0$, $\prod_{j=1}^{p_m}(\tau n \rho_{mj} + 1)^{\frac{1}{2}} / \prod_{j=1}^{p_t}(\tau n \rho_{tj} + 1)^{\frac{1}{2}} \asymp (c_{mt}\tau n)^{(p_m - p_t)/2}$.

(D2) As $n \to \infty$, $\lambda_{t0} \ll \tau n \rho_{tp_t}$, where λ_{t0} is as in (7).

Proposition 4. *Assume that* $f^*(\mathbf{y}) = N(\mathbf{y}; X_t \boldsymbol{\theta}_t^*; \phi^* I)$. *Consider* $m \neq t$ *and that Conditions B1, B2, C1, C2, D1 and D2 hold.*

(i) Let $m \in S$ *be a spurious model. Then, for any fixed* $\alpha \in (0,1)$ *and* $\kappa > 0$,

$$E_{f^*}(p(M_m \mid \mathbf{y})) \ll \max\left\{[p(M_m)/p(M_t)]^\alpha (\tau n)^{\alpha(p_t - p_m)/2}, e^{-\kappa n}, e^{-\tau n \rho_{tp_t}/2}\right\}.$$

(ii) Let $m \in S^c$ *be a non-spurious model and* λ_{tm} *as in* (7). *Then, for any fixed* $\gamma < 1$ *and* $\kappa > 0$,

$$E_{f^*}(p(M_m \mid \mathbf{y})) \ll \max\left\{e^{-\lambda_{tm}^\gamma/2}\left[\frac{p(M_m)}{p(M_t)(\tau n)^{(p_m-p_t)/2}}\right]^\gamma, e^{-\kappa n}, e^{-\tau n \rho_{tp_t}/2}\right\}.$$

3.4 pMOM prior

Proposition 5 below states that, under suitable conditions, the pMOM prior attains a rate to discard spurious models featuring a term that is essentially $(\tau n)^{3(p_m-p_t)/2}$, and hence faster than the $(\tau n)^{(p_m-p_t)/2}$ shown for Normal priors. To ease the algebra we assume that X_k has zero column means and unit variances. By Proposition 1 in Rossell and Telesca (2017) the Bayes factor under the pMOM prior in (5) is

$$B_{tm} = D_{tm}\left(1+\frac{p_m-p_t}{n-p_m}\tilde{F}_{mt}\right)^{-\frac{a_\phi+n}{2}}\frac{\prod_{j=1}^{p_m}(\tau n\rho_{mj}+1)^{\frac{1}{2}}}{\prod_{j=1}^{p_t}(\tau n\rho_{tj}+1)^{\frac{1}{2}}}, \tag{14}$$

where

$$D_{tm} = \frac{\int\int N(\boldsymbol{\theta}_t;\tilde{\boldsymbol{\theta}}_t,\phi\tilde{V}_t)\mathrm{IG}\left(\phi;\frac{a_\phi+n}{2},\frac{\tilde{s}_t}{2}\right)\prod_{j\in M_t}d(\theta_{tj}/\sqrt{\phi})d\boldsymbol{\theta}_t d\phi}{\int\int N(\boldsymbol{\theta}_m;\tilde{\boldsymbol{\theta}}_m,\phi\tilde{V}_m)\mathrm{IG}\left(\phi;\frac{a_\phi+n}{2},\frac{\tilde{s}_m}{2}\right)\prod_{j\in M_m}d(\theta_{mj}/\sqrt{\phi})d\boldsymbol{\theta}_m d\phi},$$

$d(z)=z^2/\tau$, $\tilde{V}_k^{-1}=X_k'X_k+V_k^{-1}/(\tau n)$, $\tilde{\boldsymbol{\theta}}_k=\tilde{V}_kX_k'\mathbf{y}$ and $\tilde{F}_{mt}$, $\tilde{s}_k$ and ρ_{kj} are as in (13) for the particular case $V_k=\mathrm{diag}(X_k'X_k)^{-1}$.

The Bayes factor in (14) is hence equal to that in (13) times a penalty term D_{tm} that helps penalize spurious models $m\in S$. Intuitively, this is because the posterior distribution of $d(\theta_{mj}/\sqrt{\phi})=\theta_{mj}^2/(\phi\tau)$ concentrates at 0 for truly spurious $\theta_{mj}^*=0$, at a rate that is at most σ/τ, where σ is the largest (posterior) variance in $\tilde{V}_m$. To state a simple rate, Proposition 5 assumes technical conditions E1–E5 discussed in Section S14. These can be relaxed, at the cost of a more involved rate for $E_{f^*}(p(M_m\mid\mathbf{y}))$.

Proposition 5. *Assume that $f^*(\mathbf{y})=N(\mathbf{y};X_t\boldsymbol{\theta}_t^*;\phi^*I)$. Let $m\in S$ be a spurious model and assume that Conditions B1, C1, D1 and E1–E5 hold. Then*

$$E_{f^*}(p(M_m\mid\mathbf{y}))\ll\max\left\{\left(\frac{p(M_m)}{p(M_t)}\right)^{\alpha}\left(\frac{\tau^3 n}{\sigma^2}\frac{\rho_{mp_m}}{\rho_{m1}}\right)^{-\alpha\frac{p_m-p_t}{2}},e^{-\kappa n},e^{-\tau n\rho_{tp_t}/2}\right\},$$

for any fixed $\kappa>0$ and $\alpha<1$, where σ is the largest diagonal element in $\tilde{V}_m$.

Relative to Sections 3.1–3.3, Proposition 5 features an acceleration factor τ/σ for each truly spurious variable in M_m and a term $\rho_{mp_m}^{1/2}/\rho_{m1}^{1/2}$ involving eigenvalues. If the latter is bounded and $\sigma\asymp 1/n$ (e.g. under restricted eigenvalue conditions), the acceleration is of order $(\tau n)^{p_m-p_t}$. Proposition 5 is tighter than results in Johnson and Rossell (2012), e.g. under uniform $p(M_k)$ we prove consistency when $p\ll(\tau n)^{\alpha/2}$ for any $\alpha<3$ (Section 5) whereas Johnson and Rossell (2012) required $p\ll n$.

3.5 Misspecified mean structure

So far we assumed that the data analyst poses a model $p(\mathbf{y}\mid\boldsymbol{\theta},\phi)=N(\mathbf{y};X\boldsymbol{\theta},\phi)$ and that the data-generating $f^*(\mathbf{y})=N(\mathbf{y};X\boldsymbol{\theta}^*,\phi^*I)$ lies in the considered family.

Although X may contain non-linear basis expansions, e.g. splines or tensor products, there are practically-relevant situations where either the mean or the error structure are misspecified. Proposition 6 considers the mean misspecification case. Specifically, it considers that $f^*(\mathbf{y}) = N(\mathbf{y}; W\boldsymbol{\beta}^*, \xi^* I)$ for some $n \times q$ matrix W, $\boldsymbol{\beta}^* \in \mathbb{R}^q$ and $\xi^* \geq 0$. This includes situations where one did not record truly relevant variables (X misses columns from W) or the mean of $\mathbf{y}$ depends on non-linearly in ways that are not captured by X (e.g. X assumes an additive structure, whereas W contains non-linear interactions). For simplicity we state Proposition 6 for Zellner's prior but extensions to other priors follow similar lines. Proposition 7 considers $f^*(\mathbf{y}) = N(\mathbf{y}; X_t\boldsymbol{\theta}_t^*, \phi^*\Sigma^*)$ for general Σ^*, allowing for heteroskedastic and correlated errors.

The proof strategy is as follows. The framework in Section 1 applies to any f^*, if one can bound Bayes factor tail probabilities in Lemma 1. In Propositions 6–7 it is possible to bound said tails, using that f^* has Gaussian errors and eigenvalues of Σ^*. Further extensions are possible, e.g. Propositions S1–S2 in Rossell et al. (2020) deploy Lemma 1 to the case where f^* has sub-Gaussian errors, e.g. when $\mathbf{y}$ is a binary outcome.

Proposition 6 says that the rate to discard spurious models is similar to the well-specified case (slightly sped-up by a factor $e^{\phi_t^*/\xi^*} \geq 1$). The rate for non-spurious models $m \in S^c$ vanishes exponentially in a non-centrality parameter λ_{tm}, but is exponentially slower than a certain λ_m^* obtained when using the correct mean structure. Specifically, denote by M^* the true model class $N(\mathbf{y}; W\boldsymbol{\beta}, \xi I)$ indexed by $(\boldsymbol{\beta}, \xi)$. Let $H_m = X_m(X_m'X_m)^{-1}X_m'$ be the projection matrix associated to a model M_m, and define the non-centrality parameter

$$\lambda_{tm} = (W\boldsymbol{\beta}^*)'H_t(I - H_m)H_tW\boldsymbol{\beta}^*/\xi^*. \tag{15}$$

Note that λ_{tm} extends the non-centrality parameter in (7) to the misspecified case, by projecting the true mean $W\boldsymbol{\beta}^*$ onto the column space of X_t. Similarly, let $\lambda_m^* = (W\boldsymbol{\beta}^*)'(I - H_m)W\boldsymbol{\beta}^*/\xi^*$. Denote the KL-optimal parameters under M_m by $\boldsymbol{\theta}_m^* = (X_m'X_m)^{-1}X_m'W\boldsymbol{\beta}^*$ (assuming full-rank X_m) and the optimal error variance by

$$\phi_m^* = \xi^* + \frac{1}{n}(W\boldsymbol{\beta}^*)'(I - H_m)W\boldsymbol{\beta}^*.$$

If one compared M^* and M_m, by Proposition 3 one would select M^* at an exponential rate in λ_m^*. However, under misspecification the best one can hope for is to select M_t. When comparing M_t and M_m, the Bayes factor for M_m vanishes at an exponential rate in $\lambda_{tm} \leq \lambda_m^*$, with equality if and only if $W\boldsymbol{\beta}^* = X\boldsymbol{\theta}_t^*$ (the mean is well-specified).

Proposition 6. *Let $p(\boldsymbol{\theta}_k \mid \phi_k, M_k) = N(\boldsymbol{\theta}_k; 0, \phi_k\tau n(X_k'X_k)^{-1}$ be Zellner's prior and α, κ be any constants satisfying $\alpha \in (0,1)$, $\kappa > 0$. Assume that $f^*(\mathbf{y}) = N(\mathbf{y}; W\boldsymbol{\beta}^*, \xi^* I)$. Further assume B1, B2, C1, C2 for λ_{tm} be as in (15), and $\phi_t^*/\xi^* \ll \log(\lambda_{tm})$.*

(i) Let $m \in S$. If $\log((\tau n)^{\frac{p_m - p_t}{2}} e^{\phi_t^*/\xi^*} p(M_t)/p(M_m)) \ll n - p_m$ then

$$E_{f^*}(p(M_m \mid \mathbf{y})) \ll \left[(p(M_t)/p(M_m))(\tau n)^{\frac{p_m - p_t}{2}} e^{\phi_t^*/\xi^*}\right]^{-\alpha}.$$

If $\log([p(M_m)/p(M_t)](\tau n)^{\frac{p_t - p_m}{2}} e^{\phi_t^*/\xi^*}) \gg n - p_m$ then $E_{f^*}(p(M_m \mid \mathbf{y})) \ll e^{-\kappa n}$.

(ii) Let $m \in S^c$. *If* $\lambda_{tm} + \log((\tau n)^{\frac{p_m - p_t}{2}} p(M_t)/p(M_m)) \ll n - p_q$ *then*

$$E_{f^*}(p(M_m \mid \mathbf{y})) \ll \max\left\{e^{-\lambda_{tm}^{\gamma}/2}\left[\frac{p(M_m)}{p(M_t)(1+\tau n)^{(p_m-p_t)/2}}\right]^{\gamma}, e^{-\kappa n}\right\},$$

for any fixed $\gamma < 1$, $\kappa > 0$.

Further, $\lambda_{tm} \leq \lambda_m^*$, *with equality if and only if* $W\boldsymbol{\beta}^* = X\boldsymbol{\theta}_t^*$.

As a technical remark, Proposition 6 uses the minimal assumption that $\phi_t^*/\xi^* \ll \log(\lambda_{tm})$. Since the latter grows with n, this assumption holds in standard cases where ξ^* and ϕ_t^* are constant and when ϕ_t^* decreases with n (e.g. X_t is a non-parametric basis with growing dimension and hence lower error variance as ϕ_t^* as n grows), but also allows for pathological cases where ϕ_t^*/ξ^* grows slowly with n.

3.6 Misspecified covariance structure

Consider a misspecified covariance case, i.e. $f^*(\mathbf{y}) = N(\mathbf{y}; X_t\boldsymbol{\theta}_t^*, \phi^*\Sigma^*)$ for positive-definite Σ^*. Without loss of generality constrain $\mathrm{tr}(\Sigma^*) = n$, so $\phi^* = \sum_{i=1}^n \mathrm{Var}_{f^*}(y_i)/n$ is the average variance. For simplicity we assume that ϕ^* is known, in analogy to Section 3.1. Extensions to unknown ϕ^* are possible, akin to the proof of Proposition 3.

We obtain rates that resemble the well-specified case, but there are potentially important differences related to certain eigenvalues and an adjusted non-centrality parameter $\tilde{\lambda}_{tm}$. Specifically, for any model M_k with design matrix X_k denote by $\tilde{X}_t = (I - H_k)X_t$ and by $(\underline{\omega}_{tk}, \bar{\omega}_{tk})$ the smallest and largest eigenvalues of $\tilde{X}_t'\Sigma^*\tilde{X}_t(\tilde{X}_t'\tilde{X}_t)^{-1}$. Consider the non-centrality parameter

$$\tilde{\lambda}_{tm} = (\boldsymbol{\theta}_t^*)'\tilde{X}_t'\tilde{X}_t(\tilde{X}_t'\Sigma^*\tilde{X}_t)^{-1}\tilde{X}_t'\tilde{X}_t\boldsymbol{\theta}_t^*/\phi^*, \tag{16}$$

where $\tilde{X}_t = (I - H_m)X_t$. To gain intuition, in the well-specified case $\Sigma^* = I$, then $\tilde{\lambda}_{tm}$ simplifies to λ_{tm} in (7), and $\underline{\omega}_{tm} = \bar{\omega}_{tm} = 1$. More generally, $\underline{\omega}_{tm}\tilde{\lambda}_{tm} \leq \lambda_{tm} \leq \bar{\omega}_{tm}\tilde{\lambda}_{tm}$.

Proposition 7 says that spurious models are discarded at the same rate as in the well-specified case, raised to a power $1/\bar{\omega}_{tm}$. Hence, when $\bar{\omega}_{tm}$ is large, misspecifying Σ^* can lead to a significantly slower rate. The intuition is that $\bar{\omega}_{tm}$ measures the discrepancy between the model-based least-squares covariance $(\tilde{X}_t'\tilde{X}_t)^{-1}$ and its actual sampling covariance $(\tilde{X}_t'\tilde{X}_t)^{-1}\tilde{X}_t'\Sigma^*\tilde{X}_t(\tilde{X}_t'\tilde{X}_t)^{-1}$. In contrast, non-spurious models are discarded exponentially in $\tilde{\lambda}_{tm}$ so, provided $\underline{\omega}_{tm}$ is bounded, the rate remains exponential in λ_{tm}. Relative to Proposition 6 where misspecifying the mean was guaranteed to decrease power, this need not happen when misspecifying Σ^*.

Proposition 7 requires adjusting Condition C2 into C2' below.

(C2') Let $m \in S^c$, $\tilde{\lambda}_{tm}$ as in (16) and $M_q = M_t \cup M_m$ be the model with design matrix X_q combining X_t and X_m. As $n \to \infty$, $[\underline{\omega}_{mq}/\bar{\omega}_{tq}]\log(\tilde{\lambda}_{tm}) \gg 1$ and

$$\frac{\tilde{\lambda}_{tm}}{2\log(\tilde{\lambda}_{tm})} + \frac{1}{\bar{\omega}_{tq}}\left[\frac{p_m - p_t}{2}\log(\tau n) + \log\left(\frac{p(M_t)}{p(M_m)}\right)\right] - \log p_m \gg 1.$$

The interpretation of C2' is similar to C2, albeit incorporating eigenvalues. The presence of eigenvalues can be relaxed somewhat, at the expense of obtaining slower rates in Proposition 7 (see the proof). We avoid a detailed study, but note that $n^{-1}\tilde{X}_t'\Sigma^*\tilde{X}_t$ and $n^{-1}\tilde{X}_t'\tilde{X}_t$ are sample covariance matrices. Under suitable assumptions (e.g. the rows of X are independent draws from a Normal distribution) one can show that $\underline{\omega}_{tm}/\bar{\omega}_{tm}$ are bounded by constants with high probability, see Wainwright (2019) (Chapter 6).

Proposition 7. *Assume that* $f^*(\mathbf{y}) = N(\mathbf{y}; X_t\boldsymbol{\theta}_t^*, \phi^*\Sigma^*)$ *for positive-definite* Σ^*, $tr(\Sigma^*) = n$ *and known* ϕ^*. *Let* $p(\boldsymbol{\theta}_k \mid \phi_k, M_k) = N(\boldsymbol{\theta}_k; 0, \phi^*\tau n(X_k'X_k)^{-1})$ *be Zellner's prior.*

(i) Let $m \in S$ *be a spurious model. If B1 and C1 hold then, for any fixed* $\alpha < 1$,

$$E_{f^*}(p(M_m \mid \mathbf{y})) \ll \left[\frac{[p(M_m)/p(M_t)]}{(1+\tau n)^{(p_m-p_t)/2}}\right]^{\alpha \min\{1,1/\bar{\omega}_{tm}\}}.$$

(ii) Let $m \in S^c$ *be a non-spurious model. Assume Conditions B2 and C2'. Then*

$$E_{f^*}(p(M_k \mid \mathbf{y})) \ll \max\left\{e^{-\gamma\tilde{\lambda}_{tm}/2}, \left[\frac{[p(M_m)/p(M_t)]}{(\tau n)^{(p_m-p_t)/2}}\right]^{\gamma \min\{1,1/\bar{\omega}_{tq}\}} e^{-\frac{\gamma \min\{1,\bar{\omega}_{tq}\}\tilde{\lambda}_{tm}}{2}}\right\},$$

for any fixed $\gamma < 1$, *where* $\bar{\omega}_{tq}$ *is the largest eigenvalue of* $\tilde{X}_q'\Sigma^*\tilde{X}_q(\tilde{X}_q'\tilde{X}_q)^{-1}$.

4 Normalized L_0 penalties

An L_0 criterion proceeds by selecting the model

$$\hat{k} = \arg\max_k \log p(\mathbf{y} \mid \hat{\boldsymbol{\theta}}_k, \hat{\phi}_k) - \eta_k,$$

where $(\hat{\boldsymbol{\theta}}_k, \hat{\phi}_k) = \arg\max_{\boldsymbol{\theta}_k \in \Theta_k, \phi \in \Phi} p(\mathbf{y} \mid \boldsymbol{\theta}, \phi)$ is the maximum likelihood estimator under model M_k, and η_k is a penalty that may depend on the model size p_k, n and p. For example the BIC corresponds to $\eta_k = 0.5p_k\log(n)$, the RIC to $\eta_k = p_k\log(p)$ and the EBIC to $\eta_k = 0.5p_k\log(n) + \xi\log\binom{p}{p_k}$ for some $\xi \in (0,1)$.

We give results akin to Section 3 for normalized L_0 methods. We equivalently define

$$\hat{k} = \arg\max_k \frac{h(\mathbf{y},k)}{\sum_{l=1}^K h(\mathbf{y},l)},$$

where $h(\mathbf{y},k) = p(\mathbf{y} \mid \hat{\boldsymbol{\theta}}_k, \hat{\phi}_k)e^{-\eta_k}$, and refer to $\tilde{h}(\mathbf{y},k) = h(\mathbf{y},k)/\sum_{l=1}^K h(\mathbf{y},l)$ as a normalized L_0 criterion. The idea is that, given the connection between BMS and L_0 penalties (see below), one could view $\tilde{h}(\mathbf{y},k)$ as a pseudo-posterior probability for M_k that quantifies the certainty in $\hat{k}$. Let M_t be the optimal model defined in Section 1. Akin to (2), our goal is to show that $\tilde{h}(\mathbf{y},t)$ converges to 1 in the L_1 sense by studying

$$\sum_{k\neq t} E_{f^*}\left(\tilde{h}(\mathbf{y},k)\right) \leq \sum_{k\neq t} E_{f^*}\left([1+h(\mathbf{y},t)/h(\mathbf{y},k)]^{-1}\right). \tag{17}$$

Note that $h(\mathbf{y},t)/h(\mathbf{y},k)$ is analogous to $B_{tk}p(M_t)/p(M_k)$, a product of Bayes factors and prior model probabilities. From Proposition 1 and Corollaries 1–2, (17) bounds the frequentist probability of selecting M_t, type I error and power.

This section is organized as follows. First, we discuss the connection between Zellner's prior and normalized L_0 criteria. We then show Proposition 8, our main result bounding $E_{f^*}(\tilde{h}(\mathbf{y},k))$ for an individual model, akin to Section 3 where we bounded $E_{f^*}(p(M_k \mid \mathbf{y}))$. Section 5 combines these bounds across models to obtain global rates. To see the connection between Zellner's prior and normalized L_0 criteria, in Gaussian regression simple algebra shows that

$$\frac{h(\mathbf{y},t)}{h(\mathbf{y},k)} = \left(1 + \frac{p_k - p_t}{n - p_k} F_{kt}\right)^{-\frac{n}{2}} e^{\eta_k - \eta_t}, \tag{18}$$

where F_{kt} is the F-test statistic in (12). The resemblance of (18) to Zellner's prior expression in (11) allows extending Proposition 3 to L_0 penalties.

We state two technical conditions C1"–C2" required by Proposition 8, which are trivial modifications of Conditions C1–C2 from Section 2.2.

(C1") Let $m \in S$. As $n \to \infty$, $\eta_m - \eta_t \gg 1$.

(C2") Let $m \in S^c$. As $n \to \infty$, $\frac{1}{2}\lambda_{tm}/\log(\lambda_{tm}) + \eta_m - \eta_t - \log(p_m) \gg 1$.

Condition C1" holds for the BIC, RIC and EBIC, and any penalty η_k that increases with model size p_k and diverges to infinity as $n \to \infty$. Condition C2" is also mild. For example, for the BIC it suffices that $\lambda_{tm}/[\log(\lambda_{tm})p_t \log(n)] \gg 1$, for the RIC that $\lambda_{tm}/[2\log(\lambda_{tm})p_t \log(p)] \gg 1$ and for the EBIC that $\lambda_{tm}/[\log(\lambda_{tm})p_t \log(n^{1/2}p^{\xi})] \gg 1$. See Section 2.2 for discussion why these conditions are near-minimal.

Proposition 8. *Assume that* $f^*(\mathbf{y}) = N(\mathbf{y}; X_t\boldsymbol{\theta}_t^*; \phi^* I)$. *Consider* $m \neq t$ *and that Conditions B1, B2, C1" and C2" hold.*

(i) Let $m \in S$. *If* $\eta_m - \eta_t \ll n - p_m$ *then*

$$E_{f^*}\left(\tilde{h}(\mathbf{y},m)\right) \leq e^{-(\eta_m - \eta_t)\left(1 - 4\sqrt{\frac{\eta_m - \eta_t}{n - p_m}}\right)},$$

for all $n \geq n_0$, *where* n_0 *is fixed and does not depend on* m. *If* $\eta_m - \eta_t \gg n - p_m$ *then* $E_{f^*}\left(\tilde{h}(\mathbf{y},m)\right) < e^{-\kappa n}$ *for any fixed* $\kappa > 0$ *and* $n \geq n_0$.

(ii) Let $m \in S^c$. *Then, for any fixed* $\gamma < 1$, $\kappa > 0$,

$$E_{f^*}\left(\tilde{h}(\mathbf{y},m)\right) < \max\left\{e^{-\lambda_{tm}^{\gamma}/2} e^{-\gamma(\eta_m - \eta_t)}, e^{-\kappa n}\right\},$$

for all $n \geq n_0$ *where* n_0 *is fixed and does not depend on* m.

For example, for the BIC $\eta_m = 0.5p_m \log(n)$, then $E_{f^*}(\tilde{h}(\mathbf{y}, m))$ vanishes essentially at a rate $n^{-(p_m-p_t)/2}$ for spurious models, and a faster $e^{-\lambda_{tm}/2} n^{-(p_m-p_t)/2}$ for non-spurious models. This is no surprise, the BIC is essentially identical to setting a uniform model prior and Zellner's prior dispersion to $\tau = 1$, hence one obtains the same rates. Similarly, the RIC is essentially identical to $\tau n = p^2$ and uniform $p(M_m)$, and the EBIC (for the choice $\xi = 1$) to $\tau = 1$ and Beta-Binomial $p(M_m)$. The misspecification results for Zellner's prior in Sections 3.5–3.6 also extend directly to normalized L_0 penalties.

5 Global rates for regression

We now use the model-specific bounds from Sections 3–4 to obtain global bounds. We saw that, under suitable conditions, $E_{f^*}(1 - p(M_t \mid \mathbf{y})) = E_{f^*}(P(S \mid \mathbf{y})) + E_{f^*}(P(S^c \mid \mathbf{y}))$

$$\leq \sum_{l=p_t+1}^{\bar{p}} \sum_{k \in S_l} \left[\frac{p(M_k)}{p(M_t)(\tau n)^{\frac{(p_k-p_t)}{2}}} \right]^{\alpha} + \sum_{l=0}^{\bar{p}} \sum_{k \in S_l^c} e^{-\frac{\lambda_{kt}^{\alpha}}{2}} \left[\frac{p(M_k)}{p(M_t)(\tau n)^{\frac{(p_k-p_t)}{2}}} \right]^{\alpha}, \quad (19)$$

for sufficiently large n and some fixed α, where τ is the prior dispersion. In well-specified Gaussian regression, as well as with a misspecified mean, we saw that one can essentially take $\alpha = 1$ (up to lower-order terms). For the pMOM prior in the first term of (19) one may take a larger $\alpha < 3$. Similarly, for normalized L_0 criteria,

$$E_{f^*}(1 - \tilde{h}(t, \mathbf{y})) \leq \sum_{l=p_t+1}^{\bar{p}} \sum_{k \in S_l} e^{-(\eta_k - \eta_t)\left(1 - 4\sqrt{\frac{\eta_k - \eta_t}{n - p_k}}\right)} + \sum_{l=0}^{\bar{p}} \sum_{k \in S_l^c} e^{-\frac{\lambda_{tk}^{\alpha}}{2} - \alpha(\eta_k - \eta_t)}, \quad (20)$$

where η_k is the L_0 penalty, e.g. for the BIC $\eta_k = 0.5p_k \log(n)$. The bounds in Sections 3–4 also feature terms such as $e^{-\kappa n}$ that vanish exponentially with n. We omitted these, since they are typically of a smaller order, but they can easily be plugged into (19)–(20).

This section derives simpler asymptotic expressions for (19)–(20) for the uniform, Beta-Binomial and Complexity priors in (6), and for the BIC, RIC and EBIC. We study separately spurious and non-spurious models, i.e. $E_{f^*}(p(S \mid \mathbf{y}))$ and $E_{f^*}(p(S^c \mid \mathbf{y}))$, and discuss the use of priors or L_0 penalties that are not particularly sparse. Such priors attain worse asymptotic rates to discard spurious models, but they can significantly improve finite n performance. The reason for the mismatch between asymptotic and finite n results is that $E_{f^*}(p(S^c \mid \mathbf{y}))$ is typically negligible for large n, as it vanishes exponentially under eigenvalue conditions. However, $E_{f^*}(p(S^c \mid \mathbf{y}))$ can be large for finite n, particularly when optimal model is not sparse. See Section 6 for examples.

5.1 Uniform prior, spurious models

The uniform prior sets $p(M_k)/p(M_t) = 1$. From the first term in (19), using that there are $|S_l| = \binom{p-p_t}{l-p_t}$ spurious models of size l and the geometric series, one obtains

$$E_{f^*}(P(S \mid \mathbf{y})) \leq \frac{\frac{p-p_t}{(\tau n)^{\alpha/2}} - \left(\frac{p-p_t}{(\tau n)^{\alpha/2}}\right)^{\bar{p}-p_t+1}}{1 - (p-p_t)/(\tau n)^{\alpha/2}} \asymp \frac{p-p_t}{(\tau n)^{\alpha/2}}, \quad (21)$$

for sufficiently large n. The asymptotic expression in the right-hand side of (21) holds if $p - p_t \ll (\tau n)^{\alpha/2}$, i.e. when $E_{f^*}(P(S \mid \mathbf{y}))$ converges to 0. Rates for the BIC and RIC are obtained by plugging $\tau = 1$ and $\tau n = p^2$ into (21).

Expression (21) describes the effect of the prior dispersion τ on sparsity. For example, for $\tau = 1$ then $P(S \mid \mathbf{y})$ vanishes as long as $p - p_t \ll n^{1/2}$, under Zellner and Normal priors. Under the pMOM one can handle $p \ll n^{3/2}$. Another default is $\tau = \max\{1, p^{2+a}/n\}$ for small $a > 0$ (Fernández et al., 2001), which effectively sets a diffuse prior (τ grows with n, whenever $p \gg \sqrt{n}$). Under such a diffuse prior, $p - p_t \ll (\tau n)^{\alpha/2}$ and $P(S \mid \mathbf{y})$ vanishes under Zellner's, Normal and pMOM priors, regardless of the magnitude of p.

5.2 Beta-binomial prior, spurious models

The Beta-Binomial prior sets $p(M_m)/p(M_t) = \binom{p}{p_t}/\binom{p}{p_m}$. Using simple algebra and the binomial coefficient's ordinary generating function,

$$E_{f^*}(P(S \mid \mathbf{y})) < \left[1 - \frac{(p-p_t)^{1-\alpha}}{(\tau n)^{\frac{\alpha}{2}}}\right]^{-p_t-1} - 1 \asymp \frac{(p_t+1)(p-p_t)^{1-\alpha}}{(\tau n)^{\alpha/2}},$$

where the right-hand side holds if $(\tau n)^{\alpha/2} \gg (p_t+1)(p-p_t)^{1-\alpha}$, by l'Hopital's rule. If α is arbitrarily close to 1, $P(S \mid \mathbf{y})$ vanishes as long as $p_t^{a+\epsilon}(p-p_t) \ll (\tau n)^{a/2}$ for arbitrarily large but fixed $a > 0$ and any small $\epsilon > 0$. For instance, under $\tau = 1$ one can handle $p - p_t \ll n^{a/2}$ variables, i.e. p can grow polynomially with n (provided $p_t \ll n$ grows sub-linearly in n, as in Condition B1. Despite not necessarily leading to the optimal asymptotic rate, the Beta-Binomial prior handles problems of fairly large dimension and still discard all spurious models.

One can obtain slightly tighter rates for L_0 penalties and for specific priors. For Zellner's prior and known ϕ^* Lemmas S20 and S16 give

$$E_{f^*}(P(S \mid \mathbf{y})) \preceq \frac{(p_t+1)(\bar{p}-p_t)^{a/2}\log^{3/2}((\tau n)^{1/2}(p-p_t))}{(\tau n)^{1/2}},$$

for any fixed $a > 1$, i.e. the dependence on p is now logarithmic. Similarly, for unknown ϕ^* and Zellner's prior Lemma S17 gives that

$$E_{f^*}(P(S \mid \mathbf{y})) \preceq \frac{(p_t+1)}{(\tau n)^{1/2}} e^{2[\log^{3/2}((\tau n)^{1/2}(p-p_t))]\sqrt{(p-p_t)/(n-\bar{p})}}.$$

Rates for the EBIC are obtained by plugging $\tau = 1$ into this last expression.

5.3 Complexity prior, spurious models

Here $p(M_m)/p(M_t) \asymp p^{c(p_t-p_m)}\binom{p}{p_t}/\binom{p}{p_m}$, where c is the Complexity prior's parameter in (6). Simple algebra shows that

$$E_{f^*}(P(S \mid \mathbf{y})) \preceq \sum_{l=p_t+1}^{\bar{p}} \binom{l}{p_t}\left(\frac{(p-p_t)^{1-\alpha}}{(\tau n)^{\frac{\alpha}{2}}p^c}\right)^{l-p_t} \preceq \frac{(p_t+1)(p-p_t)^{1-\alpha}}{(\tau n)^{\alpha/2}p^c}.$$

Since α is arbitrarily close to 1, $P(S \mid \mathbf{y})$ vanishes under the minimal requirement that $p_t^{1+\epsilon} \ll p^c(\tau n)^{1/2}$ for some (small) fixed $\epsilon > 0$. That is, the Complexity prior can handle almost any p and discard all spurious models, even for small $c > 0$. However, as illustrated next, c also plays a role in slowing down the rate to discard small non-spurious models, which can reduce the statistical power to detect non-zero coefficients.

5.4 Non-spurious models

Our main result is Proposition 9, describing the total posterior probability assigned to models of size $p_m < p_t$ (smaller than M_t) and to those of size $p_m \geq p_t$. The rates depend on two parameters $(\underline{\lambda}, \bar{\lambda})$ that bound uniformly the non-centrality parameters λ_{tm}. We first define $(\underline{\lambda}, \bar{\lambda})$ and explain that, by Lemma S12, in Gaussian regression both are roughly proportional to n times a beta-min parameter.

Let $\underline{\lambda} = \min_{p_m < p_t} \lambda_{tm}^{\alpha}/(p_t - p_m) \geq [n\underline{v} \min_j(\theta_j^*)^2/\phi^*]^{\alpha}$ (Lemma S12), for α as in (19), where $\underline{v}$ is the smallest eigenvalue v_{tm} across models of size $p_m < p_t$. Regarding $\bar{\lambda}$, let $S_{l,j}^c \subseteq S_l^c$ be the set of non-spurious models M_m of size $p_m = l$ that contain j truly active parameters (non-zero elements in $\boldsymbol{\theta}^*$). Let $\bar{\lambda} = \min_{j \geq p_t, m \in S_{l,j}^c} \lambda_{tm}^{\alpha}/(p_t - j)$ be an analogous quantity to $\underline{\lambda}$, when minimizing over $m \in S_{l,j}^c$. By Lemma S12, we have $\bar{\lambda} \geq [n\bar{v} \min_j(\theta_j^*)^2/\phi^*]^{\alpha}$, where $\bar{v}$ is the smallest v_{tm} across models of size $p_m \in [p_t, \bar{p}]$.

Proposition 9 requires Condition F1 below to ensure that C2 in Section 2.2 holds uniformly across $p_m < p_t$. For the uniform prior $p(M_k)$, F1 is the mild requirement that $\underline{\lambda}/2 + \log p - 0.5 \log(n\tau) \gg 1$. F1 requires $\underline{\lambda}/2 + (1-\alpha) \log p - 0.5 \log(n\tau) \gg 1$ for the Beta-Binomial prior. F1 is more stringent for the Complexity prior and for large prior dispersion τ, which are sparser priors, and hence require a stronger signal. F1 is not needed for Part (ii). There, by setting sufficiently sparse priors (large τ or c) one may discard models of size $> p_t$. In particular, one could potentially set the maximum model complexity to $\bar{p} > n$ and still attain convergence in Part (ii).

(F1) Assume that $\lim_{n\to\infty} \underline{\lambda}/2 - (\alpha(1+c)-1) \log p - 0.5 \log(n\tau) = \infty$ holds for $c = -1$ when $p(M_k)$ is the uniform prior, $c = 0$ when it is the Beta-Binomial and $c > 0$ when it is the Complexity(c) prior in (6).

Proposition 9. *Let $p(M_k)$ be either the uniform or the Complexity(c) prior in (6), where $c = 0$ is the Beta-Binomial prior. Assume that for all non-spurious $m \in S^c$*

$$E_{f^*}(p(M_m \mid \mathbf{y})) \leq e^{-\frac{\lambda_{tm}^{\alpha}}{2}} (n\tau)^{-\alpha(p_m - p_t)/2} \left(\frac{p(M_m)}{p(M_t)} \right)^{\alpha},$$

for some $\alpha < 1$ and all $n \geq n_0$, where n_0 is fixed.

(i) Assume that F1 holds. Then, for the Complexity prior

$$\lim_{n\to\infty} E_{f^*} \left(\sum_{p_m=0}^{p_t-1} P(S_l^c \mid \mathbf{y}) \right) \leq e^{-\frac{\underline{\lambda}}{2} + [p_t - 1 + \alpha(1+c)] \log p + \frac{\alpha}{2} \log(n\tau)},$$

for all $n \geq n_0$. The result for the uniform prior is obtained by setting $c = -1$.

(ii) Suppose that $\lim_{n\to\infty} \bar{\lambda}/2 + \log p_t - \log(p - p_t) = \infty$. *Then, for all* $n \geq n_0$,

$$\lim_{n\to\infty} E_{f^*}\left(\sum_{p_m=p_t}^{\bar{p}} P(S_l^c \mid \mathbf{y})\right) \leq e^{-\frac{\bar{\lambda}}{2}+p_t\log(pe)} + \frac{e^{-\bar{\lambda}/2+p_t\log p_t+\log p}}{[(n\tau)^{\alpha/2}p^{\alpha(c+1)-1}]^{\bar{p}-p_t}}.$$

If $\lim_{n\to\infty} \bar{\lambda}/2 + \log p_t - \log(p - p_t) = -\infty$, *then*

$$\lim_{n\to\infty} E_{f^*}\left(\sum_{p_m=p_t}^{\bar{p}} P(S_l^c \mid \mathbf{y})\right) \leq e^{-p_t\bar{\lambda}/2}\left(\frac{1}{n\tau}\right)^{\frac{\alpha(\bar{p}-p_t)}{2}}\left(\frac{1}{p}\right)^{\alpha(c+1)(\bar{p}-p_t)-1}.$$

The results for the uniform prior are obtained by setting $c = -1$ *above.*

6 Empirical examples

We illustrate the effect of the prior formulation and signal strength on linear regression rates with two simple studies. Section 6.1 shows simulated data under orthogonal $X'X$ and Section 6.2 a setting where all pairwise correlations are 0.5, in both cases covariates are normally distributed with zero mean and unit variance. We considered three prior formulations: Zellner's prior ($\tau = 1$) coupled with either a Complexity($c = 1$) or Beta-Binomial(1,1) priors on the model space, and the pMOM prior (default $\tau = 0.348$ from Johnson and Rossell, 2010) coupled with a Beta-Binomial(1,1). For the error variance we set $p(\phi \mid M_k) \sim \text{IG}(0.005, 0.005)$. In Section 6.1 we used the methodology in Papaspiliopoulos and Rossell (2017) to obtain exact posterior probabilities, and in Section 6.2 the Gibbs sampling algorithm from Johnson and Rossell (2012) (functions `postModeOrtho` and `modelSelection` in R package `mombf`, respectively) with 10,000 iterations (i.e. $10^4 \times p$ variable updates) after a 1,000 burnin.

6.1 Orthogonal design

We considered four scenarios and simulated 100 independent datasets under each. In Scenario 1 we set $p = 100$, $n = 105$ and $p_t = 5$ truly active variables with coefficients $\theta_j^* = 0.25, 0.5, 0.75, 1, 1.5$ for $j = 1, \ldots, p_t$. In Scenario 2 again $p = 100$, $n = 105$ but coefficients were less sparse, we set $p_t = 20$ by repeating four times each coefficient in Scenario 1, i.e. $\theta_j^* = 0.25, 0.25, 0.25, 0.25, \ldots, 1.5, 1.5, 1.5, 1.5$ for $j = 1, \ldots, p_t$. Scenarios 3-4 were identical to Scenarios 1-2 (respectively) setting $p = 500$ and $n = 510$. The true error variance was $\phi^* = 1$ under all scenarios.

Figure 1 shows marginal inclusion probabilities $P(\theta_j \neq 0 \mid \mathbf{y})$. The Zellner-Complexity prior gave the smallest inclusion probabilities to truly inactive variables ($\theta_j^* = 0$), but incurred a significant loss in power to detect truly active variables. In agreement with our theory this drop was particularly severe for $p_t = 20$, e.g. when $n = 110$ inclusion probabilities were close to 0 even for fairly large coefficients. Also as predicted by the theory the power increased for $(n, p) = (510, 500)$ under all priors, but under the Zellner-Complexity prior it remained low for $\theta_j^* = 0.25$. The MOM-Beta-Binomial prior showed a good balance between power and sparsity, although for $n = 100$ it had slightly lower power to detect $\theta_j^* = 0.25$ relative to the Zellner-Beta-Binomial.

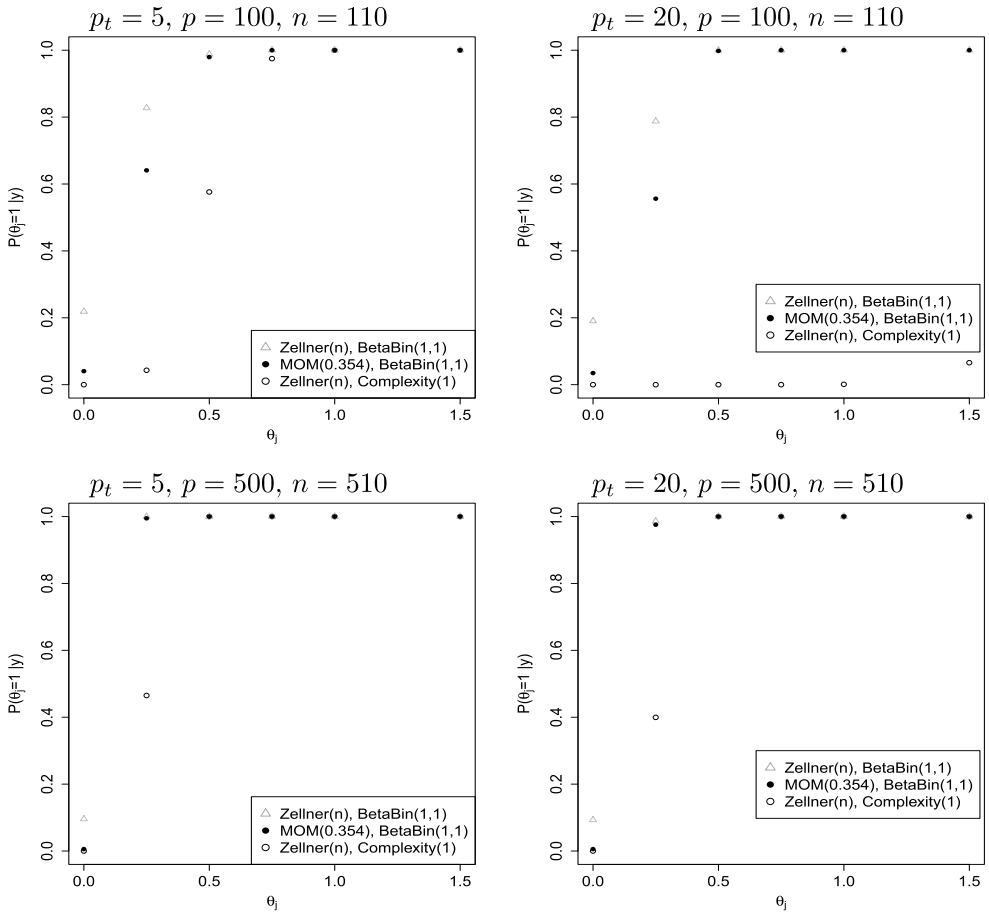

Figure 1: Average marginal inclusion probabilities under orthogonal $X'X$ and $\phi^* = 1$ for three priors: Zellner-Complexity(1), Zellner-Beta-Binomial(1,1), pMOM-Beta-Binomial(1,1). For Zellner and pMOM priors τ was set to obtain unit prior variance ($\tau = 1$, $\tau = 0.348$).

6.2 Correlated predictors

We considered normally-distributed covariates with all pairwise correlations equal to 0.5. We set $p = n$, $p_t = 10$ and considered two scenarios. In Scenario 1 $\theta_j^* = 0.5$ for all active variables $j = 1, \ldots, p_t$, whereas Scenario 2 considered weaker signals $\theta_j^* = 0.25$ again for $j = 1, \ldots, p_t$. Figure 2 shows that whichever prior achieved largest $p(M_t \mid \mathbf{y})$ depended on n and the signal strength. For large enough n all three priors discarded small non-spurious models, i.e. $\sum_{l<p_t} P(S_l^c \mid \mathbf{y})$ vanished, but the required n can be fairly large. Overall, the MOM-Beta-Binomial prior achieved a reasonable compromise between discarding spurious $m \in S$ and detecting truly active variables.

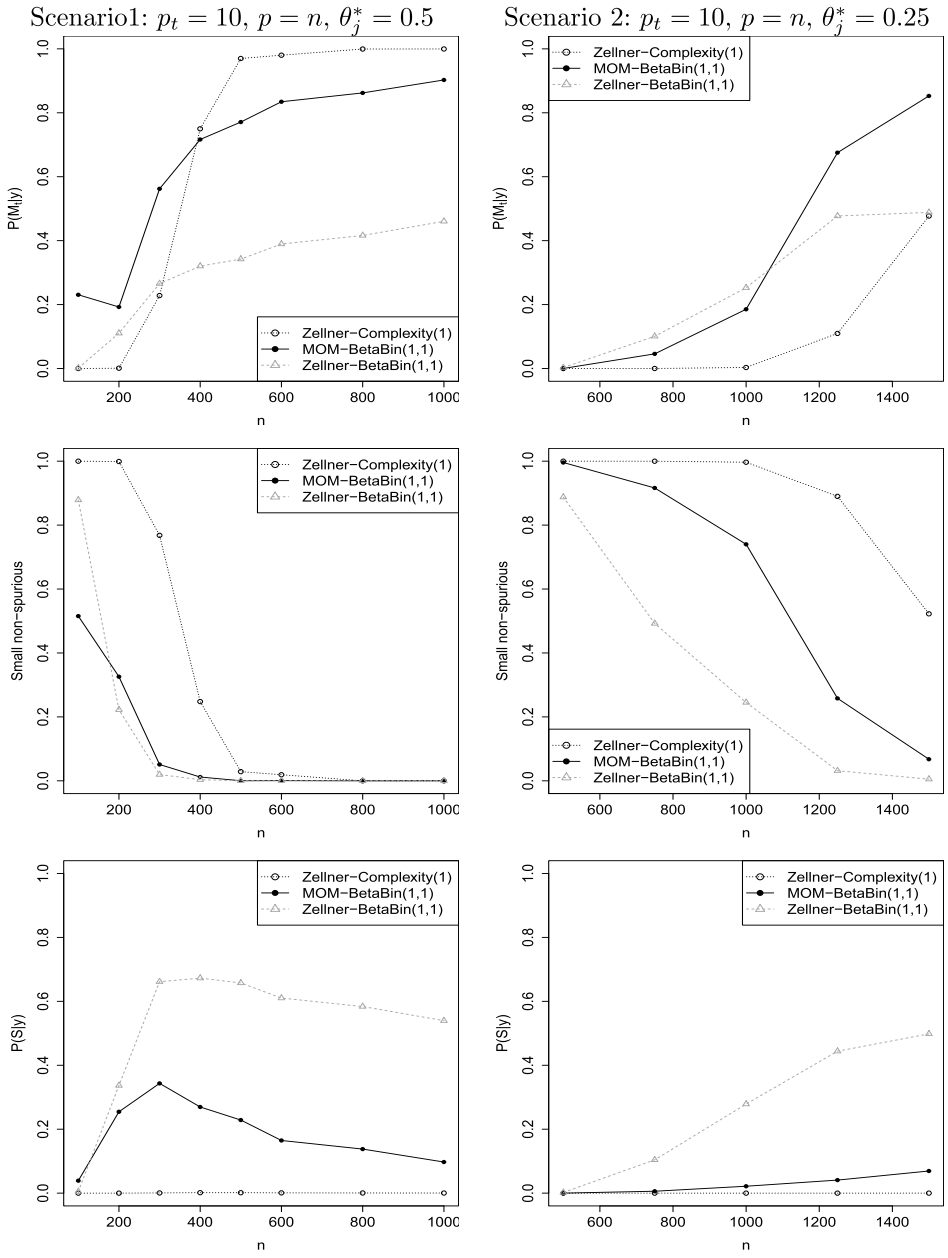

Figure 2: Linear regression simulation with pairwise correlations = 0.5. Average $p(M_t \mid \mathbf{y})$, $P(S \mid \mathbf{y})$ and $\sum_{l<p_t} P(S_l^c \mid \mathbf{y})$ under Zellner-Complexity(1), Zellner-Beta-Binomial(1,1), pMOM-Beta-Binomial(1,1) priors.

7 Discussion

We outlined a strategy to study the L_1 convergence of posterior probabilities and normalized L_0 criteria and showed that, when such convergence occurs, one can bound frequentist probabilities of correct model selection, and type I–II errors. The strategy applies generically to any model, prior and L_0 penalty, but requires non-negligible work to bound tails of Bayes factors and likelihood-ratio test statistics. Our supplementary material derives said tails for Gaussian regression, and integral bounds that may be useful for more general exponential and polynomial tails. Our rates for regression unify literature and clarify the consequences of setting sparse priors or L_0 penalties. They also clarify how convergence depends on the prior dispersion, model prior probabilities, and whether the prior is local or non-local, as well as on problem characteristics such as n, p, true sparsity p_t and the signal strength. Model misspecification also plays a role. Misspecifying the mean in (potentially non-linear) regression causes an exponential drop in power, whereas choosing the wrong error correlation can hamper type I error control.

We gave simple asymptotic expressions for popular priors and L_0 criteria. We did not study thick-tailed parameter priors, but such variations affect model selection rates only up to lower-order terms. For a wide class of local priors it is known that for spurious models $B_{mt} = O_p((\tau n)^{(-p_m - p_t)/2})$ (Dawid, 1999), which implies that L_1 convergence rates cannot be any faster. Since our obtained L_1 rates are $(\tau n)^{-\alpha(p_m - p_t)/2}$ (or tighter) for any fixed $\alpha < 1$, one cannot attain significantly faster rates with other prior families. We avoided a detailed study of eigenvalues, and referred to restricted eigenvalue conditions common in the literature. This was to highlight the main principles (the role of non-centrality parameters) and keep the results as general as possible. For a study on eigenvalues see Narisetty and He (2014) (Remarks 4–5 and Lemma 6.1), for example.

An interesting observation is that, depending on how large p is relative to n one can consider less sparse priors to detect smaller signals, which may have implications for parameter estimation. By restricting the model complexity, one can also use less sparse formulations within the set of allowed models. This is particularly relevant when the truth is non-sparse, effect sizes are small or the model's mean structure is strongly misspecified. Per our examples in this situation it can be helpful to consider strategies that exercise moderation at enforcing sparsity (e.g. the Beta-Binomial prior or the EBIC), or that do so in a data-adaptive manner (e.g. using non-local priors on parameters or empirical Bayes). Such strategies are an interesting venue for future research.

Supplementary Material

File Supplementary Material (DOI: 10.1214/21-BA1262SUPP; .pdf). Proofs and auxiliary technical results

References

Barbieri, M. and Berger, J. (2004). "Optimal predictive model selection." *The Annals of Statistics*, 32(3): 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 570

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian linear regression with sparse priors." *The Annals of Statistics*, 43(5): 1986–2018. MR3375874. doi: https://doi.org/10.1214/15-AOS1334. 567, 568, 573, 574

Chae, M., Lin, L., and Dunson, D. (2016). "Bayesian sparse linear regression with unknown symmetric error." *arXiv*, 1608.02143: 1–34. MR3994400. doi: https://doi.org/10.1093/imaiai/iay022. 567

Dawid, A. (1999). "The trouble with Bayes factors." Technical report, University College London. 567, 575, 589

Fernández, C., Ley, E., and Steel, M. (2001). "Benchmark priors for Bayesian model averaging." *Journal of Econometrics*, 100: 381–427. MR1820410. doi: https://doi.org/10.1016/S0304-4076(00)00076-2. 572, 584

Foster, D. and George, E. (1994). "The risk inflation criterion for multiple regression." *The Annals of Statistics*, 22(4): 1947–1975. MR1329177. doi: https://doi.org/10.1214/aos/1176325766. 567, 572

Gao, C., van der Vaart, A. W., and Zhou, H. H. (2015). "A general framework for Bayes structured linear models." *arXiv*, 1506.02174: 1–44. MR4152123. doi: https://doi.org/10.1214/19-AOS1909. 567

Johnson, V. and Rossell, D. (2010). "On the use of non-local prior densities for Default Bayesian Hypothesis Tests." *Journal of the Royal Statistical Society B*, 72: 143–170. MR2830762. doi: https://doi.org/10.1111/j.1467-9868.2009.00730.x. 566, 567, 586

Johnson, V. and Rossell, D. (2012). "Bayesian model selection in high-dimensional settings." *Journal of the American Statistical Association*, 24(498): 649–660. MR2980074. doi: https://doi.org/10.1080/01621459.2012.682536. 566, 567, 572, 578, 586

Martin, R., Mess, R., and Walker, S. G. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. MR3624879. doi: https://doi.org/10.3150/15-BEJ797. 573

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). "Optimal sample size for multiple testing: the case of gene expression microarrays." *Journal of the American Statistical Association*, 99(468): 990–1001. MR2109489. doi: https://doi.org/10.1198/016214504000001646. 570

Narisetty, N. and He, X. (2014). "Bayesian variable selection with shrinking and diffusing priors." *The Annals of Statistics*, 42(2): 789–817. MR3210987. doi: https://doi.org/10.1214/14-AOS1207. 567, 568, 572, 574, 589

Papaspiliopoulos, O. and Rossell, D. (2017). "Bayesian block-diagonal variable selection and model averaging." *Biometrika*, 104(2): 343–359. MR3698258. doi: https://doi.org/10.1093/biomet/asx019. 586

Petrone, S., Rousseau, J., and Scricciolo, C. (2014). "Bayes and empirical Bayes: do

they merge?" *Biometrika*, 101(2): 285–302. MR3215348. doi: https://doi.org/10.1093/biomet/ast067. 568

Rockova, V. and van der Pas, S. (2017). "Posterior concentration for Bayesian regression trees and their ensembles." *arXiv*, 1708.08734: 1–40. MR4134788. doi: https://doi.org/10.1214/19-AOS1879. 567

Rossell, D., Abril, O., and Bhattacharya, A. (2020). "Approximate Laplace approximations for scalable model selection." *arXiv*, 2012.07429: 1–72. 579

Rossell, D. and Telesca, D. (2017). "Non-local priors for high-dimensional estimation." *Journal of the American Statistical Association*, 112: 254–265. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 578

Rossell, D. (2021). "Supplementary Material of "Concentration of Posterior Model Probabilities and Normalized L_0 Criteria"." *Bayesian Analysis*. doi: https://doi.org/10.1214/21-BA1262SUPP. 568

Schwarz, G. (1978). "Estimating the dimension of a model." *Annals of Statistics*, 6: 461–464. MR0468014. 565, 572

Scott, J. and Berger, J. (2010). "Bayes and empirical Bayes multiplicity adjustment in the variable selection problem." *The Annals of Statistics*, 38(5): 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 573

Shin, M., Bhattacharya, A., and Johnson, V. (2018). "Scalable Bayesian variable selection using nonlocal prior densities in ultrahigh-dimensional settings." *Statistica Sinica*, 28(2): 1053–1078. MR3791100. 567

Wainwright, M. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press. MR3967104. doi: https://doi.org/10.1017/9781108627771. 581

Wainwright, M. J. (2009). "Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting." *IEEE Transactions on Information Theory*, 55(12): 5728–5741. MR2597190. doi: https://doi.org/10.1109/TIT.2009.2032816. 574

Yang, Y. and Pati, D. (2017). "Bayesian model selection consistency and oracle inequality with intractable marginal likelihood." *arXiv*, 1701.00311: 1–38. MR2696783. 567, 574

Yang, Y., Wainwright, M., and Jordan, M. (2016). "On the computational complexity of high-dimensional Bayesian variable selection." *The Annals of Statistics*, 44(6): 2497–2532. MR3576552. doi: https://doi.org/10.1214/15-AOS1417. 567, 574

Acknowledgments

The author thanks Gabor Lugosi and James O. Berger for helpful discussions, and the Editors and Referees for invaluable feedback in improving the exposition of this manuscript.