ABSTRACT

TANG, YIQI. Sparsity-encouraging Data-driven Priors in High-dimensional Problems. (Under the direction of Ryan Martin).

High-dimensional statistical inference problems are challenging, and, for Bayesians, the choice of prior distribution for the unknowns is highly influential and non-trivial. Under commonly assumed sparsity structures for the unknowns, we employ a novel empirical priors framework that uses the data to appropriately center the prior distribution to solve high-dimensional problems in a couple different contexts. First, we explore the theoretical and computational performance of this framework on prediction in linear regression; next, we use this empirical priors framework for estimation and variable selection in generalized linear models; finally, we consider a computational speed-up using a suitable variational approximation.

In the first project, we adopt the familiar sparse, high-dimensional linear regression model but focus on the task of prediction, where the relevant unknown is the response variable corresponding to a given set of predictor variable values. In particular, we investigate our empirical prior method's theoretical and numerical performance in the context of prediction. We show that, in certain settings, the asymptotic posterior concentration in metrics relevant to prediction quality is fast, and we establish a Bernstein–von Mises theorem ensuring that the derived prediction intervals achieve the target coverage probability asymptotically. Numerical results complement the asymptotic theory, showing that, in addition to having strong finite-sample performance in terms of prediction accuracy and uncertainty quantification, the computation time is considerably faster compared to existing Bayesian methods.

In the second project, we use the same empirical or data-driven prior framework for inference—estimation and variable selection—in sparse, high-dimensional generalized linear models, including logistic regression. For our proposed method, we prove that the posterior concentrates around the true/sparse coefficient vector at the optimal rate and, furthermore, provide conditions under which the posterior can achieve variable selection consistency. Computation of the proposed empirical Bayes posterior is simple and efficient, and, in terms of variable selection in logistic and Poisson regression, is shown to perform well in simulations compared to existing Bayesian and non-Bayesian methods.

The first two projects both utilize a Metropolis–Hastings Markov chain Monte Carlo strategy to approximately sample from the relevant posterior distribution. In the third project, we employ a novel variational approximation to our empirical prior approach in the context of high-dimensional logistic regression. Specifically, we propose an independent-Bernoulli variational approximation directly on the marginal posterior distribution of the active set

of variables, and derive a fast and efficient coordinate-ascent algorithm to find the solution. We prove that our proposed variational approximation shares the same selection consistency property that the posterior satisfies. We also how that the numerical performance of our method is on-par with state-of-the-art methods for variable selection in logistic regression.

Sparsity-encouraging Data-driven Priors in High-dimensional Problems

by
Yiqi Tang

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina
2023

APPROVED BY:

Sujit Ghosh	Amanda Lietz
Ana-Maria Staicu	Jonathan Williams

Ryan Martin
Chair of Advisory Committee

DEDICATION

To Athena, Kirk, and my parents Ningfeng and Tao.

BIOGRAPHY

Yiqi Tang was born and raised in Beijing, China. Halfway through high school, she moved to Connecticut and graduated from Miss Porter's School in 2011. She obtained her Bachelor of Science in 2015 from Davidson College double majoring in mathematics and economics. After two years of working at a small finance company, she moved to Raleigh to start her graduate studies in statistics at NC State University, first getting her Master of Statistics in 2019, and now her Doctorate in 2023. During her PhD studies, she interned for two summers at Intuit Inc. and took a year off to be a Visiting Lecturer of Statistics at Williams College 2021-2022. She will be an Assistant Professor of Statistics at Colby College in the fall of 2023.

ACKNOWLEDGEMENTS

There are many people I'd like to thank, without whom I would not be where I am today. First, I'd like to thank my advisor Dr. Ryan Martin. He has been an amazing mentor, role model, and teacher, showing me the ropes and guiding me in not only my research, but sharing his wisdom on life in academia in general. I have always appreciated his candor, and feel so lucky to have had him as my advisor. I'd also like to thank my other committee members, Dr. Ana-Maria Staicu, Dr. Sujit Ghosh, Dr. Jonathan Williams, and Dr. Amanda Lietz. I am grateful to them all for sharing their time and knowledge with me during my graduate studies and in the development of my dissertation.

Thank you to many in the Statistics department at NC State, in particular, I'd like to thank Dr. Emily Griffith for always being a resource and sounding board. Thank you to the wonderful friends I made in the department, Chelsea Robalino, Laura Wendelberger, Julia Holter, Michael McKibben, Kevin Gunn, among many others, who supported me through ups and downs over the years and made statistics fun.

Lastly, I'd like to thank my family. My mom and dad have always put me and my education first, encouraging me to travel far away to the United States to get the best education. They have instilled in me a love of learning, and sacrificed everything for me to pursue my dreams. Thank you to my daughter, Athena, for making me a mom, for bringing so much joy into our lives, and for giving me the inspiration and motivation to strive for better. Finally, thank you to my spouse, my partner, my husband, Kirk, for always being there for me, and for pushing me to do my best work and be the best version of myself.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

We live in a world with an abundance of data. Commonly termed as “big data", people generally use the phrase loosely to describe large datasets. In statistics, there are two distinct types of big data: first, where the number of observations n is large, while the number of variables p remains relatively small; and, second, where the number of observations may or may not be large, but the number of variables p is large, specifically, larger than n. High-dimensional statistics is concerned with the second of the two. There are many application areas for this type of data to occur. For example, in the social sciences such as economics or political science, studies might be done with not many participants, but each participant would answer many survey questions. In this case, $n < p$, but n and p would likely both be of moderate size, say, in the hundreds. Another example is genomic data, where there are only a number of samples (n) and many predictors (p), one for each gene expression level. The goal may be to find connections between quantitative traits and gene expression levels (linear regression), or find which genes affect the presence or absence of a particular trait (logistic regression).

Many challenges arise when the traditional “$n > p$” rule is violated, both theoretical and computational. Basic theorems that are fundamental to statistics no longer hold when p becomes larger than n. Asymptotic theory that relies on $n \to \infty$ and p fixed cannot be readily extended, as p grows with n, and in fact, $p \to \infty$ as $n \to \infty$. This has been investigated in the literature, e.g., Johnstone and Titterington (2009); Lindsay et al. (2004). Computationally,

high-dimensional problems are expensive to solve, requiring more sophisticated algorithms and greater computational power (Fan and Li 2005).

An important question is, when $p \gg n$, how does this affect our basic statistical principles? Specifically, can fundamental principles like *maximum likelihood* carry over into the $p \gg n$ setting? The answer is *No*, at least not directly, and the key observation is that, in high-dimensional settings, good estimation requires a sort of information sharing that standard methods like maximum likelihood do not possess, at least not by default. We will briefly illustrate this with a simple example. Consider the sparse normal means problem, $y_i \overset{\text{ind}}{\sim} \mathsf{N}(\theta_i, 1)$, $i = 1, 2, \ldots, n$. In other words, the data y is a n-vector and the unknown mean θ is a n-vector too; this case counts as "high-dimensional" because $p = n$. The most obvious estimator for this data would be the maximum likelihood estimator (MLE), $\hat{\theta}_{\text{MLE}} = y$. Unfortunately, this estimator is inadmissible with respect to the ℓ_2-loss, i.e., there exists other estimators $\tilde{\theta}$ such that $\mathsf{E}_\theta \|\tilde{\theta} - \theta\|^2 \leq \mathsf{E}_\theta \|\hat{\theta}_{\text{MLE}} - \theta\|^2$ for all θ, with strict inequality for some θ. This paradigm-shifting result was first shown by Stein (1956). Stein's proof was based on a construction of an estimator—now known as the *James–Stein estimator*—that satisfies the above inequalities. The James–Stein estimator $\hat{\theta}_{\text{JS}} = (1 - \frac{m-2}{\|y\|^2})y$ is shown to dominate $\hat{\theta}_{\text{MLE}}$ in l_2-loss for all $m \geq 3$, and is perhaps the first example of information sharing in the sense mentioned above. The $\|y\|^2$ term implies that the estimate of θ_j does not just depend on y_j, but on the entire y vector. This kind of information sharing initially appeared paradoxical: there is nothing apparent about the problem formulation that suggests y_1 should be relevant to estimation of θ_2, so it hardly makes sense that it is appropriate to share information in this way, let alone be necessary in a decision-theoretic sense. It is now understood that the information sharing is required because the goal is simultaneous estimation of (or inference on) the entire vector θ. With so much of modern statistical focus on these high-dimensional problems, the information sharing, shrinkage, and regularization ideas that were once thought of as paradoxical have now become a central part of our core principles.

How is this information sharing/regularization achieved in modern statistical settings? For frequentist solutions, a penalty term is added. As an example, we will consider the lasso estimator. The method, first proposed by Tibshirani (1996), is one of the most popular frequentist methods in high-dimensional analysis. For our toy problem, the lasso solution is $\hat{\theta}_{\text{lasso}} = \text{sign}(y_i)(|y_i| - \lambda)_+$, where $x_+ = \max(0, x)$ is the positive part of x. The information sharing here comes through the choice of the tuning parameter λ, which is carried out in practice by using all of the data y. The Bayesian solution typically uses a hierarchical prior that induces a certain dependence between the different θ_i's, thereby forcing the (marginal) posterior for θ_i to depend on all of y. We will also use this type of hierarchical structure in our analysis, by decomposing β into (S, β_S) where $S \subset \{1, 2, \ldots, p\}$ is the active set, the set of indices that

correspond to the locations of the signals in the β vector.

This dissertation focuses on high-dimensional regression, first on prediction in linear regression, then on variable selection and estimation in generalized linear models. In all the problems considered in this dissertation, we observe data (X, y), where X is an $n \times p$ matrix and y is an $n \times 1$ vector. There is an unknown $p \times 1$ vector of coefficients β that we wish to learn about. An initial obstacle to achieving this aim is that the model cannot be fit without some additional structure. As is common in the literature, we will assume a *sparsity* structure on the high-dimensional β vector. That is, we will assume that most of the entries in β are zero; this will be made more precise in the following chapters. There are both statistical and practical reasons for this assumption. Statistically, it is infeasible to recover the signals without this assumption. There are no proven good methods to solve problems where the sparsity assumption does not make sense for describing the observed data. Thus, the sparsity assumption is needed theoretically. Luckily, most scientific problems lend themselves to the sparsity assumption. In other words, practically speaking, sparsity usually results from the specific high-dimensional scientific application problem. As an example, in the genomic application mentioned previously, each trait is in fact only affected by a handful of gene expression levels. Of course, theoretically, we also cannot obtain good estimators if the unknown parameter is too complex, and computationally it would not be achievable either.

With a sparsity structure assumption, a plethora of methods are now available for estimating a sparse β. Frequentist methods incorporate the sparsity structure with penalized regression, e.g., lasso (Tibshirani 1996), adaptive lasso (Zou 2006), SCAD (Fan and Li 2001), and others; moreover, software is available to carry out the relevant computations easily and efficiently. But what if we want a readily-available probability distribution for uncertainty quantification? A Bayesian solution would then be the natural choice.

With a Bayesian solution, a full probability distribution would allow us to easily quantify uncertainty, but there are some challenges. In low-dimension problems, e.g., where p is fixed as $n \to \infty$, the prior does not play a substantial role since there is abundant data to offer enough information about the parameters. This is the well-known *merging of opinions* phenomenon, i.e., the posterior distributions based on any two (reasonable) prior distributions will asymptotically agree. In high-dimensional problems, on the other hand, the choice of prior matters a lot. This is because, since $p = p_n \to \infty$, there is not an accumulation of information about each individual parameter to wash out the prior effects. Since we are often lacking complete and genuine prior information about the unknowns, and since there are no "default" prior options in such cases, the natural choice is to specify prior distributions that lead to desirable posterior operating characteristics, e.g., optimal posterior concentration rates. As the reader can imagine, this is highly non-trivial. Castillo et al. (2015) and others have demonstrated

that in order to achieve the optimal concentration rates, the prior for the non-zero β coefficients must have sufficiently heavy tails, in particular, heavier than the conjugate Gaussian tails. This constraint leads to the second challenge, namely, computation of the posterior distribution. While general Markov chain Monte Carlo (MCMC) methods are available, the individual steps can be expensive and the chain can be slow to converge. Some believe these computations to be prohibitively slow for priors that include a discrete component for the zero coefficients, so they prefer continuous shrinkage priors like the horseshoe (Carvalho et al. 2010) and Dirichlet–Laplace (Bhattacharya et al. 2015).

The computational difficulties mentioned above stem from the need to work with the heavy-tailed priors that yield desired posterior concentration properties. If we could center the prior correctly, the prior tails would no longer matter. The question is, how could we appropriately center the prior? We propose the use of data to guide this selection. Inspired by the insight that prior tails would be irrelevant if the prior center was appropriately chosen, Martin et al. (2017) developed a new empirical Bayes approach for high-dimensional regression that incorporates data to both center the prior and to provide some extra regularization. This approach is powerful because it allows for conjugate priors to be used, which drastically speeds up computation, but without sacrificing on the desirable concentration rate properties enjoyed by the fully Bayesian approach with heavy-tailed priors. See, also, Martin and Walker (2019). We employ this approach in high-dimensional linear regression, focusing on prediction, and in high-dimensional generalized linear models.

In Chapter 2, we extend the empirical priors approach to the task of prediction in high-dimensional linear regression. Martin et al. (2017) have studied the theoretical properties of this approach in inference and variable selection in the high-dimensional linear regression context and investigated its numerical performance for variable selection. We continue this work by looking at prediction. We derive the predictive distribution for our empirical priors method, establish predictive rate theorems for our predictive distribution, and prove a Bernstein–von Mises theorem. We show through simulations and a real-data study that our method provides good point predictions and prediction intervals, results that are competitive across settings compared to other methods.

After tackling linear regression, we move on to another popular regression model form, namely, generalized linear models (GLM). In Chapter 3, we investigate the same empirical priors approach applied to high-dimensional generalized linear models. Instead of only focusing on high-dimensional logistic regression, as is common in the literature, we take a more general formulation, showing results that are not limited to one specific GLM model, but apply to all GLMs. We show theoretical results on variable selection consistency and posterior concentration of β to the true coefficient vector $\beta^\star$. Numerical simulations are carried out in two most

popular GLM model forms, logistic regression and Poisson regression. We demonstrate strong empirical performance of our method across different settings in these two models.

Chapter 4 extends the work of Chapter 3 in high-dimensional GLMs, but takes it in a slightly different direction. In the first two projects, a Metropolis-Hastings MCMC is used to sample the S space. This computational approach is fairly efficient for moderate p, since our empirical prior allows for a conjugate posterior with a simple form. However, as p increases, say, to the thousands, this approach is no longer computationally feasible. Chapter 4 uses the same empirical priors approach, but instead of MCMC, variational inference is used for computation. Variational methods have become quite popular in Bayesian analysis, replacing MCMC methods, especially in problems with complex models and large parameter spaces. Instead of using MCMC to approximate a marginal posterior that is not available in closed-form, variational approximation first finds a family of approximate densities, then searches for the member of the family that minimizes the Kullback–Leibler divergence to the exact posterior of interest, and finally, the exact posterior is approximated with that family member. We investigate the empirical priors method with variational inference in high-dimensional logistic regression. Existing work by Yang and Martin (2020) used a variational approximation approach with empirical priors in high-dimensional linear regression. We take a similar approach, but instead of approximating the posterior distribution with a mean-field family like they do, we propose a variational approximation on the marginal posterior for S directly, using an independent-Bernoulli model to approximate. We show that our method is simple and efficient, not only performing very well in simulations in terms of accuracy, but is also quick to compute.

CHAPTER

2

EMPIRICAL PRIORS FOR PREDICTION IN HIGH-DIMENSIONAL LINEAR REGRESSION

2.1 Introduction

Consider a linear regression model

$$y = X\beta + \sigma z, \tag{2.1}$$

where y is a $n \times 1$ vector of response variables, X is a $n \times p$ matrix of explanatory variables, β is a $p \times 1$ vector of regression parameters, $\sigma > 0$ is an unknown scale parameter, and z is a $n \times 1$ vector of independent standard normal errors. Here, our interest is in the high-dimensional setting where $p \gg n$, and our particular aim is to predict the value of a new response $\tilde{y} \in \mathbb{R}^d$ at a given $\tilde{X} \in \mathbb{R}^{d \times p}$, $d \geq 1$, an important and challenging problem in these high-dimensional scenarios.

For the task of prediction, given an estimator of β, it is conceptually straightforward to produce a point prediction of a new response. However, the regularization techniques employed by frequentist methods such as lasso (Tibshirani 1996), adaptive lasso (Zou 2006), SCAD (Fan and Li 2001) cause the estimators to have non-regular distribution theory (e.g., Pötscher and Leeb 2009), so results on uncertainty quantification, i.e., coverage properties of prediction

intervals, are few in number; but see Leeb (2006, 2009) and the references therein.

On the Bayesian side, given a full probability model, it is conceptually straightforward to obtain a predictive distribution for the new response and suggest some form of uncertainty quantification, but there are still a number of challenges. In any case, even if the first two challenges can be overcome and a predictive distribution for $\tilde{y}$ can be obtained, it is not automatic that the prediction intervals from this distribution provide valid uncertainty quantification, i.e., that the posterior 95% predictive interval will contain the to-be-observed value of $\tilde{y}$ with probability 0.95.

Our goal in this chapter is to investigate the performance of the empirical Bayes approach in Martin et al. (2017) in the context of predicting a new response. First, we review their empirical Bayes formulation in Section 2.2, adding some details about the unknown-σ case. Next, in Section 2.3, we turn to the prediction task and show that, thanks to the empirical prior's conjugacy, the corresponding empirical Bayes predictive distribution has a relatively simple form and can be easily and efficiently sampled via standard Monte Carlo techniques. Theoretical properties are investigated in Section 2.4 and, in particular, we show that our empirical Bayes predictive distribution has fast convergence rates, nearly parametric in some cases, both for in- and out-of-sample prediction settings. Moreover, under reasonable conditions, we establish a Bernstein–von Mises theorem, which implies that the derived posterior prediction intervals have the target coverage probability asymptotically. In Section 2.5 we demonstrate, in both real- and simulated-data examples, that the proposed empirical Bayes framework provides accurate point prediction, valid prediction uncertainty quantification, and fast computation across various settings compared to a number of existing methods. Finally, some concluding remarks are given in Section 2.6. We have also developed an R package, `ebreg` (Tang and Martin 2021), that provides users with tools for estimation and variable selection, as described in Martin et al. (2017), and the tools for prediction as presented here.

2.2 Empirical prior for regression

2.2.1 Known σ^2

Here we review the empirical prior approach for sparse, high-dimensional regression laid out in Martin et al. (2017). Like Castillo et al. (2015) and others, they focus on the known-σ^2 case, so we present their formulation here. Adjustments to handle the more realistic unknown-σ^2 case are described in Section 2.2.2.

Under the sparsity assumption, it is natural to decompose the high-dimensional vector β as (S, β_S), where $S \subseteq \{1, 2, \ldots, p\}$ is the configuration of β, i.e., the set of indices corresponding

to non-zero/active coefficients, and β_S is the $|S|$-vector of non-zero values; here $|S|$ denotes the cardinality of the finite set S. This decomposition suggests a hierarchical model with a marginal prior for S and a conditional prior for β_S, given S.

For the marginal prior for S, we take the mass function

$$\pi(S) = \tbinom{p}{|S|}^{-1} q_n(|S|), \quad S \subset \{1, 2, \ldots, p\}, \quad |S| \leq R, \tag{2.2}$$

where q_n is a mass function on $\{0, 1, \ldots, R\}$, which we take to be

$$q_n(s) \propto (c\, p^a)^{-s}, \quad s = 0, 1, \ldots, R, \tag{2.3}$$

with $R = \text{rank}(X)$ and (a, c) some hyperparameters to be specified; see Section 2.5. This corresponds to a truncated geometric prior on the configuration size and a uniform prior on all configurations of the given size; see, also, Castillo et al. (2015).

There is an assumption hidden in the definition (2.2) that deserves comment. The prior does not support all possible configurations, only those of size no more than $R \leq n \ll p$. The rationale for this restriction is that the true value of β, namely $\beta^\star$, is assumed to be sparse, i.e., the true configuration, $S^\star = S_{\beta^\star}$, is of size much smaller than n, so there is no reason not to incorporate an assumption like "$|S^\star| \leq n$" into the prior. Since $R \leq n$, the assumption "$|S^\star| \leq R$" encoded in (2.2) is generally stronger. However, $R = n$ is typical, e.g., if the rows of X are iid p-vectors with positive definite covariance matrix, in which case "$|S^\star| \leq R$" and "$|S^\star| \leq n$" are equivalent. Moreover, nothing significant changes about the theory if $|S^\star| \leq R < n$; see, e.g., Abramovich and Grinshtein (2010, Theorem 3). But if it happens that $R < |S^\star| < n$, then the compatibility condition described in Castillo et al. (2015) fails, so even an oracle who knows which variables are important would be unable to give a fully satisfactory solution to the problem. Since one inevitably must assume that $|S^\star| \leq R$ to establish any good properties, we opt to include such an assumption in the prior construction, via (2.2) and (2.3).

The empirical or data-dependent element comes in the conditional prior for β_S, given S. That is, set

$$\beta_S \mid S \sim \mathsf{N}_{|S|}\big(\hat{\beta}_S, \sigma^2 \gamma^{-1} (X_S^\top X_S)^{-1}\big), \tag{2.4}$$

where X_S is the $n \times |S|$ submatrix of X with only the columns corresponding to the configuration S, $\hat{\beta}_S$ is the least squares estimate based on design matrix X_S, and $\gamma > 0$ is a precision parameter to be specified. Except for being centered on the least squares estimator, this closely resembles Zellner's g-prior (e.g., Zellner 1986). Again, the idea behind a data-dependent centering is to remove the influence of the prior tails on the posterior concentration, which requires use of the data. See Martin and Walker (2014, 2019) and Martin et al. (2017) for more on this point.

Putting the two pieces together, we have the following empirical prior for β:

$$\Pi_n(d\beta) = \sum_S \pi(S) \mathsf{N}_{|S|}\big(d\beta_S \mid \hat{\beta}_S, \sigma^2\gamma^{-1}(X_S^\top X_S)^{-1}\big) \otimes \delta_{0_{S^c}}(d\beta_{S^c}), \tag{2.5}$$

where $\delta_{0_{S^c}}$ denotes a Dirac point-mass distribution at the origin in the $|S^c|$-dimensional space, and $\mu \otimes \nu$ is the product of two measures μ and ν. This is a spike-and-slab prior where the spikes are point masses and the slabs are conjugate normal densities, which have nice computational properties but are centered at a convenient estimator to eliminate the thin-tail effect on the posterior concentration rate.

Next we combine this prior with the likelihood in almost the usual way. That is, for a constant $\alpha \in (0,1)$, define the corresponding empirical Bayes posterior Π^n for β as

$$\Pi^n(d\beta) \propto L_n(\beta)^\alpha \, \Pi_n(d\beta), \tag{2.6}$$

where

$$L_n(\beta) = \mathsf{N}_n(y \mid X\beta, \sigma^2 I) \propto \exp\{-\tfrac{1}{2\sigma^2}\|y - X\beta\|^2\}, \tag{2.7}$$

is the likelihood, with $\|\cdot\|$ the ℓ_2-norm on $\mathbb{R}^n$. The power α is unusual, but the role it plays is to flatten out the posterior, effectively discounting the data slightly. Martin et al. (2017) describe this as a regularization that prevents the posterior from chasing the data too closely, and similar discounting ideas have been used for robustness purposes in certain misspecified models (e.g., Grünwald and van Ommen 2017; Holmes and Walker 2017; Syring and Martin 2018). In our present context, the α discount is a technical device to help the posterior adapt to the unknown sparsity (see Martin and Walker 2019). For those readers uncomfortable with the power likelihood in (2.6), an equivalent representation is as a genuine Bayesian update

$$\Pi^n(d\beta) \propto L_n(\beta) \Pi_n^{\text{reg}}(d\beta),$$

where

$$\Pi_n^{\text{reg}}(d\beta) \propto L_n(\beta)^{-(1-\alpha)} \, \Pi_n(d\beta)$$

is a version of the above empirical prior with an extra data-driven regularization, penalizing those β values that fit the data too well when $L_n(\beta)$ is large. In any case, we recommend taking $\alpha \approx 1$ in applications so there is no practical difference between our proposal and a closer-to-genuine Bayes posterior with $\alpha = 1$. The end result is a posterior distribution, Π^n, for β that depends on α, γ, and, in this case, the known σ^2.

In terms of basic first-order properties of Π^n, such as asymptotic concentration rates, the results are not sensitive to the choice of α and γ. That is, the same concentration rates are

obtained for all $\alpha \in (0,1)$ and all $\gamma > 0$. However, for higher-order properties, such as coverage of posterior credible sets, some conditions on α and γ are required. As discussed in Section 2.4, we will require $\alpha + \gamma \leq 1$, so if α is close to 1, then γ must be close to 0. The apparent impact of $\gamma \approx 0$ is that the conditional prior for β_S, given S, in (2.4) is wide and "non-informative" in a traditional sense. But, for example, if we treat the rows of X as iid samples with mean 0 and covariance matrix Σ, then for any fixed S, we have that $(X_S^\top X_S)^{-1} = n^{-1}\widehat{\Sigma}_S^{-1}$, where Σ_S is the submatrix corresponding to configuration S and the "hat" indicates an estimate based on the sample in X. So the prior for β_S, given S, has a data-driven center and variance roughly $O(n^{-1})$. Therefore, choosing $\gamma \approx 0$ as suggested by the theory also has some practical intuition: it effectively prevents the conditional prior for β_S, given S, from being "too informative."

An important practical advantage of this formulation is that Π^n is relatively simple. Indeed, the conditional posterior distribution for β_S, given S, is just

$$\pi^n(\beta_S \mid S) = \mathsf{N}_{|S|}\big(\beta_S \mid \hat{\beta}_S, \tfrac{\sigma^2}{\alpha+\gamma}(X_S^\top X_S)^{-1}\big). \tag{2.8}$$

For variable selection-related tasks, the marginal posterior for the configuration, S, is the relevant object, and a closed-form expression is available:

$$\pi^n(S) \propto \pi(S)\big(\tfrac{\gamma}{\alpha+\gamma}\big)^{|S|/2} \exp\{-\tfrac{\alpha}{2\sigma^2}\|y - \hat{y}_S\|^2\},$$

where $\hat{y}_S$ is the fitted response based on the least squares fit to (y, X_S). Martin et al. (2017) propose a Metropolis–Hastings procedure to sample from S and, if β samples are also desired, then one can augment the S sampler by sampling from the conditional posterior for β_S, given S, along the way.

2.2.2 Unknown σ^2

For the realistic case where the error variance is unknown, there are different strategies one can employ. The simplest strategy, taken in Martin et al. (2017), is to construct an estimator, $\hat{\sigma}^2$, and plug it in to the known-σ^2 formulas above. They used a lasso-driven estimator, discussed in Reid et al. (2016), in their numerical examples, and their method had very good performance. But variance estimates post-selection can be unreliable (e.g., Hong et al. 2018), which can impact other posterior summaries, such as credible regions, so we want to consider an alternative based on a prior distribution for σ^2.

Consider an inverse gamma prior for σ^2, with density

$$\pi(\sigma^2) = b_0^{a_0}\Gamma(a_0)^{-1}(\sigma^2)^{-(a_0+1)}e^{-b_0/\sigma^2}, \quad \sigma^2 > 0,$$

where a_0 and b_0 are fixed shape and scale parameters, respectively. Incorporating this into the prior formulation described above, expanding the likelihood in (2.7) as

$$L_n(\beta,\sigma^2) = \mathsf{N}_n(y \mid X\beta, \sigma^2 I) \propto (\sigma^2)^{-1/2} \exp\{-\tfrac{1}{2\sigma^2}\|y - X\beta\|^2\},$$

to include σ^2, and combining the two as in (2.6), the following properties of the posterior distribution are easy to verify. First, the conditional posterior for β_S, given S and σ^2 is exactly as in (2.8); second, the conditional posterior distribution for σ^2, given S, is again inverse gamma with shape $= a_0 + \frac{\alpha n}{2}$ and scale $= b_0 + \frac{\alpha}{2}\|y - \hat{y}_S\|^2$; and, finally, the marginal posterior for S is

$$\pi^n(S) \propto \pi(S)\big(\tfrac{\gamma}{\alpha+\gamma}\big)^{|S|/2}\big\{b_0 + \tfrac{\alpha}{2}\|y - \hat{y}_S\|^2\big\}^{-(a_0+\alpha n/2)}. \tag{2.9}$$

Therefore, the MCMC strategy described above to evaluate the posterior can proceed immediately with this alternative expression for π^n.

2.3 Empirical Bayes predictive distribution

Given the empirical Bayes posterior defined above, either for known or unknown σ^2, we can immediately obtain a corresponding predictive distribution. Consider a pair $(\tilde{X}, \tilde{y})$ where $\tilde{X} \in \mathbb{R}^{d\times p}$ is a given matrix of explanatory variable values at which we seek to predict the corresponding response $\tilde{y} \in \mathbb{R}^d$.

If σ^2 were known, or if a plug-in estimator is used, then the conditional posterior predictive distribution of $\tilde{y}$, given S, is familiar, and given by

$$f^n_{\tilde{X}}(\tilde{y} \mid S) = \mathsf{N}_d\big(\tilde{y} \mid \tilde{X}_S\hat{\beta}_S, \sigma^2 I_d + \tfrac{\sigma^2}{\alpha+\gamma}\tilde{X}_S(X_S^\top X_S)^{-1}\tilde{X}_S^\top\big).$$

To obtain the predictive distribution for $\tilde{y}$, we simply need to integrate out S with respect to its posterior, i.e.,

$$f^n_{\tilde{X}}(\tilde{y}) = \sum_S \pi^n(S) f^n_{\tilde{X}}(\tilde{y} \mid S). \tag{2.10}$$

Of course, one cannot evaluate this sum because there are too many terms, but it is possible to use samples of S from the marginal posterior, $\pi^n(\cdot)$, to get a Monte Carlo approximation of the predictive density $f^n_{\tilde{X}}(\tilde{y})$ at some set values of $\tilde{y}$. Alternatively, one can augment the aforementioned MCMC algorithm to sample $\tilde{y}$ from $f^n_{\tilde{X}}(\tilde{y} \mid S)$ along the Markov chain; see below. Having a sample from the predictive distribution is advantageous when it comes to creating posterior credible sets for prediction. For example, in the $d = 1$ case, a 95% posterior prediction interval can be found by computing quantiles of the sample taken from the predictive

distribution.

Very little changes when the inverse gamma prior for σ^2 is adopted. Indeed, the predictive density $f^n_{\tilde{X}}(\tilde{y} \mid S)$ is just the density for a d-variate Student-t distribution, with $2a_0 + \alpha n$ degrees of freedom, location $\tilde{X}_S \hat{\beta}_S$, and scale matrix

$$\frac{b_0 + (\alpha/2)\|y - \hat{y}_S\|^2}{a_0 + \alpha n/2} \Big(I_d + \frac{1}{\alpha + \gamma} \tilde{X}_S (X_S^\top X_S)^{-1} \tilde{X}_S^\top \Big). \tag{2.11}$$

From here, sampling from the predictive (2.10) can proceed exactly like before, with straightforward modifications to accommodate the Student-t instead of normal shape.

For computation, Martin et al. (2017) recommend a simple Metropolis–Hastings scheme based on the marginal posterior distribution for S. If the focus was strictly on variable selection tasks, so that only the posterior distribution of S were relevant, then a shotgun stochastic search strategy could also be taken, as in Liu et al. (2021), which avoids sampling on the complex S-space. But here our focus is on prediction, so we want the samples of S so that we can readily sample from the conditional predictive distribution of $\tilde{y}$, given S, along the way. Specifically, if $q(S' \mid S)$ is a proposal function, then a single iteration of our Metropolis–Hastings sampler goes as follows:

1. Given a current state S', sample $S_{\text{tmp}} \sim q(\cdot \mid S')$.
2. Set $S = S_{\text{tmp}}$ with probability

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

   where π^n is as in (2.9); otherwise, set $S = S'$.
3. Sample $\tilde{y}$ from a d-variate Student-t distribution with $2a_0 + \alpha n$ degrees of freedom, location $\tilde{X}_S \hat{\beta}_S$, and scale matrix (2.11) depending on the given S.

Repeating this process T times, we obtain an approximate sample $\{\tilde{y}^{(t)} : t = 1, \ldots, T\}$ from the predictive distribution (2.10) corresponding to a given $d \times p$ matrix $\tilde{X}$ of covariates at which prediction is desired. In our implementation, we use a symmetric proposal distribution $q(S \mid S')$, i.e., one that samples S uniformly from those models that differ from S' in exactly one position, which simplifies the acceptance probability above since the q-ratio is identically 1.

Of course, the quality of samples from the predictive distribution depends on that of the samples from the marginal posterior distribution for S. It helps that there is a closed-form expression for $\pi^n(\cdot)$, but the configuration space is still very large and complicated, making it virtually impossible for a Markov chain to do a complete exploration in any reasonable amount of time. Fortunately, a complete exploration of the space is not necessary. The theory presented

in Martin et al. (2017)—see, also, Appendix 2.7.1—says that $\pi^n(S)$ will tend to be largest at/near the true configuration. So with a warm start, based on, say, the lasso configuration, the proposed MCMC algorithm quickly explores the subspace of plausible configurations and, in turn, leads to high-quality predictions. Justification for this claim is based on the strong empirical performance in Section 2.5 below, in Martin et al. (2017), and in other settings, e.g., Lee et al. (2019), Liu and Martin (2019), Martin and Ning (2020), and Liu et al. (2023).

2.4 Asymptotic properties

2.4.1 Setup

The goal here is to explore the asymptotic properties of the empirical Bayes predictive distribution defined above. In particular, in Section 2.4.2 we bound the rate at which the posterior distribution Π^n in (2.6) concentrates on β vectors that lead to accurate prediction of a new observation which, in certain cases, despite the high dimensionality, is close to the parametric root-n rate. Also, in Section 2.4.3, we investigate distributional approximations of the predictive distribution and corresponding uncertainty quantification properties. Some of our results that follow rely on details presented in Martin et al. (2017) so, for the reader's convenience, we summarize the relevant points in Appendix 2.7.1.

To fix notation, etc., let $\beta^\star$ denote the true p-vector of regression coefficients with configuration $S_{\beta^\star} = \{j : \beta_j^\star \neq 0\}$. Since $p \gg n$, quality estimation is hopeless without some underlying low-dimensional structure, and here the relevant low-dimensional structure is *sparsity*, i.e., the size $|S_{\beta^\star}|$ of the true configuration is small relative to n or, more generally, to the rank R. Rates in such problems are determined by the triple $(n, p, s^\star)$, where $s^\star$ is a sequence controlling the sparsity, increasing slowly with n. Verzelen (2012) splits the entire class of problems into two cases, namely, *high-* and *ultra-high-dimensional*, based on whether $s^\star \log(p/s^\star)$ is small or large relative to n. He shows that, for prediction-related tasks, for true $\beta^\star$'s that are $s^\star$-sparse, the optimal rate ε_n satisfies

$$\varepsilon_n^2 = \min\{n^{-1} s^\star \log(p/s^\star), 1\}.$$

Here the phase transition between high- and ultra-high-dimensional problems is clear, in particular, that there is a limit to how accurate predictions can be in the latter case. Since we aim for statements like "the Hellinger distance between the true density and the predictive density in (2.10) is $\lesssim \varepsilon_n$" (e.g., Corollary 1), and such conclusions are not meaningful if $\varepsilon_n \not\to 0$, we will focus exclusively on Verzelen's ordinary high-dimensional case. That is, we make the

following *standing assumption:*

$$(n, p, s^\star) \text{ satisfies } s^\star \log(p/s^\star) = o(n) \text{ as } n \to \infty. \tag{2.12}$$

In other settings, like in Theorem 2 and its corollary below, where it is necessary to separate $\beta^\star$ from $X\beta^\star$, additional assumptions about X and $S^\star$ are required. Fortunately, as we discuss in more detail below, it is known that such assumptions are not unreasonable in settings that satisfy (2.12).

As is typically done in theoretical analyses of the $p \gg n$ problem (e.g., Castillo et al. 2015), we work in the case of known error variance. Moreover, for simplicity and consistency with the existing literature, we assume that the rank R of X equals n. Finally, in what follows, we will write "$\mathsf{E}_{\beta^\star}$" to denote expectation with respect to the n-vector y defined in (2.1) with true regression coefficient $\beta^\star$.

2.4.2 Concentration rates

For ease of presentation, in Theorem 1 and its corollary, we focus on $d = 1$, so the $\tilde{X}$ matrix can be replaced by a (column) p-vector $\tilde{x}$. The results hold more generally, but they are rather cumbersome to present, so we defer those details to Appendix 2.7.2.

For a given $\tilde{x} \in \mathbb{R}^p$ and particular values β and $\beta^\star$, let $h_{\tilde{x}}(\beta^\star, \beta)$ denote the Hellinger distance between $\mathsf{N}(\tilde{x}^\top\beta, \sigma^2)$ and $\mathsf{N}(\tilde{x}^\top\beta^\star, \sigma^2)$. Following Guhaniyogi and Dunson (2015), define an unconditional Hellinger distance

$$h(\beta^\star, \beta) = \Big\{ \int h_{\tilde{x}}^2(\beta^\star, \beta)\, Q_n(d\,\tilde{x}) \Big\}^{1/2},$$

where Q_n is the empirical distribution of those p-vectors that fill the rows of X. Then the following theorem establishes the asymptotic concentration rate of the proposed empirical Bayes posterior relative to the prediction-focused metric $h(\beta^\star, \beta)$.

Theorem 1. *Let Π^n be the empirical Bayes posterior defined in* (2.6)*, and let $s^\star$ be a sequence such that* (2.12) *holds. Then there exists constants $G, M > 0$ such that*

$$\sup_{\beta^\star : |S_{\beta^\star}| = s^\star} \mathsf{E}_{\beta^\star} \Pi^n(\{\beta \in \mathbb{R}^p : h(\beta^\star, \beta) > M\varepsilon_n\}) \lesssim e^{-Gn\varepsilon_n^2} \to 0,$$

where $\varepsilon_n^2 = n^{-1} s^\star \log(p/s^\star) \to 0$.

Proof. Without loss of generality, assume $\sigma^2 = 1$. Let $k_{\tilde{x}}(\beta^\star, \beta)$ denote the Kullback–Leibler

divergence of $\mathsf{N}(\tilde{x}^\top\beta, 1)$ from $\mathsf{N}(\tilde{x}^\top\beta^\star, 1)$. Then we have

$$h^2_{\tilde{x}}(\beta^\star, \beta) \leq 2\, k_{\tilde{x}}(\beta^\star, \beta) = |\tilde{x}^\top(\beta - \beta^\star)|^2.$$

From the above inequality, upon taking expectation over $\tilde{x} \sim Q_n$, we get

$$h^2(\beta^\star, \beta) \leq n^{-1}\|X(\beta - \beta^\star)\|^2.$$

Therefore,

$$\{\beta : h^2(\beta^\star, \beta) > M^2\varepsilon_n^2\} \subseteq \{\beta : \|X(\beta - \beta^\star)\|^2 > M^2 n\varepsilon_n^2\},$$

and it follows from Theorem 1 in Martin et al. (2017)—see Appendix 2.7.1—that, for suitable M, the expected Π^n-probability of the right-most event above is exponentially small, uniformly in $s^\star$-sparse $\beta^\star$. □

The conditions here are different from those in Jiang (2007) and Guhaniyogi and Dunson (2015), but it may help to compare the rates obtained. Note that, beyond sparsity, no assumptions are made in Theorem 1 above on the magnitude of $\beta^\star$, whereas the latter two papers assume $\|\beta^\star\|_1 = O(1)$ which requires either (a) $s^\star$ grows slowly and non-zero signals vanish slowly, or (b) $s^\star$ grows not-so-slowly and the non-zero signals vanish rapidly. The more realistic case is (a), so suppose $s^\star \asymp (\log n)^k$ for some $k > 0$. If p is polynomial in n, i.e., $p \asymp n^K$ for any $K > 0$, then we have

$$\varepsilon_n \asymp n^{-1/2}(\log n)^{(k+1)/2},$$

which is nearly the parametric root-n rate. And if p is sub-exponential in n, i.e., if $\log p \asymp n^r$ for $r \in (0,1)$, then ε_n is $n^{-(1-r)/2}$ modulo logarithmic terms which, again, is close to the parametric rate when r is small. In any case, if the analogy between the sparse normal means problem and the regression problem considered here holds up in the context of prediction, the minimax rate results in Mukherjee and Johnstone (2015) suggest that the rate ε_n in Theorem 1 cannot be significantly improved. Of course, posterior concentration rates in terms of $\|X(\beta - \beta^\star)\|$ have been established for other models, such as horseshoe (e.g., Ghosh and Chakrabarti 2015; van der Pas et al. 2017a, 2014), so results like that in Theorem 1 would apply for those methods as well.

The next result connects the rate in Theorem 1 to the posterior predictive density $f_x^n(\tilde{y})$ defined in (2.10). Following the discussion above, this predictive density convergence rate is very fast, close to the root-n rate in some cases.

Corollary 1. *Let $f^\star_{\tilde{x}}(\tilde{y}) = \mathsf{N}(\tilde{y} \mid \tilde{x}^\top\beta^\star, \sigma^2)$ denote the true distribution of the new observation, for a given x, and let $H(f^\star_{\tilde{x}}, f^n_{\tilde{x}})$ denote the Hellinger distance between this and the predictive density*

$f^n_{\tilde{x}}$ *in* (2.10)*. Under the conditions of Theorem 1,*

$$\sup_{\beta^\star:|S_{\beta^\star}|=s^\star} \mathsf{E}_{\beta^\star} \int H^2(f^\star_{\tilde{x}}, f^n_{\tilde{x}}) Q_n(d\tilde{x}) \lesssim \varepsilon_n^2.$$

Proof. Since $f^n_{\tilde{x}}(\tilde{y}) = \int \mathsf{N}(\tilde{y} \mid \tilde{x}^\top\beta, \sigma^2)\Pi^n(d\beta)$, using convexity of H^2, Jensen's inequality, and Fubini's theorem we get

$$\begin{aligned}\int H^2(f^\star_{\tilde{x}}, f^n_{\tilde{x}}) Q_n(d\tilde{x}) &\leq \int\int h^2_{\tilde{x}}(\beta^\star,\beta)\Pi^n(d\beta)Q_n(d\tilde{x}) \\ &= \int h^2(\beta^\star,\beta)\Pi^n(d\beta).\end{aligned}$$

For M and ε_n as in Theorem 1, if we set $A_n = \{\beta : h(\beta^\star,\beta) \leq M\varepsilon_n\}$, then the right-hand side above equals

$$\int_{A_n} h^2(\beta^\star,\beta)\Pi^n(d\beta) + \int_{A_n^c} h^2(\beta^\star,\beta)\Pi^n(d\beta). \tag{2.13}$$

The first term is bounded by a constant times ε_n^2 by definition of A_n. And since Hellinger distance is no more than 2, the second term is bounded by a constant times $\Pi^n(A_n^c)$. Theorem 1 above shows that $\mathsf{E}_{\beta^\star}\Pi^n(A_n^c)$ is exponentially small, definitely smaller than ε_n^2. The claim follows since both terms in (2.13) are of order ε_n^2 or smaller. □

The particular metric in Theorem 1 measures the prediction quality for new $\tilde{x}$'s which are already in the current sample, that is, averaging over $\tilde{x} \sim Q_n$. In other words, Theorem 1 considers a sort of in-sample prediction quality. Intuitively, however, we expect that similar conclusions could be made for out-of-sample predictions, provided that the new x value does not differ too much from the rows in the observed X. Indeed, we are considering large n so, if the rows of X are sampled from a distribution $Q^\star$, then we can expect Q_n to be a decent approximation of $Q^\star$, so results that involve averaging over Q_n cannot be drastically different from the corresponding results with averaging over $Q^\star$. Our simulation experiments investigate exactly this situation—new $\tilde{x}$ is an independent sample from the distribution that generated the original X—and the mean square prediction error results confirm the above intuition.

Formal out-of-sample prediction results are possible, but require additional assumptions about the design matrix. Here we revert back to the general case of d-dimensional prediction, with $d \geq 1$, characterized by a $d \times p$ matrix $\tilde{X}$ at which prediction is desired. Intuitively, if $\tilde{X}$ is genuinely new, then we have no direct measurements of $\tilde{X}\beta^\star$ as we would for in-sample prediction, so we cannot hope for quality prediction without accurate estimation of $\beta^\star$. Since the response only provides direct information about the mean $X\beta^\star$, estimation of $\beta^\star$ requires

disentangling $\beta^\star$ from $X\beta^\star$. Towards this, following Castillo et al. (2015, Def. 2.3) and Arias-Castro and Lounici (2014, Eq. 11), define the "smallest scaled sparse singular value of X of size s" as

$$\underline{\kappa}(s;X) = n^{-1/2} \inf_{\beta\in\mathbb{R}^p : 0<|S_\beta|\leq s} \frac{\|X\beta\|}{\|\beta\|}, \quad s=0,1,\ldots,p. \tag{2.14}$$

Arias-Castro and Lounici (2014) showed that a sparse $\beta^\star$, with $|S_{\beta^\star}| = s^\star$, is identifiable from a model with design matrix X if and only if $\underline{\kappa}(2s^\star;X) > 0$. So one must assume at least that in order to estimate $\beta^\star$ and to accurately predict at an arbitrary $\tilde{X}$.

In Theorem 2 below, we will need a complementary "largest sparse singular value" for the $d \times p$ matrix at which prediction is desired. In particular, set

$$\overline{\kappa}(s;\tilde{X}) = d^{-1/2} \sup_{\beta\in\mathbb{R}^p : 0<|S_\beta|\leq s} \frac{\|\tilde{X}\beta\|}{\|\beta\|}, \quad s=0,1,\ldots,p.$$

The theorem and its corollary characterize the posterior and predictive distribution concentration rates, respectively, in terms of the $\underline{\kappa}$ and $\overline{\kappa}$ quantities. Following their statements, we describe what these rates look like in some relevant cases.

Theorem 2. *Let Π^n be the empirical Bayes posterior defined in* (2.6) *and $s^\star$ a sequence such that* (2.12) *holds and $\underline{\kappa}(Cs^\star;X) > 0$ for $C > 2$. For fixed $d \geq 1$, let $\tilde{X}$ be a $d \times p$ matrix at which prediction is desired. Then there exists positive constants (C', G, M), with $C' > 2$, such that*

$$\sup_{\beta^\star : |S_{\beta^\star}| = s^\star} \mathsf{E}_{\beta^\star} \Pi^n\Big(\Big\{\beta\in\mathbb{R}^p : h_x(\beta^\star,\beta) > \frac{M\overline{\kappa}(C's^\star;\tilde{X})\varepsilon_n}{\underline{\kappa}(Cs^\star;X)}\Big\}\Big) \lesssim e^{-Gn\varepsilon_n^2} \to 0,$$

where $\varepsilon_n^2 = n^{-1}s^\star\log(p/s^\star) \to 0$.

Proof. Again, without loss of generality, we assume $\sigma^2 = 1$. Set

$$\Delta_n = \frac{M\overline{\kappa}(C's^\star;\tilde{X})\varepsilon_n}{\underline{\kappa}(Cs^\star;X)}.$$

From the total probability formula, the Π^n-probability in question is bounded by

$$\Pi^n(\{\beta\in\mathbb{R}^p : h_x(\beta^\star,\beta) > \Delta_n, |S_\beta| \leq C''s^\star\}) + \Pi^n(\{\beta\in\mathbb{R}^p : |S_\beta| > C''s^\star\}).$$

Theorem 2 in Martin et al. (2017) shows that, for some $C'' > 1$, the expected value of the second term in the above display is exponentially small; see Appendix 2.7.1 for a precise statement. So it suffices to focus on the first term. If β is such that $|S_\beta| \leq C''s^\star$, then $|S_{\beta-\beta^\star}| \leq C's^\star$, where

$C' = C'' + 1 > 2$. So by definition of $\overline{\kappa}$,

$$h^2_{\tilde{X}}(\beta^\star, \beta) \le \|\tilde{X}(\beta - \beta^\star)\|^2 \le \overline{\kappa}^2(C' s^\star; \tilde{X}) \|\beta - \beta^\star\|^2.$$

Therefore,

$$\{h_{\tilde{X}}(\beta^\star, \beta) > \Delta_n\} \cap \{|S_\beta| \le C'' s^\star\} \implies \|\beta - \beta^\star\| > \frac{M \varepsilon_n}{\underline{\kappa}(C s^\star; X)},$$

and Theorem 3 in Martin et al. (2017)—see Appendix 2.7.1—establishes that the expected Π^n-probability of the latter event is exponentially small. □

Like in the in-sample prediction setting, this posterior concentration rate result in Theorem 2 for out-of-sample prediction can be immediately converted into a rate for the predictive density $f^n_x(\tilde{y})$ defined in (2.10) at the given x.

Corollary 2. *For a fixed $d \times p$ matrix $\tilde{X}$ at which prediction is desired, let $f^\star_{\tilde{X}}(\tilde{y}) = \mathsf{N}_d(\tilde{y} \mid \tilde{X}^\top \beta^\star, \sigma^2 I_d)$ denote the true distribution of the new observation and $H(f^\star_{\tilde{X}}, f^n_{\tilde{X}})$ the Hellinger distance between this and the predictive density $f^n_{\tilde{X}}$ in* (2.10)*. Under the assumptions of Theorem 2,*

$$\sup_{\beta^\star : |S_{\beta^\star}| = s^\star} \mathsf{E}_{\beta^\star} H^2(f^\star_{\tilde{X}}, f^n_{\tilde{X}}) \lesssim \frac{\overline{\kappa}^2(C' s^\star; \tilde{X}) \varepsilon_n^2}{\underline{\kappa}^2(C s^\star; X)}.$$

Proof. The proof proceeds exactly like that of Corollary 1 but for not needing to integrate over $\tilde{X}$ and applying Theorem 2 instead of Theorem 1. □

Of course, the ε_n in the above two results is the same as in Theorem 1 and its corollary, so can still be close to the parametric root-n rate. However, the rate in these latter results depends in a non-trivial way on the design matrix X and the new $\tilde{X}$ at which prediction is desired, through the $\underline{\kappa}$ and $\overline{\kappa}$ quantities. A relevant case to consider is that where the rows of X and of $\tilde{X}$ are iid Gaussian p-vectors with mean 0, variance 1, and correlation matrix Σ. For a number of relevant correlation structures in Σ, e.g., independence, first-order autoregressive, block diagonal, etc., it is known that $\underline{\kappa}(s; X)$ is bounded away from 0, with high probability (as a function of the random X), for all $s \lesssim (n/\log p)^{1/2}$; see Castillo et al. (2015, Examples 7–8) and Arias-Castro and Lounici (2014, Sec. 2.6). So, at least for these cases, the $\underline{\kappa}$ term in the denominator is not expected to affect the rate in Theorem 2 or its corollary. Next, for the numerator, we can bound $\overline{\kappa}(s; \tilde{X}) \le d^{-1/2} \|\tilde{X}_S\|_F$, where S is a subset of size $|S| \le s$ and $\|\cdot\|_F$ denotes the Frobenius norm. And since the entries of $\tilde{X}$ are $O_p(1)$, we can conclude that $\overline{\kappa}(s; \tilde{X}) = O_p(s^{1/2})$. Therefore, putting all this together, the upper bound obtained in Corollary 2 is $\lesssim s^\star \varepsilon_n^2$. And for small $s^\star$, e.g., a power of $\log n$ like discussed above, the overall rate is still roughly root-n. To put this "$s^\star \varepsilon_n^2$" quantity in perspective, consider an oracle case where the configuration $S^\star$, of size $s^\star$, is known.

Then the mean square error for the oracle least squares estimator of the mean response, $\tilde{X}_{S^\star}\hat{\beta}_{S^\star}$, given X and $\tilde{X}$, is

$$\mathsf{E}_{\beta^\star}\|\tilde{X}_{S^\star}\hat{\beta}_{S^\star} - \tilde{X}_{S^\star}\beta^\star_{S^\star}\|^2 = \mathrm{tr}\{\tilde{X}_{S^\star}(X_{S^\star}^\top X_{S^\star})^{-1}\tilde{X}_{S^\star}^\top\}.$$

Plugging in the approximation, $n^{-1}X_{S^\star}^\top X_{S^\star} \approx \Sigma_{S^\star}$, where the latter is the submatrix of Σ corresponding to the variables identified in $S^\star$, it is easy to check that the trace is n^{-1} times a chi-square random variable with $d s^\star$ degrees of freedom, i.e., is $O_p(s^\star n^{-1})$, as a function of $(X, \tilde{X})$. Therefore, since the $s^\star\varepsilon_n^2$ upper bound in Corollary 2 is of the same order as the oracle mean square error for estimating the mean response at $\tilde{X}$, our rates cannot be significantly improved.

2.4.3 Uncertainty quantification

Beyond prediction accuracy, one would also want the posterior predictive distribution to be calibrated in the sense that a $100(1-\zeta)\%$ prediction interval, for $\zeta \in (0, \frac{1}{2})$, has coverage probability $1-\zeta$, at least approximately. That is, one may ask if the predictive distribution above provides valid uncertainty quantification. To first build some intuition, recall the predictive density in (2.10):

$$f_{\tilde{X}}^n(\tilde{y}) = \sum_S \pi^n(S)\,\mathsf{N}_d\big(\tilde{y} \mid \tilde{X}_S\hat{\beta}_S, \sigma^2 I_d + \tfrac{\sigma^2}{\alpha+\gamma}\tilde{X}_S(X_S^\top X_S)^{-1}\tilde{X}_S^\top\big).$$

If we happen to have $\mathsf{E}_{\beta^\star}\pi^n(S^\star) \to 1$, then

$$f_{\tilde{X}}^n(\tilde{y}) \approx \mathsf{N}_d\big(\tilde{y} \mid \tilde{X}_{S^\star}\hat{\beta}_{S^\star}, \sigma^2 I_d + \tfrac{\sigma^2}{\alpha+\gamma}\tilde{X}_{S^\star}(X_{S^\star}^\top X_{S^\star})^{-1}\tilde{X}_{S^\star}^\top\big), \tag{2.15}$$

and one will recognize the right-hand side as roughly the *oracle* predictive distribution, the one based on knowledge of the correct configuration $S^\star$. The only difference between this predictive distribution and the standard fixed-model version found in textbooks is the factor $(\alpha+\gamma)^{-1}$. We prefer our predictive density to be at least as wide as the oracle, which suggests choosing (α, γ) such that $\alpha+\gamma \leq 1$, maybe strictly less than 1. With this choice, we expect the posterior prediction intervals to approximately achieve the nominal frequentist coverage probability.

To make the above heuristics precise, write the posterior distribution for β as

$$\Pi^n(B) = \sum_S \pi^n(S)\big\{\mathsf{N}_{|S|}\big(\hat{\beta}_S, \tfrac{\sigma^2}{\alpha+\gamma}(X_S^\top X_S)^{-1}\big) \otimes \delta_{0_{S^c}}\big\}(B), \quad B \subseteq \mathbb{R}^p.$$

For the given $d \times p$ matrix $\tilde{X}$ at which prediction is desired, set $\psi = \tilde{X}\beta$. Then the derived

posterior distribution for ψ is

$$\Pi_\psi^n(A) := \sum_S \pi^n(S)\, \mathsf{N}_d\big(A \mid \hat{\psi}_S, \sigma^2 V_S\big), \quad A \subseteq \mathbb{R}^d, \tag{2.16}$$

where

$$\hat{\psi}_S = \tilde{X}_S \hat{\beta}_S \quad \text{and} \quad V_S = (\alpha+\gamma)^{-1} \tilde{X}_S (X_S^\top X_S)^{-1} \tilde{X}_S^\top.$$

Then we have the following Bernstein–von Mises theorem, similar to that in Martin and Ning (2020), which formally establishes a Gaussian approximation to the marginal posterior distribution of ψ, which will almost immediately justify (2.15).

Theorem 3. *Write $d_{\text{TV}}(P,Q) = \sup_A |P(A) - Q(A)|$ for the total variation distance between probability measures P and Q. Let $\mathsf{N}_d(\hat{\psi}_{S^\star}, \sigma^2 V_{S^\star})$ denote the oracle posterior for ψ based on knowledge of the true configuration $S^\star$. If $\mathsf{E}_{\beta^\star} \pi^n(S^\star) \to 1$, then*

$$\mathsf{E}_{\beta^\star} d_{\text{TV}}\big(\Pi_\psi^n, \mathsf{N}_d(\hat{\psi}_{S^\star}, \sigma^2 V_{S^\star})\big) \to 0.$$

Proof. Define $D_n(A) = |\Pi_\psi^n(A) - \mathsf{N}_d(A \mid \hat{\psi}_{S^\star}, \sigma^2 V_{S^\star})|$ for Borel sets $A \subseteq \mathbb{R}^d$. Since $|\sum_i a_i| \le \sum_i |a_i|$, we immediately get the following upper bound:

$$D_n(A) \le \sum_S \pi^n(S) \big| \mathsf{N}_d(A \mid \hat{\psi}_S, \sigma^2 V_S) - \mathsf{N}_d(A \mid \hat{\psi}_{S^\star}, \sigma^2 V_{S^\star}) \big|.$$

The absolute difference above is 0 when $S = S^\star$ and bounded by 2 otherwise, so

$$d_{\text{TV}}\big(\Pi_\psi^n, \mathsf{N}_d(\hat{\psi}_{S^\star}, \sigma^2 V_{S^\star})\big) \le 2 \sum_{S \neq S^\star} \pi^n(S) = 2\{1 - \pi^n(S^\star)\}.$$

After taking expectation of both sides, the upper bound vanishes by assumption. □

Our Bernstein–von Mises theorem enjoys a relatively simple proof thanks to two things: the conjugate prior and the assumption that $\mathsf{E}_{\beta^\star} \pi^n(S^\star) \to 1$. First, as a result of using a conjugate (empirical) prior for β_S, given S, the marginal posterior distribution for ψ is exactly a mixture of Gaussians as in (2.16). Therefore, if the posterior for S concentrates on $S^\star$, then the result follows immediately. If not for this conjugacy, e.g., if one used the Laplace-type priors from Castillo et al. (2015), then a proof of result like that in Theorem 3 would be much more involved, even with the posterior for S concentrating on $S^\star$. Second, the "selection consistency" property, namely, $\mathsf{E}_{\beta^\star} \pi^n(S^\star) \to 1$, is one that has been extensively studied in the Bayesian literature on high-dimensional regression. Beyond the conditions needed for the rate results in Section 2.4.2, to be able to detect the correct set of variables, the non-zero coefficients need to be sufficiently

large. More formally, these considerations lead to a so-called "beta-min condition" (e.g., Arias-Castro and Lounici 2014; Bühlmann and van de Geer 2011; Castillo et al. 2015), namely,

$$\min_{j \in S_{\beta^\star}} |\beta_j^\star| \gtrsim \Big\{ \frac{\sigma^2 \log p}{n \underline{\kappa}^2 (C|S_{\beta^\star}|; X)} \Big\}^{1/2}. \tag{2.17}$$

Note that the definition of "sufficiently large" here depends on $\underline{\kappa}$ and, therefore, implicitly assumes that this quantity is positive. Under the beta-min condition, Martin et al. (2017) establish selection consistency for the empirical Bayes posterior Π^n under investigation here; see Appendix 2.7.1 for a precise statement. It is intuitively clear that a condition like (2.17) is needed in order to identify the correct $S^\star$, and it is apparently not too strong of an assumption given that existing methods are able to recover $S^\star$ empirically for a wide range of examples. We will have more to say about the beta-min condition in the context of our main result of this section, Corollary 4 below, on the coverage probability of credible intervals derived from the posterior predictive distribution.

The Gaussian approximation to the marginal posterior Π_ψ^n should not be a big surprise, given its relatively simple form in (2.16) and the strong selection consistency property discussed above. Of critical importance for us here is that Theorem 3 implies a corresponding Bernstein–von Mises theorem for the predictive density.

Corollary 3. *Under the conditions of Theorem 3,*

$$\mathsf{E}_{\beta^\star} d_{\text{TV}}\big(f_{\tilde{X}}^n, \mathsf{N}_d(\hat{\psi}_{S^\star}, \sigma^2(I_d + V_{S^\star})\big) \to 0.$$

Proof. The posterior and oracle predictive distributions are convolutions of $\mathsf{N}_d(0, \sigma^2 I_d)$ with Π_ψ^n and $\mathsf{N}_d(\hat{\psi}_{S^\star}, \sigma^2 V_{S^\star})$, respectively. So the claim follows from Theorem 3 and general results on information loss, namely, Lemma B.11 and Equation (B.14) in Ghosal and van der Vaart (2017). □

The conclusion of Corollary 3 is that the predictive distribution will closely resemble the oracle predictive distribution based on knowledge of $S^\star$. To visualize this, we carry out a small simulation study where data y is generated from model (2.1), with $\sigma = 1$, under the "very-sparse" settings described in Section 2.5.2 below. In particular, we have $n = 100$ and $p = 250$, but there are only $|S^\star| = 5$ non-zero coefficients, each of magnitude $A = 4$. The X matrix has rows that are iid p-variate normal with mean 0, variance 1, and a first-order autoregressive correlation structure with parameter $r = 0.5$. The goal is to predict a new $\tilde{y}$ of dimension $d = 1$, where the corresponding $\tilde{x}$ vector is an independent draw from the same p-variate normal distribution that generated X. Figure 2.1 shows a histogram of a sample drawn from our posterior predictive distribution with the corresponding oracle predictive density function

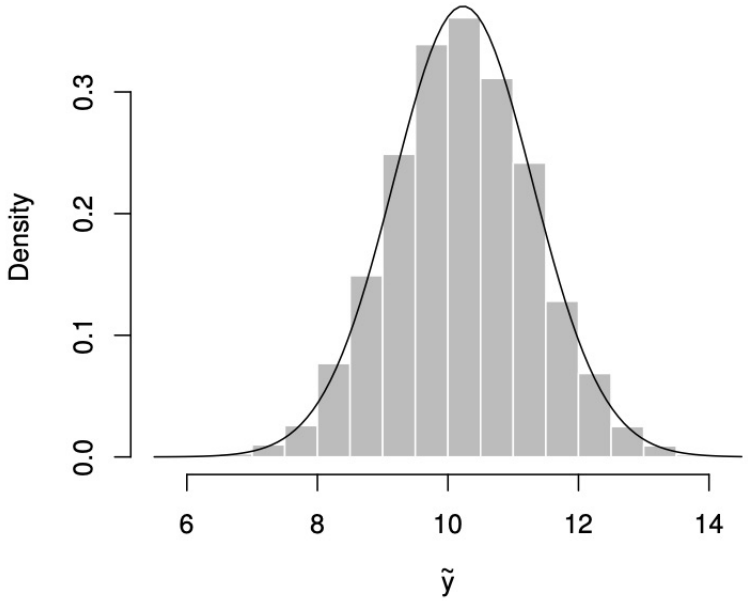

Figure 2.1: Histogram of a Monte Carlo sample drawn from the posterior predictive distribution, f_x^n in (2.10), and the corresponding oracle predictive density function overlaid.

overlaid. These computations were actually done in an unknown-σ^2 scenario, so that the oracle is a shifted and scaled Student-t density. That the two distributions match very closely confirms Corollary 3, but it is perhaps surprising that the accuracy of the normal approximation kicks in with only $n = 100$, even for a relatively high-dimensional setting.

The Bernstein–von Mises result in Corollary 3 is in terms of a strong, total variation distance, which implies that central $1-\zeta$ regions derived from the posterior predictive distribution are approximately

$$\big\{\tilde{y} \in \mathbb{R}^d : (\tilde{y} - \hat{\psi}_{S^\star})^\top (I_d + V_{S^\star})^{-1} (\tilde{y} - \hat{\psi}_{S^\star}) \leq \sigma^2 \chi^2_{\zeta/2}(d)\big\},$$

where $\chi^2_{\zeta/2}(d)$ is the upper-$\zeta/2$ quantile of $\mathsf{ChiSq}(d)$. For the special case $d = 1$, and in the original notation, the posterior prediction interval for $\tilde{y}$, at a covariate (column) vector $\tilde{x}$, is approximately

$$x^\top \hat{\beta}_{S^\star} \pm z_{\zeta/2}\, \sigma \big\{1 + (\alpha+\gamma)^{-1} \tilde{x}_{S^\star}^\top (X_{S^\star}^\top X_{S^\star})^{-1} \tilde{x}_{S^\star}\big\}^{1/2},$$

where $z_{\zeta/2}$ is the upper-$\zeta/2$ quantile of the standard normal distribution. Clearly, this is a $100(1-\zeta)\%$ frequentist prediction interval for a new $\tilde{y}$ drawn from model (2.1), with the new covariate vector $\tilde{x}$, provided that $\alpha + \gamma \leq 1$. Moreover, as this limiting credible interval matches that of an oracle who knows $S^\star$, the size cannot be improved.

Corollary 4. *Under the conditions of Theorem 3, if $\alpha + \gamma \leq 1$, then prediction intervals for a new*

$\tilde{y}$ drawn from model (2.1)*, with the new covariate vector $\tilde{x} \in \mathbb{R}^p$, derived from the empirical Bayes posterior predictive distribution* (2.10) *attain the nominal frequentist coverage probability and are of optimal size asymptotically.*

To conclude this section, we give a few remarks to put the result in Corollary 4 in perspective. First, to our knowledge, there are no uncertainty quantification results for the Bayes solution in the high-dimensional regression setting based on the horseshoe prior, although the developments in van der Pas et al. (2017b) would suggest that results along this line are possible under appropriate conditions. More recently, Belitser and Ghosal (2020) investigated the uncertainty quantification properties of an empirical Bayes solution closely related to that in Martin et al. (2017) adopted here, but not in the context of prediction. Roughly, they show that suitably constructed empirical Bayes credible balls for the full β vector approximately achieve the nominal frequentist coverage probability. For example, in their Corollary 4, for any credibility level ζ, they identify an ℓ_2-norm ball

$$B_n(\zeta) = \{\beta \in \mathbb{R}^p : \|\beta - \hat{\beta}\| \leq M_\zeta \, \hat{\rho}_n^\star\},$$

where $\hat{\beta}$ is the empirical Bayes posterior mean, $\hat{\rho}_n^\star$ is a function depending on data, on features of $\beta^\star$, and on other quantities, and M_ζ is a constant, such that

$$\sup_{\beta^\star} \mathsf{P}_{\beta^\star}\{B_n(\zeta) \not\ni \beta^\star\} \leq \zeta,$$

with the supremum taken over a class of vectors $\beta^\star$ that satisfy an *excessive bias restriction*. The specific details are not relevant for the present discussion, so suffice it to say that the excessive bias restriction rules about $\beta^\star$ are difficult to detect, such as sparse vectors whose non-zero entries are relatively small. Therefore, sparsity plus the beta-min condition (2.17) implies the excessive bias restriction. Consequently, if $\tilde{x}$ is a new p-vector at which prediction is required, one can naively convert the above credible ball into a credible interval for the mean response $\mu = \mu(\tilde{x})$ at $\tilde{x}$, i.e.,

$$\{\mu \in \mathbb{R} : |\mu - \tilde{x}^\top \hat{\beta}| \leq M_\zeta \hat{\rho}_n^\star / \|\tilde{x}\|\},$$

and this, in turn, can be suitably enlarged for predicting an independent response at $\tilde{x}$. Although this prediction interval would asymptotically achieve the target coverage probability under the weaker excessive bias restriction, it still has some theoretical and practical disadvantages. First, a number of the ingredients in the credible ball $B_n(\zeta)$ and in the corresponding prediction interval are either unspecified (e.g., depend on "sufficiently large" constants) or depend on features of the unknown $\beta^\star$, so putting the theory into practice is not straightforward. Second, computation of this alternative prediction interval is more expensive than our

proposed approach based on (2.10) because the high-dimensional β cannot be integrated out. Third, because of their indirect/naive construction, it is likely that the prediction intervals described above are less efficient than those in Corollary 4, which are optimal. So even though our uncertainty quantification results here make stronger assumptions than the weakest of those appearing in other contexts in the literature, our conclusions are stronger and our implementation is easier. Plus, our simulation experiments show strong empirical performance across various settings.

2.5 Numerical results

2.5.1 Methods

Here we investigate the performance of our empirical Bayes prediction method compared to existing methods. Our method, which we denote by *EB*, is as described in Section 2.3, with the following hyperparameter settings:

- the complexity prior q_n for the size $|S|$ in (2.3) uses $a = 0.05$ and $c = 1$;
- the posterior construction in Section 2.2 uses $\gamma = 0.005$ in (2.4) and $\alpha = 0.99$ in (2.6);
- and the inverse gamma prior for σ^2 described in Section 2.2.2 has shape and scale parameters $a_0 = 0.01$ and $b_0 = 4$, respectively.

The latter two are the default settings in the R package `ebreg`. This is compared to predictions based on the horseshoe, denoted by *HS*, using the default Jeffreys prior for σ^2 implemented in the `horseshoe` package (van der Pas et al. 2016). That package does not return samples from the predictive distribution, but these are easy to obtain from the (β, σ^2) samples. In our simulations, both the EB and HS methods return 5000 posterior predictive samples after a burn-in of 1000. The aforementioned methods give full predictive distributions, which yield both point predictions and prediction intervals. We also compare with point predictions obtained from lasso and adaptive lasso using the R packages `lars` (Hastie and Efron 2013) and `parcor` (Kraemer and Schaefer 2014), respectively.

2.5.2 Simulated data experiments

The theory presented in Section 2.4 focuses on the high- but not ultra-high-dimensional setting described in Verzelen (2012), i.e., where $(n, p, s^\star)$ are such that $s^\star \log(p/s^\star)$ is small compared to n. So in our examples presented below, we consider settings, common in medical and social science applications, with p ranging from just slightly larger than n to several times larger than

n. To keep the focus on the effect of the ambient and effective dimensions, i.e., p and $s^\star$, on prediction performance, throughout these experiments we take $n = 100$ fixed, and vary $(p, s^\star)$ accordingly. In limited runs with larger sample sizes, i.e., $n \in \{200, 300, 400\}$, we found that the relative comparisons between methods was comparable, so these results are not presented here. We investigate the prediction error, the coverage rates, and length of prediction intervals produced by the proposed method compared to the above competitors. In particular, we consider univariate predictions ($d = 1$) in an "out-of-sample" context where the new $\tilde{x}$ vector is independently sampled from that same p-variate normal distribution described above, and the goal is to predict a new response $\tilde{y} = \tilde{x}^\top \beta^\star + \sigma z$ from the model (2.1).

The take-away message here is that EB is generally better than HS across a range of settings, in terms of both prediction accuracy and uncertainty quantification. In particular, first, in the out-of-sample mean square prediction error comparisons, there are cases where the two methods have comparable performance and a number of cases where EB is significantly better, but no cases where HS is significantly better. Second, the prediction intervals for the two methods both generally are within an acceptable range of the target 95% prediction coverage probability but, with the exception of a few cases at $p = 500$ in the very-sparse setting (Table 2.3), the EB intervals tend to be shorter than the HS intervals. On top of its strong statistical performance, our EB method is more efficient computationally, as it generally finishes 5 times faster than the implementation of HS in the corresponding R package.

Very-sparse settings

We select $s^\star = 5$ specific β_j's to be non-zero, with the rest being zero. In particular, we place non-zero values at positions 3, 4, 15, 22, and 25 in the p-vector β. This configuration captures a number of different features: 3 and 4 are adjacent, 4 and 15 have a large gap between them, and 22 and 25 are close neighbors but not adjacent. All of the non-zero β_j's take value A, where $A \in \{2, 4, 8\}$. The rows of the design matrix, X, are p-variate normal with zero mean, unit variance, and first-order autoregressive dependence structure, with pairwise correlation equals $r^{|i-j|}$ and $r \in \{0.2, 0.5, 0.8\}$. For each (A, r) pair, we consider $p \in \{125, 250, 500\}$, which yields a total of 27 different settings. MSPEs are shown in Table 2.1 and the prediction interval coverage probabilities and mean lengths are presented in Tables 2.2 and 2.3, respectively. Summaries of the standard errors are provided in the respective table captions, and based on 250 runs.

EB performs very well across all the settings in terms of MSPE, with HS performing similarly, and both generally beating lasso and adaptive lasso. The standard errors generally range between 0.08 and 0.98, with an average of 0.18. Based on these standard errors, there are no significant differences in prediction performance between EB and HS, at least not in terms of MSPE. However, there is a substantial difference in terms of computation time. EB is based

Table 2.1: Comparison of mean square prediction error (MSPE) for the four different methods across various very sparse settings—of dimension $p \in \{125, 250, 500\}$, signal size $A \in \{2, 4, 8\}$, and correlation $r \in \{0.2, 0.5, 0.8\}$—as described in the text. The standard errors are generally between 0.08 and 0.98.

p		$A=2$			$A=4$			$A=8$		
		$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8
125	EB	0.86	1.13	0.99	1.15	0.91	0.94	0.98	0.99	1.22
	HS	0.89	1.21	1.07	1.20	1.01	1.05	1.03	1.07	1.31
	Lasso	1.18	1.51	1.38	1.57	1.48	1.48	1.66	1.51	1.90
	Alasso	0.93	1.15	1.05	1.70	1.45	1.76	5.20	3.53	8.99
250	EB	1.05	1.10	1.12	1.07	1.24	0.87	1.15	1.07	0.96
	HS	1.12	1.19	1.17	1.12	1.34	0.98	1.20	1.12	1.00
	Lasso	1.31	1.50	1.52	1.38	1.69	1.35	1.31	1.35	1.33
	Alasso	1.07	1.14	1.23	1.50	1.92	2.07	3.47	4.14	5.60
500	EB	0.97	0.93	0.93	0.93	1.16	1.08	1.01	1.02	1.17
	HS	1.00	1.08	0.99	1.09	1.15	1.05	1.12	1.04	1.19
	Lasso	1.28	1.37	1.44	1.49	1.35	1.25	1.43	1.55	1.44
	Alasso	0.98	1.00	1.12	1.35	1.44	2.14	4.54	3.83	6.52

on a two-groups or spike-and-slab model formulation, generally believed to be too expensive to compute. But contrary to this popular belief, EB's run-time is about 20% of that for HS, consistent with the claims made by Martin and Ning (2020) in a different context.

For the 95% prediction interval comparisons, we compare EB and HS to that of an *oracle* who knows the true configuration $S^\star$. Of course, the coverage probabilities for the oracle prediction interval are exactly 0.95, but, according to Table 2.2, both EB and HS have prediction coverage probability within an acceptable range of the target 95% level. And in terms of interval lengths, Table 2.3 reveals that both EB and HS are comparable in efficiency to the oracle prediction intervals. This confirms the claims made about EB prediction intervals in Section 2.4. There are, however, a few cases in the $p = 500$ row of Table 2.3 where HS gives significantly shorter interval lengths. Obviously, the $p = 500$ case is generally harder, but it is not clear if this performance difference is just a coincidence or if there is something special about these cases. A theoretical investigation into the uncertainty quantification properties of the horseshoe prior clearly is needed.

Table 2.2: Comparison of coverage probability for the two different 95% prediction intervals across various very sparse settings—of dimension $p \in \{125, 250, 500\}$, signal size $A \in \{2, 4, 8\}$, and correlation $r \in \{0.2, 0.5, 0.8\}$—as described in the text. Standard errors are between 0.01 and 0.02.

p		$A=2$			$A=4$			$A=8$		
		$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8
125	EB	0.96	0.94	0.96	0.95	0.96	0.97	0.97	0.96	0.94
	HS	0.96	0.92	0.96	0.93	0.96	0.94	0.96	0.94	0.92
250	EB	0.95	0.96	0.95	0.96	0.94	0.96	0.94	0.96	0.95
	HS	0.95	0.96	0.94	0.96	0.92	0.96	0.94	0.94	0.95
500	EB	0.95	0.95	0.96	0.96	0.94	0.96	0.96	0.93	0.94
	HS	0.95	0.94	0.95	0.93	0.94	0.96	0.95	0.92	0.93

Less-sparse settings

Let $\Delta \in (0,1)$ control the density of the signals, so that $s^\star = \Delta p$. Obviously, if the density is too large, then the problem will eventually turn ultra-high-dimensional, so we focus on a limited range of small Δ values such that $\Delta p \log(\Delta^{-1}) \leq 0.8n$, where the 0.8 factor keeps the problem at a practically reasonable level of difficulty. (We tried other more extreme cases, e.g., $p = 500$ and $\Delta = 0.1$, and found that all the methods performed poorly in terms of prediction, so apparently that setting is just too difficult.) Here we repeat the same comparisons as in the previous section, adding comparisons across different values of Δ; of course, the range of Δ for which the problem is not ultra-high-dimensional depends on p. Here we found that that adaptive lasso predictions were, across the board, not competitive, so we removed it from the comparisons.

From Table 2.4, EB is outperforming the other methods in almost all of these settings in terms of MSPE. Our method consistently provides low MSPE values, typically between 1 and 1.5, with the exception of a few settings when $p = 500$. Excluding the $p = 500$ settings, the standard errors lie between 0.08 and 0.35, with most below or around 0.10; the mean is 0.15, and the median is 0.13. Taking these standard errors into consideration, EB and HS perform similarly when the number of signals is relatively low, in particular, at (p, Δ) pairs $(125, 0.05)$, $(125, 0.1)$, and $(250, 0.05)$. But in the moderate density cases, namely, (p, Δ) equal $(125, 0.2)$ and $(250, 0.1)$, the EB predictions are significantly better. The $p = 500$ case is more challenging and we find that results for all three methods are rather unstable. Plots of the raw prediction error magnitudes (not shown) are long-tailed, especially HS and lasso, which explains the larger MSPEs and corresponding standard errors. While there is too much instability in the $p = 500$

Table 2.3: Comparison of mean length for the three different 95% prediction intervals across various very sparse settings—of dimension $p \in \{125, 250, 500\}$, signal size $A \in \{2, 4, 8\}$, and correlation $r \in \{0.2, 0.5, 0.8\}$—as described in the text. Standard errors are between 0.01 and 0.03.

p		$A=2$			$A=4$			$A=8$		
		$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8
125	EB	4.12	4.13	4.15	4.11	4.14	4.16	4.13	4.10	4.17
	HS	4.15	4.16	4.17	4.15	4.18	4.16	4.15	4.13	4.16
	Oracle	4.06	4.07	4.07	4.05	4.09	4.08	4.07	4.04	4.07
250	EB	4.13	4.15	4.18	4.12	4.14	4.13	4.15	4.12	4.14
	HS	4.14	4.17	4.15	4.13	4.14	4.09	4.16	4.12	4.12
	Oracle	4.07	4.09	4.09	4.06	4.08	4.05	4.09	4.07	4.05
500	EB	4.12	4.14	4.17	4.08	4.15	4.15	4.10	4.12	4.18
	HS	4.10	4.08	4.11	4.05	4.09	4.07	4.05	4.09	4.10
	Oracle	4.07	4.08	4.08	4.04	4.08	4.05	4.04	4.06	4.09

case to make any conclusive claims of significant differences, the EB method appears to be "less unstable" compared to the others, which is a positive sign. And just like in the very-sparse settings discussed above, the run-time for EB is a small fraction of that for HS.

Tables 2.5 and 2.6 report the prediction coverage probabilities and mean lengths, respectively, for the 95% prediction intervals based on EB and HS. Most of the coverage probabilities are close to the target 95%, however, we do see that as p and/or Δ increase, a few of the coverage probabilities moved slightly further away from 95%. For example, when $p = 500$, $r = 0.5$, and $A = 8$, EB has its lowest probability of 88%. A few more of EB's coverage probabilities are lower than HS's, although there are also a few instances where EB's coverage is better than HS's. Most of these differences are not statistically significant, however. The lengths of the EB prediction intervals are consistently shorter than those of HS, and in all settings where $p \geq 250$, the difference between lengths of EB and HS prediction intervals are significant.

2.5.3 Real data application

Following the example used in Bhadra et al. (2019a), we use the same real-world data set to examine how our method performs. This pharmacogenomics data set is publicly available in the NCI-60 database, and can be accessed via the R package `mixOmics` (Le Cao et al. 2016), dataset `multidrug`. The expression level of 12 different human ABC transporter genes are predicted using compound concentration levels. To keep our analysis on par with that in

Table 2.4: Comparison of MSPEs for the three different methods across various less-sparse settings—of dimension $p \in \{125, 250, 500\}$, density Δ, signal size $A \in \{2, 4, 8\}$, and correlation $r \in \{0.2, 0.5, 0.8\}$—as described in the text. The standard errors are between 0.08 and 0.35, with the exception of some in the $p = 500$ case.

p	Δ		$A=2$			$A=4$			$A=8$		
			$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8
125	0.05	EB	1.05	1.16	1.07	1.00	1.02	0.90	1.14	0.94	1.09
		HS	1.11	1.25	1.16	1.09	1.11	0.99	1.24	1.02	1.14
		Lasso	1.41	1.62	1.46	1.44	1.59	1.55	1.77	1.71	1.75
125	0.1	EB	1.06	0.86	1.13	1.03	1.27	1.16	1.16	1.16	1.23
		HS	1.17	1.03	1.32	1.26	1.56	1.37	1.35	1.36	1.44
		Lasso	1.69	1.54	1.72	2.00	1.97	1.98	2.04	2.16	2.38
125	0.2	EB	1.43	1.20	1.27	1.37	1.47	1.39	1.58	1.54	1.33
		HS	2.53	2.04	2.04	2.38	2.08	1.98	2.28	2.15	1.96
		Lasso	3.34	2.36	2.23	3.05	2.59	2.31	3.20	2.89	2.52
250	0.05	EB	0.99	1.01	1.37	1.22	1.14	1.13	1.25	1.03	1.02
		HS	1.11	1.29	1.60	1.54	1.38	1.27	1.56	1.19	1.29
		Lasso	1.34	1.51	1.63	1.91	1.39	1.51	1.84	1.52	1.46
250	0.1	EB	1.28	1.16	1.11	1.32	1.47	1.50	1.33	1.48	1.27
		HS	1.90	1.86	2.24	2.17	2.37	2.36	2.04	2.50	2.11
		Lasso	3.69	1.94	1.42	3.57	2.57	2.07	3.34	2.57	1.80
500	0.05	EB	1.52	1.18	1.53	1.46	1.28	1.40	1.57	1.70	1.35
		HS	2.97	1.69	3.02	2.38	1.61	2.20	6.44	2.12	2.29
		Lasso	9.71	2.65	1.70	20.69	2.61	1.75	41.23	2.92	1.89

Bhadra et al. (2019a), we only predict with the 853 compounds that have no missing values. The data set includes 60 samples, which we randomly split into a training and testing set of 75% and 25%, respectively. Thus, in this regression scenario, $n = 45$ and $p = 853$. Each random training and testing split is performed 20 times, and we calculate the average out-of-sample MSPE for these 20 trials, shown in Table 2.7.

For the 12 different transporter genes, our empirical Bayes method obtained marginally better out-of-sample MSPE in three of the genes (A2, A8, and A12) than those from the other methods implemented, while being very comparable with the other methods on the other 9 genes as response variables. For gene B1, all four methods have significantly smaller MSPE values than for the other 11 genes, with EB having the largest. A closer look at the B1 case reveals that the estimated regression coefficients based on, e.g., lasso, are all very small, even the non-zero values. So small, in fact, that EB tends to select no variables; consequently, its predictions on the testing set are based simply on the mean response from the training set,

Table 2.5: Comparison of coverage probability for the two different 95% prediction intervals across various less-sparse settings—of dimension $p \in \{125, 250, 500\}$, density Δ, signal size $A \in \{2, 4, 8\}$, and correlation $r \in \{0.2, 0.5, 0.8\}$—as described in the text. All standard errors are between the values of 0.008 and 0.021, with an average of 0.015.

p	Δ		$A=2$			$A=4$			$A=8$		
			$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8
125	0.05	EB	0.95	0.94	0.94	0.96	0.94	0.97	0.94	0.96	0.96
		HS	0.96	0.94	0.94	0.95	0.94	0.97	0.94	0.96	0.94
125	0.1	EB	0.94	0.97	0.95	0.96	0.93	0.95	0.94	0.95	0.94
		HS	0.95	0.97	0.94	0.95	0.93	0.95	0.94	0.95	0.94
125	0.2	EB	0.92	0.92	0.93	0.94	0.91	0.94	0.92	0.90	0.94
		HS	0.91	0.96	0.94	0.92	0.94	0.97	0.92	0.93	0.95
250	0.05	EB	0.95	0.96	0.92	0.95	0.94	0.95	0.91	0.96	0.94
		HS	0.96	0.94	0.93	0.92	0.95	0.94	0.91	0.95	0.94
250	0.1	EB	0.93	0.96	0.94	0.92	0.91	0.91	0.92	0.93	0.95
		HS	0.95	0.96	0.93	0.94	0.94	0.91	0.94	0.91	0.94
500	0.05	EB	0.92	0.94	0.95	0.95	0.94	0.93	0.91	0.88	0.95
		HS	0.96	0.96	0.94	0.98	0.96	0.94	0.94	0.93	0.90

which is apparently not so effective. Therefore, in cases where the signals are quite small, there is perhaps some advantage to using a continuous shrinkage-style prior like the horseshoe compared to a discrete selection-style prior like that proposed here.

2.6 Discussion

In this chapter, we apply a recently proposed empirical Bayes method to the context of prediction in sparse high-dimensional linear regression settings. The key idea is to let the data inform the prior center so that the tail of the prior distribution have little influence on the posterior concentration properties. This allows for faster computation—since conjugate Gaussian priors can be used, leading to closed-form marginalization and dimension reduction—without sacrificing on posterior concentration rates. In the context of prediction, being able to formulate the Bayesian model using conjugate priors means that the predictive distribution can be written (almost) in closed-form; it allows for some analytical integration, yielding a relatively easy to compute posterior predictive distribution for the purpose of constructing prediction intervals, etc. We also extended the theoretical results presented in Martin et al. (2017) to obtain posterior concentration rates relevant to the prediction problem, for both in- and out-of-sample cases,

Table 2.6: Comparison of mean length for the three different 95% prediction intervals across various less-sparse settings—of dimension $p \in \{125, 250, 500\}$, density Δ, signal size $A \in \{2, 4, 8\}$, and correlation $r \in \{0.2, 0.5, 0.8\}$—as described in the text. All standard errors are between the values of 0.01 and 0.63, with an average of 0.04.

p	Δ		$A=2$			$A=4$			$A=8$		
			$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8	$r=0.2$	0.5	0.8
125	0.05	EB	4.15	4.12	4.17	4.12	4.09	4.16	4.13	4.16	4.17
		HS	4.23	4.17	4.22	4.19	4.15	4.21	4.19	4.22	4.20
		Oracle	4.11	4.07	4.10	4.08	4.05	4.09	4.09	4.12	4.10
125	0.1	EB	4.14	4.14	4.20	4.13	4.15	4.20	4.15	4.16	4.17
		HS	4.50	4.53	4.52	4.50	4.51	4.49	4.53	4.50	4.49
		Oracle	4.22	4.23	4.21	4.22	4.23	4.20	4.24	4.24	4.19
125	0.2	EB	4.20	4.23	4.43	4.17	4.24	4.50	4.19	4.19	4.38
		HS	5.59	5.59	5.67	5.59	5.63	5.73	5.63	5.53	5.54
		Oracle	4.59	4.58	4.58	4.56	4.60	4.65	4.58	4.54	4.56
250	0.05	EB	4.11	4.16	4.25	4.12	4.11	4.22	4.15	4.14	4.21
		HS	4.42	4.56	4.55	4.49	4.47	4.49	4.50	4.48	4.49
		Oracle	4.19	4.24	4.25	4.21	4.18	4.22	4.24	4.22	4.21
250	0.1	EB	4.20	4.23	4.41	4.20	4.22	4.42	4.17	4.23	4.40
		HS	5.65	5.63	6.04	5.58	5.59	5.55	5.55	5.48	5.61
		Oracle	4.59	4.58	4.56	4.60	4.56	4.57	4.57	4.58	4.58
500	0.05	EB	4.26	4.20	4.54	4.51	4.20	4.39	4.31	4.19	4.44
		HS	6.53	5.53	6.68	5.76	5.47	5.54	7.09	5.44	5.50
		Oracle	4.63	4.56	4.58	4.62	4.56	4.56	4.56	4.55	4.61

and established a Bernstein–von Mises theorem the sheds light on the empirical Bayes posterior's potential for valid uncertainty quantification in prediction. All these desirable features are confirmed by the results in real- and simulated-data examples.

While the computations here using the basic MCMC on the S-space are relatively fast, it is worth asking if speed can be further improved if some margin of approximation is allowed. One option is to use a *variational approximation* (e.g., Blei et al. 2017) to Π^n. Such an approach, using a point mass–Gaussian mixture mean-field approximation, was presented in Ray et al. (2020) but for the case where the posterior being approximated is based on independent (and data-free) Laplace priors for β_S, given S, as described in Castillo et al. (2015). Given that the empirical prior formulation considered here is itself based on point mass–Gaussian type mixtures, the corresponding posterior is in some sense "closer" to this mean-field approximation. So, in addition to the substantial decrease in computation time, some improved performance is to be expected. Work on this has been done (Yang and Martin 2020) and the theoretical and empirical

Table 2.7: Mean square prediction error for the four methods averaged over 20 random training/testing splits of the data as described in Section 2.5.3. The rows correspond to different response variables being predicted. The standard errors range from 0.00 to 0.12, with an average of 0.03.

Response	EB	HS	Lasso	Alasso
A1	0.93	0.93	0.97	1.00
A2	0.93	0.94	1.09	0.99
A3	0.93	0.93	0.97	1.07
A4	0.93	0.93	0.88	1.00
A5	0.93	0.96	0.92	0.94
A6	0.93	0.93	0.98	0.93
A7	0.93	0.93	1.07	0.92
A8	0.93	0.94	1.06	1.01
A9	0.93	0.92	0.82	0.92
A10	0.93	0.93	0.99	0.93
A12	0.93	0.95	1.04	1.04
B1	0.73	0.54	0.61	0.42

results obtained confirm these expectations, and our work in Chapter 4 is along those lines, with a slightly different formulation on the variational family considered.

An interesting question is if this empirical Bayes methodology can be extended to handle sparse, high-dimensional generalized linear models, such as logistic regression. In the Gaussian setting considered here, the notion of prior centering is quite natural and relatively simple to arrange, but the idea itself is not specific to Gaussian models. Chapter 3 carries out this non-trivial extension to GLM, where we obtain good theoretical and numerical results on variable selection and estimation. Results like those obtained here for prediction, can likely be established in that more general context too, and has yet to be explored.

2.7 Technical details

2.7.1 Summary of results from Martin et al. (2017)

For the reader's convenience, here we summarize four results from Martin et al. (2017), in the context of estimation and variable selection, relevant to our present investigation about prediction. These are not simply restatements of the results in that paper, however, here we have refined the conditions and also strengthened some of the conclusions.

The first result concerns the concentration rate properties of the posterior distribution, Π^n,

based on the empirical prior described in Section 2.2, in the known-σ case, with respect to a metric that focuses on estimation of the mean response.

Proposition 1. *Let Π^n be the empirical Bayes posterior defined in* (2.6)*, and let $s^\star$ be a sequence such that* (2.12) *holds. Then there exists positive constants G and M such that*

$$\sup_{\beta^\star : |S_{\beta^\star}| = s^\star} \mathsf{E}_{\beta^\star} \Pi^n(\{\beta \in \mathbb{R}^p : \|X(\beta - \beta^\star)\| > M\varepsilon_n\}) \lesssim e^{-Gn\varepsilon_n^2} \to 0,$$

where $\varepsilon_n^2 = n^{-1} s^\star \log(p/s^\star) \to 0$.

A small difference between Proposition 1 and the statement of Theorem 1 in Martin et al. (2017) is the exponential upper bound. Their proof actually establishes the exponential bound, but they did not include this detail in their statement. The exponential bound adds value, however, in applications like Corollary 1 in Section 2.4.2 above.

The above result focuses on a metric that is relevant to estimating the mean response which, of course, is relevant for our prediction context here. But the metric's dependence on the design matrix X makes "in-sample" prediction most natural. For prediction at generic predictor variable settings, rates can be obtained if we focus on a different metric, namely, $\|\beta - \beta^\star\|$, which is directly related to estimation of the regression coefficients.

Concentration rate results in this stronger metric require additional assumptions about X. In particular, the sparse singular value defined in Equation (2.14) above needs to be positive for some configurations a small factor larger than the true $S^\star$. As this is only slightly stronger than what is required for a sparse $\beta^\star$ to be identifiable (Arias-Castro and Lounici 2014), it is not too much to ask in order to achieve accurate estimation. The following summarizes Theorem 3 in Martin et al. (2017). Just like with Proposition 1, here we present the result with an exponential upper bound, which is part of the proof of their original theorem, but not its statement.

Proposition 2. *Let Π^n be the empirical Bayes posterior defined in* (2.6) *and $s^\star$ a sequence such that* (2.12) *holds and $\underline{\kappa}(C s^\star; X) > 0$ for $C > 2$. Then there exists positive constants G and M such that*

$$\sup_{\beta^\star : |S_{\beta^\star}| = s^\star} \mathsf{E}_{\beta^\star} \Pi^n\Big(\Big\{\beta \in \mathbb{R}^p : \|\beta - \beta^\star\| > \frac{M\varepsilon_n}{\underline{\kappa}(C s^\star; X)}\Big\}\Big) \lesssim e^{-Gn\varepsilon_n^2} \to 0,$$

where $\varepsilon_n^2 = n^{-1} s^\star \log(p/s^\star) \to 0$.

Next is a basic but essential result pertaining to the posterior's ability to identify the important variables. The following proposition says that the posterior tends not to overfit, i.e., does not include unnecessary variables. In other words, while the ambient dimension of the posterior is very high, its effective dimension is not too much larger than that of an oracle who

knows the correct configuration. One more time, the exponential upper bound here does not appear in the statement of Theorem 2 in Martin et al. (2017), but their proof does establish this.

Proposition 3. *Let Π^n be the empirical Bayes posterior defined in* (2.6)*, and let $s^\star$ be a sequence such that* (2.12) *holds. Then there exists constants $C > 1$ and $G > 0$ such that*

$$\sup_{\beta^\star : |S_{\beta^\star}| = s^\star} \mathsf{E}_{\beta^\star} \Pi^n(\{\beta \in \mathbb{R}^p : |S_\beta| > C s^\star\}) \lesssim e^{-Gn\varepsilon_n^2} \to 0,$$

where $\varepsilon_n^2 = n^{-1} s^\star \log(p/s^\star) \to 0$.

Finally, the strongest of the results relevant to structure learning is the following posterior variable selection consistency theorem. This proposition says that the posterior probability assigned to the true configuration approaches 1, which implies that any variable selection procedure derived from the full posterior, e.g., based on marginal variable inclusion probabilities or the maximum *a posteriori* configuration, will asymptotically identify exactly the correct set of variables.

Proposition 4. *Let Π^n be the empirical Bayes posterior defined in* (2.6) *and $s^\star$ a sequence such that* (2.12) *holds and $\underline{\kappa}(C s^\star; X) > 0$ for $C > 2$. If $\beta^\star$ is such that $|S_{\beta^\star}| = s^\star$ and*

$$\min_{j \in S^\star} |\beta_j^\star| \geq \frac{M\sigma}{\underline{\kappa}(C s^\star; X)} \Big(\frac{2 \log p}{n} \Big)^{1/2},$$

for some constant $M > 1$, then $\mathsf{E}_{\beta^\star}\{\Pi^n(\beta : S_\beta = S_{\beta^\star})\} \to 1$ as $n \to \infty$.

The statement here differs slightly from that in Martin et al. (2017), because there are some minor mistakes in their formulation. A proof of this version of the theorem can be found in the supplementary material at `https://arxiv.org/abs/1406.7718v5`.

2.7.2 Predicting a d-dimensional response, $d > 1$

In Section 2.4 we mentioned that, although we focus there on the case of $d = 1$ for the sake of simplicity, suitable versions of the results hold for any fixed $d \geq 1$. Here we describe how those results can be extended to the more general case.

Let $\tilde{X}$ denote a generic $d \times p$ matrix, with $d > 1$, at which predictions are desired. For coefficient vectors β and $\beta^\star$, write $h_{\tilde{X}}(\beta^\star, \beta)$ for the conditional Hellinger distance between $\mathsf{N}_d(\tilde{X}\beta, \sigma^2 I)$ and $\mathsf{N}_d(\tilde{X}\beta^\star, \sigma^2 I)$, and like in Section 2.4.2, define

$$h(\beta^\star, \beta) = \Big\{ \int h_{\tilde{X}}^2(\beta^\star, \beta) \, Q_n^d(d\tilde{X}) \Big\}^{1/2},$$

where Q_n^d is the distribution of matrices $\tilde{X}$ obtained by randomly sampling (without replacement) d rows from the original X matrix, i.e.,

$$Q_n^d = \binom{n}{d}^{-1} \sum_{T \subset \{1,\ldots,n\}:|T|=d} \delta_{X[T,\cdot]},$$

δ is the Dirac point mass distribution, and $X[T,\cdot]$ denotes the submatrix of X obtained by keeping only the rows corresponding to indices in T. The claim in Theorem 1 pertains to a sort of marginal Hellinger distance, of which the quantity $h(\beta^\star,\beta)$ above is a generalization. The crux of the proof of Theorem 1 is in bounding this Hellinger distance in terms of a corresponding Kullback–Leibler divergence which has a simple form in this Gaussian setting. If $k_{\tilde{X}}$ is the analogous conditional Kullback–Leibler divergence of $\mathsf{N}_d(\tilde{X}\beta,\sigma^2 I)$ from $\mathsf{N}_d(\tilde{X}\beta^\star,\sigma^2 I)$, then we immediately get

$$h_{\tilde{X}}(\beta^\star,\beta) \leq 2k_{\tilde{X}}(\beta^\star,\beta).$$

Since the Kullback–Leibler divergence is additive for independent joint distributions, if $\tilde{X}$ is a realization from Q_n^d, i.e., if $\tilde{X} = X[T,\cdot]$ for a random chosen subset T of indices of size d, then we have

$$k_{\tilde{X}}(\beta^\star,\beta) = \sum_{i\in T} k_{x_i}(\beta^\star,\beta) = \sum_{i\in T} |x_i^\top(\beta-\beta^\star)|^2,$$

where x_i denotes a row of X (treated as a column vector). A key point is that this depends on $\tilde{X} = X[T,\cdot]$ only through T. So since there are $\binom{n-1}{d-1}$ such T of size d that contain each row of X, averaging over T gives

$$\int h_{\tilde{X}}^2(\beta^\star,\beta)\, Q_n^d(d\tilde{X}) \leq 2\binom{n}{d}^{-1} \sum_T \sum_{i\in T} |x_i^\top(\beta-\beta^\star)|^2 = \frac{d}{n}\sum_{i=1}^n |x_i^\top(\beta-\beta^\star)|^2.$$

The right-hand side is $d\,n^{-1}\|X(\beta-\beta^\star)\|^2$, and since d is fixed, the posterior concentration rate in terms of $h(\beta^\star,\beta)$ is no slower than that in terms of $\|X(\beta-\beta^\star)\|$, and the latter we know; see Appendix 2.7.1. So, the conclusion of Theorem 1 holds for any $d \geq 1$.

Corollary 1 also holds for any $d \geq 1$, with modifications like those described above when $d > 1$. Our reason for focusing on the $d = 1$ case in the main text is that the notation for and interpretation of the Q_n^d-type in-sample prediction is cumbersome.

CHAPTER

3

EMPIRICAL PRIORS FOR INFERENCE & VARIABLE SELECTION IN HIGH-DIMENSIONAL GENERALIZED LINEAR MODELS

3.1 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 et al. 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. 2019a, 2017, 2019b; Carvalho et al. 2010; van der Pas et al.

2017b) and those based on spike-and-slab mixture priors (Belitser and Ghosal 2020; Castillo et al. 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 et al. 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 et al. 2019) and the theory focuses mostly on variable selection; the one exception, Jeong and Ghosal (2021) gave very general results on posterior concentration rates in high-dimensional GLMs but did not address the model selection question or efficient implementations of the Bayesian solutions they studied. The goal of the present chapter is to offer a Bayesian (or at least Bayesian-like) solution to the high-dimensional GLM problem, one that has both strong theoretical support—optimal posterior concentration rates and model selection consistency—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 et al. (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 et al. 2017; Martin and Tang 2020), there is strong general theory in Martin and Walker (2019) and promising results in various applications, including Liu and Martin (2019); Liu et al. (2023); Martin (2019).

The goal here is to develop the aforementioned empirical Bayes strategy for the case of high-dimensional GLMs. In Section 3.2, we introduce the set up of the GLM problem and review

the empirical Bayes approach for linear regression. In Section 3.3, 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 3.4, we offer theoretical support for our proposed solution. In particular, we have two 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 et al. (2019), that ensure our marginal posterior for the configuration will concentrate on 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 new proof techniques are needed compared to Martin et al. (2017); Martin and Walker (2019). Section 3.5 investigates the numerical performance of the proposed method in variable selection in logistic and Poisson regression compared to existing methods. Finally, some concluding remarks are given in Section 3.6; proofs are presented in the Appendix.

3.2 Setup and background

3.2.1 Problem setup

Suppose the observables $y_1, \ldots, y_n$ 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 $\beta \in \mathbb{R}^p$ into the model via the relationship

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

where h is a given bijection called the *link function*. The so-called "canonical" link is $h = \dot{b}^{-1}$ and, in this case, the above relationship simplifies to $\eta_i = x_i^\top \beta$. 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 $f_\beta(y_i \mid x_i)$ for the density/mass function determined by the parameter vector β. Then the likelihood and log-likelihood functions are, respectively,

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

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 $\hat{\beta}$ is the solution to the equation

$$\dot{\ell}_n(\hat{\beta}) = 0 \iff \{y - \dot{b}(X\hat{\beta})\}^\top \mathrm{diag}\{\dot{\xi}(X\beta)\} X = 0, \tag{3.1}$$

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

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

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

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

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

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

a key point being that this quantity is bounded as a function of β; similarly, in Poisson regression

with the canonical log link, the W matrix has diagonal entries

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

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 β. To side-step this obstacle, we shall assume, as is common, that the GLM is *sparse* in the sense that most of the β 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 β 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 β as (S, β_S), where S is a set of indices that corresponds to the "active" coefficients in β and β_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 $L_n(S, \beta_S)$ and $\ell_n(S, \beta_S)$, respectively, the configuration-specific MLE is $\hat{\beta}_S$, the observed Fisher information is $J_n(S, \hat{\beta}_S)$, etc. Throughout we will write $|S|$ for the cardinality of a configuration S, $\beta^\star$ 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 $S(\beta) = \{j : \beta_j \neq 0\}$ to denote the configuration corresponding to a given coefficient vector.

All of the results in Narisetty et al. (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 $\hat{\beta}_S$ is *not* estimating $\beta^\star_S$. Instead, $\hat{\beta}_S$ is estimating the minimizer of the Kullback–Leibler divergence which, in the present case, is a solution to the equation

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

Let $\beta^\dagger_S$ denote this solution. Note that, first, this notation is slightly misleading, since there is no "full vector" $\beta^\dagger$ of which $\beta^\dagger_S$ is the S-specific sub-vector; instead, there is a different $|S|$-vector $\beta^\dagger_S$ for each S. Second, if S is a superset of $S^\star$, then $\beta^\dagger_S = \beta^\star_S$; this explains why Narisetty et al. (2019) do not need $\beta^\dagger_S$ in their superset-only analysis. Lemmas 1–2 in Appendix 3.7.1 below generalize Lemmas A1 and A3 in Narisetty et al. (2019) to cover the case where $\hat{\beta}_S$ is compared to $\beta^\dagger_S$ rather than $\beta^\star_S$.

3.2.2 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 3.7.3 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 et al. (2017) and in Martin and Tang (2020).

The normal linear model assumes that $y_i \sim \mathsf{N}(x_i^\top \beta, \sigma^2)$, 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 β as the pair (S, β_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 β_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

$$f_n(s) \propto p^{-as}, \quad s = 1, 2, \ldots, s_{\max},$$

   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 hyperparameter $a > 0$ plays a crucial role in the posterior and we will discuss this further below.

2. The conditional prior for β_S, given S, is where the data enters into the prior. To avoid the sub-optimal behavior of the the thin-tailed Gaussian prior while simultaneously retaining the computational convenience of the conjugate Gaussian form, Martin et al. (2017) suggested

$$(\beta_S \mid S) \sim \mathsf{N}_{|S|}\big(\hat{\beta}_S, \gamma (X_S^\top X_S)^{-1}\big), \quad S \subset \{1, 2, \ldots, p\},$$

   where $\hat{\beta}_S$ is the S-specific least-squares estimator, hence the data-dependence, and $\gamma > 0$ is a tuning parameter to be specified.

This empirical prior gets 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 β. In particular, the corresponding marginal posterior for the configuration $S = S(\beta)$ can be used for model selection purposes. Computation is simple and fast, via efficient using a 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 et al. (2017) demonstrate theoretically that the posterior Π^n achieves the adaptive, minimax optimal asymptotic concentration rate at the true/sparse $\beta^\star$. In other words, there is no other posterior distribution—genuinely Bayesian or otherwise—that can concentrate around the true $\beta^\star$ 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 chapter, but for the case of high-dimensional GLMs, so more details about the construction and the results will be given below.

3.3 Empirical Bayes for high-dimensional GLMs

3.3.1 Prior distribution

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

$$\pi_n(S) \propto \tbinom{p}{|S|}^{-1} p^{-a|S|}, \quad S \subset \{1, 2, \ldots, p\} \text{ such that } |S| \leq s_n, \tag{3.2}$$

where $a > 0$ 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 3.2.2. Second, the data-driven conditional prior for β_S, given S, is

$$(\beta_S \mid S) \sim \Pi_{n,S} := \mathsf{N}_{|S|}\big(\hat{\beta}_S, \gamma J_n(S, \hat{\beta}_S)^{-1}\big), \tag{3.3}$$

where $\gamma > 0$ is a constant to be specified. Above, the prior center is $\hat{\beta}_S$, the data-driven maximum likelihood estimator for the given S and, similarly, $J_n(S, \hat{\beta}_S)$ is the corresponding data-driven observed Fisher information matrix. So this empirical conditional prior has a similar form as that in Section 3.2.2 for the linear model case, the specific pieces that go into it are just different

here in the GLM setting.

As indicated above, the conditional prior for β_S, given S, is data-driven in the sense that both the prior mean vector and covariance matrix depend on data y through $\hat{\beta}_S$. Also, recall that the diagonal entries of the information matrix $J_n(S, \hat{\beta}_S)$ 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 $\beta \sim \Pi_n$ is given by

$$\Pi_n(d\beta) = \sum_S \pi_n(S) \mathsf{N}_{|S|}(d\beta_S \mid \hat{\beta}_S, \gamma J_n(S, \hat{\beta}_S)^{-1}) \otimes \delta_{0_{S^c}}(d\beta_{S^c}), \tag{3.4}$$

where the sum is over all configurations S supported by the prior mass function π_n and $\delta_{0_{S^c}}(d\beta_{S^c})$ denotes a Dirac point mass differential term at the origin for the component β_{S^c}. The empirical prior depends on the hyperparameters (a, 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.

3.3.2 Posterior distribution

If L_n denotes the GLM's likelihood function, and Π_n the empirical prior in (3.4) with the mixture form, then the proposed posterior distribution for β is defined as

$$\Pi^n(d\beta) = \frac{L_n(\beta)^\alpha \, \Pi_n(d\beta)}{\int_{\mathbb{R}^p} L_n(\vartheta)^\alpha \, \Pi_n(d\vartheta)}, \quad \beta \in \mathbb{R}^p, \tag{3.5}$$

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 $\widetilde{\Pi}_n(d\beta) \propto L_n(\beta)^{-(1-\alpha)} \Pi_n(d\beta)$, and get exactly the same posterior. In our opinion, the version in (3.5) 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 et al. (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 2020). 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 van Ommen 2017; Grünwald and Mehta 2020), and/or uncertainty quantification (e.g., Martin and Ning 2020; Martin and Tang 2020).

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

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

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:

$$\pi^n(S) \propto \pi(S)(1 + \alpha\gamma)^{-|S|/2} \, L_n(S, \hat{\beta}_S)^\alpha, \quad \text{all large } n. \tag{3.6}$$

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 et al. 2016; Shun and McCullagh 1995). This class is described in Section 3.4 below, and a precise result on the Laplace approximation's accuracy is given in Lemma 4 in Appendix 3.7.2. Details on posterior computation are given next.

3.3.3 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 (3.6) 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 (3.6).

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 as in Equation (3.6). 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 models $S^{(1)}, \ldots, S^{(M)}$ 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 $\hat{\beta}_S$, the Fisher information $J_n(S, \hat{\beta}_S)$, 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 β_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

$$\pi^n(j) := \pi^n(\{S : S \ni j\}) \approx \frac{1}{M} \sum_{m=1}^{M} 1\{S^{(m)} \ni j\}, \tag{3.7}$$

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; more on this in Section 3.5 below.

If there is also interest in estimation of/inference on the coefficients β, then one can easily augment the above Monte Carlo scheme by inserting a rejection sampling step where, given S, the corresponding coefficient β_S is drawn from the corresponding conditional posterior. Similarly, if 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 from the posited model given $(S, \beta_S, \tilde{X})$ is made.

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). With the SSS approach, 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 most likely due to SSS having to compute the posterior probabilities for all the neighboring models.

3.4 Asymptotic properties

3.4.1 Setup and conditions

Narisetty et al. (2019) and Cao and Lee (2020) investigate certain asymptotic properties of their proposed posterior for β, 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 β concentrates around the true $\beta^\star$ at (nearly) the optimal rate (e.g., Rigollet 2012), adaptive to the unknown sparsity level $|S(\beta^\star)|$.

Below are two conditions crucial to the developments here and in Narisetty et al. (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 $K > 0$, $\lambda > 0$, $w \in [0,1]$, and $w' \in (w,1]$ 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

$$\lambda \le \min_{S:|S|\le|S^\star|+s_n} \lambda_{\min}\{n^{-1}J_n(S,\beta_S^\dagger)\} \le \Lambda_{|S^\star|+s_n} \le K^2(n/\log p)^w, \tag{3.8}$$

where

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

and

$$s_n = O\big((n/\log p)^{(1-w')/2}\big), \quad n \to \infty. \tag{3.9}$$

Condition 3. The power $\alpha > 0$ in (3.5) is strictly less than 1. Also:

(a) the prior hyperparameter $\gamma > 0$ in (3.3) satisfies

$$\gamma = O(\Lambda^2_{2s_n}), \quad n \to \infty, \tag{3.10}$$

where Λ_s and s_n are as defined above, and

(b) the prior hyperparameter $a > 0$ in (3.2) is such that $p^{a-\kappa} > s_n$, where $\kappa > 1$ is the constant specified in Lemma 3 and s_n is as in (3.9).

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 et al. (2019) and Cao and Lee (2020). The upper bound weakens and generalizes the "bounded eigenvalue condition" in, e.g., Bondell and Reich (2012). Since $J_n(S, \beta_S^\dagger) = X_S^\top W(S, \beta_S^\dagger) X_S$, 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 (3.8) holds with $w = 0$. In such cases, this gives the weakest constraint (3.9) on the support of the marginal prior for S, since we can take $w' = 0$ 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 $W(S, \beta_S^\dagger)$. Since the diagonal entries are typically increasing functions of their arguments, the bound is $w(\|X_S \beta_S^\dagger\|_\infty)$. Since the focus is on large configurations, and since $\|X_S \beta_S^\dagger\|_\infty = \|X_S \beta_S^\star\|_\infty = \|X\beta^\star\|_\infty$ for $S \supset S^\star$, the upper bound in (3.8) is only slightly stronger than assuming "$w(\|X\beta^\star\|_\infty) \lesssim (n/\log p)^w$." 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 $\beta^\star$ to satisfy

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

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 w in (3.8) is determined by X, so it is not impossible to determine w and to set γ in (3.10) accordingly. For example, if X is such that $s \mapsto \Lambda_s$ is uniformly bounded, then (3.8) holds with $w = 0$ and then γ can be taken as a constant too. When $w > 0$, the corresponding γ is a diverging sequence in n. Recall that the primary part of the $(\beta_S \mid S)$ covariance matrix is $J_n(S, \hat{\beta}_S)^{-1}$, which itself is $O(n^{-1})$. So, multiplying this by a $\gamma \to \infty$ still allows the prior mass to be concentrating around the data-driven center $\hat{\beta}_S$, just slower. Finally, as argued following the statement of Lemma 3 in Appendix 3.7.1, the constant κ can be chosen

arbitrarily close to 1, so "$a > \kappa$," which is implied by Condition 3(b), is just a little stronger than "$a > 1$." Indeed, since s_n is growing strictly slower than $n^{1/2}$, it suffices to take $a - \kappa$ greater than $\frac{1}{2}(\log n)/(\log p)$, which is never more than $\frac{1}{2}$.

Below we present two different types of results. The first type concerns generally how the posterior distribution Π^n for the coefficient vector β concentrates its mass around the sparse true value $\beta^\star$. The second type concerns how the marginal posterior mass function π^n for S concentrates around its true value $S^\star$, where $S^\star = S(\beta^\star)$ is the configuration corresponding to the true $\beta^\star$. Proofs are in Appendix 3.8.

Although the Gaussian linear model is a GLM, it is possible to derive equivalent results for this model directly and under weaker conditions (Martin et al. 2017). So the machinery presented below is intended only for the genuinely more complex GLM setting, e.g., binary regression with logistic or probit links.

3.4.2 Posterior concentration results

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

$$H_n(\beta^\star, \beta) = \Big[\frac{1}{n} \sum_{i=1}^n \int \{p_\beta(y_i \mid x_i)^{1/2} - p_{\beta^\star}(y_i \mid x_i)^{1/2}\}^2 \, dy_i \Big]^{1/2}. \tag{3.11}$$

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

$$\sup_{\beta^\star : |S(\beta^\star)| \le s_n} \mathsf{E}_{\beta^\star} \Pi^n(\{\beta : H_n(\beta^\star, \beta) > M \varepsilon_n(\beta^\star)\}) \to 0, \quad n \to \infty, \tag{3.12}$$

where $\varepsilon_n^2(\beta^\star) = n^{-1} s^\star \log p$, *with* $s^\star = |S(\beta^\star)|$ *and* $M > 0$ *a sufficiently 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 et al. (2017), and Belitser and Ghosal (2020). The difference between the rate in Theorem 4 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 $|S(\beta^\star)|$. 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 4 above and an effective-dimension bound like in Theorem 5 below together imply a rate in terms of other distances, including the ℓ_2-distance on β. 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 5. *Under Conditions 1–3, the posterior* Π^n *defined above satisfies*

$$\sup_{\beta^\star : |S(\beta^\star)| \leq s_n} \mathsf{E}_{\beta^\star} \Pi^n \Big(\Big\{ \beta : \|\beta - \beta^\star\|_2^2 > \frac{M \varepsilon_n^2(\beta^\star)}{\phi^2(K|S(\beta^\star)|, W(\beta^\star))} \Big\} \Big) \to 0, \quad n \to \infty, \tag{3.13}$$

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

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

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 $X\beta$, not on β 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 β 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 (3.13) 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 β concentrates at the (near) optimal rate for sparse $\beta^\star$ true vector *suggests* that the posterior for S is concentrating on the true $S^\star = S(\beta^\star)$, but this is not a consequence of Theorem 4. The asymptotic behavior of the S-posterior must be investigated directly. The first such result concerns the "effective dimension" of the posterior distribution for β is

not much larger than that of $\beta^\star$. In other words, π^n concentrates on S with $|S|$ that are smaller than a multiple of $|S^\star|$, where $S^\star = S(\beta^\star)$.

Theorem 5. *Under Conditions 1–3, for any* $C > (1-\alpha\kappa/a)^{-1} > 1$,

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

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

That the constant C above is greater than 1 follows from the fact that $a > \kappa$ and $\alpha < 1$. Again, the take-away message here is that the posterior distribution for β is concentrating on a space that is genuinely low-dimensional and, in particular, is of dimension not much greater than $|S^\star|$. Theorem 5 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 6. *Under Conditions 1–3,*

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

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

Virtually the same argument used to prove Theorem 6 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 (3.14).

Theorem 7. *Under Conditions 1–3,* $\pi^n\{S(\beta^\star)\} \to 1$ *with* $\mathsf{P}_{\beta^\star}$*-probability* $\to 1$ *for all* $\beta^\star$ *such that* $c|S(\beta)^\star| \leq s_n$ *and*

$$\min_{j\in S(\beta^\star)} \beta_j^{\star 2} \geq c\, n^{-1}|S(\beta^\star)|\Lambda_{c|S(\beta^\star)|}\log p, \tag{3.14}$$

for some constant $c > 1$.

The condition (3.14) is exactly the same as in Narisetty et al. (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 $w > 0$.

3.5 Numerical results

3.5.1 Methods and metrics

Our focus here is on variable selection performance of our proposed empirical Bayes method compared to existing methods in a couple different GLM contexts. There is a plethora of literature (e.g., Bhadra et al. 2019b; Cao and Lee 2020; Narisetty et al. 2019; Wei and Ghosal 2020) that focuses on high-dimensional logistic regression, or a Bernoulli model with logit link function. In fact, most of the papers that we are familiar with focus exclusively on logistic regression. Methods that can be applied to other GLMs, and the corresponding numerical comparisons, are far less common. For this reason, we decided to demonstrate our method's numerical performance in both *logistic* regression and *Poisson* regression (with log link).

We start with some details about the competitor methods we consider, including lasso, adaptive lasso, SCAD, MCP, and horseshoe. 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 et al. (2017) and Bhadra et al. (2019b). 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 β, and since our interest is in 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. 2023), but these results are not reported here as the horseshoe was computationally restrictive in terms of run-time when running simulations with a large number of replications.

We implement the proposed empirical Bayes solution using the Metropolis–Hastings strategy in Section 3.3.3. 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 (3.7). 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 (3.7) 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.

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. These metrics are defined as

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP}+\text{FN}} \\ \text{TNR} &= \frac{\text{TN}}{\text{TN}+\text{FP}} \\ \text{MCC} &= \frac{\text{TP}\times\text{TN}-\text{FP}\times\text{FN}}{\{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})\}^{1/2}}, \end{aligned}$$

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 most relevant of the metrics. On the other hand, if the two errors cost roughly the same, then MCC would be a more relevant metric.

3.5.2 Simulation studies

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 $\beta^\star$ 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 $\exp(x_i^\top \beta^\star)/\{1+\exp(x_i^\top \beta^\star)\}$, for $i = 1, \ldots, n$. The settings are determined by the triple (p, s, r), so there are altogether eight simulation settings here. There are 100 replications performed at each of the eight settings, and the variable selection results for the

methods and metrics described above are summarized in Table 3.1.

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 four of the eight settings but, as is common, it tends to over-select as evidenced by its low TNR. Horseshoe has the lowest TPR in all 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 the highest MCC in four of the eight settings, while MCP has the highest MCC in the other four. Adaptive lasso has the lowest MCC across all the settings.

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 $\exp(x_i^\top \beta^\star)$. All other settings remain the same as in the logistic regression simulations above, with the same (p, r, s) combinations. The results for these Poisson regression simulations are summarized in Table 3.2.

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 perform better than all other methods across all the settings for all three metrics. Comparing the methods, SCAD and MCP are not as competitive in these Poisson regression settings. We were not able to include results for the horseshoe due to its restrictively long runtime here.

3.6 Conclusion

In this chapter, we extend the empirical or data-driven prior specification strategy first proposed by Martin et al. (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

Table 3.1: Comparison of TPR, TNR, and MCC for the two EB methods with different cutoffs and the five other methods across various settings in logistic regression.

p	$\|S\|$	r	Metric	EB1	EB2	HS	lasso	alasso	SCAD	MCP
			TPR	1.000	0.998	0.833	1.000	0.643	1.000	1.000
200	4	0	TNR	0.973	0.995	1.000	0.953	0.865	0.959	0.986
			MCC	0.669	0.907	0.905	0.645	0.208	0.584	0.787
			TPR	0.995	0.980	0.800	1.000	0.828	1.000	1.000
200	4	0.2	TNR	0.977	0.997	1.000	0.964	0.851	0.955	0.983
			MCC	0.708	0.936	0.886	0.703	0.289	0.560	0.746
			TPR	0.884	0.779	0.283	0.973	0.579	0.979	0.944
200	8	0	TNR	0.972	0.995	1.000	0.901	0.851	0.953	0.985
			MCC	0.723	0.812	0.502	0.548	0.210	0.668	0.821
			TPR	0.824	0.700	0.298	0.949	0.751	0.951	0.893
200	8	0.2	TNR	0.971	0.995	1.000	0.931	0.815	0.954	0.985
			MCC	0.670	0.760	0.526	0.619	0.274	0.656	0.792
			TPR	0.980	0.948	0.583	0.990	0.810	0.998	0.998
400	4	0	TNR	0.990	0.998	1.000	0.972	0.930	0.974	0.992
			MCC	0.750	0.908	0.733	0.605	0.317	0.540	0.750
			TPR	0.950	0.875	0.555	0.995	0.863	0.990	0.985
400	4	0.2	TNR	0.993	0.999	1.000	0.978	0.921	0.969	0.989
			MCC	0.793	0.892	0.732	0.664	0.314	0.499	0.698
			TPR	0.518	0.374	0.050	0.908	0.413	0.924	0.811
400	8	0	TNR	0.977	0.991	1.000	0.940	0.934	0.960	0.987
			MCC	0.448	0.411	0.132	0.489	0.162	0.538	0.670
			TPR	0.543	0.351	0.091	0.916	0.659	0.920	0.829
400	8	0.2	TNR	0.990	0.997	1.000	0.961	0.917	0.966	0.989
			MCC	0.572	0.499	0.221	0.595	0.269	0.563	0.709

Table 3.2: Comparison of TPR, TNR, and MCC for the two EB methods with different cutoffs and the five 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

sense that it balances the strong theoretical performance that is necessary to justify its use in applications with the computational simplicity and efficiency necessary for it 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 theoretical techniques to successfully overcome them here in this chapter. 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 also quite general and can be applied beyond the most common logistic regression setting.

Our numerical investigations in this chapter focused exclusively on variable selection, 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 point estimation of and/or uncertainty quantification about the coefficient vector β. 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 much 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.

3.7 Technical preliminaries

3.7.1 Likelihood-related properties

First, we present here two key results from Narisetty et al. (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. The first result establishes an important continuity property for the observed information matrix, and the second a convergence rate for the MLE uniformly over configurations that are not "too complex." There is nothing particularly special about logistic regression, so the results presented here cover arbitrary GLMs.

Also, for our purposes, it is not enough to consider the true $\beta^\star$ exclusively. So, the results presented below cover the case of general configurations S that are not supersets of $S(\beta^\star)$, i.e., where $\beta_S^\dagger$ is needed in place of $\beta_S^\star$. Fortunately, the arguments they used to prove their Lemmas A1 and A3 go through almost word-for-word when replacing $\beta_S^\star$ with $\beta_S^\dagger$ where appropriate. Finally, none of the considerations that follow make sense if $S = \emptyset$, 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 $S \neq \emptyset$, so $|S| \geq 1$.

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

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

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

Lemma 2. *Under Conditions 1–2,*

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

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

Proof. Just a quick sketch to explain the more general case under consideration here. The same argument in Narisetty et al. (2019) establishes that

$$\mathsf{P}_{\beta^*}\big\{\|\hat{\beta}_S - \beta_S^\dagger\|^2 \gtrsim n^{-1}|S|\Lambda_{|S|}\log p\big\} \leq p^{-2|S|},$$

for all S with $1 \leq |S| \leq s_n$. There are $\binom{p}{s} \leq p^s$ many configurations of size s, so the union bound gives

$$\mathsf{P}_{\beta^*}\big\{\|\hat{\beta}_S - \beta_S^\dagger\|^2 \gtrsim n^{-1}|S|\Lambda_{|S|}\log p \ \text{ for some } S\big\} \lesssim \sum_{s=1}^{s_n} p^s\, p^{-2s} \lesssim p^{-1} = o(1).$$

The upper bound is $o(1)$, which proves the claim. □

Lemma A1–A3 in Narisetty et al. (2019) give an important bound on the log-likelihood difference. This is a version of Lemma 3 in Lee and Cao (2021), which was quoted as below in Equation A7 of Cao and Lee (2020).

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

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

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

The constant κ is important in what follows, so it deserves some explanation. Since the entries of y are exponential family, they are sub-Gaussian. This implies existence of a moment-generating function and, moreover, a Gaussian-like upper bound on that moment-generating function. If $\mu = \mu(\beta^\star)$ and $\Sigma = \Sigma(\beta^\star)$ are the (true) mean vector and (true) covariance matrix of y, respectively, then the standardized random vector $z = \Sigma^{-1/2}(y - \mu)$ is still sub-Gaussian. This implies that there exists a $\delta > 0$ such that $\mathsf{E}_{\beta^\star} \exp(u^\top z) \le \exp\{\frac{1}{2}(1+\delta)\|u\|^2\}$, for any fixed vector $u \in \mathbb{R}^n$ in the column space of $\Sigma^{1/2} X_S$ for any S with $|S| \le |S^\star| + s_n$. So δ depends on the true distribution of y. The central limit theorem implies that, if u is a unit vector, then $u^\top z$ is asymptotically normal, which suggest that, in an asymptotic analysis, δ can be made arbitrarily small. The κ eluded to above is $\kappa = 1 + \delta$, so if we are letting $n \to \infty$, then it suffices to let κ be any number strictly greater than 1.

3.7.2 Marginal likelihood

Here we provide justification for the approximation (3.6) of the marginal posterior π^n at configuration S. Recall that the marginal posterior satisfies

$$\pi^n(S) = \frac{\pi_n(S)\, m_y(S)}{\sum_R \pi_n(R)\, m_y(R)},$$

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

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

So the goal is to lower- and upper-bound the integral $m_y(S)$. 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

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

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

$$1 \le \frac{m_y(S)}{(1+\alpha\gamma)^{-|S|/2}\, L_n(S,\hat\beta_S)^\alpha} \le 1 + e^{-C|S|\Lambda_{|S|}\log p},$$

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

$$\frac{\pi^n(S)}{\pi^n(S^\star)} \leq 2\,\frac{\pi_n(S)}{\pi_n(S^\star)}(1+\alpha\gamma)^{-(|S|-|S^\star|)/2}\,\frac{L_n(S,\hat\beta_S)^\alpha}{L_n(S^\star,\hat\beta_{S^\star})^\alpha}. \tag{3.15}$$

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

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

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

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

which agrees with the familiar Laplace approximation expression used in (3.6). Similarly, for an upper bound on the marginal likelihood, we get

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

where, again, the first term in the above bound agrees with the Laplace approximation expression in (3.6). For $\beta_S \in A_S^c$, we have $\ell_n(S,\beta_S) - \ell_n(S,\hat\beta_S) \leq -C\,n\,r_n^2(S)$, so

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

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

$$\frac{m_y(S)}{m_y(S^\star)} \leq \frac{(1+\alpha\gamma)^{-|S|/2}\{1+e^{-Cnr_n^2(S)}\}}{(1+\alpha\gamma)^{-|S^\star|/2}} \frac{L_n(S,\hat{\beta}_S)^\alpha}{L_n(S^\star,\hat{\beta}_{S^\star})^\alpha}.$$

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

3.7.3 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 $\beta = (S, \beta_S)$, and $\Pi^n = \Pi^{n,\alpha}$ the corresponding posterior that uses power $\alpha \in (0,1)$ on the likelihood function. Also, let $\varepsilon_n = \varepsilon_n(\beta^\star)$ be a generic deterministic sequence that satisfies $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$ and may depend on the true $\beta^\star$.

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

$$\pi_n(S^\star) \gtrsim e^{-Bn\varepsilon_n^2}, \quad \text{for all large } n, \tag{3.16}$$

$$\mathsf{P}_{\beta^\star}\big[\Pi_{n,S^\star}\{\mathscr{L}_n(S^\star)\} > e^{-Dn\varepsilon_n^2}\big] \to 1, \quad n \to \infty, \tag{3.17}$$

where

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

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

$$\sum_S \pi_n(S) \int [\mathsf{E}_{\beta^\star}\{\pi_{n,S}(\beta_S)^m\}]^{1/m} \, d\beta_S \lesssim e^{Gn\varepsilon_n}, \quad n \text{ large.} \tag{3.18}$$

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 (3.11); see Appendix 3.8.1 below for details.

3.8 Proofs

3.8.1 Proof of Theorem 4

The proof proceeds by first checking that the local and global prior conditions, as described above, are met under the conditions stated in Theorem 4 above. Then Theorem 2 of Martin and Walker (2019) implies the Hellinger rate result in (3.12). Since Theorem 5 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* $\varepsilon_n(\beta^\star) = (n^{-1}|S(\beta^\star)|\log p)^{1/2}$.

Proof. Fix $\beta^\star$ and set $s^\star = |S(\beta^\star)|$. The first part (3.16) 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 (3.16) holds with $B = 1 + a > 0$.

Next, for (3.17), the data-dependent neighborhood $\mathscr{L}_n(S^\star)$ is given by

$$\mathscr{L}_n(S^\star) = \{\beta_{S^\star} : \ell_n(S^\star, \beta_{S^\star}) - \ell_n(S^\star, \hat{\beta}_{S^\star}) > -d\, n\varepsilon_n^2\}.$$

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

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

where $\tilde{\beta}_{S^\star} \in \mathscr{L}_n(S^\star)$ satisfies $\|\tilde{\beta}_{S^\star} - \hat{\beta}_{S^\star}\| \leq \|\beta_{S^\star} - \hat{\beta}_{S^\star}\|$. Applying Lemma 1 first with $\tilde{\beta}_{S^\star}$ and then with $\hat{\beta}_{S^\star}$ gives

$$J_n(n, \tilde{\beta}_{S^\star}) \leq c_n\, J_n(S^\star, \hat{\beta}_{S^\star}), \quad \text{with probability} \to 1, \tag{3.19}$$

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

$$\mathscr{L}_n(S^\star) \supset \{\beta_{S^\star} : \gamma^{-1}(\beta_{S^\star} - \hat{\beta}_{S^\star})^\top J_n(S^\star, \hat{\beta}_{S^\star})(\beta_{S^\star} - \hat{\beta}_{S^\star}) < \delta_n\},$$

where $\delta_n := 2d\, n\varepsilon_n^2 / c_n\gamma \gg p^{-1}$. 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,

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

The lower bound on $\Pi_n\{\mathscr{L}_n(S^\star)\}$ is $\geq e^{-Dn\varepsilon_n^2} = e^{-Ds^\star \log p}$ for some $D > 0$, as was to be shown. The prior probability bound holds surely, so (3.17) follows from the "probability $\to 1$" conclusion in (3.19). □

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

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

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

Lemma 2 establishes the MLE bounds

$$\|\hat\beta_S - \beta_S^\dagger\|^2 \lesssim n^{-1}|S|\Lambda_{|S|}\log p, \quad \text{uniformly in } S \text{ with } |S| \lesssim |S^\star|, \tag{3.20}$$

and, therefore, with probability tending to 1,

$$(1-\zeta_n)J_n(S,\beta_S^\dagger) \leq J_n(S,\hat\beta_S) \leq (1+\zeta_n)J_n(S,\beta_S^\dagger).$$

Let $\mathscr{E}_n$ denote the event in (3.20); since $\mathsf{P}_{\beta^\star}(\mathscr{E}_n) \to 1$, we will restrict attention to cases where $\mathscr{E}_n$ holds in what follows. Then

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

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

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

Write $\beta_S - \hat{\beta}_S = (\beta_S - \beta_S^\dagger) + (\beta_S^\dagger - \hat{\beta}_S)$ and then expand the quadratic form in the above display to get

$$(\beta_S - \hat{\beta}_S)^\top \Sigma_S^{-1} (\beta_S - \hat{\beta}_S) \geq (\beta_S - \beta_S^\dagger)^\top \Sigma_S^{-1} (\beta_S - \beta_S^\dagger) + 2(\beta_S - \beta_S^\dagger)^\top \Sigma_S^{-1} (\hat{\beta}_S - \beta_S^\dagger).$$

Then it is easy to check that

$$\pi_{n,S}(\beta_S) \leq \left(\tfrac{1+\zeta_n}{1-\zeta_n}\right)^{|S|/2} e^{|(\beta_S - \beta_S^\dagger)^\top \Sigma_S^{-1} (\hat{\beta}_S - \beta_S^\dagger)|} \, \mathsf{N}_{|S|}(\beta_S \mid \beta_S^\dagger, \Sigma_S).$$

Apply the Cauchy–Schwarz inequality to the quadratic form in the exponent above:

$$|(\beta_S - \beta_S^\dagger)^\top \Sigma_S^{-1} (\hat{\beta}_S - \beta_S^\dagger)| \leq \|\Sigma_S^{-1/2} (\beta_S - \beta_S^\dagger)\| \, \|\Sigma_S^{-1/2} (\hat{\beta}_S - \beta_S^\dagger)\|.$$

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

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

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

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

This is constant in data y, so the expectation in (3.18) can be ignored—and m can be arbitrarily close to 1. Moreover, the integral in (3.18) over β_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

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

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 those low-complexity S's, with $|S| \leq C s^\star$, is upper bounded by $e^{G s^\star \log p} = e^{G n \epsilon_n^2}$, which completes verification of (3.18). □

3.8.2 Proof of Theorem 6

By (3.15), the marginal posterior mass function π^n satisfies

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

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

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

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

3.8.3 Proof of Theorem 5

Those S with $|S| > C|S^\star|$ that are proper supersets of $S^\star$ have already been covered in the proof of Theorem 6 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.,

$$L_n(S,\hat{\beta}_S) \le L_n(S^+,\hat{\beta}_{S^+}).$$

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

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

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

3.8.4 Proof of Theorem 7

Take any fixed $\beta^\star$ that meets the stated conditions, and set $S^\star = S(\beta^\star)$ and $s^\star = |S^\star|$. Define $S \in \mathscr{S} := \{S : |S| \le C s^\star \text{ and } S \not\supseteq S^\star\}$, where $C > 1$ is as in the statement of Theorem 5. The key point is that there is at least one important variable omitted in the models $S \in \mathscr{S}$. Let $\rho_n^2 = \rho_n^2(\beta^\star) = n^{-1} v \log p$ denote the lower bound on $\beta_j^{\star 2}$ for $j \in S^\star$, as defined in (3.14), where $v = v(s^\star) = c\, s^\star \Lambda_{c\,s^\star}$. In their Supplementary Materials, Narisetty et al. (2019) showed that, with probability tending to 1,[1]

$$\ell_n(S,\hat\beta_S) - \ell_n(S^\star,\hat\beta_{S^\star}) \le -F|S\setminus S^\star| n\rho_n^2, \quad \text{uniformly over } S \in \mathscr{S},$$

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

$$\begin{aligned}
\sum_{S\in\mathscr{S}} \pi^n(S) &\lesssim \sum_{S\in\mathscr{S}} \frac{\pi^n(S)}{\pi^n(S^\star)} \\
&\le \sum_{S\in\mathscr{S}} \frac{\pi_n(S)}{\pi_n(S^\star)} (1+\alpha\gamma)^{-(|S|-|S^\star|)/2}\, e^{\alpha\{\ell_n(S,\hat\beta_S)-\ell_n(S^\star,\hat\beta_{S^\star})\}} \\
&\le \sum_{S\in\mathscr{S}} \frac{\binom{p}{s^\star}}{\binom{p}{|S|}} (1+\alpha\gamma)^{-(|S|-s^\star)/2} p^{-a(|S|-s^\star)-\alpha F v(|S|-s^\star)},
\end{aligned}$$

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

[1] The result that they *stated*, i.e., $\ell_n(S,\hat\beta_S) - \ell_n(S^\star,\hat\beta_{S^\star}) \lesssim -n\rho_n^2$, 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.

size of $S \cap S^\star$. That is,

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

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

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

we get

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

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

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

We have that $\alpha F v > 1 + a$ 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

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

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

CHAPTER

4

EMPIRICAL PRIORS FOR VARIABLE SELECTION IN HIGH-DIMENSIONAL LOGISTIC REGRESSION: A VARIATIONAL APPROXIMATION

4.1 Introduction

For the classic logistic regression, consider $y_1, y_2, \ldots, y_n$, independently distributed observations where each y_i has density function

$$f(y_i \mid x_i) = \exp\{y_i x_i^\top \beta - \log[1 + \exp(x_i^\top \beta)]\},$$

where x_i denotes row i of design matrix X, and β is a $p \times 1$ vector of coefficients. We only consider high-dimensional cases, where $p > n$.

In high-dimensional settings, it is common for the number of variables p can greatly exceed the number of observations n. In the empirical investigations in Chapters 2 and 3, we focused on cases with dimension moderate p, which was largely due to computational constraints. As

both n and p increase, MCMC becomes increasingly expensive, to the point that it can become infeasible. For this reason, there is interest in developing other more computationally efficient approximations to the posterior distribution.

Variational inference is a popular alternative to traditional MCMC. Instead of integration, the problem is converted to optimization, and allows for fast computation without sacrificing (too much) on the theoretical properties; for a detailed summary of variational inference, see Blei et al. (2017). With our empirical priors approach, we can also employ variational inference to aid in our computations. Previous work has been done in Yang and Martin (2020) to construction variational approximations to posteriors driven by empirical priors in high-dimensional linear regression. In this chapter, we focus on high-dimensional generalized linear models, same as Chapter 3, but specifically on logistic regression, and present a novel variational approximation to our empirical priors posterior distribution applied directly to the marginal posterior of the active set S. Variational inference has been applied to high-dimensional logistic regression by other authors, e.g., Ray et al. (2020), Zhang et al. (2019), Jaakkola and Jordan (2000), and Guoqiang (2022). The novelty of our proposed method is two-fold: first, instead of a more traditional Gaussian or Laplace prior, we use an empirically-centered prior, the benefits of which have been discussed extensively in the previous chapters; second, our use of the data-driven prior gives us access to a relatively simple expression for the marginal posterior mass function of the configuration S. This expression, unfortunately, does not directly allow for inference on S, so we propose a variational approximation directly on this marginal posterior. Rather than the somewhat complicated variational approximation with the mean-field family (mixture of a Gaussian and point-mass distribution) as is commonly done in the literature, we propose a simple independent-Bernoulli approximation to the marginal posterior for S, which yields a much simpler and transparent approximation. This transparency allows us to easily establish, in Theorem 8 below, that the proposed variational approximation shares the same strong selection consistency property as the marginal posterior for S that it is approximating. There are results of this type for high-dimensional linear regression (e.g., Guoqiang 2022; Huang et al. 2016; Ormerod et al. 2017), and also some general results for other brands of variational approximations (Ohn and Lin 2021), but we are not aware of any results in the literature on selection consistency for variational approximations in high-dimensional logistic regression. Our simulations show that this method performs well compared to other existing methods, both Bayesian and frequentist, and that this method produces results that approximate the inclusion probabilities from the MCMC method in Chapter 3 well. Our method is efficient and fast even with large n and p, settings that are not feasible for the MCMC method.

The rest of this chapter is organized as follows. First, we give some background on variational inference and how it is commonly formulated in high-dimensional linear regression in the

literature in Section 4.2. Then, in Sections 4.3.1 and 4.3.2, we re-introduce the setup of the problem and our empirically-centered prior and its associated posterior. In Section 4.4, we propose our novel variational approximation for said posterior, and in Section 4.4.1 detail the coordinate ascent variational inference (CAVI) algorithm for computation and its derivation. We show numerical results of our method compared with other methods in Section 4.5, including comparisons with our EB-MCMC method from the previous chapter. Finally, in Section 4.6 offers a discussion of our method as well as directions for further work.

4.2 Background on variational inference

Here we present the framework for variational inference. For simplicity, we focus this presentation on the high-dimensional linear regression context. The key ideas and principles remain roughly the same for other problems, e.g., high-dimensional logistic regression.

Consider a linear regression model $y = X\beta + \varepsilon$, where y is a $n \times 1$ vector of observed data, X is a $n \times p$ matrix, β is a $p \times 1$ vector of coefficients, and ε is a $p \times 1$ vector of error. It is standard to decompose this high-dimensional problem of finding β into a hierarchical problem with (S, β_S), where $S \subset \{1, 2, \ldots, p\}$ is the set of indices that correspond to active signals and β_S is the subset of β for a particular configuration S.

Take S to be a binary $p \times 1$ vector, where $S_j = 1$ if variable j is in the active set and $S_j = 0$ if it is not, then, following convention in the literature (e.g., Yang and Martin (2020), Ray and Szabó (2020)), a variational approximation of the following form is typical,

$$q_\theta(S, \beta) = \prod_{j=1}^{p} q_{j,\theta}(\beta_j \mid S_j) q_{j,\theta}(S_j),$$

where

$$q_{j,\theta}(S_j) = \begin{cases} \phi_j, & S_j = 1 \\ 1 - \phi_j, & S_j = 0 \end{cases}$$

$$q_{j,\theta}(\beta_j \mid S_j) = \begin{cases} \mathsf{N}(\beta_j \mid \mu_j, \tau_j^2), & S_j = 1 \\ \delta_0(\beta_j), & S_j = 0. \end{cases}$$

Here, the variational parameter θ consists of three p-vectors, $\theta = (\mu, \tau^2, \phi)$. The β_j's are taken to be independently distributed from a mixture of Gaussian and point-mass, i.e., $\beta_j \sim \phi_j \mathsf{N}(\mu_j, \tau_j^2) +$

$(1-\phi_j)\delta_0$. Collecting all the β_j's together, we have a family of approximate densities

$$\mathscr{L} = \Big\{ \bigotimes_{j=1}^{p} \{\phi_j \mathsf{N}(\mu_j, \tau_j^2) + (1-\phi_j)\delta_0\} : \mu_j \in \mathbb{R}, \tau_j^2 > 0, \phi_j \in [0,1] \Big\}.$$

This is known as a mean-field family, and is commonly used in variational inference, as is pointed out in the overview work of Blei et al. (2017). It allows for a relatively accurate approximation that is still easy to compute, thanks to the independence of the β_j's.

Once the variational family is proposed, a quantity known as the Evidence Lower Bound (ELBO) is calculated,

$$K(\theta) = \mathsf{E}_{(S,\beta)\sim q_\theta} \log\{\tilde{\pi}^n(S,\beta)/q_\theta(S,\beta)\},$$

and with an optimization algorithm such as coordinate ascent or gradient ascent, the optimal variational parameter is obtained by $\hat{\theta} = \arg\max_\theta K(\theta)$ and the original posterior distribution can now be approximated with $q_{\hat{\theta}}$.

This approximation has a number of drawbacks. The variational parameter θ is of dimension $3p$, which is huge space to optimize over. The algorithm needs to approximate over a large number of dimensions, and as one could imagine, the more dimensions there are, the more difficult it is to obtain an accurate approximation. This is the primary motivation behind our proposed variational approximation on S itself, yielding a much simpler approximation with only a p-vector ϕ in $[0,1]^p$ as our variational parameter.

4.3 High-dimensional logistic regression

4.3.1 Setup

The likelihood function for our logistic regression is as follows

$$L_n(\beta) = \prod_{i=1}^{n} f(y_i \mid x_i) = \prod_{i=1}^{n} \exp\{y_i x_i^\top \beta - \log[1+\exp(x_i^\top \beta)]\}.$$

Of course, the log-likelihood function is

$$\ell_n(\beta) = \sum_{i=1}^{n} y_i x_i^\top \beta - \log[1+\exp(x_i^\top \beta)]. \tag{4.1}$$

In the high-dimensional context, logistic regression has been investigated with various methods, both frequentist (e.g., Fan and Lv 2011; Tibshirani 1996; Zou 2006) and Bayesian (e.g., Piironen et al. 2017), for a more detailed account, see Section 3.1 in the previous chapter.

For our analysis, recall from the previous chapter that we decompose the high-dimensional problem into a hierarchical problem with $\beta = (S, \beta_S)$, where S is the active set that indicates where the signals are located in the β vector, and β_S is the $|S|$-vector that correspond to a specific S configuration. We also define $s = |S|$ to be the size of the active set. Following the decomposition of $\beta \to (S, \beta_S)$, we will also use $L_n(S, \beta_S)$ to denote the likelihood function corresponding to a specific S, and $l_n(S, \beta_S)$ its log-likelihood.

4.3.2 Empirical prior and its posterior

Recall the empirical prior and posterior distributions for logistic regression in Chapter 3. Here, we briefly present them again for completeness and ease of reading.

We first define the empirical prior for S

$$\pi_n(S) = \binom{p}{s}^{-1} p^{-as},$$

where $a > 0$ is a hyperparameter and $s = |S|$ is the cardinality of the active set, i.e. the number of active coefficients. Then, the conditional prior for β_S given S is empirically-driven, centered around the S-specific MLE $\hat{\beta}_S$, is defined as

$$(\beta_S \mid S) \sim \mathsf{N}_{|S|}(\hat{\beta}_S, \gamma(X_S^\top W(\hat{\beta}_S) X_S)^{-1}),$$

where $\gamma > 0$ is another hyperparameter to be specified. Recall that $W(\cdot)$ is a diagonal matrix that depends on covariates, with the diagonal entry being bounded in $[0, 1]$. Then the overall empirical prior Π_n for the p-vector β, which depends on data through the MLEs $\hat{\beta}_S$, is

$$\Pi_n(d\beta) = \sum_S \pi_n(S) \mathsf{N}_{|S|}(d\beta_S \mid \hat{\beta}_S, \gamma(X_S^\top W(\hat{\beta}) X_S)^{-1}) \times \delta_0(d\beta_{S^c}), \quad \beta \in \mathbb{R}^p.$$

Following the previous two chapters and Martin et al. (2017), the posterior distribution combines the likelihood and prior in almost the usual Bayesian way, with a fractional power $\alpha \in (0, 1)$ added to the likelihood

$$\Pi^n(d\beta) \propto L_n(\beta)^\alpha \Pi_n(d\beta),$$

where α is a constant to be specified. In practice, we take α to be very close to 1 ($\alpha = 0.99$), that is, our posterior is not so different in practice from a genuine Bayesian posterior.

In principle, we can integrate out the β_S to obtain a marginal posterior on S (see Section 3.3.2 in Chapter 3). Unlike the linear model case, this marginal posterior for S is not available in

closed-form, so we employ Laplace's method to get the following simple approximation:

$$\pi^n(S) \propto \pi_n(S)(1+\alpha\gamma)^{-|S|/2} L_n(S, \hat{\beta}_S)^{\alpha}. \tag{4.2}$$

With this approximation, we can now use Metropolis–Hastings to (approximately) sample from this posterior, obtaining inclusion probabilities for each variable. This is the strategy we utilized in the previous two chapters, specifically using a symmetric proposal distribution that allows the algorithm to move in the S space, see Sections 3.3.3 for details. Our previous MCMC strategy moves from state S to a new S' that is different from S in only one position of S. This strategy works well with moderate p, but as p increases, this strategy becomes incredibly inefficient, and would need to run a long time to be able to sufficiently move around the p-dimensional S space. This is our motivation for developing a different strategy, one that is much more efficient without losing on accuracy.

4.4 A new variational approximation

Thanks to Laplace approximation, our marginal posterior for S has a very nice form (up to normalization). We take advantage of the simple expression in Equation (4.2) by applying our variational approximation directly to this marginal posterior. That is, instead of what is typically done in the literature with the variational approximation applied to the joint posterior for (S, β_S), we can look directly at the posterior for S, simplifying the structure of the problem.

Recall that the configuration S can be interpreted as both a subset of $\{1, 2, \ldots, p\}$ and a binary p-vector $S = (S_1, \ldots, S_p)$ with $S_j = 1$ indicating the that variable j is active and $S_j = 0$ indicating that it is inactive. Taking the latter interpretation, a natural choice of approximation to the marginal posterior of S is an independent Bernoulli model. That is, our proposed approximate marginal mass function for S is

$$q_\phi(S) = \prod_{j=1}^{p} \phi_j^{S_j} (1-\phi_j)^{1-S_j},$$

where $\phi = (\phi_1, \phi_2, ..., \phi_p)^\top \in [0,1]^p$ is the variational parameter to be determined. The goal is to find the value ϕ that minimizes the Kullback–Leibler divergence of π^n from q_ϕ. In other words, we aim to minimize the objective function

$$K(\phi) = \mathsf{E}_{S \sim q_\phi}\Big[\log \frac{q_\phi(S)}{\pi^n(S)}\Big]. \tag{4.3}$$

That our variational approximation is simpler than those commonly found in the liter-

ature on high-dimensional Bayesian methods makes it more transparent in some ways. In particular, it is not immediately clear that the posterior approximation would inherit the statistical properties (e.g., asymptotic concentration rate) enjoyed by the posterior distribution itself. Considerable effort has been spent to prove that, in certain cases, the variational approximations do, in fact, inherit at least some of the posterior's desirable properties; see, e.g., Alquier and Ridgway (2020); Ray and Szabó (2020); Ray et al. (2020); Zhang and Gao (2020). One property that the aforementioned references do not establish is *selection consistency* of the variational approximation. One possible reason for this gap in the existing literature is that the general tool used to transfer properties of the posterior distribution to the variational approximation requires certain exponential bounds on the posterior behavior that have yet to be established for the selection consistency-related properties. Since our post-marginalization approximation is simpler, we ought to be able to attack the problem directly and, indeed, below we demonstrate that our variational approximation $q_{\hat{\phi}}$ can achieve selection consistency, just like our marginal posterior π^n.

To set the scene, let $\beta^\star$ denote the true coefficient (a p-vector with $p = p_n \to \infty$) and let $S^\star$ denote the true configuration. Recall that $\beta^\star$ and $S^\star$ are actually sequences indexed by n, e.g., $S^\star$ can be interpreted as a binary p_n-vector with the only constraint being a limit on how many 1's it can contain; this is made explicit in the statement of Theorem 7 in Chapter 3. Recall, also, that under the conditions of the aforementioned theorem, we have that $\pi^n(S^\star) \to 1$ as $n \to \infty$ in $\mathsf{P}_{\beta^\star}$-probability; consequently, $\sum_{S \neq S^\star} \pi^n(S) \to 0$ in $\mathsf{P}_{\beta^\star}$-probability. Now, write the objective function as $K = K_n$ to make the (previously implicit) dependence on n and $p = p_n$ explicit. We first show that $K_n(\phi)$ does not converge to ∞ as $n \to \infty$ for all $\phi \in [0,1]^p$. Take $\phi^\star$ such that $\phi^\star_j = S^\star_j$, i.e., $\phi^\star$ is just $S^\star$ interpreted as a binary indicator vector. Then it is easy to check that $K_n(\phi^\star) \to 0$, so $K_n(\phi) \not\to \infty$ uniformly in ϕ. This means that K_n can asymptotically distinguish between the quality of different ϕ values. Writing the objective function as

$$\begin{aligned} K_n(\phi) &= \sum_S q_\phi(S) \log \frac{q_\phi(S)}{\pi^n(S)} \\ &= q_\phi(S^\star) \log \frac{q_\phi(S^\star)}{\pi^n(S^\star)} + \sum_{S \neq S^\star} q_\phi(S) \log \frac{q_\phi(S)}{\pi^n(S)}, \end{aligned}$$

makes the following observation clear: the only way to prevent $K_n(\phi) \to \infty$—which we know is possible by the argument above—is if $q_\phi(S) \to 0$ for all $S \neq S^\star$. But this implies that $q_\phi(S^\star) \to 1$ in $\mathsf{P}_{\beta^\star}$-probability as $n \to \infty$. The above remarks must apply to $\phi = \hat{\phi}$, a minimizer of K_n, and the desired result follows. We have therefore proved the following selection consistency theorem for the proposed variational approximation.

Theorem 8. *Under the conditions of Theorem 7 in Chapter 3, focused on the logistic regression*

case, the proposed variational approximation shares the same selection consistency property as the marginal posterior π^n that it is approximating. That is,

$$q_{\hat{\phi}}(S^\star) \to 1 \quad \textit{in } \mathsf{P}_{\beta^\star}\textit{-probability} \quad \textit{as } n \to \infty.$$

Note that, if π^n was independent in the sense that the joint mass function for S factored as the product of the S_j marginal mass functions, then it would follow immediately that $\hat{\phi}_j$ equals $\pi^n(S_j = 1)$ for all $j = 1, \ldots, p$. Of course, the marginal posterior for S under π^n surely will not factor in this way, but in the asymptotic limit, where π^n concentrates all of its mass on $S^\star$, this factorization does hold. Therefore, the $\hat{\phi}_j$ values ought to be close to $\pi^n(S_j = 1)$, what we referred to previously as the *inclusion probability*. Since the inclusion probability is what was used to carry out variable selection in Chapter 3, and since $\hat{\phi}_j$ is roughly approximating the inclusion probability, we can expect that the variable selection performance by our proposed variational approximation here is comparable to that of the MCMC-based method presented in the previous chapter—but with much faster computation. We investigate the proximity of $\hat{\phi}$ to the inclusion probabilities produced by MCMC in Section 4.5.2.

4.4.1 Implementation

Unfortunately, directly minimizing (4.3) is a major challenge for non-linear models like logistic regression and other GLMs. Note that the log-likelihood term depends on S in a very complex, non-linear way, so it is difficult—if not impossible—to evaluate the expected value of the log-likelihood with respect to $S \sim q_\phi$ analytically. As an alternative, we follow Ray et al. (2020) and introduce an additional parameter η and use it to bound the log-likelihood function:

$$\ell_n(S, \beta_S) \geq \sum_{i=1}^{n} \Big\{ \log \frac{\exp(\eta_i)}{1+\exp(\eta_i)} - \frac{\eta_i}{2} + \Big(y_i - \frac{1}{2}\Big) M_i(S,\beta) - \frac{\tanh(\eta_i/2)}{4\eta_i} [M_i(S,\beta)^2 - \eta_i^2] \Big\}, \quad (4.4)$$

where $M_i(S,\beta) = \sum_{j \in S} x_{ij}\beta_j$. This lower bound on the log-likelihood leads to an upper bound on the original objective function $\phi \mapsto K(\phi)$ in (4.3), so the new proposed strategy is to minimize this upper bound.

Since $\hat{\beta}$ is the MLE, $L_n(S, \hat{\beta}_S) \geq L_n(S, \tilde{\beta}_S)$, where $\tilde{\beta}_S$ is the coefficient estimator from another method, such as lasso or SCAD. Denoting the right-hand-side of Equation (4.4) as $g_n(S, \beta_S)$ and plugging in Equation (4.2), our objective function becomes

$$\mathsf{E}_{S \sim q_\phi}\Big[\log \frac{q_\phi(S)}{\pi^n(S)}\Big] = \mathsf{E}_{S \sim q_\phi}\Big[\log \frac{q_\phi(S)}{\pi_n(S)(1+\alpha\gamma)^{-|S|/2} L_n(S, \hat{\beta}_S)^\alpha}\Big]$$

$$\leq \mathsf{E}_{S\sim q_\phi}\Big[\log q_\phi(S) - \log \pi_n(S) + \frac{|S|}{2}\log(1+\alpha\gamma) - \alpha g_n(S,\tilde{\beta}_S)\Big],$$

where $\pi_n(S)$ is the complexity prior on S defined in Section 4.3.2. We develop a coordinate-ascent variational inference (CAVI) algorithm to find its solution.

Our CAVI algorithm searches for the maximizer of $\mathsf{E}_{S\sim q_\phi}\Big[-\log \frac{q_\phi(S)}{\pi^n(S)}\Big]$. Thanks to the introduction of the free parameter $\eta = (\eta_1, \eta_2, \ldots, \eta_n)^\top$, we have a closed-form equation for the lower bound approximation.

$$\begin{aligned}
\mathsf{E}_{S\sim q_\phi}\Big[-\log\frac{q_\phi(S)}{\pi^n(S)}\Big] &\geq \mathsf{E}_{S\sim q_\phi}\Big[-\log q_\phi(S) + \log \pi_n(S) - \frac{|S|}{2}\log(1+\alpha\gamma) + \alpha g_n(S,\tilde{\beta}_S)\Big] \\
&= -\mathsf{E}_{S\sim q_\phi}\Big[\log q_\phi(S)\Big] + \mathsf{E}_{S\sim q_\phi}\Big[\log \pi_n(S)\Big] - \mathsf{E}_{S\sim q_\phi}\Big[\frac{|S|}{2}\log(1+\alpha\gamma)\Big] \\
&\quad + \mathsf{E}_{S\sim q_\phi}\Big[\alpha g_n(S,\tilde{\beta}_S)\Big]
\end{aligned}$$

We will look at each of the four components of this lower bound. The first component, $\mathsf{E}_{S\sim q_\phi}\Big[\log(q_\phi(S))\Big]$, can be evaluated directly,

$$\begin{aligned}
\mathsf{E}_{S\sim q_\phi}\Big[\log(q_\phi(S))\Big] &= \mathsf{E}_{S\sim q_\phi}\Big\{\sum_{j=1}^{p}\Big[S_j\log(\phi_j) + (1-S_j)\log(1-\phi_j)\Big]\Big\} \\
&= \sum_{j=1}^{p}\Big[\phi_j\log(\phi_j) + (1-\phi_j)\log(1-\phi_j)\Big]
\end{aligned}$$

The third component is also straightforward,

$$\mathsf{E}_{S\sim q_\phi}\Big[\frac{|S|}{2}\log(1+\alpha\gamma)\Big] = \sum_{j=1}^{p}\frac{1}{2}\phi_j\log(1+\alpha\gamma)$$

The second component, $\mathsf{E}_{S\sim q_\phi}\Big[\log \pi_n(S)\Big]$, cannot be evaluated closed-form, so we will use a lower-bound,

$$\begin{aligned}
\mathsf{E}_{S\sim q_\phi}\Big[\log \pi_n(S)\Big] &= \mathsf{E}_{S\sim q_\phi}\Big[-\log\binom{p}{|S|} - a|S|\log(p)\Big] \\
&\geq \mathsf{E}_{S\sim q_\phi}\Big\{-|S|[1+\log(p)] - a|S|\log(p)\Big\} \\
&= -\sum_{j=1}^{p}\phi_j\Big[1+\log(p) + a\log(p)\Big]
\end{aligned}$$

The fourth component requires us to evaluate the two expected values $\mathsf{E}_{S\sim q_\phi}M_i(S,\tilde{\beta})$ and

$\mathsf{E}_{S\sim q_\phi} M_i(S,\tilde\beta)^2$, where $M_i(S,\tilde\beta)=\sum_{j\in S} x_{ij}\tilde\beta_j=\sum_{j=1}^p S_j x_{ij}\tilde\beta_j$. This gives

$$\mathsf{E}_{S\sim q_\phi} M_i(S,\tilde\beta)=\mathsf{E}_{S\sim q_\phi}\Big[\sum_{j=1}^p S_j x_{ij}\tilde\beta_j\Big]=\sum_{j=1}^p \phi_j x_{ij}\tilde\beta_j$$

and

$$\begin{aligned}
\mathsf{E}_{S\sim q_\phi} M_i(S,\tilde\beta)^2 &= \mathsf{V}_{S\sim q_\phi}[M_i(S,\tilde\beta)]+\mathsf{E}_{S\sim q_\phi}[M_i(S,\tilde\beta)]^2\\
&= \mathsf{V}_{S\sim q_\phi}\Big(\sum_{j=1}^p S_j x_{ij}\tilde\beta_j\Big)+\Big(\sum_{j=1}^p \phi_j x_{ij}\tilde\beta_j\Big)^2\\
&= \sum_{j=1}^p \mathsf{V}_{S_j\sim\mathsf{Ber}(\phi_j)}(S_j x_{ij}\tilde\beta_j)+\Big(\sum_{j=1}^p \phi_j x_{ij}\tilde\beta_j\Big)^2\\
&= \sum_{j=1}^p \phi_j(1-\phi_j)x_{ij}^2\tilde\beta_j^2+\Big(\sum_{j=1}^p \phi_j x_{ij}\tilde\beta_j\Big)^2
\end{aligned}$$

Putting everything together,

$$\begin{aligned}
\mathsf{E}_{S\sim q_\phi}\Big[-\log\frac{q_\phi(S)}{\pi^n(S)}\Big] \geq & -\sum_{j=1}^p \phi_j[1+\log(p)+a\log(p)-0.5\log(1+\alpha\gamma)]\\
&+\alpha\sum_{i=1}^n\Big\{\log\frac{\exp(\eta_i)}{1+\exp(\eta_i)}-\frac{\eta_i}{2}+\Big(y_i-\frac{1}{2}\Big)\sum_{j=1}^p \phi_j x_{ij}\tilde\beta_j\\
&-\frac{1}{4\eta_i}\tanh\frac{\eta_i}{2}\Big[\sum_{j=1}^p \phi_j(1-\phi_j)x_{ij}^2\tilde\beta_j^2+\Big(\sum_{j=1}^p \phi_j x_{ij}\tilde\beta_j\Big)^2-\eta_i^2\Big)\Big]\Big\}\\
&-\sum_{j=1}^p\Big[\phi_j\log(\phi_j)+(1-\phi_j)\log(1-\phi_j)\Big].
\end{aligned}$$

Taking the derivative of the above with respect to ϕ_j, and setting $\omega_j=\log(\frac{\phi_j}{1-\phi_j})$, we have the following update equation for the $(t+1)$-th iteration of ϕ_j,

$$\begin{aligned}
\omega_j^{(t+1)} = & \alpha\tilde\beta_j\sum_{i=1}^n\Big(y_i-\frac{1}{2}\Big)x_{ij}-\frac{\alpha\tilde\beta_j}{4}\sum_{i=1}^n\frac{1}{\eta_i^{(t)}}\tanh\Big(\frac{\eta_i^{(t)}}{2}\Big)\Big(x_{ij}^2\tilde\beta_j+2x_{ij}\sum_{k\neq j}\phi_k^{(t)}x_{ik}\tilde\beta_k\Big)\\
&+\frac{1}{2}\log(1+\alpha\gamma)-(a+1)\log(p)-1 \qquad (4.5)
\end{aligned}$$

$$\phi_j^{(t+1)}=\frac{\exp(\omega_j^{(t+1)})}{1+\exp(\omega_j^{(t+1)})} \qquad (4.6)$$

Algorithm 1 CAVI for variational empirical Bayes

Input: data (X, y); a fixed estimator $\tilde{\beta}$ based on another method such as SCAD or Lasso; an initial $\phi^{(0)}$ for the initial value of ϕ; a stopping threshold ϵ; a maximum number of iterations max.iter.

Initialize η with Equation (4.7) and $\phi^{(0)}$.
Calculate $\omega^{(0)}$, $\omega^{(1)}$ by Equation (4.5) and $\phi^{(1)}$ based on Equation (4.6).
Set $t = 1$.
while $\max_j |H(\phi_j^{(t)}) - H(\phi_j^{(t-1)})| > \varepsilon$ and $t <$ max.iter **do**
    for $j = 1, 2, \ldots, p$ **do**
        Update $\omega_j^{(t+1)}$ by Equation (4.5)
        Compute $\phi_j^{(t+1)}$ based on Equation (4.6)
    end for
    Update η with Equation (4.7)
    $t = t + 1$
end while
Return $\hat{\phi} = \phi^{(t)}$

Our free parameter η is updated with

$$\eta_i^{(t+1)} = \mathsf{E}_q[M_i(S)^2]^{1/2} = \Big\{\sum_{j=1}^{p} \phi_j^{(t)}(1-\phi_j^{(t)}) x_{ij}^2 \tilde{\beta}_j^2 + \Big(\sum_{j=1}^{p} \phi_j^{(t)} x_{ij} \tilde{\beta}_j\Big)^2\Big\}^{1/2} \tag{4.7}$$

We provide details of our CAVI algorithm in Algorithm 1. Inputs for our algorithm include the data X and observables y, a fixed estimator $\tilde{\beta}$, an initial value for ϕ, a stopping threshold ε, and a maximum number of iterations to ensure the algorithm would stop even without converging. The fixed estimator could be based on any other reliable method—both lasso and SCAD work fairly well—that is also not difficult to compute. In practice, we took our $\tilde{\beta}$ from SCAD, with a small caveat—the form of Equation (4.5) necessitates that zero entries in $\tilde{\beta}$ will not be updated iteratively, so we take the zero entries from SCAD and add a small amount of noise to them. The initial value for ϕ could be also set to match $\tilde{\beta}$, but in practice, we have found that $\phi^{(0)} = (0.5, \ldots, 0.5)^\top$ works equally well.

The stopping criterion evaluates the difference of ϕ values between consecutive iterations, and the algorithm stops when the difference is below a threshold ε, specifically, following Ray et al. (2020), Yang and Martin (2020), and Huang et al. (2016), we look at the maximum entropy criterion. We stop our algorithm when $\max_j |H(\phi_j^{(t)}) - H(\phi_j^{(t-1)})| > \varepsilon$, where $H : [0,1] \to \mathbb{R}$ is defined as $H(z) = -z\log_2(z) - (1-z)\log_2(1-z)$.

For variable selection, the solution $\hat{\phi}$ from Algorithm 1 offers an approximation of inclusion probabilities for each β_j to be included in the true active set $S^\star$. For point estimation, we can

then take the indices j that satisfy $\hat{\phi}_j \geq 0.5$, and calculate the MLE $\hat{\beta}$ based on the model with the chosen set of indices.

4.5 Results

We demonstrate the efficiency and accuracy of our method with numerical simulations. We compare our method, EB-VI, with a number of state-of-the-art methods for high-dimensional logistic regression. We first compare our method with other Bayesian methods including variational Bayes with Laplace prior (Ray and Szabó (2020)) and with Gaussian prior (cite), and R packages varbvs, SkinnyGibbs, BinaryEMVS, BhGLM, and rstanarm (citations). We also compare our method to a few other popular frequentist and Bayesian methods, including horseshoe, lasso, adaptive lasso, SCAD, and MCP. Lastly, we compare our EB-VI to the method discussed in Chapter 3, denoted here as EB-MCMC, especially looking at the difference between the solution $\hat{\phi}$ obtained in EB-VI and the inclusion probabilities from EB-MCMC, to investigate whether our variational approximation effectively solves the original problem of interest.

4.5.1 Comparisons with other methods

For the comparisons with other state-of-the-art Bayesian methods, we consider the same simulation settings as Ray and Szabó (2020). We look at five different simulation settings, with 200 runs each. In all five test settings, the entries of the design matrix X are simulated from an independent Normal distribution with mean 0 and varying standard deviation σ, i.e., $X_{ij} \sim \mathsf{N}(0, \sigma^2)$. The number of active signals s and the size of the signals A both vary in the five tests. Test 1 through 3 take $n = 250$, $p = 500$, and Test 4 and Test 5 are higher-dimensional, with $n = 2500$, $p = 5000$. All tests place the nonzero signals at the beginning of the true coefficient vector. The specific settings are summarized below:

Test 1. $n = 250$, $p = 500$, $\sigma = 0.25$, $s = 5$, and $A = 4$

Test 2. $n = 250$, $p = 500$, $\sigma = 2$, $s = 10$, and $A = 6$

Test 3. $n = 250$, $p = 500$, $\sigma = 0.5$, $s = 15$, and $A \sim \text{Unif}(-2, 2)$

Test 4. $n = 2500$, $p = 5000$, $\sigma = 0.5$, $s = 25$, and $A = 2$

Test 5. $n = 2500$, $p = 5000$, $\sigma = 1$, $s = 10$, and $A \sim \text{Unif}(-1, 1)$

We compare two metrics, True Positive Rate (TPR) and False Discovery Rate (FDR), defined as, respectively,

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{and} \quad \text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}},$$

Table 4.1: Comparison of TPR and FDR for select Bayesian methods across five test settings. Rows 2–6 of both the TPR and FDR panels are taken from Table 3 in Ray and Szabó (2020).

	Algorithm	Test 1	Test 2	Test 3	Test 4	Test 5
TPR	**EB-VI**	0.96 ± 0.10	1.00 ± 0.00	0.30 ± 0.12	1.00 ± 0.00	0.31 ± 0.09
	VB(Lap)	0.99 ± 0.06	1.00 ± 0.00	0.51 ± 0.11	1.00 ± 0.00	0.40 ± 0.28
	VB(Gauss)	1.00 ± 0.01	1.00 ± 0.02	0.54 ± 0.11	1.00 ± 0.00	0.85 ± 0.06
	varbvs	1.00 ± 0.00	1.00 ± 0.00	0.68 ± 0.11	1.00 ± 0.00	0.87 ± 0.06
	SkinnyGibbs	0.98 ± 0.06	1.00 ± 0.02	0.51 ± 0.12	-	-
	BinEMVS	0.99 ± 0.03	1.00 ± 0.00	0.58 ± 0.11	-	-
FDR	**EB-VI**	0.03 ± 0.08	0.03 ± 0.05	0.03 ± 0.08	0.00 ± 0.01	0.01 ± 0.07
	VB(Lap)	0.49 ± 0.11	0.00 ± 0.02	0.41 ± 0.14	0.01 ± 0.02	0.03 ± 0.05
	VB(Gauss)	0.63 ± 0.07	0.09 ± 0.13	0.52 ± 0.12	0.81 ± 0.02	0.95 ± 0.01
	varbvs	0.93 ± 0.01	0.08 ± 0.08	0.83 ± 0.03	0.93 ± 0.00	0.91 ± 0.01
	SkinnyGibbs	0.80 ± 0.03	0.11 ± 0.11	0.71 ± 0.07	-	-
	BinEMVS	0.43 ± 0.14	0.19 ± 0.10	0.63 ± 0.10	-	-

where TP is the number of truely positive signals identified by the algorithm, FN is the number of signals that the algorithm failed to identify, and FP is the number of variables that are actually noise but falsely selected by the algorithm.

The results are shown in Table 4.1, where rows 2 through 6 of both the TPR and FDR values are taken from Table 3 of Ray and Szabó (2020), and row 1 is from our newly proposed method. Our method does well in locating the signals in Tests 1, 2, and 4, as evidenced by the high TPR values. Tests 3 and 5 have the most challenging settings for signal discovery, as the TPR values are the lowest for all methods. In terms of FDR, our method does very well, obtaining very low FDR values in all five tests and outperforming the other methods in four of the five. Our EB-VI method is also very efficient, finishing each single run in around a second for the first three settings with $n = 250$, $p = 500$, on-par with the time required for the variational Bayesian with Laplace prior (VB-Laplace) method of Ray and Szabó (2020); for the higher-dimensional tests 4 and 5, our method generally took 2–3 minutes for a single run, slightly slower than the VB-Laplace method.

We also compare our EB-VI against the methods discussed in Chapter 3, and use the same simulations settings as the logistic regression simulations from the previous chapter. The methods, metrics, and configurations are laid out in detail in Section 3.5.2. These results are shown in Table 4.2. The right-most column includes results for our new method, EB-VI, while the other columns are taken from Table 3.1. We see that EB-VI does fairly well across these settings for all three metrics. Our method has very high True Negative Rate (TNR) across all

settings and in four out of the eight total settings, EB-VI has the highest MCC value. As the sparsity level decreases (the settings with $|S| = 8$), it is harder to identify all the signals, the TPR values for our method suffers a little, but is still on-par with other methods, and generally similar to the TPR values from the EB-MCMC method (denoted EB1 and EB2 in Table 4.2).

4.5.2 Comparisons with EB-MCMC

We discussed the theoretical connections in Section 4.4 between the solution from our variational approximated empirical priors method EB-VI and the inclusion probabilities obtained with MCMC from our empirical priors method EB-MCMC from Chapter 3. Here, we demonstrate that the newly proposed EB-VI method is approximating the MCMC inclusion probabilities well numerically using simulations. To compare the $\hat{\phi}$ against the inclusion probabilities, denoted by $\hat{\pi}^n$, we will look at the following metric,

$$D = \sqrt{\frac{1}{p}\mathsf{E}\Big(\|\hat{\pi}^n - \hat{\phi}\|^2\Big)}, \tag{4.8}$$

where the expectation is with respect to data. Note that the $\hat{\phi}$ here is not exactly the minimizer to the Kullback-Leibler divergence of π^n from q_ϕ, but the solution to the CAVI algorithm detailed in Algorithm 1. $\hat{\pi}^n$ is the inclusion probabilities for the empirical priors posterior distribution, as approximated by MCMC with $M = 10,000$ samples.

Table 4.3 provides a comparison of the distance D across select settings, each setting with 100 runs. The true coefficient vector $\beta^\star = (A, \ldots, A, 0, \ldots, 0)^\top$, with signal size A and the number of true signals s. The entries X_{ij}'s in the design matrix X are independently generated with a standard Normal distribution, i.e., $X_{ij} \overset{\text{iid}}{\sim} \mathsf{N}(0,1)$.

As is evident from Table 4.3, the distances between $\hat{\phi}$ and $\hat{\pi}^n$ are small across the different settings tested, which is another indication that numerically, our CAVI algorithm is approximating the inclusion probabilities from MCMC well. In terms of runtime, as one would expect, EB-VI is much faster than EB-MCMC. For a dataset that takes EB-MCMC around 30 seconds to run, it would take EB-VI roughly between 0.05 to 0.1 seconds.

4.6 Discussion

In this chapter, we propose a novel variational approximation directly on the marginal posterior for S for variable selection in high-dimensional logistic regression. Following chapter 3, we use an empirically-centered prior, which yields a marginal posterior S that has a simple form after Laplace approximation. Our simple independent-Bernoulli approximation for S

shrinks the variational parameter space from what is typically $3p$ to just $[0,1]^p$, streamlining the CAVI algorithm for computations. Thanks to its relative simplicity, we are able to prove (see Theorem 8) that our proposed variational approximation shares the same selection consistency property as the marginal posterior it is approximating. Simulations show that our method is efficient and produces good results, on-par with existing state-of-the-art methods.

We explored the relationship between the solution of our variational approximation and inclusion probabilities obtained with MCMC through simulations, and confirmed that the variational method does well in approximating the posterior distribution empirically. Theoretical properties of the relationship between the variational approximation and the original posterior of interest has yet to be explored, and would be an area of focus for further research. More work can also be done on proving other consistency results for our variational approximation. Finally, the proposed methodology can be extended beyond the logistic regression case to approximate the marginal posterior distribution for other high-dimensional generalized linear models.

Table 4.2: Comparison of TPR, TNR, and MCC for select frequentist and Bayesian methods across various settings. Columns 1–7 are taken from Table 3.1 in the previous chapter. Column 8 are results from the newly-proposed EB-VI.

p	$\|S\|$	r	Metric	EB1	EB2	HS	lasso	alasso	SCAD	MCP	**EB-VI**
			TPR	1.000	0.998	0.833	1.000	0.643	1.000	1.000	0.990
200	4	0	TNR	0.973	0.995	1.000	0.953	0.865	0.959	0.986	0.999
			MCC	0.669	0.907	0.905	0.645	0.208	0.584	0.787	0.976
			TPR	0.995	0.980	0.800	1.000	0.828	1.000	1.000	0.995
200	4	0.2	TNR	0.977	0.997	1.000	0.964	0.851	0.955	0.983	0.999
			MCC	0.708	0.936	0.886	0.703	0.289	0.560	0.746	0.979
			TPR	0.884	0.779	0.283	0.973	0.579	0.979	0.944	0.649
200	8	0	TNR	0.972	0.995	1.000	0.901	0.851	0.953	0.985	0.999
			MCC	0.723	0.812	0.502	0.548	0.210	0.668	0.821	0.780
			TPR	0.824	0.700	0.298	0.949	0.751	0.951	0.893	0.655
200	8	0.2	TNR	0.971	0.995	1.000	0.931	0.815	0.954	0.985	0.999
			MCC	0.670	0.760	0.526	0.619	0.274	0.656	0.792	0.788
			TPR	0.980	0.948	0.583	0.990	0.810	0.998	0.998	0.990
400	4	0	TNR	0.990	0.998	1.000	0.972	0.930	0.974	0.992	1.000
			MCC	0.750	0.908	0.733	0.605	0.317	0.540	0.750	0.983
			TPR	0.950	0.875	0.555	0.995	0.863	0.990	0.985	0.983
400	4	0.2	TNR	0.993	0.999	1.000	0.978	0.921	0.969	0.989	1.000
			MCC	0.793	0.892	0.732	0.664	0.314	0.499	0.698	0.983
			TPR	0.518	0.374	0.050	0.908	0.413	0.924	0.811	0.449
400	8	0	TNR	0.977	0.991	1.000	0.940	0.934	0.960	0.987	1.000
			MCC	0.448	0.411	0.132	0.489	0.162	0.538	0.670	0.617
			TPR	0.543	0.351	0.091	0.916	0.659	0.920	0.829	0.485
400	8	0.2	TNR	0.990	0.997	1.000	0.961	0.917	0.966	0.989	1.000
			MCC	0.572	0.499	0.221	0.595	0.269	0.563	0.709	0.674

Table 4.3: The distance D as defined in Equation (4.8) between the EB-VI solution and the EB-MCMC inclusion probabilities for various settings.

(n, p, s, A)	(100, 200, 4, 3)	(100, 200, 6, 3)	(100, 200, 4, 6)	(200, 400, 4, 3)
D	0.080	0.098	0.046	0.052

REFERENCES

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

Abramovich, F. and Grinshtein, V. (2016). Model selection and minimax estimation in generalized linear models. *IEEE Transactions on Information Theory*, 62(6):3721–3730.

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

Alquier, P. and Ridgway, J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475 – 1497.

Arias-Castro, E. and Lounici, K. (2014). Estimation and variable selection with exponential weights. *Electronic Journal of Statistics*, 8(1):328–354.

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

Belitser, E. and Ghosal, S. (2020). Empirical Bayes oracle uncertainty quantification for regression. *The Annals of Statistics*, 48(6):3113 – 3137.

Belitser, E. and Nurushev, N. (2020). Needles and straw in a haystack: Robust confidence for possibly sparse sequences. *Bernoulli*, 26(1):191–225.

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

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019b). Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: a review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

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(500):1610–1624.

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(3):277–279.

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

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(8):807.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018.

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(4):2069–2101.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Li, R. (2005). Statistical challenges with high dimensionality. In *Proceedings of the international Congress of Mathematicians*.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory*, 57(8):5467–5484.

Fang, X. and Ghosh, M. (2023). High-dimensional properties for empirical priors in linear regression with unknown error variance. *Statistical Papers*, pages 1–26.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

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

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Ghosh, P. and Chakrabarti, A. (2015). Posterior concentration properties of a general class of shrinkage estimators around nearly black vectors. Unpublished manuscript, `arXiv:1412.8161`.

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2023). *rstanarm: Bayesian applied regression modeling via Stan.* R package version 2.21.4.

Grünwald, P. and van Ommen, T. (2017). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. *Bayesian Analysis*, 12(4):1069–1103.

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(1):2040–2119.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, 110(512):1500–1514.

Guoqiang, L. (2022). A variational inference method for Bayesian variable selection. Unpublished manuscript, `arXiv:2211.11383`.

Hastie, T. and Efron, B. (2013). *lars: Least Angle Regression, Lasso and Forward Stagewise*. R package version 1.2.

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

Holmes, C. C. and Walker, S. G. (2017). Assigning a value to a power likelihood in a general Bayesian model. *Biometrika*, 104(2):497–503.

Hong, L., Kuffner, T. A., and Martin, R. (2018). On overfitting and post-selection uncertainty assessments. *Biometrika*, 105(1):221–224.

Huang, X., Wang, J., and Liang, F. (2016). A variational algorithm for Bayesian variable selection. Unpublished manuscript, `arXiv:1602.07640`.

Jaakkola, T. S. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10:25–37.

Jeong, S. and Ghosal, S. (2021). Posterior contraction in sparse generalized linear models. *Biometrika*, 108(2):367–379.

Jiang, W. (2007). Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. *The Annals of Statistics*, 35(4):1487–1511.

Johnstone, I. M. and Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4237–4253.

Kraemer, N. and Schaefer, J. (2014). *parcor: Regularized estimation of partial correlation matrices*. R package version 0.2-6.

Le Cao, K.-A., Rohart, F., Gonzalez, I., and Dejean, S. (2016). *mixOmics: Omics Data Integration Project*. R package version 6.1.1.

Lee, K. and Cao, X. (2021). Bayesian group selection in logistic regression with application to MRI data analysis. *Biometrics*, 77(2):391–400.

Lee, K., Lee, J., and Lin, L. (2019). Minimax posterior convergence rates and model selection consistency in high-dimensional DAG models based on sparse Cholesky factors. *The Annals of Statistics*, 47(6):3413–3437.

Leeb, H. (2006). The distribution of a linear predictor after model selection: unconditional finite-sample distributions and asymptotic approximations. In *Optimality*, volume 49 of *IMS Lecture Notes Monograph Series*, pages 291–311. Institute of Mathematical Statistics, Beachwood, OH.

Leeb, H. (2009). Conditional predictive inference post model selection. *The Annals of Statistics*, 37(5B):2838–2876.

Lindsay, B. G., Kettenring, J., and Siegmund, D. O. (2004). A Report on the Future of Statistics. *Statistical Science*, 19(3):387 – 413.

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

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

Liu, C., Yang, Y., Bondell, H., and Martin, R. (2021). Bayesian inference in high-dimensional linear models using an empirical correlation-adaptive prior. *Statistica Sinica*, 31(4):2051–2072.

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

Martin, R., Mess, R., and Walker, S. G. (2017). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*, 23(3):1822–1847.

Martin, R. and Ning, B. (2020). Empirical priors and coverage of posterior credible sets in a sparse normal mean model. *Sankhya A*, 82(2):477–498.

Martin, R. and Tang, Y. (2020). Empirical priors for prediction in sparse high-dimensional linear regression. *Journal of Machine Learning Research*, 21(144):1–30.

Martin, R. and Walker, S. G. (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics*, 8(2):2188–2206.

Martin, R. and Walker, S. G. (2019). Data-driven priors and their posterior concentration rates. *Electronic Journal of Statistics*, 13(2):3049–3081.

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

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125.

Mukherjee, G. and Johnstone, I. M. (2015). Exact minimax estimation of the predictive density in sparse Gaussian models. *The Annals of Statistics*, 43(3):937–961.

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(527):1205–1217.

Ohn, I. and Lin, L. (2021). Adaptive variational Bayes: Optimality, computation and applications. Unpublished manuscript, `arXiv:2109.03204`.

Ormerod, J. T., You, C., and Müller, S. (2017). A variational Bayes approach to variable selection. *Electronic Journal of Statistics*, 11(2):3549–3594.

Piironen, J., Vehtari, A., et al. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051.

Pötscher, B. M. and Leeb, H. (2009). On the distribution of penalized maximum likelihood estimators: the LASSO, SCAD, and thresholding. *Journal of Multivariate Analysis*, 100(9):2065–2082.

Ray, K. and Szabó, B. (2020). Variational Bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117(539):1270–1281.

Ray, K., Szabó, B., and Clara, G. (2020). Spike and slab variational Bayes for high dimensional logistic regression. *Advances in Neural Information Processing Systems*, 33:14423–14434.

Reid, S., Tibshirani, R., and Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67.

Rigollet, P. (2012). Kullback-Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2):639–665.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(4):749–760.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 197–206, Berkeley and Los Angeles. University of California Press.

Syring, N. and Martin, R. (2018). Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486.

Syring, N. and Martin, R. (2023). Gibbs posterior concentration rates under sub-exponential type losses. *Bernoulli*, 29(2):1080–1108.

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

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

van der Pas, S., Scott, J., Chakraborty, A., and Bhattacharya, A. (2016). *horseshoe: Implementation of the Horseshoe Prior*. R package version 0.1.0.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017a). Adaptive posterior contraction rates for the horseshoe. *Electronic Journal of Statistics*, 11(2):3196–3225.

van der Pas, S., Szabó, B., and van der Vaart, A. (2017b). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274. With a rejoinder by the authors.

van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2014). The horseshoe estimator: posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8(2):2585–2618.

Verzelen, N. (2012). Minimax risks for sparse regressions: ultra-high dimensional phenomenons. *Electronic Journal of Statistics*, 6:38–90.

Walker, S. and Hjort, N. L. (2001). On Bayesian consistency. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(4):811–821.

Walker, S. G., Lijoi, A., and Prünster, I. (2005). Data tracking and the understanding of Bayesian consistency. *Biometrika*, 92(4):765–778.

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

Yang, Y. and Martin, R. (2020). Variational approximations of empirical Bayes posteriors in high-dimensional linear models. Unpublished manuscript, `arXiv:2007.15930`.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques*, volume 6 of *Studies in Bayesian Econometrics and Statistics*, pages 233–243. North-Holland, Amsterdam.

Zhang, C.-X., Xu, S., and Zhang, J.-S. (2019). A novel variational Bayesian method for variable selection in logistic regression models. *Computational Statistics & Data Analysis*, 133:1–19.

Zhang, F. and Gao, C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180 – 2207.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.