Electronic Journal of Statistics
Vol. 18 (2024) 3212–3246
ISSN: 1935-7524
https://doi.org/10.1214/24-EJS2274

Empirical Bayes inference in sparse high-dimensional generalized linear models

Yiqi Tang

Department of Statistics, Colby College, Waterville, ME, US,
e-mail: ytang@colby.edu

Ryan Martin

Department of Statistics, North Carolina State University, Raleigh, NC, US,
e-mail: rgmarti3@ncsu.edu

Abstract: High-dimensional linear models have been widely studied, but the developments in high-dimensional generalized linear models, or GLMs, have been slower. In this paper, we propose an empirical or data-driven prior leading to an empirical Bayes posterior distribution which can be used for estimation of and inference on the coefficient vector in a high-dimensional GLM, as well as for variable selection. We prove that our proposed posterior concentrates around the true/sparse coefficient vector at the optimal rate, provide conditions under which the posterior can achieve variable selection consistency, and prove a Bernstein–von Mises theorem that implies asymptotically valid uncertainty quantification. Computation of the proposed empirical Bayes posterior is simple and efficient, and is shown to perform well in simulations compared to existing Bayesian and non-Bayesian methods in terms of estimation and variable selection.

MSC2020 subject classifications: Primary 62C12, 62E20, 62J12.
Keywords and phrases: Data-dependent prior, logistic regression, model selection, Poisson log-linear model, posterior asymptotics.

Received May 2023.

1. Introduction

Generalized linear models, or GLMs, which include normal, logistic, and Poisson regression as important special cases, are essential tools for data analysis in all quantitative fields; see, e.g., McCullagh and Nelder (1989) for a thorough introduction. In modern applications, it is common for the number of predictor variables, p, to greatly exceed the sample size, n; this is the so-called "$p \gg n$" problem. For example, logistic regression for presence/absence of a trait, with gene expression levels as covariates is one such problem. By now there is an enormous body of literature on the $p \gg n$ problem in the case of normal linear regression. Popular methods include the lasso and its variants (Hastie, Tibshirani and Friedman, 2009). Bayesian efforts in the normal linear regression problem can be split into two categories: those based on shrinkage priors such as

the *horseshoe* (Bhadra et al., 2017, 2019a,b; Carvalho, Polson and Scott, 2010; van der Pas, Szabó and van der Vaart, 2017) and those based on spike-and-slab mixture priors (Belitser and Ghosal, 2020; Castillo, Schmidt-Hieber and van der Vaart, 2015; Castillo and van der Vaart, 2012; George and McCullogh, 1993). On the non-Bayesian side, a number of these methods have been extended from the normal linear model to other GLMs, e.g., the R package `glmnet` (Friedman, Hastie and Tibshirani, 2010) offers a comprehensive lasso-based toolkit, but the Bayesian developments in this direction are still limited; the methods tend to be tailored to logistic regression (Cao and Lee, 2020; Narisetty, Shen and He, 2019) and the theory focuses mostly on variable selection; the one exception, Jeong and Ghosal (2021) gave results on posterior concentration rates in high-dimensional GLMs but did not address model selection or implementations of the Bayesian solutions they studied. The main goal of the present paper is to offer a Bayesian (or at least Bayesian-like) solution to the high-dimensional GLM problem, having both strong theoretical support and an efficient numerical implementation that is not tailored to any one specific GLM.

A challenge for Bayesian inference in high-dimensional models is that the priors for which posterior computations are relatively simple generally do not produce good theoretical posterior concentration properties and, vice versa, the priors with theoretical justification make posterior computations difficult and expensive. For example, in normal linear regression, a computationally simple prior is a mixture of conjugate mean-zero normal priors on the coefficients, each component corresponding to a subset of active variables, but it has been shown (Castillo and van der Vaart, 2012) that the thin tails of the normal can lead to sub-optimal theoretical properties. A theoretically better choice of prior is one with heavier, Laplace-type tails, but this added complexity translates to higher computational cost. To overcome this obstacle, Martin, Mess and Walker (2017) proposed the idea of using the data to properly center the prior. The motivation is that the tails of the prior should not matter if the prior is strategically centered, so then the computationally simpler conjugate normal priors could still be used. Centering the model-specific conjugate normal priors on the corresponding least-squares estimators makes the approach *empirical Bayes*, in a certain sense, and the previous authors show that the corresponding empirical Bayes posterior has optimal asymptotic concentration properties and has strong empirical performance compared to existing Bayesian and non-Bayesian methods. In other words, the double-use of data—in the prior and in the likelihood—does not hurt the method's performance in any way; in fact, one could argue that the double-use of data actually helps. Beyond the normal linear model (Martin, Mess and Walker, 2017; Martin and Tang, 2020), there is strong general theory in Martin and Walker (2019) and promising results in applications, including Liu and Martin (2019); Liu, Martin and Shen (2023); Martin (2019).

The goal here is to develop the aforementioned empirical Bayes strategy for the case of high-dimensional GLMs. In Section 2, we introduce the set up of the GLM problem and review the empirical Bayes approach for linear regression. In Section 3, we present our empirical Bayes GLM, including the particular choice of data-driven prior, the corresponding empirical Bayes posterior, and

our proposed computational strategy. The key challenge in the present GLM case compared to previous efforts in the linear model setting is that the models are sufficiently complicated that there is no conjugacy and, therefore, no posterior computations can be done in closed form. Here the "informativeness" of the data-driven prior allows for some simple and accurate approximations. In Section 4, we offer theoretical support for our proposed solution. In particular, we have two basic kinds of posterior concentration results: those for the GLM coefficients, which are relevant to estimation, and those for the so-called configuration, or active set, which are relevant to variable selection. In the former case, we give sufficient conditions for the posterior to concentrate around the true (sparse) coefficient vector at rates equivalent to those established in, e.g., Jeong and Ghosal (2021), which agree with the minimiax optimal rates in the linear model setting. In the latter case, we give sufficient conditions (e.g., on the size of the smallest non-zero coefficient), comparable to those in Narisetty, Shen and He (2019), that ensure our marginal posterior for the configuration will concentrate on a set that contains exactly the true set of active variables. While the results we obtain are similar to those found elsewhere in the literature, it is important to emphasize that, to our knowledge, there is no single Bayesian method for which both of these kinds of posterior concentration properties have been established. And, again, the lack of conjugacy and closed-form expressions for (parts of) the posterior distribution creates some challenges compared to the linear model case, so some relatively new proof techniques—in-probability versus in-expectation bounds—are needed compared to Martin, Mess and Walker (2017) and Martin and Walker (2019). Beyond posterior concentration, we offer a Bernstein–von Mises theorem, comparable to that for linear models in Castillo, Schmidt-Hieber and van der Vaart (2015), which establishes a large-sample Gaussian approximation of the posterior and, among other things, implies that the marginal posterior uncertainty quantification (e.g., credible intervals) are asymptotically valid. Section 5 investigates the numerical performance of the proposed method in terms of estimation and variable selection in logistic and Poisson regression compared to existing methods. Finally, some concluding remarks are given in Section 6; proofs are presented in the Appendix.

2. Setup and background

2.1. Problem setup

Suppose $y_1, \ldots, y_n$ are independent, where y_i has density/mass function

$$f_{\eta_i}(y_i) \propto \exp\{y_i\eta_i - b(\eta_i)\}, \quad i = 1, \ldots, n,$$

indexed by the real-valued natural parameters $\eta_1, \ldots, \eta_n$, where b is a known, strictly convex function, i.e., $\ddot{b}(\eta) > 0$ for all η. The interpretation of b is that the expected value and variance of y_i are $\dot{b}(\eta_i)$ and $\ddot{b}(\eta_i)$, respectively. If the response y_i has an associated vector of predictor variables $x_i \in \mathbb{R}^p$, for $i = 1, \ldots, n$, then

introduce a coefficient vector $\theta \in \mathbb{R}^p$ into the model via the relationship

$$(h \circ \dot{b})(\eta_i) = x_i^\top \theta, \quad i = 1, \ldots, n,$$

where h is a given bijection called the *link function*. The "canonical" link is $h = \dot{b}^{-1}$ and, in this case, the above simplifies to $\eta_i = x_i^\top \theta$. But there are cases when non-canonical link functions h are used and our theory here allows for this. For example, in the Bernoulli data case, our results apply to both logistic regression (canonical link) and probit regression (non-canonical link).

Write $f_\theta(y_i \mid x_i)$ for the density/mass function determined by the parameter vector θ. Then the likelihood and log-likelihood functions are, respectively,

$$L_n(\theta) = \prod_{i=1}^n f_\theta(y_i \mid x_i) \quad \text{and} \quad \ell_n(\theta) = \log L_n(\theta).$$

To ease the notation, set $\xi(\eta) = (h \circ \dot{b})(\eta)$; if h is the canonical link, then ξ is the identity mapping. Then the maximum likelihood estimator (MLE) $\hat{\theta}$ is the solution to the equation

$$\dot{\ell}_n(\hat{\theta}) = 0 \iff \{y - \dot{b}(X\hat{\theta})\}^\top \mathrm{diag}\{\dot{\xi}(X\theta)\}X = 0, \tag{1}$$

where $\mathrm{diag}(\cdot)$ denotes the diagonal matrix determined by its vector argument, and $\dot{\xi}(X\theta)$ is the application of $\dot{\xi}$ to each entry in the vector $X\theta$; if h is the canonical link, then $\dot{\xi} \equiv 1$. The negative second derivative of the log-likelihood function is a matrix

$$J_n(\theta) = -\ddot{\ell}_n(\theta) = X^\top W(\theta) X,$$

where $W(\theta)$ is a diagonal matrix with entries

$$W_{ii}(\theta) = w(x_i^\top \theta) = \dot{u}(x_i^\top \theta)\dot{\xi}(x_i^\top \theta), \quad i = 1, \ldots, n,$$

where $u = h^{-1}$ is the inverse link function. When h is the canonical link, we get $W_{ii}(\theta) = \ddot{b}(x_i^\top \theta)$. The observed Fisher information matrix is $J_n(\hat{\theta}) = X^\top W(\hat{\theta}) X$, the negative Hessian of the log-likelihood function evaluated at $\hat{\theta}$, which is positive definite. For example, in binary regression with the canonical logit link, the W matrix has diagonal entries

$$W_{ii}(\theta) = \frac{\exp(x_i^\top \theta)}{\{1 + \exp(x_i^\top \theta)\}^2}, \quad i = 1, \ldots, n,$$

which is bounded as a function of θ. Similarly, in Poisson regression with the canonical log link, the W matrix has diagonal entries

$$W_{ii}(\theta) = \exp(x_i^\top \theta), \quad i = 1, \ldots, n.$$

Our interest is in high-dimensional cases where the number of predictor variables p exceeds the sample size n. In such cases, the direct model fitting as described above cannot be done; intuitively, the data is not informative enough

to reliably learn the very high-dimensional parameter θ. To side-step this obstacle, we shall assume, as is common, that the GLM is *sparse* in the sense that most of the θ coefficients are 0 (or at least negligible). Then there is an "active set" of the predictor variables corresponding to the non-zero coefficient values, but this is unknown because θ itself is unknown. To avoid overusing the term "model", here we will call the unknown active set of variables a *configuration* and denote it generically by S. Since the configuration is unknown, it makes sense to decompose the unknown θ as (S, θ_S), where S is a set of indices that corresponds to the "active" coefficients in θ and θ_S is the vector of coefficients that correspond to a configuration S. Then, the above notation can be adjusted in a natural way. That is, the likelihood and log-likelihood functions can be written as $L_n(S, \theta_S)$ and $\ell_n(S, \theta_S)$, respectively, the configuration-specific MLE is $\hat{\theta}_S$, the observed Fisher information is $J_n(S, \hat{\theta}_S)$, etc. Throughout we will write $|S|$ for the cardinality of a configuration S, $\theta^\star$ for the true coefficients, and $S^\star$ for the true configuration; in some cases, we will write $s = |S|$ for the configuration size and, naturally, $s^\star = |S^\star|$ for the true sparsity level. Also, with a slight abuse of this notation, we will occasionally write $S(\theta) = \{j : \theta_j \neq 0\}$ to denote the configuration corresponding to a given coefficient vector.

All of the results in Narisetty, Shen and He (2019) and in Cao and Lee (2020) are established by focusing on configurations S that are supersets of the true $S^\star$. Since our asymptotic analysis goes beyond those in the aforementioned references, we will also need to consider configurations that are not supersets of $S^\star$, so some generalizations are in order. The key observation is that, if $S \not\supset S^\star$, then $\hat{\theta}_S$ is *not* estimating $\theta^\star_S$. Instead, $\hat{\theta}_S$ is estimating the minimizer of the Kullback–Leibler divergence which, in our case, is a solution to the equation

$$\{\dot{b}(X\theta^\star) - \dot{b}(X_S\theta_S)\}^\top \mathrm{diag}\{\dot{\xi}(X_S\theta_S)\}X_S = 0.$$

Let $\theta^\dagger_S$ denote this solution. Note that, first, this notation is slightly misleading, since there is no "full vector" $\theta^\dagger$ of which $\theta^\dagger_S$ is the S-specific sub-vector; instead, there is a different $|S|$-vector $\theta^\dagger_S$ for each S. Second, if S is a superset of $S^\star$, then $\theta^\dagger_S = \theta^\star_S$; this explains why Narisetty, Shen and He (2019) do not need $\theta^\dagger_S$ in their superset-only analysis. Lemmas 1–2 in Appendix A.1 below generalize Lemmas A1 and A3 in Narisetty, Shen and He (2019) to cover the case where $\hat{\theta}_S$ is compared to $\theta^\dagger_S$ rather than $\theta^\star_S$.

2.2. *Empirical priors*

A very general construction of empirical or data-driven priors and the corresponding posterior concentration rate theory is given in Martin and Walker (2019). There are elements of their general formulation used here in this application, but it is not necessary to review these for our purposes; see Appendix A.3 for some of the relevant technical details. For this reason, we focus our review here on the high-dimensional linear regression case investigated in Martin, Mess and Walker (2017) and in Martin and Tang (2020).

The normal linear model assumes that $y_i \sim \mathsf{N}(x_i^\top \theta, \sigma^2)$, independent, for $i = 1, \ldots, n$; for the discussion here, we take σ to be known, but the case where σ is unknown and assigned a prior has been considered in Martin and Tang (2020) and Fang and Ghosh (2023). As above, the idea is to reinterpret the high-dimensional coefficient vector θ as the pair (S, θ_S), the configuration and configuration-specific parameters. Then the natural Bayesian approach to this would be to specify the prior hierarchically: first introduce a marginal prior for S, then a conditional prior for θ_S, given S.

1. For the marginal prior for S, the previous authors suggest the sparsity-encouraging prior mass function $\pi(S) \propto \binom{p}{|S|}^{-1} f_n(|S|)$, i.e., a rapidly decaying marginal prior mass function f_n for the configuration size $|S|$ times a conditional uniform prior for S of a given size $|S|$. Here we consider the *complexity prior* that takes
$$f_n(s) \propto p^{-\beta s}, \quad s = 1, 2, \ldots, s_{\max},$$
where $s_{\max}$ is a specified maximum complexity, which could be just the trivial choice p or something smaller, such as $s_{\max} = \text{rank}(X)$. The hyper-parameter $\beta > 0$ plays a crucial role and is discussed further below.
2. The conditional prior for θ_S, given S, is where the data enters into the prior. To avoid the sub-optimal behavior of the thin-tailed Gaussian prior while simultaneously retaining the computational convenience of the conjugate Gaussian form, Martin, Mess and Walker (2017) suggested
$$(\theta_S \mid S) \sim \mathsf{N}_{|S|}\big(\hat{\theta}_S, \gamma (X_S^\top X_S)^{-1}\big), \quad S \subset \{1, 2, \ldots, p\},$$
where $\hat{\theta}_S$ is the S-specific least squares estimator—which makes the prior data-dependent—and $\gamma > 0$ is a tuning parameter to be specified.

This empirical prior is combined with the data-driven likelihood basically according to Bayes's theorem, resulting in a (data-dependent) probability distribution Π^n that can be used for making inference on θ. In particular, the corresponding marginal posterior for the configuration $S = S(\theta)$ can be used for model selection purposes. Computation is simple and fast, via an efficient Metropolis–Hastings-style Markov chain Monte Carlo procedure, thanks to the conjugacy of the data-centered empirical prior. An associated R package `ebreg` (Tang and Martin, 2021) is also available.

The data-driven prior distribution, among other things, implies that this is not a genuinely "Bayesian" solution, so we cannot expect that it automatically inherits the good properties that Bayesian solutions typically enjoy. But Martin, Mess and Walker (2017) demonstrate theoretically that the posterior Π^n achieves the adaptive, minimax optimal asymptotic concentration rate at the true/sparse $\theta^\star$. In other words, there is no other posterior distribution—genuinely Bayesian or otherwise—that can concentrate around the true $\theta^\star$ any faster. They also established that, under suitable conditions, the aforementioned marginal posterior distribution for the configuration concentrates its mass

asymptotically on the true configuration, thus providing consistent model selection. We establish similar properties here in this paper, but for the case of high-dimensional GLMs, so more details about the construction and the results will be given below.

3. Empirical Bayes for high-dimensional GLMs

3.1. Prior distribution

As before, write the structured, high-dimensional coefficient vector θ as (S, θ_S), where the configuration S is a subset of $\{1, 2, \ldots, p\}$ and θ_S is a configuration-specific parameter that respects the structure determined by S. Based on this decomposition, we follow Martin, Mess and Walker (2017) and proceed with specification of the (empirical) prior Π_n for θ hierarchically. First, the marginal prior for S has mass function

$$\pi_n(S) \propto \tbinom{p}{|S|}^{-1} p^{-\beta|S|}, \quad S \subset \{1, 2, \ldots, p\} \text{ such that } |S| \leq s_n, \tag{2}$$

where $\beta > 0$ is a constant to be specified and s_n is a deterministic, diverging sequence. This is the same as the marginal prior for S in Section 2.2. Second, the data-driven conditional prior for θ_S, given S, is

$$(\theta_S \mid S) \sim \Pi_{n,S} := \mathsf{N}_{|S|}\big(\hat{\theta}_S, \gamma J_n(S, \hat{\theta}_S)^{-1}\big), \tag{3}$$

where $\gamma > 0$ is a constant to be specified. Above, the prior center is $\hat{\theta}_S$, the data-driven maximum likelihood estimator for the given S and, similarly, $J_n(S, \hat{\theta}_S)$ is the corresponding data-driven observed Fisher information matrix. So this empirical conditional prior has a similar form as that in Section 2.2 for the linear model case, but the specific pieces that go into it are just different here in the GLM setting.

As indicated above, the conditional prior for θ_S, given S, is data-driven in the sense that both the prior mean vector and covariance matrix depend on data y through $\hat{\theta}_S$. Also, recall that the diagonal entries of the information matrix $J_n(S, \hat{\theta}_S)$ are growing like $O(n)$, so the prior covariance matrix is rather small. This is counter-intuitive when the prior center is a fixed constant, but makes perfect sense when the prior mean is data-driven. That is, we "believe in" the data-driven prior center so the prior ought to be relatively tightly concentrated there; plus, if the prior covariance matrix were large, then there would be no point in/benefit to the data-driven prior centering.

To summarize, the sparsity-encouraging empirical prior $\theta \sim \Pi_n$ is given by

$$\Pi_n(d\theta) = \sum_S \pi_n(S)\, \mathsf{N}_{|S|}(d\theta_S \mid \hat{\theta}_S, \gamma J_n(S, \hat{\theta}_S)^{-1}) \otimes \delta_{0_{S^c}}(d\theta_{S^c}), \tag{4}$$

where the sum is over all configurations S supported by the prior mass function π_n and $\delta_{0_{S^c}}(d\theta_{S^c})$ denotes a Dirac point mass differential term at the origin

for the component θ_{S^c}. The empirical prior depends on the hyperparameters (β, γ, s_n) which will be discussed in more detail below. As is common, the theory offers some guidance on how to choose these hyperparameters, but not enough to fully determine their values.

3.2. Posterior distribution

If L_n denotes the GLM's likelihood, and Π_n the empirical prior in (4) with the mixture form, then the proposed posterior distribution for θ is defined as

$$\Pi^n(d\theta) = \frac{L_n(\theta)^\alpha \, \Pi_n(d\theta)}{\int_{\mathbb{R}^p} L_n(\vartheta)^\alpha \, \Pi_n(d\vartheta)}, \quad \theta \in \mathbb{R}^p, \tag{5}$$

where $\alpha \in (0,1)$ is a fixed constant that can be chosen arbitrarily close to 1.

Our proposed use of a power-likelihood in the Bayesian updating might make some readers uncomfortable, so some remarks on this are in order. First, with our use of a data-driven prior, the lines between the "likelihood part" and "prior part" in Bayes's formula have been blurred. So, one can easily adjust the above formula so that it is the ordinary likelihood L_n combined with a slightly modified empirical prior $\widetilde{\Pi}_n(d\theta) \propto L_n(\theta)^{-(1-\alpha)} \, \Pi_n(d\theta)$, and get exactly the same posterior. In our opinion, the version in (5) is preferred because it is more transparent. Second, the original motivation for choosing $\alpha < 1$ (e.g., Martin and Walker, 2014) was to prevent the posterior from tracking the data too closely as a result of its double-use of the data; this is discussed extensively in, e.g., Walker and Hjort (2001) and Walker, Lijoi and Prünster (2005). Relatively recent evidence suggests that a choice of $\alpha < 1$ is not necessary for good posterior concentration properties (e.g., Belitser and Ghosal, 2020; Belitser and Nurushev). But what is needed to accommodate $\alpha = 1$ adds significant technical complications without any benefits in terms of faster rates, etc. Indeed, in the case of GLMs, Jeong and Ghosal (2021) pointed out that there are non-trivial differences in the strength of their theoretical results for $\alpha = 1$ versus $\alpha < 1$; in particular, much stronger conditions are required in some cases to get the same concentration rates using $\alpha = 1$ compared to $\alpha < 1$. Finally, there may even be some practical benefits to the use of power-likelihoods, e.g., in terms of robustness (e.g., Miller and Dunson, 2019; Syring and Martin, 2018, 2023), "safety" (e.g., Grünwald and Mehta, 2020; Grünwald and van Ommen, 2017), and/or uncertainty quantification (e.g., Martin and Ning, 2020; Martin and Tang, 2020).

Back to the task at hand, thanks to the parametrization $\theta = (S, \theta_S)$ and the hierarchical prior, a marginal posterior distribution for S is available, i.e.,

$$\begin{aligned} \pi^n(S) &= \frac{\pi_n(S) \int_{\mathbb{R}^{|S|}} L_n(S,\theta_S)^\alpha \, \mathsf{N}_{|S|}(\theta_S \mid \hat{\theta}_S, \gamma J_n(S,\hat{\theta}_S)^{-1}) \, d\theta_S}{\sum_R \pi_n(R) \int_{\mathbb{R}^{|R|}} L_n(R,\theta_R)^\alpha \, \mathsf{N}_{|R|}(\theta_R \mid \hat{\theta}_R, \gamma J_n(R,\hat{\theta}_R)^{-1}) \, d\theta_R} \\ &\propto \pi_n(S) \int_{\mathbb{R}^{|S|}} L_n(S,\theta_S)^\alpha \, \mathsf{N}_{|S|}(\theta_S \mid \hat{\theta}_S, \gamma J_n(S,\hat{\theta}_S)^{-1}) \, d\theta_S, \end{aligned}$$

where $x \mapsto \mathsf{N}(x \mid \mu, \sigma^2)$ denotes a $\mathsf{N}(\mu, \sigma^2)$ density. Although there is generally no closed-form expression for the last integral above, a formal Laplace approximation gives a nice, simple expression:

$$\pi^n(S) \propto \pi(S)\,(1+\alpha\gamma)^{-|S|/2}\,L_n(S,\hat{\theta}_S)^{\alpha}, \quad \text{all large } n. \tag{6}$$

Cao and Lee (2020) also employ a Laplace approximation, though their expression is different because their prior is different. Of course, it is too much to expect that this approximation be accurate simultaneously across *all* configurations, but we do not need such a strong result. It is enough that this approximation be accurate over a class of relatively small configurations (e.g., Barber, Drton and Tan, 2016; Shun and McCullagh, 1995). This class is described in Section 4 below, and a precise result on the Laplace approximation's accuracy is given in Lemma 4 in Appendix A.2. Details on posterior computation are given next.

3.3. Computation

If variable selection is the goal, then focus is on the marginal posterior for S. We propose to use the Laplace approximation of $\pi^n(S)$ in (6) in a simple Metropolis–Hastings Markov chain Monte Carlo scheme. Note that this does not require that we can evaluate the normalizing constant implicit in (6).

Given a proposal function $q(S' \mid S)$, one iteration of the Metropolis–Hastings algorithm is as follows:

1. Given a current state S, sample $S' \sim q(\cdot \mid S)$.
2. Go to the new state S' with probability

$$\min\Big\{1, \frac{\pi^n(S)}{\pi^n(S')}\frac{q(S' \mid S)}{q(S \mid S')}\Big\},$$

where $\pi^n(S)$ is defined in (6). Otherwise, stay in the current state S.

We use a proposal distribution that is symmetric, one that samples S' uniformly from those that differ from S in exactly one position. This simplifies our computations as the q-ratio above is simply 1. This process is repeated M times, which yields a sample of configurations $S^{(1)}, \ldots, S^{(M)}$ from our posterior distribution $\pi^n(S)$; this is after a burn-in period that excludes the first 20% of the samples generated. This is relatively efficient, since the likelihood-based ingredients—the MLE $\hat{\theta}_S$, the Fisher information $J_n(S, \hat{\theta}_S)$, etc.—only need to be evaluated for the configurations S that are selected in the MCMC.

For variable selection, a relevant quantity is the *inclusion probability* associated with each candidate variable $j = 1, \ldots, p$ that could be included. The simplest way to express this is as the posterior probability that coefficient θ_j attached to variable X_j is non-zero. With a slight abuse of notation, we will refer to the inclusion probability for variable j as $\pi^n(j)$, which is

$$\pi^n(j) := \pi^n(\{S : S \ni j\}) \approx \frac{1}{M}\sum_{m=1}^{M} 1\{S^{(m)} \ni j\}, \tag{7}$$

where the right-hand side is the Monte Carlo approximation, the proportion of those configurations $S^{(m)}$'s drawn that include variable j. From here, a natural variable selection procedure would include all those variables for which the inclusion probability exceeds a specified threshold; see Section 5 below.

If there is also interest in estimation of/inference on the coefficients θ, then one can easily augment the above Monte Carlo scheme by inserting a rejection sampling step where, given S, the corresponding coefficient θ_S is drawn from the corresponding conditional posterior. Thanks to the empirical prior structure, a simpler alternative would be to average the S-specific MLEs with respect to the the posterior distribution π^n of S, which is akin to the exponential aggregation in, e.g., Rigollet and Tsybakov (2012). Similarly, if the goal is prediction of a new response $\tilde{y}$ associated with a new set of covariate values $\tilde{X}$, then a third step is added where in a draw is made from the posited model given $(S, \theta_S, \tilde{X})$.

When S is the sole focus, alternatively, a shotgun stochastic search (SSS) algorithm can be employed, as in Algorithm 1 of Cao and Lee (2020). In SSS, models that are neighbors of a selected model are evaluated, and then the next chosen model is sampled from these neighbors proportional to their posterior probabilities, repeating the process. This is more efficient than the aforementioned Monte Carlo strategy in effectively exploring the configuration space. In practice, however, we found that our method with $M = 10^4$ posterior samples (with a 20% burn-in), computes significantly faster; this is likely due to SSS having to compute posterior probabilities for all the neighboring models.

4. Asymptotic properties

4.1. Setup and conditions

Narisetty, Shen and He (2019) and Cao and Lee (2020) investigate certain asymptotic properties of their proposed posterior for θ, but they focus exclusively on (a) logistic regression and (b) results concerning the marginal posterior π^n for the configuration S. Here we extend the analysis beyond the logistic regression case to arbitrary GLMs as described above, with arbitrary link functions, and establish conditions under which our proposed posterior distribution Π^n for θ concentrates around the true $\theta^\star$ at (nearly) the optimal rate (e.g., Rigollet, 2012), adaptive to the unknown sparsity level $|S(\theta^\star)|$.

Below are two conditions crucial to the developments here and in Narisetty, Shen and He (2019) and Cao and Lee (2020). These can be roughly classified as conditions on the dimension of the problem and on the design matrix X. The third condition concerns the hyperparameters in our empirical prior.

Condition 1. $p = p_n \to \infty$ and $\log p = o(n)$ as $n \to \infty$.

Condition 2. There exists $K > 0$, $\lambda > 0$, $\tau \in [0, 1]$, and $\tau' \in (\tau, 1]$ such that

(a) the entries in X are bounded in absolute value by K

(b) if $\lambda_{\min}$ and $\lambda_{\max}$ are operators that return the smallest and largest eigenvalues of their arguments, respectively, then

$$\lambda \leq \min_{S:|S|\leq|S^\star|+s_n} \lambda_{\min}\{n^{-1}J_n(S,\theta_S^\dagger)\} \leq \Lambda_{|S^\star|+s_n} \leq K^2(n/\log p)^\tau, \quad (8)$$

where

$$\Lambda_k = \max_{S:|S|\leq k} \lambda_{\max}\{n^{-1}J_n(S,\theta_S^\dagger)\}, \quad k = 1,2,\ldots,$$

and

$$s_n = O\big((n/\log p)^{(1-\tau')/2}\big), \quad n \to \infty. \quad (9)$$

Condition 3. The power $\alpha > 0$ in (5) is strictly less than 1. Also:

(a) the prior hyperparameter $\gamma > 0$ in (3) satisfies

$$\gamma = O(\Lambda^2_{2s_n}), \quad n \to \infty, \quad (10)$$

where Λ_s and s_n are as defined above, and

(b) the prior hyperparameter $\beta > 0$ in (2) is such that $p^{\beta-\kappa} > s_n$, where $\kappa > 1$ is the constant specified in Lemma 3 and s_n is as in (9).

Conditions 1 and 2(a) are standard in the high-dimensional inference literature. The lower bound in Condition 2(b) is a type of restricted eigenvalue condition, similar to those assumed in Narisetty, Shen and He (2019) and Cao and Lee (2020). The upper bound weakens and generalizes the "bounded eigenvalue condition" in, e.g., Bondell and Reich (2012). Since $J_n(S,\theta_S^\dagger) = X_S^\top W(S,\theta_S^\dagger)X_S$, it is clear that if $n^{-1}X_S^\top X_S$ has bounded eigenvalues, and if the entries of W are uniformly bounded, as would be the case in Examples 5–8 in Jeong and Ghosal (2021), including logistic regression, then the upper bound in (8) holds with $\tau = 0$. Such cases correspond to the weakest constraint (9) on the support for S, since we can take $\tau' = 0$ too. More generally, if $n^{-1}X_S^\top X_S$ has bounded eigenvalues, then $\Lambda_{|S|}$ is bounded by the maximum entry on the diagonal of $W(S,\theta_S^\dagger)$. Since the diagonal entries are typically increasing functions of their arguments, the bound is $w(\|X_S\theta_S^\dagger\|_\infty)$. Since the focus is on large configurations, and since $\|X_S\theta_S^\dagger\|_\infty = \|X_S\theta_S^\star\|_\infty = \|X\theta^\star\|_\infty$ for $S \supset S^\star$, the upper bound in (8) is only slightly stronger than assuming "$w(\|X\theta^\star\|_\infty) \lesssim (n/\log p)^\tau$". The Poisson log-linear model is one of the most challenging examples, where $h = \dot{b}^{-1}$ and ξ is the identity, so $w(\eta) = e^\eta$. For our Condition 2 to be met in the Poisson case, we would roughly need $\theta^\star$ to satisfy

$$\|X\theta^\star\|_\infty \lesssim \log\{(n/\log p)^\tau\} = O(\log n).$$

When p is polynomial in n, this restriction is equivalent to that in Remark 2 of Jeong and Ghosal (2021); when $\log p$ is a small power of n, the above restriction is stronger than theirs. But our sometimes-stronger condition here allows for a more extensive asymptotic analysis, i.e., results on the marginal posterior for S.

For Condition 3(a), the constant τ in (8) is determined by X, so, theoretically, it is not impossible to determine τ and to set γ in (10) accordingly. For example,

if X is such that $s \mapsto \Lambda_s$ is uniformly bounded, then (8) holds with $\tau = 0$ and then γ can be taken as a constant too. When $\tau > 0$, the corresponding γ is a diverging sequence in n. Recall that the primary part of the $(\theta_S \mid S)$ covariance matrix is $J_n(S, \hat{\theta}_S)^{-1}$, which itself is $O(n^{-1})$. So, multiplying this by $\gamma \to \infty$ still allows the prior mass to be concentrating around the data-driven center $\hat{\theta}_S$, just slower. Finally, as argued in Narisetty, Shen and He (2019), the constant κ can be chosen arbitrarily close to 1, so "$\beta > \kappa$," which is implied by Condition 3(b), is just a little stronger than "$\beta > 1$." Indeed, since s_n is growing strictly slower than $n^{1/2}$, it suffices to take $\beta - \kappa$ greater than $\frac{1}{2}(\log n)/(\log p) \leq \frac{1}{2}$.

Below we present two different types of results. The first type concerns generally how the posterior distribution Π^n for the coefficient vector θ concentrates its mass around the sparse true value $\theta^\star$. The second type concerns how the marginal posterior mass function π^n for S concentrates around its true value $S^\star$, where $S^\star = S(\theta^\star)$ is the configuration corresponding to the true $\theta^\star$. The proofs, presented in Appendix B, follow those in Martin, Mess and Walker (2017) and Martin and Walker (2014) relatively closely. The key difference is that, here, we cannot get closed-form expressions for the posterior, so suitable in-expectation bounds as in the above references are out of reach. The proofs here are based on in-probability bounds, which are more flexible, so our general strategies here might be applicable in other situations too.

Although the Gaussian linear model is a GLM, it is possible to derive equivalent results for this model directly and under weaker conditions (Martin, Mess and Walker, 2017). So the machinery presented below is intended only for the genuinely more complex GLM setting.

4.2. Posterior concentration results

As a first result, we consider concentration of the full posterior for θ around the true, sparse coefficient vector $\theta^\star$ under a statistically universal metric, namely, the Hellinger distance between joint distributions of y determined by the true $\theta^\star$ and by a generic θ. More specifically, if $p_\theta(y \mid x)$ denotes the distribution of (scalar) y, given covariate vector x and coefficient vector θ, then define the (expected) Hellinger distance H_n as

$$H_n(\theta^\star, \theta) = \Big[\frac{1}{n}\sum_{i=1}^n \int \{p_\theta(y_i \mid x_i)^{1/2} - p_{\theta^\star}(y_i \mid x_i)^{1/2}\}^2 \, dy_i\Big]^{1/2}. \tag{11}$$

This is just the expected squared Hellinger distance between marginals, where expectation is with respect to the empirical distribution of the rows x_i in the matrix X. Note that H_n depends on X.

Theorem 1. *Under Conditions 1–3, the posterior Π^n defined above satisfies*

$$\sup_{\theta^\star : |S(\theta^\star)| \leq s_n} \mathsf{E}_{\theta^\star} \Pi^n(\{\theta : H_n(\theta^\star, \theta) > M\varepsilon_n(\theta^\star)\}) \to 0, \quad n \to \infty, \tag{12}$$

where $\varepsilon_n^2(\theta^\star) = n^{-1} s^\star \log p$, with $s^\star = |S(\theta^\star)|$ and $M > 0$ a large constant.

This is the same rate established in Theorem 2 of Jeong and Ghosal (2021) for a class of Bayesian posterior distributions based on priors that do not depend on data. This is also effectively the minimax optimal rate, i.e., $\{n^{-1}s^\star \log(p/s^\star)\}^{1/2}$, in sparse, high-dimensional linear regression that is attained by Abramovich and Grinshtein (2010), Abramovich and Grinshtein (2016), Arias-Castro and Lounici (2014), Martin, Mess and Walker (2017), and Belitser and Ghosal (2020). The difference between the rate in Theorem 1 and the minimax optimal rate is negligible since $p \gg s^\star$ and, consequently, $\log(p/s^\star) \sim \log p$. It is also important to recognize that the rate achieved is optimal corresponding to the unknown, true sparsity level $|S(\theta^\star)|$. Since the method itself has no knowledge of this sparsity level, we say that the minimax optimal rate is achieved *adaptively*.

Admittedly, the Hellinger rate is not so easily interpretable, but rates under different metrics are possible. For a more interpretable result (modulo more complicated conditions), we can appeal to the arguments given in the proof of Theorem 3 in Jeong and Ghosal (2021). Their proof shows that a Hellinger rate like in Theorem 1 above and an effective-dimension bound like in Theorem 2 below together imply a rate in terms of other distances, including the ℓ_2-distance on θ. Their argument is not for a specific kind of posterior distribution, so what works for them in their case works equally well for us here. The following corollary makes our claims precise.

Corollary 1. *Under Conditions 1–3, the posterior* Π^n *defined above satisfies*

$$\sup_{\theta^\star : |S(\theta^\star)| \leq s_n} \mathsf{E}_{\theta^\star} \Pi^n \Big(\Big\{ \theta : \|\theta - \theta^\star\|_2^2 > \frac{M \varepsilon_n^2(\theta^\star)}{\phi^2(K|S(\theta^\star)|, W(\theta^\star))} \Big\} \Big) \to 0, \quad n \to \infty, \tag{13}$$

where $\varepsilon_n^2(\theta^\star) = n^{-1}|S(\theta^\star)| \log p$, $M > 0$ *and* $K > 0$ *are constants, and*

$$\phi(s, W) = \inf_{\theta : 1 \leq |S(\theta)| \leq s} \frac{\|W^{1/2} X \theta\|_2}{n^{1/2} \|\theta\|_2},$$

denotes the smallest s-sparse singular value of $X^\top W X$.

The appearance of an additional term—the sparse singular value—depending on X is expected since the response y depends directly on $X\theta$, not on θ itself. This is easy to see in the linear model case where the Hellinger distance is proportional to the ℓ_2-norm between fitted values. So to strip the X away and investigate the posterior concentration directly in terms of θ requires some conditions on X, which are baked into the effect the ϕ term has on the rate. For example, if the ϕ term in (13) is bounded away from 0, which amounts to a condition on X, then that term can be absorbed into the constant M and the ℓ_2-rate agrees with the Hellinger rate above. In any case, the result here is the same as that proved in Jeong and Ghosal (2021), so the reader interested in details about ϕ can refer to their discussion.

That the posterior for θ concentrates at the (near) optimal rate for sparse $\theta^\star$ true vector *suggests* that the posterior for S is concentrating on the true $S^\star = S(\theta^\star)$, but this is not a consequence of Theorem 1. The asymptotic behavior of

the S-posterior must be investigated directly. The first such result concerns the "effective dimension" of the posterior distribution for θ is not much larger than that of $\theta^\star$. In other words, π^n concentrates on S with $|S|$ that are smaller than a multiple of $|S^\star|$, where $S^\star = S(\theta^\star)$.

Theorem 2. *Under Conditions 1–3, for any* $C > (1 - \alpha\kappa/\beta)^{-1} > 1$,

$$\sum_{S:|S|>C|S^\star|} \pi^n(S) \to 0, \quad \textit{in } \mathsf{P}_{\theta^\star}\textit{-probability},$$

for all $\theta^\star$ *such that* $|S^\star| \leq s_n$.

That the constant C above is greater than 1 follows from the fact that $\beta > \kappa$ and $\alpha < 1$. Again, the take-away message here is that the posterior distribution for θ is concentrating on a space that is genuinely low-dimensional and, in particular, is of dimension not much greater than $|S^\star|$. Theorem 2 also *suggests* that π^n is concentrating on $S^\star$, but this is not a direct consequence. Towards this, we have one more result which states that π^n tends to not over-fit, i.e., it tends to avoid supersets of $S^\star$.

Theorem 3. *Under Conditions 1–3,*

$$\sum_{S:S\supset S(\theta^\star)} \pi^n(S) \to 0 \quad \textit{in } \mathsf{P}_{\theta^\star}\textit{-probability},$$

for all $\theta^\star$ *such that* $|S(\theta^\star)| \leq s_n$.

Virtually the same argument used to prove Theorem 3 can be used to conclude that π^n will not concentrate on any S that contains an unimportant variable. To ensure that π^n does not concentrate on models that exclude at least one important variable, it is necessary to assume that the non-zero coefficients attached to the important variables are not too small. The intuition is that, if an important variable is "just barely" important, the data may not be informative enough to detect it given the relatively strong penalty on model complexity. The following result imposes a version of the familiar *beta-min condition* (Bühlmann, 2011; Bühlmann and van de Geer, 2011) to ensure that the signals are large enough to be detected; see (14).

Theorem 4. *Under Conditions 1–3,* $\pi^n\{S(\theta^\star)\} \to 1$ *with* $\mathsf{P}_{\theta^\star}$*-probability* $\to 1$ *for all* $\theta^\star$ *such that* $c|S(\theta)^\star| \leq s_n$ *and*

$$\min_{j\in S(\theta^\star)} \theta_j^{\star 2} \geq cn^{-1}|S(\theta^\star)|\Lambda_{c|S(\theta^\star)|}\log p, \tag{14}$$

for some constant $c > 1$.

The condition (14) is exactly the same as in Narisetty, Shen and He (2019) and Cao and Lee (2020). It is also equivalent to the beta-min condition for lasso variable selection consistency when $w = 0$, and is slightly stronger when $w > 0$.

4.3. Posterior distribution approximation

Aside from knowing that the posterior distribution for θ concentrates around the true $\theta^\star$ sufficiently fast, there is interest in knowing the approximate form or shape of that limiting posterior. In particular, this helps to demonstrate that uncertainty quantification about θ (e.g., credible sets) derived from the posterior distribution is at least asymptotically valid in a frequentist sense. Along these lines, under the same conditions as in the previous theorems, Theorem 5 below establishes a *Bernstein–von Mises* result comparable to that in Corollary 2 of Castillo, Schmidt-Hieber and van der Vaart (2015) for a high-dimensional linear regression model. To our knowledge, this is the first posterior normality result in the literature on sparse, high-dimensional GLMs.

Towards this, define the Gaussian approximation

$$\Psi^n(d\theta) = \mathsf{N}_{|S^\star|}\big(d\theta_{S^\star} \mid \hat{\theta}_{S^\star}, \varrho\, J_n(S^\star, \hat{\theta}_{S^\star})^{-1}\big) \otimes \delta_{0^{S^{\star c}}}(d\theta_{S^{\star c}}), \tag{15}$$

where the variance multiplier ϱ depends only on the inputs (α, γ) as follows:

$$\varrho = \gamma(1 + \alpha\gamma)^{-1}.$$

This is effectively the Gaussian posterior approximation that the data analyst would use if he/she *knew* the configuration $S^\star$ and, in particular, the marginal posterior credible intervals for θ_j, with $j \in S^\star$, would virtually agree with the classical likelihood-based confidence intervals returned from, say, R's `glm` function for large n; more on these points below. Theorem 5 below states that our proposed posterior distribution Π^n asymptotically resembles the near-oracle Gaussian approximation (15) in the total variation sense.

Theorem 5. *Under Conditions 1–3, for Ψ^n as in (15),*

$$\mathsf{E}_{\theta^\star}\, d_{\text{TV}}\big(\Pi^n, \Psi^n\big) \to 0,$$

for any $\theta^\star$ such that (14) holds and

$$n^{-1}|S(\theta^\star)|^3 \Lambda_{|S(\theta^\star)|} \log p \to 0. \tag{16}$$

Recall that Condition 3 requires that, in most cases, $\gamma = \gamma_n$ is a diverging sequence, i.e., $\gamma \to \infty$ as $n \to \infty$. In that case, $\varrho = \varrho_n$ also depends on n and satisfies $\varrho \to \alpha^{-1}$ as $n \to \infty$. Even if γ is allowed to be a constant, we can choose that constant to make ϱ as close to α^{-1} as we like. Since $\alpha \in (0, 1)$ is a fixed constant that can be (and is) taken arbitrarily close to 1, the limiting multiplier α^{-1} need only be slightly greater than 1. So, Ψ^n in (15) is effectively the *oracle* Gaussian approximation and, likewise, our posterior uncertainty quantification is asymptotically both valid and effectively efficient.

Condition (16) amounts to assuming that the true $\theta^\star$ is a bit lower complexity compared to what was needed for posterior concentration rates, etc. This restriction is not too surprising, since the result is closely tied to the accuracy of Laplace's approximation, which is only expected in relatively low-complexity

cases. Indeed, Shun and McCullagh (1995) show that Laplace approximations are expected to be accurate for integrals over spaces of dimension $o(n^{1/3})$, which is precisely the range of $|S(\theta^\star)|$ that would satisfy (16); see, also, Barber, Drton and Tan (2016). The restriction (16) can also be compared to the corresponding result in Castillo, Schmidt-Hieber and van der Vaart (2015). For one thing, in the case of a Gaussian likelihood, there is already a strong, built-in nudge towards a Gaussian posterior, so it makes sense that these authors could prove a result comparable to that in Theorem 5 under weaker conditions. They also adopt a "small lambda regime"—referring to the rate parameter in their Laplace prior for non-zero coefficients—which offers additional flexibility to handle more complex configurations. All this being said, we make no claims that (16) is necessary for our proposed posterior to enjoy a Bernstein–von Mises theorem—improvements on some of the bounds used in our analysis may be available.

5. Numerical results

5.1. Methods and metrics

Our focus here is on variable selection and estimation performance of our proposed empirical Bayes method compared to existing methods in two different GLM contexts. There is a plethora of literature (e.g., Bhadra et al., 2019a; Cao and Lee, 2020; Narisetty, Shen and He, 2019; Wei and Ghosal, 2020) that focuses on high-dimensional logistic regression, but a dearth of literature on Bayesian methods—with numerical implementations—that can handle GLMs besides logistic regression. For this reason, we showcase our method's performance in both *logistic* regression and *Poisson* regression (with canonical log link).

We start with some details about the competitor methods we consider, including lasso, adaptive lasso, SCAD, MCP, horseshoe, and skinnyGibbs. Most of these methods were implemented via R packages: lasso and adaptive lasso with `glmnet` and SCAD and MCP through `ncvreg`. No R package is available for the horseshoe in GLMs, so we relied instead on STAN. For logistic regression, the horseshoe method is coded using the `rstan` package, and hyperparameters were chosen based on recommendation of Piironen and Vehtari (2017) and Bhadra et al. (2019a). Specifically, we use the regularized horseshoe prior with $c^2 \sim \mathsf{InvGamma}(2, 8)$, and τ determined based on the number of effective nonzero coefficients. The MCMC for the horseshoe was run with two chains and 5,000 posterior draws for each chain. Posterior samples were obtained for the coefficient vector θ, and for the task of variable selection, we follow the default procedure that deems a variable as "active" if its 95% posterior credible interval does not include zero. For Poisson regression, there was not as much guidance in the literature on implementation; we used the `rstanarm` package (Goodrich et al., 2022), but these results are not reported here as the horseshoe was computationally restrictive in terms of runtime when running simulations with a large number of replications. SkinnyGibbs was implemented with R package `skinnybasad`, directly obtained from the authors. Since this method was devel-

oped exclusively for variable selection in logistic regression, we do not present results for skinnyGibbs in the Poisson regression setting.

We implement the proposed empirical Bayes solution using the Metropolis–Hastings strategy in Section 3.3. We also tried the shotgun stochastic search as discussed there, but we found that this produced similar results to Monte Carlo with no appreciable gain in computational efficiency; so we opted for the simpler of the two. After a burn-in period in our sampling scheme, we return $M = 10^4$ posterior samples of the configuration S, from which we can evaluate the inclusion probabilities as defined in (7). For a variable selection procedure based on our empirical Bayes solution, we propose

$$\hat{S} = \hat{S}(t) = \{j : \pi^n(j) > t\},$$

where $\pi^n(j)$ is the (Monte Carlo approximation of) inclusion probability in (7) and $t \in (0, 1)$ is a user-specified threshold. We investigate the performance of our proposed method with two different choices of the threshold t, namely, $t = 0.1$ and $t = 0.5$; we refer to these below as EB1 and EB2, respectively. The latter threshold 0.5 corresponds to selecting the so-called "median probability model" advocated for in, e.g., Barbieri and Berger (2004). The former threshold 0.1 is designed to select a larger subset of variables and is motivated by Bayesian decision-theoretic considerations where a Type II error (missing an important variable) is taken to be 9 times as costly as a Type I error.

For comparing the performance of the different variable selection methods, we report three different metrics: sensitivity (TPR), specificity (TNR), and Matthew's correlation coefficient (MCC). Sensitivity, or true positive rate, allows us to see how well the method does in identifying true signals or genuinely non-zero coefficients; specificity, or true negative rate, shows the method's ability to correctly identify the noise or genuinely zero coefficients; and MCC looks at all the four categories of the confusion matrix and combines them into one single metric. An MCC of 1 means the method perfectly distinguished signal from noise, while an MCC of 0 means the signal discovery was no better than random guessing. These metrics are defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad \mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})\}^{1/2}},$$

where, TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively, and for MCC, we adopt the convention $0/0 \equiv 0$. The relevance of these metrics depends on the data analyst's priorities. For example, if Type II errors are more costly than Type I, then TPR would be the relevant metric. On the other hand, if the two errors cost roughly the same, then MCC would be a more relevant metric. MCC has been shown to be more reliable than other measures that are commonly used, including area under the receiver operating characteristic curve, balanced accuracy, bookmaker informedness, and markedness (Chicco and Jurman, 2023).

For logistic regression, we focused on variable selection since skinnyGibbs, the most competitive method in the Bayesian landscape, is designed only for logistic regression variable selection. For Poisson regression, in addition to variable selection, we also compare estimation performance of our method to the others (lasso, adaptive lasso, SCAD, and MCP). Neither horseshoe nor skinnyGibbs are included in the Poisson regression comparisons—horseshoe is too restrictive in terms of runtime and no version of skinnyGibbs is currently available for GLMs besides logistic regression. For estimation, the metric we consider is mean squared error (MSE).

5.2. Simulation studies

5.2.1. Logistic regression

We fixed the sample size at $n = 100$ and considered two different values of p, namely, $p = 200$ and $p = 400$. The true coefficient vector $\theta^\star$ is set to have its first s many components equal to 3 and the rest set to 0; here the cardinality s can take two values, $s = 4$ and $s = 8$. The rows of the design matrix X are randomly simulated from a multivariate normal with mean 0, variance 1, and covariance matrix Σ, where $\Sigma_{ij} = r^{|i-j|}$ corresponds to a first-order autoregressive correlation structure. The correlation parameter r takes values $r = 0$ and $r = 0.2$. The response variables $y_1, \ldots, y_n$ are generated independently, where y_i has a Bernoulli distribution with success probability $\exp(x_i^\top \theta^\star)/\{1 + \exp(x_i^\top \theta^\star)\}$, for $i = 1, \ldots, n$. The settings are determined by the triple (p, s, r), so there are altogether eight simulation settings. There are 500 replications performed at each of the eight settings, and the variable selection results for the methods and metrics described above are summarized in Table 1.

Our results showed that our EB method performs comparably to the other methods. First, if the data analyst places higher priority on correctly identifying the active variables, then TPR would be his/her preferred metric. In this case, EB1—with a smaller cutoff $t = 0.1$, consistent with the higher priority on finding active variables—performs fairly well in terms of TPR compared to the other five methods. Lasso has the highest TPR in two of the eight settings but, as is common, it tends to over-select as evidenced by its low TNR. SCAD has the highest TPR in the other six settings. Horseshoe has the lowest TPR in all but one of the settings. This is because the standard/default strategy recommends using 95% credible intervals, which is too conservative; a lower credibility level should be used if a less conservative selection procedure is desired. Second, if the data analyst's priorities are more balanced, i.e., aiming for parsimonious models with good overall performance, then MCC would be the go-to metric and he/she would prefer the more balanced EB2 with a larger cutoff $t = 0.5$. Here, EB2 has an MCC comparable with the other methods. Adaptive lasso has the lowest MCC in six of the settings, whereas SkinnyGibbs has the highest across all eight settings. This is not unexpected, given that skinnyGibbs is designed specifically for variable selection in logistic regression.

TABLE 1
Comparison of TPR, TNR, and MCC for the two EB methods with different cutoffs and the six other methods across various settings in logistic regression.

p	$\|S\|$	r	Metric	EB1	EB2	HS	lasso	alasso	SCAD	MCP	skinny
			TPR	0.980	0.966	0.871	0.998	0.701	1.000	0.999	0.997
200	4	0	TNR	0.971	0.992	1.000	0.953	0.854	0.962	0.986	0.999
			MCC	0.676	0.859	0.927	0.630	0.228	0.597	0.783	0.980
			TPR	0.962	0.926	0.786	0.995	0.806	0.999	0.996	0.985
200	4	0.2	TNR	0.975	0.994	1.000	0.966	0.847	0.955	0.983	0.999
			MCC	0.691	0.856	0.878	0.708	0.272	0.561	0.744	0.973
			TPR	0.746	0.643	0.276	0.959	0.579	0.962	0.925	0.813
200	8	0	TNR	0.940	0.983	1.000	0.895	0.851	0.948	0.982	0.998
			MCC	0.519	0.618	0.498	0.528	0.206	0.640	0.790	0.865
			TPR	0.747	0.648	0.287	0.953	0.705	0.951	0.888	0.714
200	8	0.2	TNR	0.956	0.988	1.000	0.932	0.837	0.952	0.984	0.998
			MCC	0.570	0.663	0.514	0.631	0.268	0.644	0.780	0.806
			TPR	0.922	0.897	0.582	0.993	0.741	1.000	1.000	0.980
400	4	0	TNR	0.989	0.996	1.000	0.973	0.924	0.974	0.991	0.998
			MCC	0.728	0.841	0.728	0.618	0.265	0.543	0.744	0.920
			TPR	0.926	0.860	0.568	0.994	0.836	0.998	0.995	0.952
400	4	0.2	TNR	0.993	0.998	1.000	0.980	0.931	0.971	0.990	0.998
			MCC	0.795	0.860	0.737	0.685	0.336	0.514	0.718	0.899
			TPR	0.490	0.407	0.057	0.896	0.468	0.917	0.843	0.498
400	8	0	TNR	0.974	0.987	1.000	0.941	0.930	0.962	0.987	0.996
			MCC	0.406	0.401	0.145	0.493	0.196	0.546	0.694	0.591
			TPR	0.548	0.438	0.092	0.931	0.617	0.922	0.842	0.565
400	8	0.2	TNR	0.983	0.992	1.000	0.958	0.921	0.965	0.989	0.996
			MCC	0.513	0.501	0.226	0.580	0.260	0.564	0.714	0.646

In terms of runtime, we compared the methods by fixing $n = 100$, $|S| = 4$, and $r = 0$, but varying $p \in \{200, 300, 400, 500, 600, 700, 800\}$. Each method was run 5 times to generate an average runtime. Figure 1 plots the runtime of our EB method and the skinnyGibbs method as p increases. This comparison might be rather striking, so some remarks are in order. It is not possible for a Markov chain to explore the S-space thoroughly in an acceptable amount of time when p is even moderately large. Our proposed method is not designed to thoroughly explore the S-space; instead, the goal is to do a careful-but-admittedly-incomplete exploration focused on the high-probability configurations so that it can provide reliable variable selection. As our results show, apparently this latter goal can be accomplished competitively with far shorter runtimes.

5.2.2. Poisson regression

The data X is generated the same way as in the logistic regression settings, with one minor change—the common standard deviation across the rows of X is set at 0.3 instead of 1. This is to ensure that the Poisson variables generated are not exponentially large. The response variables $y_1, \ldots, y_n$ are generated independently, with y_i having a Poisson distribution with rate $\exp(x_i^\top \theta^\star)$. All

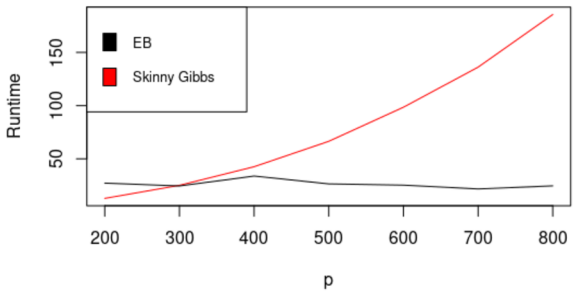

FIG 1. *A comparison of the runtimes (in seconds) of EB and skinnyGibbs with fixed* $n = 100, |S| = 4, r = 0$ *and varying* $p = \{200, 300, 400, 500, 600, 700, 800\}$.

TABLE 2
Comparison of TPR, TNR, and MCC for the two EB methods with different cutoffs and the four other methods across various settings in Poisson regression.

p	$\|S\|$	r	Metric	EB1	EB2	lasso	alasso	SCAD	MCP
			TPR	1.000	1.000	1.000	0.998	0.975	0.950
200	4	0	TNR	0.999	1.000	0.893	0.949	0.983	0.987
			MCC	0.973	0.998	0.395	0.547	0.879	0.906
			TPR	1.000	1.000	1.000	1.000	0.973	0.903
200	4	0.2	TNR	1.000	1.000	0.907	0.964	0.995	0.998
			MCC	0.992	0.999	0.425	0.620	0.905	0.893
			TPR	1.000	1.000	0.979	0.983	0.598	0.494
200	8	0	TNR	0.997	1.000	0.856	0.915	0.981	0.989
			MCC	0.969	0.998	0.437	0.555	0.549	0.514
			TPR	1.000	1.000	0.983	0.958	0.514	0.421
200	8	0.2	TNR	0.998	1.000	0.898	0.946	0.983	0.988
			MCC	0.980	0.998	0.508	0.625	0.512	0.470
			TPR	1.000	1.000	1.000	1.000	0.990	0.975
400	4	0	TNR	1.000	1.000	0.926	0.955	0.995	0.999
			MCC	0.991	1.000	0.350	0.445	0.868	0.936
			TPR	1.000	1.000	1.000	0.998	0.945	0.863
400	4	0.2	TNR	1.000	1.000	0.940	0.971	0.977	0.978
			MCC	0.994	0.999	0.392	0.533	0.863	0.850
			TPR	1.000	1.000	0.979	0.971	0.494	0.299
400	8	0	TNR	0.999	1.000	0.914	0.942	0.986	0.992
			MCC	0.985	1.000	0.417	0.499	0.434	0.322
			TPR	1.000	1.000	0.960	0.946	0.415	0.318
400	8	0.2	TNR	0.998	1.000	0.936	0.960	0.989	0.993
			MCC	0.973	0.999	0.465	0.552	0.401	0.357

other settings remain the same as in the logistic regression simulations above, with the same (p, r, s) combinations. The Poisson regression simulation results for variable selection are summarized in Table 2, where 100 replications were run at each simulation setting.

We see that EB does very well compared to the other methods, and even better comparatively than in the logistic regression settings. EB2 has the highest MCC value in all eight configurations, and in fact, both EB1 and EB2 per-

TABLE 3
Comparison of mean squared error (MSE) values for the EB method and the four other methods across various settings in Poisson regression.

p	$\lvert S\rvert$	r	EB	lasso	alasso	SCAD	MCP
200	4	0	0.117	2.642	1.365	2.064	2.375
200	4	0.2	0.094	1.677	0.813	6.536	7.714
200	8	0	0.196	10.143	9.906	43.698	48.180
200	8	0.2	0.469	16.201	17.839	56.941	61.823
400	4	0	0.207	3.925	2.105	2.184	3.195
400	4	0.2	0.100	2.177	1.176	9.638	10.754
400	8	0	4.691	16.920	14.248	59.989	59.021
400	8	0.2	7.419	17.451	14.531	60.560	64.833

form better than all other methods across all the settings for all three metrics. Comparing the methods, SCAD and MCP are not as competitive in Poisson regression as in logistic regression.

For estimation, we use the same simulation settings as above, and compare the (empirical) mean squared error (MSE) values, $\mathsf{E}_{\theta^\star}\|\hat{\theta} - \theta^\star\|^2$, for the various methods. The results summarized in Table 3 show that our proposed method is an incredibly strong performer across all the settings.

6. Conclusion

In this paper, we extend the empirical or data-driven prior specification strategy first proposed by Martin, Mess and Walker (2017) to the case of high-dimensional GLMs and investigate its theoretical and practical performance. In particular, we show that the proposed solution is ideal in the sense that it balances the strong theoretical performance that is necessary to justify its use in applications with the computational simplicity and efficiency necessary to be applicable in these problems. The balance comes from the data-driven prior centering: we enjoy the computational advantage of a relatively simple, thin-tailed prior without subjecting ourselves to the theoretical sub-optimality that results from thin-tailed priors with fixed centers. Compared to the linear models previously investigated, a challenge here in the GLM context is that there is no conjugacy in the prior and, therefore, no closed-form expressions for any of the posterior features. These challenges affect both the theory and computation, but we have used some new techniques to successfully overcome them here in this paper. While there are other methods available in the literature that have good empirical performance, and others that have powerful theoretical results, our contribution here is unique in the sense that our solution achieves both. The solution is general and can be applied beyond logistic regression setting.

Our numerical investigations in this paper focused on variable selection and point estimation, but the method itself is capable of answering other questions. In a follow-up work it would be interesting to explore the performance of the proposed method in the context of uncertainty quantification about the co-

efficient vector θ. In the high-dimensional linear model setting, work has been done in Martin and Tang (2020) to look at how the method does in prediction—similar work with a focus on point prediction and uncertainty quantification of prediction can be carried out in this high-dimensional GLM context. On the theoretical side, there are some new techniques employed here in the proofs, namely, relying on in-probability bounds as opposed to bounds in expectation. The latter have their advantages, but the former are more flexible. We expect that this added flexibility would be useful in other cases where, as is common, the priors would not be exactly conjugate.

Appendix A: Technical preliminaries

A.1. Likelihood-related properties

First, we present here two key results from Narisetty, Shen and He (2019), namely, Lemmas A1 and A3, respectively, both concerning asymptotic properties of the likelihood function and MLEs in the specific case of high-dimensional logistic regression with Bernoulli response variables. Lemmas 1 and 2 suitably generalize these two results.

It is not enough for us to focus on the true $\theta^\star$ exclusively, so the results presented below cover general configurations S that may or may not be supersets of $S(\theta^\star)$, i.e., where $\theta_S^\dagger$ is needed in place of $\theta_S^\star$. Fortunately, the arguments Narisetty, Shen and He (2019) used to prove their Lemmas A1 and A3 go through almost word-for-word when replacing $\theta_S^\star$ with $\theta_S^\dagger$ where appropriate. Finally, none of the considerations that follow make sense if $S = \varnothing$, e.g., if there is no parameter, then it does not make sense to ask about properties of the information matrix or about consistency of the MLE. This detail is not relevant when only considering S that are supersets of $S^\star$, but our analysis requires consideration of general S. So, without loss of generality, wherever relevant, we restrict attention to $S \neq \varnothing$, so $|S| \geq 1$.

The first result establishes an important continuity property for the observed information matrix.

Lemma 1. *Under Conditions 1–2, for any fixed constant $c > 0$, there exists $\zeta_n \to 0$ such that*

$$(1 - \zeta_n)\, J_n(S, \theta_S) \leq J_n(S, \theta_S^\dagger) \leq (1 + \zeta_n)\, J_n(S, \theta_S),$$

for any (S, θ_S) such that $|S|\, \|\theta_S - \theta_S^\dagger\|^2 = o(1)$.

Note, for a given S, the corresponding $\zeta_n = \zeta_{n,S}$ sequence is of the order

$$\zeta_n \sim \{n^{-1}|S|^2 \Lambda_{|S|} \log p\}^{1/2},$$

so the choice of ζ_n that holds uniformly over all the sufficiently low-complexity configurations is of the order $\{n^{-1} s_n^2 \Lambda_{s_n} \log p\}^{1/2}$.

The next result is a convergence rate for the MLE uniformly over configurations that are not "too complex." This corresponds to Lemma A3 in Narisetty, Shen and He (2019) covering the case where the GLM's score function has sub-Gaussian tails. The critical sub-exponential case is covered in Lee, Chae and Martin (2024), based on key results in Spokoiny (2017).

Lemma 2. *Under Conditions 1–2,*

$$\max_{S:|S|=s} \|\hat{\theta}_S - \theta_S^\dagger\|^2 = O_P(n^{-1} s \Lambda_s \log p), \quad n \to \infty,$$

uniformly over all s with $1 \leq s \leq s_n$

The following result is an important consequence of the two lemmas above. The sub-Gaussian case is based on Lemma 3 in Lee and Cao (2021), quoted as Equation A7 of Cao and Lee (2020), both based primarily on the analysis in Narisetty, Shen and He (2019). The sub-exponential case is substantially more involved, and addressed in Lee, Chae and Martin (2024).

Lemma 3. *Under Conditions 1–2, there exists a constant $\kappa > 1$ such that*

$$\ell_n(S, \hat{\theta}_S) - \ell_n(S^\star, \hat{\theta}_{S^\star}) \leq \kappa(\log p)(|S| - |S^\star|), \quad \textit{with } \mathsf{P}_{\theta^\star}\textit{-probability} \to 1,$$

uniformly over S with $S \supset S^\star$ and $|S| \leq s_n$.

A.2. Marginal likelihood

Here we provide justification for the approximation (6) of the marginal posterior π^n at configuration S. Recall that the marginal posterior satisfies

$$\pi^n(S) = \frac{\pi_n(S)\, m_n(S)}{\sum_{S'} \pi_n(S)\, m_n(S)},$$

where $m_n(S)$ is the marginal likelihood for configuration S:

$$m_n(S) = \int_{\mathbb{R}^{|S|}} \underbrace{L_n(S, \theta_S)^\alpha \, \mathsf{N}_{|S|}(\theta_S \mid \hat{\theta}_S, \gamma J_n(S, \hat{\theta}_S)^{-1})}_{= g(\theta_S), \text{ say}} \, d\theta_S.$$

So the goal is to lower- and upper-bound the integral $m_n(S)$. Aside from providing justification for our simple computational strategy, the result in Lemma 4 below will be useful in the proofs of our main results below. In particular, we will have a need to bound

$$\frac{\pi^n(S)}{\pi^n(S^\star)} = \frac{\pi_n(S)\, m_n(S)}{\pi_n(S^\star)\, m_n(S^\star)}$$

Lemma 4. *Under Conditions 1–2, the marginal likelihood $m_y(S)$ satisfies*

$$1 \leq \frac{m_n(S)}{(1+\alpha\gamma)^{-|S|/2}\, L_n(S, \hat{\theta}_S)^\alpha} \leq 1 + e^{-C|S|\Lambda_{|S|}\log p},$$

with $\mathsf{P}_{\theta^\star}$-probability tending to 1, for a constant $C > 0$. Consequently,

$$\frac{\pi^n(S)}{\pi^n(S^\star)} \leq 2\,\frac{\pi_n(S)}{\pi_n(S^\star)}(1+\alpha\gamma)^{-(|S|-|S^\star|)/2}\,\frac{L_n(S,\hat\theta_S)^\alpha}{L_n(S^\star,\hat\theta_{S^\star})^\alpha}. \tag{17}$$

Proof. Thanks to Lemma 2, it is safe to assume that the MLEs $\hat\theta_S$ are all within a small neighborhood of their respective targets $\theta^\dagger_S$. Split the marginal likelihood integral into two parts according to $\mathbb{R}^{|S|} = A_S \cup A_S^c$, where $A_S = \{\theta_S : \|\theta_S - \hat\theta_S\|^2 \leq r_n^2(S)\}$, and $r_n^2(S) = n^{-1}|S|\Lambda_{|S|}\log p$. For $\theta_S \in A_S$, by Taylor's theorem and Lemma 1, we get

$$\begin{aligned}
\ell_n(S,\theta_S) - \ell_n(S,\hat\theta_S) &\geq -\tfrac{1}{2}(1+\zeta_n)(\theta_S-\hat\theta_S)^\top J_n(S,\hat\theta_S)(\theta_S-\hat\theta_S)\\
\ell_n(S,\theta_S) - \ell_n(S,\hat\theta_S) &\leq -\tfrac{1}{2}(1-\zeta_n)(\theta_S-\hat\theta_S)^\top J_n(S,\hat\theta_S)(\theta_S-\hat\theta_S).
\end{aligned}$$

The first of the above two displays gives a lower bound on the marginal likelihood,

$$\begin{aligned}
m_n(S) &\geq \int_{A_S} g(\theta_S)\,d\theta_S\\
&\geq \{1+\alpha\gamma/(1+\zeta_n)\}^{-|S|/2}\,L_n(S,\hat\theta_S)^\alpha\\
&\geq (1+\alpha\gamma)^{-|S|/2}\Big\{\frac{1+\alpha\gamma/(1+\zeta_n)}{1+\alpha\gamma}\Big\}^{-|S|/2}\,L_n(S,\hat\theta_S)^\alpha\\
&\geq (1+\alpha\gamma)^{-|S|/2}\,L_n(S,\hat\theta_S)^\alpha,
\end{aligned}$$

which agrees with the familiar Laplace approximation expression used in (6). Similarly, for an upper bound on the marginal likelihood, we get

$$m_n(S) = \Big(\int_{A_S} + \int_{A_S^c}\Big) g(\theta_S)\,d\theta_S \leq (1+\alpha\gamma)^{-|S|/2} L_n(S,\hat\theta_S)^\alpha + \int_{A_S^c} g(\theta_S)\,d\theta_S,$$

where, again, the first term in the above bound agrees with the Laplace approximation expression in (6). For $\theta_S \in A_S^c$, we have $\ell_n(S,\theta_S) - \ell_n(S,\hat\theta_S) \leq -Cnr_n^2(S)$, so

$$\int_{A_S^c} g(\theta_S)\,d\theta_S \leq L_n(S,\hat\theta_S)^\alpha e^{-\alpha Cn\delta_n^2(S)}\,\Pi_{n,S}(A_S^c) \leq L_n(S,\hat\theta_S)^\alpha e^{-\alpha Cnr_n^2(S)}.$$

The prior probability of A_S^c has a non-trivial, exponentially small upper bound, but it is no smaller than the other exponentially small term in the above display, so its inclusion does not improve the overall bound. Since the above arguments all hold with probability tending to 1, this completes the proof of the lemma's first claim. For the second claim, we consider the ratio

$$\frac{m_n(S)}{m_n(S^\star)} \leq \frac{(1+\alpha\gamma)^{-|S|/2}\{1+e^{-Cnr_n^2(S)}\}}{(1+\alpha\gamma)^{-|S^\star|/2}}\,\frac{L_n(S,\hat\theta_S)^\alpha}{L_n(S^\star,\hat\theta_{S^\star})^\alpha}.$$

Then the second claim follows since the term in curly braces above is bounded by 2, uniformly in S. □

A.3. Empirical priors and posterior concentration

We briefly describe the general framework in Martin and Walker (2019) for establishing posterior concentration rates with empirical priors in high-dimensional problems. They put forward sufficient conditions on the empirical prior, one they called *local* and the other *global*. In what follows, let $\Pi_n = (\pi_n, \Pi_{n,S})$ be a general empirical prior for $\Theta = (S, \Theta_S)$, and $\Pi^n = \Pi^{n,\alpha}$ the corresponding posterior that uses power $\alpha \in (0,1)$ on the likelihood function. Also, let $\varepsilon_n = \varepsilon_n(\theta^\star)$ be a generic deterministic sequence that satisfies $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$ and may depend on the true $\theta^\star$.

Local Prior Condition. Given ε_n, there exists constants $B, D > 0$ such that

$$\pi_n(S^\star) \gtrsim e^{-Bn\varepsilon_n^2}, \quad \text{for all large } n, \tag{18}$$

$$\mathsf{P}_{\theta^\star}\big[\Pi_{n,S^\star}\{\mathcal{L}_n(S^\star)\} > e^{-Dn\varepsilon_n^2}\big] \to 1, \quad n \to \infty, \tag{19}$$

where

$$\mathcal{L}_n(S^\star) = \{\theta \in \mathbb{R}^p : L_n(S^\star, \theta_{S^\star}) \geq e^{-dn\varepsilon_n^2} L_n(S^\star, \hat{\theta}_{S^\star})\}, \quad d > 0,$$

Global Prior Condition. Given ε_n, there exists $G > 0$ and $m > 1$ such that

$$\sum_S \pi_n(S) \int [\mathsf{E}_{\theta^\star}\{\pi_{n,S}(\theta_S)^m\}]^{1/m}\, d\theta_S \lesssim e^{Gn\varepsilon_n}, \quad n \text{ large}. \tag{20}$$

Under these conditions, a convergence rate in terms of Hellinger distance between joint distributions follows; see Theorem 2 in Martin and Walker (2019). In regression cases like the GLMs under consideration here, the Hellinger rate in terms of joint distributions implies the same rate for the root average squared conditional Hellinger distances in (11); see Appendix B.1 below for details.

Appendix B: Proofs

B.1. Proof of Theorem 1

The proof proceeds by first checking that the local and global prior conditions, as described above, are met under the conditions stated in Theorem 1 above. Then Theorem 2 of Martin and Walker (2019) implies the Hellinger rate result in (12). Since Theorem 2 holds independently and under the same conditions as the theorem we are currently proving, we can assume, where relevant, that S is such that $|S| \leq C|S^\star|$.

Lemma 5. *Under Conditions 1–3, our proposed empirical prior satisfies the local prior condition as described above, with* $\varepsilon_n(\theta^\star) = (n^{-1}|S(\theta^\star)| \log p)^{1/2}$.

Proof. Fix $\theta^\star$ and set $s^\star = |S(\theta^\star)|$. The first part of the local prior condition is easy to check, with $n\varepsilon_n^2 = s^\star \log p$. Indeed, using the inequality $\binom{p}{s} \leq p^s$ we get

$$\pi_n(S^\star) = \frac{(p^{-a})^{s^\star}}{\binom{p}{s^\star}} \geq e^{-(1+a)s^\star \log p} = e^{-(1+a)n\varepsilon_n^2},$$

so the bound in (18) holds with $B = 1 + a > 0$.

Next, for (19), the data-dependent neighborhood $\mathcal{L}_n(S^\star)$ is given by

$$\mathcal{L}_n(S^\star) = \{\theta_{S^\star} : \ell_n(S^\star, \theta_{S^\star}) - \ell_n(S^\star, \hat\theta_{S^\star}) > -dn\varepsilon_n^2\}.$$

Since ℓ_n is concave, this is a bounded neighborhood of $\hat\theta_{S^\star}$. It is a relatively small neighborhood too, since the log-likelihood is of order n; this means that Lemma 1 applies to any $\theta_{S^\star} \in \mathcal{L}_n(S^\star)$ and to $\hat\theta_{S^\star}$. For $\theta_{S^\star} \in \mathcal{L}_n(S^\star)$, Taylor's theorem implies

$$\ell_n(S^\star, \theta_{S^\star}) - \ell_n(S^\star, \hat\theta_{S^\star}) = -\tfrac{1}{2}(\theta_{S^\star} - \hat\theta_{S^\star})^\top J_n(S^\star, \tilde\theta_{S^\star})(\theta_{S^\star} - \hat\theta_{S^\star}),$$

where $\tilde\theta_{S^\star} \in \mathcal{L}_n(S^\star)$ satisfies $\|\tilde\theta_{S^\star} - \hat\theta_{S^\star}\| \le \|\theta_{S^\star} - \hat\theta_{S^\star}\|$. Applying Lemma 1 first with $\tilde\theta_{S^\star}$ and then with $\hat\theta_{S^\star}$ gives

$$J_n(n, \tilde\theta_{S^\star}) \le c_n\, J_n(S^\star, \hat\theta_{S^\star}), \quad \text{with probability} \to 1, \tag{21}$$

where $c_n = (1 - \zeta_n)(1 + \zeta_n) \to 1$. Therefore,

$$\mathcal{L}_n(S^\star) \supset \{\theta_{S^\star} : \gamma^{-1}(\theta_{S^\star} - \hat\theta_{S^\star})^\top J_n(S^\star, \hat\theta_{S^\star})(\theta_{S^\star} - \hat\theta_{S^\star}) < \delta_n\},$$

where $\delta_n := 2dn\varepsilon_n^2/c_n\gamma \gg p^{-1}$. The prior probability of the event on the right-hand side of the above display is the probability that $Z < \delta_n$, where $Z \sim \mathsf{ChiSq}(s^\star)$. Therefore,

$$\begin{aligned}
\Pi_{n,S^\star}\{\mathcal{L}_n(S^\star)\} &= \mathsf{P}(Z < \delta_n) \\
&= \frac{1}{2^{s^\star/2}\Gamma(s^\star/2)} \int_0^{\delta_n} z^{s^\star/2-1} e^{-z/2}\, dz \\
&\ge \frac{e^{-\delta_n/2}}{2^{s^\star/2}\Gamma(s^\star/2)} \int_0^{\delta_n} z^{s^\star/2-1}\, dz \\
&= \frac{2e^{-\delta_n/2}\delta_n^{s^\star/2}}{s^\star 2^{s^\star/2}\Gamma(s^\star/2)}.
\end{aligned}$$

The lower bound on $\Pi_n\{\mathcal{L}_n(S^\star)\}$ is $\ge e^{-Dn\varepsilon_n^2} = e^{-Ds^\star \log p}$ for some $D > 0$, as was to be shown. The prior probability bound holds surely, so (19) follows from the "probability $\to 1$" conclusion in (21). □

Lemma 6. *Under Conditions 1–3, our proposed empirical prior satisfies the global prior condition as described above, with* $\varepsilon_n(\theta^\star) = (n^{-1}|S(\theta^\star)| \log p)^{1/2}$.

Proof. The empirical prior density $\pi_{n,S}$ is given by

$$\pi_{n,S}(\theta_S) = (2\pi)^{-|S|/2}|\gamma^{-1} J_n(S, \hat\theta_S)|^{1/2} \exp\{-\tfrac{1}{2\gamma}(\theta_S - \hat\theta_S)^\top J_n(S, \hat\theta_S)(\theta_S - \hat\theta_S)\}.$$

Lemma 2 establishes the MLE bounds

$$\|\hat\theta_S - \theta_S^\dagger\|^2 \lesssim n^{-1}|S|\Lambda_{|S|} \log p, \quad \text{uniformly in } S \text{ with } |S| \lesssim |S^\star|, \tag{22}$$

and, therefore, with probability tending to 1,

$$(1-\zeta_n)J_n(S,\theta_S^\dagger) \le J_n(S,\hat\theta_S) \le (1+\zeta_n)J_n(S,\theta_S^\dagger).$$

Let $\mathcal{E}_n$ denote the event in (22); since $\mathsf{P}_{\theta^\star}(\mathcal{E}_n) \to 1$, we will restrict attention to cases where $\mathcal{E}_n$ holds in what follows. Then

$$\begin{aligned}\pi_{n,S}(\theta_S) \le (2\pi)^{-|S|/2}(1+\zeta_n)^{|S|/2}|\gamma^{-1}J_n(S,\theta_S^\dagger)|^{1/2} \\ \times \exp\{-\tfrac{1-\zeta_n}{2\gamma}(\theta_S-\hat\theta_S)^\top J_n(S,\theta_S^\dagger)(\theta_S-\hat\theta_S)\}.\end{aligned}$$

If $\Sigma_S = \gamma(1-\zeta_n)^{-1}J_n(S,\theta_S^\dagger)^{-1}$, then the upper bound can be simplified as

$$\pi_{n,S}(\theta_S) \le \big(\tfrac{1+\zeta_n}{1-\zeta_n}\big)^{|S|/2}(2\pi)^{-|S|/2}|\Sigma_S|^{-1/2}\exp\{-\tfrac{1}{2}(\theta_S-\hat\theta_S)^\top\Sigma_S^{-1}(\theta_S-\hat\theta_S)\}.$$

Write $\theta_S-\hat\theta_S = (\theta_S-\theta_S^\dagger)+(\theta_S^\dagger-\hat\theta_S)$ and then expand the quadratic form in the above display to get

$$(\theta_S-\hat\theta_S)^\top\Sigma_S^{-1}(\theta_S-\hat\theta_S) \ge (\theta_S-\theta_S^\dagger)^\top\Sigma_S^{-1}(\theta_S-\theta_S^\dagger)+2(\theta_S-\theta_S^\dagger)^\top\Sigma_S^{-1}(\hat\theta_S-\theta_S^\dagger).$$

Then it is easy to check that

$$\pi_{n,S}(\theta_S) \le \big(\tfrac{1+\zeta_n}{1-\zeta_n}\big)^{|S|/2}e^{|(\theta_S-\theta_S^\dagger)^\top\Sigma_S^{-1}(\hat\theta_S-\theta_S^\dagger)|}\,\mathsf{N}_{|S|}(\theta_S \mid \theta_S^\dagger,\Sigma_S).$$

Apply Cauchy–Schwarz to the quadratic form in the exponent above gives

$$|(\theta_S-\theta_S^\dagger)^\top\Sigma_S^{-1}(\hat\theta_S-\theta_S^\dagger)| \le \|\Sigma_S^{-1/2}(\theta_S-\theta_S^\dagger)\|\,\|\Sigma_S^{-1/2}(\hat\theta_S-\theta_S^\dagger)\|.$$

By Condition 2 and Lemma 2, the second term above can be bounded as

$$\|\Sigma_S^{-1/2}(\hat\theta_S-\theta_S^\dagger)\|^2 \le \gamma^{-1}n\Lambda_{|S|}\,\|\hat\theta_S-\theta_S^\dagger\|^2 \lesssim \gamma^{-1}\,|S|\,\Lambda_{|S|}^2\,\log p.$$

By Condition 3(a), this is upper bounded by a constant times $|S|\log p$. Let $t_S^2 \sim |S|\log p$ denote that upper bound. Then the empirical prior density is bounded as

$$\pi_{n,S}(\theta_S) \le \big(\tfrac{1+\zeta_n}{1-\zeta_n}\big)^{|S|/2}e^{t_S\|\Sigma_S^{-1/2}(\theta_S-\theta_S^\dagger)\|}\,\mathsf{N}_{|S|}(\theta_S \mid \theta_S^\dagger,\Sigma_S).$$

This is constant in data y, so the expectation in (20) can be ignored—and m can be arbitrarily close to 1. Moreover, the integral in (20) over θ_S can now be upper bounded by the moment generating function of a chi distribution, with $|S|$ degrees of freedom, evaluated at t_S. This moment generating function does not have a convenient closed-form expression—it involves the confluent hypergeometric function—but since t_S is large (proportional to $\log p$) for all S, we can apply the standard asymptotic approximation of the chi distribution's moment generating function (Abramowitz and Stegun, 1966, Ch. 13) to get

$$\int e^{t_S\|\Sigma_S^{-1/2}(\theta_S-\theta_S^\dagger)\|}\,\mathsf{N}_{|S|}(\theta_S \mid \theta_S^\dagger,\Sigma_S)\,d\theta_S \lesssim e^{G|S|\log p},$$

for some constant $G > 0$. Multiplying by $\{(1+\zeta_n)/(1-\zeta_n)\}^{|S|/2}$ does not affect the bound. Averaging the bound $e^{G|S|\log p}$ over low-complexity S's, with $|S| \lesssim s^\star$, is upper bounded by $e^{Gs^\star\log p} = e^{Gn\varepsilon_n^2}$, which proves (20). □

B.2. Proof of Theorem 3

By (17), the marginal posterior mass function π^n satisfies

$$\frac{\pi^n(S)}{\pi^n(S^\star)} \lesssim \frac{\pi_n(S)}{\pi_n(S^\star)}(1+\alpha\gamma)^{-(|S|-|S^\star|)/2}\,\exp[\alpha\{\ell_n(S,\hat\theta_S)-\ell_n(S^\star,\hat\theta_{S^\star})\}],$$

where the constant baked into "$\lesssim$" is 2. By Lemma 3, with probability converging to 1, the exponential term is uniformly upper-bounded in S with $S \supset S^\star$ and $|S| \le s_n$ by $\exp\{\alpha\kappa(\log p)(|S|-|S^\star|)\}$. Then the prior mass ratio satisfies

$$\frac{\pi_n(S)}{\pi_n(S^\star)} \lesssim \frac{\binom{p}{|S^\star|}}{\binom{p}{|S|}}\, p^{-a(|S|-|S^\star|)},$$

so summing the (limiting) upper bound over all $S \supset S^\star$ gives

$$\begin{aligned}
\sum_{S:S\supset S^\star}\pi^n(S) &\lesssim \sum_{S:S\supset S^\star}\frac{\pi^n(S)}{\pi^n(S^\star)} \\
&= \sum_{s=|S^\star|+1}^{s_n}\frac{\binom{p}{|S^\star|}\binom{p-|S^\star|}{p-s}}{\binom{p}{s}}\{(1+\alpha\gamma)^{-1/2}\}^{s-|S^\star|}p^{-(\beta-\alpha\kappa)(s-|S^\star|)} \\
&\le \sum_{s=|S^\star|+1}^{s_n} s^{s-|S^\star|}p^{-(\beta-\alpha\kappa)(s-|S^\star|)} \\
&\le \sum_{s=|S^\star|+1}^{s_n}(s_n p^{-(\beta-\alpha\kappa)})^{s-|S^\star|}.
\end{aligned}$$

Since $s_n < p^{\beta-\alpha\kappa}$ by Condition 3, the dominating series converges and, therefore, the tail of that same series must form a divergent sequence as $|S^\star| \to \infty$, which proves the claim.

B.3. Proof of Theorem 2

Those S with $|S| > C|S^\star|$ that are proper supersets of $S^\star$ have already been covered in the proof of Theorem 3 above. So it suffices to consider S that are large but exclude some important variables. Define the mapping $S \to S^+ = S \cup S^\star$. The only part of the posterior π^n that depends on S itself—not just on $|S|$—is the likelihood component, and the likelihood is increasing in complexity, i.e.,

$$L_n(S,\hat\theta_S) \le L_n(S^+,\hat\theta_{S^+}).$$

So, if $\mathcal{S} = \{S : C|S^\star| < |S| \le s_n \text{ and } S \not\supset S^\star\}$, then we can proceed as follows:

$$\sum_{S\in\mathcal{S}}\pi^n(S) \lesssim \sum_{S\in\mathcal{S}}\frac{\pi^n(S)}{\pi^n(S^\star)}$$

$$
\begin{aligned}
&\leq \sum_{S\in\mathcal{S}} \frac{\pi_n(S)}{\pi_n(S^\star)}(1+\alpha\gamma)^{-(|S|-|S^\star|)/2}\, e^{\alpha\{\ell_n(S^+,\hat\theta_{S^+})-\ell_n(S^\star,\hat\theta_{S^\star})\}} \\
&\leq p^{\alpha\kappa|S^\star|} \sum_{s>C|S^\star|} s_n^{s-|S^\star|} p^{-(\beta-\alpha K)(s-|S^\star|)} \\
&= p^{\alpha\kappa C|S^\star|}(s_n p^{-a})^{(C-1)|S^\star|} \times O(1).
\end{aligned}
$$

Since $s_n \ll p^\beta$ and $\beta > \alpha\kappa C/(C-1)$ by definition of C, the upper bound vanishes, proving the claim.

B.4. Proof of Theorem 4

Take any fixed $\theta^\star$ that meets the stated conditions, and set $S^\star = S(\theta^\star)$ and $s^\star = |S^\star|$. Define $S \in \mathcal{S} := \{S : |S| \leq Cs^\star \text{ and } S \not\supseteq S^\star\}$, where $C > 1$ is as in the statement of Theorem 2. The key point is that there is at least one important variable omitted in the models $S \in \mathcal{S}$. Let $\rho_n^2 = \rho_n^2(\theta^\star) = n^{-1} v \log p$ denote the lower bound on $\theta_j^{\star 2}$ for $j \in S^\star$, as defined in (14), where $v = v(s^\star) = cs^\star \Lambda_{cs^\star}$. In their Supplementary Materials, Narisetty, Shen and He (2019) showed that, with probability tending to 1,[1]

$$
\ell_n(S,\hat\theta_S) - \ell_n(S^\star,\hat\theta_{S^\star}) \leq -F|S\setminus S^\star| n\rho_n^2, \quad \text{uniformly over } S \in \mathcal{S},
$$

for a constant $F > 0$. Then we get the bound

$$
\begin{aligned}
\sum_{S\in\mathcal{S}} \pi^n(S) &\lesssim \sum_{S\in\mathcal{S}} \frac{\pi^n(S)}{\pi^n(S^\star)} \\
&\leq \sum_{S\in\mathcal{S}} \frac{\pi_n(S)}{\pi_n(S^\star)}(1+\alpha\gamma)^{-(|S|-|S^\star|)/2}\, e^{\alpha\{\ell_n(S,\hat\theta_S)-\ell_n(S^\star,\hat\theta_{S^\star})\}} \\
&\leq \sum_{S\in\mathcal{S}} \frac{\binom{p}{s^\star}}{\binom{p}{|S|}}(1+\alpha\gamma)^{-(|S|-s^\star)/2} p^{-\beta(|S|-s^\star)-\alpha F v(|S|-s^\star)},
\end{aligned}
$$

where we used the fact that $|S\setminus S^\star| \geq |S| - s^\star$. Since the summands only depend on $|S|$, the sum over $S \in \mathcal{S}$ can be simplified by first choosing the overall size of the model, then choosing the size of $S \cap S^\star$. That is,

$$
\sum_{S\in\mathcal{S}}(\cdots) = \sum_{s=0}^{Cs^\star} \sum_{t=0}^{s\wedge(s^\star-1)} \binom{s^\star}{t}\binom{p-s^\star}{s-t}(\cdots),
$$

where t indexes the size of $S \cap S^\star$. Note that t can be at most $s^\star - 1$ since S is not allowed to be a superset of $S^\star$. Plugging in the expression for $(\cdots)$ and using the bound

$$
\frac{\binom{s^\star}{t}\binom{p-s^\star}{s-t}\binom{p}{s^\star}}{\binom{p}{s}} \leq s^{s-t} p^{s^\star - t},
$$

[1]The result that they *stated*, i.e., $\ell_n(S,\hat\theta_S) - \ell_n(S^\star,\hat\theta_{S^\star}) \lesssim -n\rho_n^2$, is incorrect—the difference should depend on how close S is to $S^\star$. But the result that they *proved* is the one stated here.

we get

$$\sum_{S\in\mathcal{S}} \pi^n(S) \le \sum_{s=0}^{Cs^\star} \sum_{t=0}^{s\wedge(s^\star-1)} (\phi p^{-\beta})^{s-s^\star} s^{s-t} (p^{1-\alpha J v})^{s^\star-t},$$

where $\phi = (1+\alpha\gamma)^{-1/2}$. Split the outer sum on the right-hand side above into two pieces: $s \le s^\star - 1$ and $s \ge s^\star$. For the first sum,

$$\begin{aligned}\sum_{s=0}^{s^\star-1} \sum_{t=0}^{s} (\phi p^{-\beta})^{s-s^\star} s^{s-t} (p^{1-\alpha F v})^{s^\star-t} &= \sum_{s=0}^{s^\star-1} \Big(\frac{p^\beta}{s\phi}\Big)^{s^\star-s} \sum_{t=0}^{s} (s p^{1-\alpha F v})^{s^\star-t} \\ &\lesssim \sum_{s=0}^{s^\star-1} (\phi^{-1} p^{1+\beta-\alpha F v})^{s^\star-s} \\ &\lesssim \phi^{-1} p^{1+\beta-\alpha F v}.\end{aligned}$$

We have that $\alpha F v > 1+\beta$ because $v = v(s^\star)$ is or can be made large: if $s^\star \to \infty$ then $v \to \infty$ or, otherwise, the constant $c > 1$ baked into v can be chosen sufficiently large. Since ϕ^{-1} is linear in γ, which is at most polynomial in n, the negative power of p dominates so the bound is $o(1)$. Similarly, for the second sum

$$\begin{aligned}\sum_{s=s^\star}^{Cs^\star} \sum_{t=0}^{s^\star-1} (\phi p^{-\beta})^{s-s^\star} s^{s-t} (p^{1-\alpha F v})^{s^\star-t} &= \sum_{s=s^\star}^{Cs^\star} \Big(\frac{p^\beta}{s\phi}\Big)^{s^\star-s} \sum_{t=0}^{s^\star-1} (s p^{1-\alpha F v})^{s^\star-t} \\ &\lesssim s^\star p^{1-\alpha F v} \sum_{s=s^\star}^{Cs^\star} \Big(\frac{s^\star\phi}{p^\beta}\Big)^{s-s^\star} \\ &\lesssim s^\star p^{1-\alpha F v},\end{aligned}$$

where the last "$\lesssim$" follows because $p^\beta \gg s^\star$ and $\phi < 1$. By the same argument as given for the first summation, the remaining term is $o(1)$, so we can conclude that $\sum_{S\in\mathcal{S}} \pi^n(S) \to 0$ with probability tending to 1, which proves the claim.

B.5. Proof of Theorem 5

To start, write the posterior distribution Π^n as

$$\Pi^n(d\theta) = \sum_S \pi^n(S)\, \Pi_S^n(d\theta_S) \otimes \delta_{0_{S^c}}(d\theta_{S^c}),$$

where Π_S^n is the conditional posterior of θ_S, given S, having a density with respect to Lebesgue measure on the corresponding $|S|$-dimensional Euclidean space. Then the total variation distance between Π^n and the Gaussian approximation Ψ^n in (15) is

$$d_{\text{TV}}\big(\Pi^n, \Psi^n\big) \le \sum_S \pi^n(S)\, d_{\text{TV}}\big(\Pi_S^n \otimes \delta_{0_{S^c}},\, \Psi^n\big)$$

$$= \pi^n(S^\star)\, d_{\text{TV}}\big(\Pi^n_{S^\star} \otimes \delta_{0_{S^{\star c}}},\, \Psi^n\big) + \sum_{S \neq S^\star} \pi^n(S)\, d_{\text{TV}}\big(\Pi^n_S \otimes \delta_{0_{S^c}},\, \Psi^n\big).$$

Since the total variation distance is bounded by 2 and $\sum_{S \neq S^\star} \pi^n(S) \to 0$ in $\mathsf{P}_{\theta^\star}$-probability by Theorem 4, the latter term is $o_{\mathsf{P}_{\theta^\star}}(1)$. It remains to show the same for $d_{\text{TV}}\big(\Pi^n_{S^\star} \otimes \delta_{0_{S^{\star c}}},\, \Psi^n\big)$. Since both distributions concentrate on the same $S^\star$ configuration, we can drop the "$\otimes$" terms both above and in (15). That is, the goal is simply to bound

$$d_{\text{TV}}\big\{\Pi^n_{S^\star},\, \mathsf{N}_{|S^\star|}\big(\hat{\theta}_{S^\star},\, \varrho\, J_n(S^\star, \hat{\theta}_{S^\star})^{-1}\big)\big\}.$$

The posterior $\Pi^n_{S^\star}$ has a density $\pi^n_{S^\star}$ with respect to Lebesgue measure; similarly, the normal distribution in the above display has a density with respect to Lebesgue measure, which we will denote by $\psi^n_{S^\star}$. Following the argument used to establish the Laplace approximation result in Lemma 4, it suffices to focus our attention here on integrals over θ_S in a sufficiently small neighborhood of the MLE $\hat{\theta}_S$. Once localized, using the form of the proposed empirical prior, we can apply the continuity property in Lemma 1 to get (locally) pointwise lower and upper bounds, respectively, on the density $\pi^n_{S^\star}$, i.e.,

$$\underbrace{\frac{L_n^\alpha(S^\star, \hat{\theta}_{S^\star})}{m_n(S^\star)\,\{1 + \alpha\gamma(1 \pm \zeta_n)\}^{|S^\star|/2}}}_{=G_n^\pm} \underbrace{\mathsf{N}\big(\theta_{S^\star} \mid \hat{\theta}_{S^\star},\, \tfrac{\gamma}{1+\alpha\gamma(1\pm\zeta_n)} J_n(S^\star, \hat{\theta}_{S^\star})^{-1}\big)}_{=g_n^\pm(\theta_{S^\star})},$$

where $m_n(S^\star)$ denotes the marginal likelihood under configuration $S^\star$ and $\zeta_n = \zeta_{n,S^\star}$ is the vanishing sequence identified in Lemma 1; more on ζ_n below. Denote the lower and upper bounds as $\underline{\pi}^n_{S^\star}$ and $\overline{\pi}^n_{S^\star}$, respectively. If we can show that both the lower and upper bounds have vanishing L_1-distance to the Gaussian approximation, then we are done. And since both of the bounds have the same form, we will focus here on the upper bound $\overline{\pi}^n_{S^\star}$, the one where "$\pm = -$." A simple triangle inequality-type argument gives the bound

$$|\overline{\pi}^n_{S^\star}(\cdot) - \psi^n_{S^\star}(\cdot)| \leq G_n^- |g_n^-(\cdot) - \psi^n_{S^\star}(\cdot)| + |G_n^- - 1| \psi^n_{S^\star}(\cdot),$$

and, consequently,

$$d_{\text{TV}}(\overline{\pi}^n_{S^\star}, \psi^n_{S^\star}) \leq G_n^-\, d_{\text{TV}}(g_n^-, \psi^n_{S^\star}) + |G_n^- - 1|.$$

It follows from Lemma 4 that the $G_n^- \to 1$ in $\mathsf{P}_{\theta^\star}$-probability as $n \to \infty$. So it remains to bound the total variation distance in the above display, which is between two Gaussians with a common mean and very similar variances. A special case of the general result in Theorem 1.8 of Arbas, Ashtiani and Liaw (2023)—see, also, Devroye, Mehrabian and Reddad (2023)—gives that

$$d_{\text{TV}}(g_n^-, \psi^n_{S^\star}) \lesssim \Big| \varrho \div \frac{\gamma}{1 + \alpha\gamma(1 - \zeta_n)} - 1 \Big|\, |S^\star|^{1/2}.$$

It is easy to check that the upper bound is of the order $\zeta_n |S^\star|^{1/2}$, so we need to consider how quickly ζ_n is vanishing. Recall that, following the statement of

Lemma 1, it was noted that, for a particular configuration, in this case $S = S^\star$, the choice of ζ_n was of the order $\{n^{-1}|S^\star|^2\Lambda_{|S^\star|}\log p\}^{1/2}$. That means

$$d_{\text{TV}}(g_n^-, \psi_{S^\star}^n) \lesssim \{n^{-1}|S^\star|^3\Lambda_{|S^\star|}\log p\}^{1/2}.$$

By the additional configuration size condition (16), the upper bound above is vanishing. This proves that $d_{\text{TV}}\big(\Pi^n, \Psi^n\big) \to 0$ in $\mathsf{P}_{\theta^\star}$-probability and, since it is also uniformly bounded, the theorem's statement (in term of expectations) follows from the dominated convergence theorem.

Acknowledgments

The authors thank the Editor and anonymous Associate Editor and reviewers for their helpful feedback. Special thanks to Jeyong Lee and Minwoo Chae for spotting some technical issues in a previous version of the manuscript, and to Naveen Narisetty for sharing his skinny Gibbs codes.

Funding

This work was partially supported by the U. S. National Science Foundation, under grants DMS–1737933, DMS–1811802, and SES–205122.

References

Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. *Electronic Journal of Statistics* **4** 932–949. MR2721039

Abramovich, F. and Grinshtein, V. (2016). Model selection and minimax estimation in generalized linear models. *IEEE Transactions on Information Theory* **62** 3721–3730. MR3506759

Abramowitz, M. and Stegun, I. A. (1966). *Handbook of Mathematical Functions*. Dover, New York. MR0208797

Arbas, J., Ashtiani, H. and Liaw, C. (2023). Polynomial time and private learning of unbounded Gaussian mixture models. In *Proceedings of the 40th International Conference on Machine Learning. ICML'23*. JMLR.org.

Arias-Castro, E. and Lounici, K. (2014). Estimation and variable selection with exponential weights. *Electronic Journal of Statistics* **8** 328–354. MR3195119

Barber, R. F., Drton, M. and Tan, K. M. (2016). Laplace approximation in high-dimensional Bayesian regression. In *Statistical Analysis for High-Dimensional Data* 15–36. Springer. MR3616262

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Ann. Statist.* **32** 870–897. MR2065192

Belitser, E. and Ghosal, S. (2020). Empirical Bayes oracle uncertainty quantification for regression. *The Annals of Statistics* **48** 3113–3137. MR4185802

Belitser, E. and Nurushev, N. Needles and straw in a haystack: Robust confidence for possibly sparse sequences. *Bernoulli* **26** 191–225. MR4036032

Bhadra, A., Datta, J., Polson, N. G. and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis* **12** 1105–1131. MR3724980

Bhadra, A., Datta, J., Polson, N. G. and Willard, B. (2019a). Lasso meets horseshoe: A survey. *Statistical Science* **34** 405–427. MR4017521

Bhadra, A., Datta, J., Li, Y., Polson, N. G. and Willard, B. (2019b). Prediction risk for the horseshoe regression. *Journal of Machine Learning Research (JMLR)* **20** Paper No. 78, 39. MR3960932

Bondell, H. D. and Reich, B. J. (2012). Consistent high-dimensional Bayesian variable selection via penalized credible regions. *Journal of the American Statistical Association* **107** 1610–1624. MR3036420

Bühlmann, P. (2011). Comments on 'Regression shrinkage and selection via the lasso: A retrospective'. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** 277-279.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Springer Series in Statistics*. Springer, Heidelberg. MR2807761

Cao, X. and Lee, K. (2020). Variable Selection Using Nonlocal Priors in High-Dimensional Generalized Linear Models With Application to fMRI Data Analysis. *Entropy* **22** 807. MR4220303

Carvalho, C. M., Polson, N. G. and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97** 465–480. MR2650751

Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics* **43** 1986–2018. MR3375874

Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40** 2069–2101. MR3059077

Chicco, D. and Jurman, G. (2023). The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining* **16** 4.

Devroye, L., Mehrabian, A. and Reddad, T. (2023). The total variation distance between high-dimensional Gaussians with the same mean. arXiv: 1810.08693.

Fang, X. and Ghosh, M. (2023). High-dimensional properties for empirical priors in linear regression with unknown error variance. *Statistical Papers*, to appear. MR4693352

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33** 1–22.

George, E. I. and McCullogh, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88** 881–889.

Goodrich, B., Gabry, J., Ali, I. and Brilleman, S. (2022). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.21.3.

Grünwald, P. D. and Mehta, N. A. (2020). Fast rates for general unbounded

loss functions: from ERM to generalized Bayes. *The Journal of Machine Learning Research* **21** 2040–2119. MR4095335

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis* **12** 1069–1103. MR3724979

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning*, 2nd ed. Springer-Verlag, New York. MR1851606

Jeong, S. and Ghosal, S. (2021). Posterior contraction in sparse generalized linear models. *Biometrika* **108** 367–379. MR4259137

Lee, K. and Cao, X. (2021). Bayesian group selection in logistic regression with application to MRI data analysis. *Biometrics* **77** 391–400. MR4307642

Lee, J., Chae, M. and Martin, R. (2024). Advances in Bayesian model selection consistency for high-dimensional generalized linear models. In preparation.

Liu, C. and Martin, R. (2019). An empirical G-Wishart prior for sparse high-dimensional Gaussian graphical models. arXiv:1912.03807.

Liu, C., Martin, R. and Shen, W. (2023). Empirical priors and posterior concentration in a piecewise polynomial sequence model. *Statistica Sincica*, to appear; arXiv:1712.03848.

Martin, R. (2019). Empirical priors and posterior concentration rates for a monotone density. *Sankhyā A.* **81** 493–509. MR4043484

Martin, R., Mess, R. and Walker, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli* **23** 1822–1847. MR3624879

Martin, R. and Ning, B. (2020). Empirical priors and coverage of posterior credible sets in a sparse normal mean model. *Sankhya A* **82** 477–498. MR4136243

Martin, R. and Tang, Y. (2020). Empirical Priors for Prediction in Sparse High-dimensional Linear Regression. *Journal of Machine Learning Research* **21** 1–30. MR4138128

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics* **8** 2188–2206. MR3273623

Martin, R. and Walker, S. G. (2019). Data-driven priors and their posterior concentration rates. *Electronic Journal of Statistics* **13** 3049–3081. MR4010592

McCullagh, P. M. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London. MR3223057

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association* **114** 1113–1125. MR4011766

Narisetty, N. N., Shen, J. and He, X. (2019). Skinny Gibbs: a consistent and scalable Gibbs sampler for model selection. *Journal of the American Statistical Association* **114** 1205–1217. MR4011773

Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*

11 5018–5051. MR3738204

Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics* **40** 639–665. MR2933661

Rigollet, P. and Tsybakov, A. B. (2012). Sparse estimation by exponential weighting. *Statist. Sci.* **27** 558–575. MR3025134

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)* **57** 749–760. MR1354079

Spokoiny, V. (2017). Penalized maximum likelihood estimation and effective dimension. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques* **53** 389–429. MR3606746

Syring, N. and Martin, R. (2018). Calibrating general posterior credible regions. *Biometrika* **106** 479–486. MR3949316

Syring, N. and Martin, R. (2023). Gibbs posterior concentration rates under sub-exponential type losses. *Bernoulli* **29** 1080–1108. MR4550216

Tang, Y. and Martin, R. (2021). ebreg: Implementation of the empirical Bayes method R package version 0.1.3.

van der Pas, S., Szabó, B. and van der Vaart, A. (2017). Uncertainty Quantification for the Horseshoe (with Discussion). *Bayesian Analysis* **12** 1221–1274. MR3724985

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society. Series B. Statistical Methodology* **63** 811–821. MR1872068

Walker, S. G., Lijoi, A. and Prünster, I. (2005). Data tracking and the understanding of Bayesian consistency. *Biometrika* **92** 765–778. MR2234184

Wei, R. and Ghosal, S. (2020). Contraction properties of shrinkage priors in logistic regression. *Journal of Statistical Planning and Inference* **207** 215–229. MR4066132