Bayesian methods for variable selection

Anupreet Porwal

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:
Abel Rodríguez, Chair
Adrian E. Raftery, Chair
Adrian Dobra

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Bayesian methods for variable selection

Anupreet Porwal

Co-Chairs of the Supervisory Committee:
Professor Abel Rodríguez
Department of Statistics

Professor Adrian E. Raftery
Department of Statistics and Sociology

Choosing a statistical model and accounting for uncertainty about this choice are important parts of the scientific process and are required for common statistical tasks such as parameter estimation, interval estimation, statistical inference, point prediction and interval prediction. A canonical example is the variable selection problem in a linear regression model. Many ways of doing this have been proposed, including Bayesian and penalized regression methods. Each of these proposals has advantages and disadvantages, and the trade-offs are not always well understood.

In this dissertation, we first compare 21 popular existing methods via an extensive simulation study based on a wide range of real datasets. We found that three adaptive Bayesian model averaging (BMA) methods performed best across all the statistical tasks. Subsequently, we also investigate the effect of model space priors on model inference under the BMA framework. For this study, we consider eight reference model space priors used in the literature and three adaptive parameter priors recommended by the previous study.

Additionally, we proposed a novel objective prior based on Power-expected-posterior priors for generalized linear models that relies on a Laplace expansion of the likelihood of the imaginary training sample. We investigate both asymptotic and finite-sample properties

of the procedures, showing that they are both asymptotically and intrinsically consistent, and that their performance is superior to that of alternative approaches in the literature especially for heavy tailed versions of the priors.

Finally, we propose a framework that unifies the two Bayesian paradigms of inducing sparsity namely (mixture of) g-priors and continuous shrinkage priors. The mixture of g-priors use a single shrinkage parameter across all predictors included in the model, incorporate correlation structure of covariates into the priors and allows for model selection, however suffers from Conditional Lindley Paradox (CLP). Continuous shrinkage priors like the Horseshoe prior, on the other hand, allow for a different shrinkage parameter for each coefficient however does not perform model selection. We propose global local-g priors that borrows strength of the two paradigms and allows differential shrinkage across predictors while performing model selection. Additionally, we propose Dirichlet Process (DP) block-g priors that allows combining g-priors and Bayesian non-parametric tools to incorporate correlation structure in the priors as well as adaptively identify and cluster predictors with varying degrees of relevance using a different shrinkage parameter for different clusters. We show empirically and theoretically that our proposed priors avoid CLP while still performing competitively or superior to existing methods in terms of model selection, parameter estimation and prediction.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

This dissertation stands as a testament to the collective efforts of all those who supported and believed in me. While their names are mentioned here, their influence resonates throughout every page, every idea, and every discovery.

First and foremost, I want to express my deepest gratitude to my PhD advisors, Prof. Abel Rodríguez and Prof. Adrian E. Raftery, for their unwavering support, meaningful insights, and continuous encouragement over the past five years. Adrian, thank you for sparking my interest in the field, broadening my perspective, and igniting my passion for model selection. Abel, I am grateful for the freedom you granted me to explore and pursue my ideas, and for welcoming me as your first PhD student at UDub.

I extend my sincere appreciation to my esteemed committee members, Prof. Adrian Dobra and Prof. Youngjun Choe, for their constructive critiques and valuable suggestions that have greatly enhanced the quality of this work. Additionally, I want to express my gratitude to Prof. Thomas Richardson for allowing me to serve as a Teaching Assistant for his econometrics course, for being my informal advisor during challenging times, and being a true friend.

Turning to my undergraduate advisors, I extend my heartfelt thanks to Prof. Amit Mitra, Prof. Sharmistha, and Prof. Debasis Kundu. Your guidance and passion for Statistics ignited my interest and encouraged me to pursue graduate studies. A special acknowledgement goes to Prof. Piyush Rai, whose dedication to Bayesian statistics single-handedly made me a Bayesian and whose belief in my research capabilities were pivotal.

I owe a debt of gratitude to my cherished friends, who have consistently supported me and provided the energy to persist and move forward. To my cohort, Anna and Sarah, your energy

is infectious; Shane and Michael, thank you to help me be my true self; Zhaoqi and Incheoul, great office mates indeed; and Alan, your study sessions and camaraderie mean the world to me. To my alternative cohort, Kristof, Heather, Shelly, and Katina, your companionship has brightened many nights; Peter, your positivity lifts my spirits; and Bryan and Daphne, exceptional GSRs. I also extend my appreciation to Sara, Seth, Nick, Steve, James, Shreya, Ronak, Ellen, Jess Kunke, Medha, Erin, Alana, Alex Kokot, Vydhourie, Nilanjana, Thang, Greg, Ziyu, Federico, Carlos, and Diego. From India, my pillars of strength, Ashish, Shiva, Varad, Aakash Ghosh, Shubham, Aakash Gupta, Pranav Sharda, Archit, Aakriti, and Janak, your unwavering support means the world. A special thanks to Krishna Pranav, Aldrin Domer, and Bryce Van Vleet for countless zoom focus sessions and your invaluable emotional support. Finally, to my constants, roommates Siddarth Iyer and Srini Iyer, you are treasures.

Beyond my dissertation projects, my PhD journey was enriched by interdisciplinary research opportunities that shaped my perspective. Thank you, Prof. Yuanjin Zhou, for the consulting class and ADAP project, allowing me to contribute to meaningful research for older adults with Dementia. To my collaborators at Merck, Himel, Erina, and Vladimir Svetnik, I am grateful for the chance to work on cutting-edge multi-omics research. Finally, thank you, Fan Chen and Jiexing Wu at Google; I am excited for our future collaborations.

I offer a heartfelt thanks to the department staff for their unwavering support and their ability to brighten our days with their smiles. Lastly, but most importantly, my family: Mummy and Papa, your sacrifices, dedication, and blessings have enabled me, a small-town boy from India, to dream big and achieve a PhD. To my brother and sister-in-law, your unwavering support and inspiration have been critical. I hope I continue to make you proud. To my nephew Anush, I am so lucky to be your uncle, I hope you take our family name to new heights and continue to fill our lives with happiness.

DEDICATION

to my parents

Chapter 1
INTRODUCTION

This dissertation discusses and compares existing methods for variable selection develops new Bayesian methodologies for modeling sparsity. In this chapter, we provide a brief motivation and a review of the variable selection methods that exist in the literature. We then discuss the main contributions of this dissertation to the model selection literature.

1.1 Motivation

Choosing a statistical model and accounting for uncertainty about this choice are important parts of the scientific process, and are required for common statistical tasks such as parameter estimation, interval estimation, statistical inference, point prediction and interval prediction. Accounting for model uncertainty, rather than selecting a single statistical model, improves predictive performance and robustness in estimation and inference of model parameters (Raftery et al., 1997).

One canonical example of model uncertainty is that of variable selection in linear regression model. Given an n-dimensional continuous response variable, $\boldsymbol{Y}$, and a set of p potential predictor variables $\boldsymbol{X} = (X_1, \ldots, X_p) \in \mathcal{R}^{n \times p}$, the aim is to do statistical analysis of the data when it is not known in advance which of the 2^p possible models is most appropriate. Consider a binary indexing vector $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)$ for the model space which indicates which explanatory variables are part of a model $\mathcal{M}_{\boldsymbol{\gamma}}$. Under $\mathcal{M}_{\boldsymbol{\gamma}}$, the linear regression model can then be expressed as:

$$\mathcal{M}_{\boldsymbol{\gamma}} : \boldsymbol{Y} = \mathbf{1}_n \alpha + \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, where $\mathcal{N}$ denotes the multivariate normal distribution, and $\boldsymbol{X}_{\boldsymbol{\gamma}}$ is a $n \times p_{\boldsymbol{\gamma}}$ matrix where each column is centered around its mean and $p_{\boldsymbol{\gamma}}$ denotes the

number of explanatory variables in the model $\mathcal{M}_\gamma$. Other examples of model uncertainty include the choice of functional forms of the variables, choice of variables and link functions in generalized linear models (GLMs), and the choice of error distribution, for instance to account for potential outliers. In this dissertation, we focus on developing and evaluating methods for model selection in presence of model uncertainty for linear models and GLMs.

1.2 Background

Historically, one approach to deal with model uncertainty has been to determine the variables in a model using subject matter expertise, however this often leaves open questions due to lack of expertise, and a data-based approach is desired for at least some of the variables. Another approach is to always include all the candidate variables, but this can lead to poor statistical performance when there are many such variables. Many of the early statistical approaches were stepwise methods, in which variables were sequentially added or removed on the basis of significant tests, but these have not been found to have good theoretical or empirical properties (Miller, 2002; Freedman, 1983).

In the past thirty years, many satisfactory methods have been proposed. Most of these are either Bayesian techniques or penalized likelihood-based approaches. Many of the Bayesian techniques are some form of Bayesian model averaging (BMA) (Leamer, 1978; Raftery, 1988; George and McCulloch, 1993; Madigan and Raftery, 1994). BMA provides a formal way to account for model uncertainty in statistical inference. Several reviews of the BMA literature are available (see for e.g., Kass and Raftery, 1995; Hoeting et al., 1999; Fernandez et al., 2001; Wasserman, 2000; Clyde and George, 2004; Forte et al., 2018). The basic idea of BMA is that it uses prior probabilities for the models considered, and Bayes' theorem to deal with model uncertainty by calculating their posterior probabilities.

Assuming that there is one true model among the set of 2^p candidate models, the posterior probability of a model $\mathcal{M}_\gamma$ is

$$P(\mathcal{M}_\gamma|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|\mathcal{M}_\gamma)P(\mathcal{M}_\gamma)}{\sum_{\gamma'} P(\boldsymbol{Y}|\mathcal{M}_{\gamma'})P(\mathcal{M}_{\gamma'})},$$

where $P(\mathcal{M}_{\boldsymbol{\gamma}})$ is the prior model probability of $\mathcal{M}_{\boldsymbol{\gamma}}$, and $P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}})$ is the marginal likelihood of the model after integrating out parameters with respect to the prior $\pi_{\boldsymbol{\gamma}}$, namely:

$$P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}) = \int \mathcal{N}(\boldsymbol{Y}|\mathbf{1}_n\alpha + \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2 I_n)\pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \alpha, \sigma^2)d\boldsymbol{\beta}_{\boldsymbol{\gamma}}d\alpha d\sigma^2.$$

Under BMA, we can express the predictive distribution of a quantity of interest, Δ, such as a parameter or an observable future quantity, as a weighted average of its predictive distributions under the different candidate models:

$$P(\Delta|\boldsymbol{Y}) = \sum_{\boldsymbol{\gamma}} P(\Delta|\mathcal{M}_{\boldsymbol{\gamma}})P(\mathcal{M}_{\boldsymbol{\gamma}}|\boldsymbol{Y}),$$

where the posterior model probabilities $P(\mathcal{M}_{\boldsymbol{\gamma}}|\boldsymbol{Y})$ serve as averaging weights. In the case where Δ is a regression coefficient, the resulting posterior distribution, $P(\Delta|\boldsymbol{Y})$, is a mixture of a point mass at 0 and a continuous density.

BMA has several appealing good theoretical properties (Raftery and Zheng, 2003). BMA point estimators and predictions minimize mean squared error (MSE); BMA estimation and prediction intervals are calibrated, and BMA predictive distributions have optimal performance in the log score sense (Madigan and Raftery, 1994). These properties hold on average over the prior distribution, extending similar results for Bayesian estimation (Rubin and Schenker, 1986), but the results are somewhat robust to this assumption (Mattei, 2020). Used in this way, as a distribution of parameter values over which performance is averaged, the prior distribution has been referred to as the world distribution (Jeffreys, 1961), the practical distribution (Raftery and Zheng, 2003), or the effect-size distribution (Park et al., 2010), and analysis using this concept has been called "empirical frequentist" (Berger, 2021).

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

of parameter priors and computational algorithms. We focus on effect of model space priors on model uncertainty in Chapter 3.

For the parameter prior distribution in linear regression, several default choices have been proposed. Among the first was the Jeffrey-Zellner-Siow (JZS) Cauchy prior, with a standard Jeffreys prior for the intercept and error variance (Jeffreys, 1961; Zellner and Siow, 1980). Another early prior was the Zellner g-prior (Zellner, 1986a) of the form

$$\begin{aligned} \pi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\alpha, \sigma^2, g) &\sim \mathcal{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|0, g\sigma^2(\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}), \\ \pi_{\boldsymbol{\gamma}}(\alpha, \sigma) &\propto \sigma^{-1}, \end{aligned} \tag{1.2}$$

where $\boldsymbol{X}_{\boldsymbol{\gamma}}$ is the $n \times p_{\boldsymbol{\gamma}}$ matrix consisting of the covariates $\boldsymbol{X}_j$ for which $\gamma_j = 1$ (Fernandez et al., 2001). The prior variance of the regression parameters is controlled by the user-specified value g, and the effective prior sample size is n/g, where n is the sample size.

Various choices of g have been proposed (Forte et al., 2018). Zellner proposed using $g = n$, corresponding to a prior sample size of 1; this has been called the Unit Information Prior (UIP) (Kass and Wasserman, 1995). Another choice is $g = 1$, corresponding to a prior sample size of n (van Zwet, 2019). One justification for this choice is that studies have sample sizes designed to have the power to detect effects of known sizes, so that the prior and sampling variances are similar. An intermediate choice is $g = \sqrt{n}$ (Fernandez et al., 2001), with a prior sample size of $\sqrt{n}$; this has been found to work well in high-dimensional settings (Young et al., 2014). The Benchmark prior where $g = \max\{n, p^2\}$ has also been recommended (Fernandez et al., 2001); it combines the consistency properties of the UIP with the good small sample performance of the Risk Inflation factor (RIC)(Foster et al., 1994).

The UIP can also be approximated by the Bayesian Information Criterion (BIC) (Schwarz, 1978a; Raftery, 1995). The Akaike Information Criterion (AIC) can be used as the basis for an approximation to the posterior model probabilities under a prior that is similar to Zellner's g-prior with $g = 1$, i.e. with an equivalent prior sample size of n (Akaike, 1983; Burnham and Anderson, 2002).

An alternative is not to use a specified g, but instead to estimate g from the data. This can be done in an empirical Bayes way, either for each model separately (Hansen and Yu, 2003), or globally (Clyde and George, 2000; George and Foster, 2000). It can also be done in a more fully Bayesian way, by specifying a prior on g, such as the hyper-g approach (Liang et al., 2008).

A different type of prior used in BMA is the non-local prior (NLP) (Johnson and Rossell, 2012a; Rossell and Telesca, 2017), which removes mass close to zero. The spike and slab method approximates the zero values of lower-dimensional models with continuous distributions around zero (George and McCulloch, 1993, 1997).

It is important to stress that above priors for variable selection place positive probability on specific coefficients being exactly zero. An alternative Bayesian approach to sparsity that is popular in the literature is to use continuous shrinkage priors. Under these priors, the coefficients β_j are modeled independently using absolutely continuous distributions around 0, designed to differentially shrink coefficients of different magnitude (see, e.g., Bhadra et al., 2019 for a comprehensive review). These continuous shrinkage priors, often called global-local shrinkage priors, can often be reduced to scale mixture of normals. Under such representation, they involve a global shrinkage parameter controlling the overall sparsity and a local shrinkage parameter per coefficient controlling the local shrinkage of outlier signals. A general form of global-local shrinkage priors is given by

$$\beta_j \sim \mathcal{N}(0, \sigma^2\lambda_j^2) \quad \lambda_j \sim f(\lambda_j|\tau) \quad \tau \sim g(\tau), \tag{1.3}$$

which coupled with the Jeffreys' prior on (α, σ^2), i.e., $\pi(\alpha, \sigma^2) \propto \sigma^{-2}$. When $\lambda_j \sim \mathcal{C}^+(0, \tau)$ and $\tau \sim \mathcal{C}^+(0, 1)$ where $\mathcal{C}^+(0, a)$ denotes the half-Cauchy distribution with scale parameter a, it leads to a popular continuous shrinkage prior called the Horseshoe prior (Carvalho et al., 2010). Other examples include the Bayesian Lasso (Park and Casella, 2008), the Normal-Gamma prior (Brown and Griffin, 2010), the Dirichlet-Laplace prior (Bhattacharya et al., 2015), global-local shrinkage priors (Polson and Scott, 2012), the Beta-prime prior (Bai and Ghosh, 2018), the tail-adaptive shrinkage prior (Lee et al., 2020) and the Horseshoe-pit prior

(Denti et al., 2021). Continuous shrinkage priors tend to have computational advantages and are very effective in predictive settings. However, because they place probability zero on any one value of the parameter space, variable selection under these priors can be performed only by thresholding the posterior distributions of the model coefficients. While ad-hoc techniques have been devised for this purpose (e.g., see Li and Pati, 2017), thresholding tends to work best in settings where enough prior information is available to establish practical significance. Additionally, continuous shrinkage priors assume β_j are independent apriori and do not take correlation in to account in design of the prior.

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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathcal{R}^p}{\arg\min}(\boldsymbol{Y} - \alpha\mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \alpha\mathbf{1}_n - \boldsymbol{X}\boldsymbol{\beta}) + h_\lambda(\boldsymbol{\beta}). \tag{1.4}$$

The estimates from these techniques can also be viewed as maximum-a-posteriori (MAP) estimates under a prior of the form $p(\boldsymbol{\beta}) \propto \exp\{-h_\lambda(\boldsymbol{\beta})/\sigma^2\}$. The LASSO (Least absolute shrinkage and selection operator) (Tibshirani, 1996) was the first and remains perhaps the most widely used technique in this class, where the penalty takes the form $h_\lambda(\boldsymbol{\beta}) = \lambda\sum_{j=1}^{p}|\beta_j|$, and constrains the l_1 norm of the parameter vector. The popularity of the LASSO is due to factors that include the computational efficiency of the least angle regression (LARS) and coordinate ascent algorithms that can be used to estimate it (Efron et al., 2004; Friedman et al., 2007), its ability to provide a sparse estimate of $\boldsymbol{\beta}$, and the oracle property, namely that the LASSO will asymptotically find a superset of the correct predictors (Tibshirani, 2011).

However, LASSO also suffers from several known issues. The oracle property ensures only that the true predictors will be asymptotically part of the selected model, but not the converse, so that there can be many false positive selections, even asymptotically. LASSO also tends to over-shrink the true signals in the observed data and hence produce biased estimates (Casella et al., 2010). It can also be unstable in the presence of highly correlated

covariates. As pointed out by Holmes (Holmes, 2011), "In the presence of strong correlations between predictors with differing effect sizes, frequentist sparsity approaches, including the lasso, will tend to select a single variable within a group of collinear predictors, discarding the others in the pursuit of sparsity. However, the choice of the particular predictor might be highly variable and by selecting one we may ignore weak (but important) predictors which are highly correlated with stronger predictors."

The LASSO has a constant rate of penalization for all coefficients which can cause excessive shrinking of non-zero components. Some of the other popular penalty methods vary in the shape or rate of penalty applied to the coefficients. The smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010) methods involve a non-convex penalty which is constant for smaller coefficients and decreases to 0 for larger coefficients. The elastic net (Zou and Hastie, 2005) involves a convex combination of ridge and lasso penalties, encouraging grouping effects among strongly correlated variables, and thus addresses one concern mentioned above for the LASSO. Like the LASSO, these methods threshold some coefficients to zero, leading to simultaneous variable selection and estimation.

A common issue with penalized likelihood approaches is the lack of uncertainty quantification, since variable selection is an outcome of the constrained optimization problem and not a probabilistic statement of inclusion (Park and Casella, 2008; Fouskakis et al., 2015; Womack et al., 2014). As a result, the zeros induced may not be the same zeros that one would get from a full variable selection approach (Hans, 2009). They also do not provide a way to account for model uncertainty. EM Variable Selection (EMVS) (Ročková and George, 2014) and the Spike-and-Slab LASSO (SS LASSO) (Ročková and George, 2018) are two methods that synthesize ideas from BMA and penalized likelihood. In principle they could quantify uncertainty, but that has not yet been implemented in the associated software.

In practice, it is not clear which of the many proposed methods to use. Among penalized likelihood methods, LASSO probably remains the most used, perhaps because it was the first one proposed, there is a well-defined software package to implement it (the `glmnet` R

package), and it is fast (Friedman et al., 2010). Among Bayesian methods there is less clarity, and the relative performance of Bayesian and penalized likelihood methods is also not clear.

1.3 Contribution and outline of the dissertation

Many methods have been proposed for statistical analysis using linear regression models in the presence of model uncertainty. When the model is known in advance and only its parameters have to be estimated, there is consensus on how to perform statistical analysis, using either a frequentist or Bayesian approach. When the model is to be determined as part of the analysis, however, it is less clear which method to use.

In Chapter 2, we fill this gap by comparing 21 of the most popular methods for variable selection in linear models through an extensive set of simulation studies based closely on real datasets that span a range of situations encountered in practical data analysis. We compared the methods across the five statistical tasks focused on estimation, prediction and inference. We recommend three adaptive Bayesian model averaging (BMA) methods that are computationally efficient and performed best across all statistical tasks.

In Chapter 3, we focus on understanding the effect of model space priors on variable selection in linear regression models for the statistical tasks of parameter estimation, interval estimation, inference, point and interval prediction. We consider eight reference model space priors used in the literature and three adaptive parameter priors recommended in Chapter 2. We carry out an extensive simulation study based on 14 real datasets representing a range of situations encountered in practice. We found that beta-binomial model space priors specified in terms of the prior probability of model size performed best on average across various statistical tasks and datasets, outperforming priors that were uniform across models.

Unlike the previous two chapters, in chapter 4, we focus on model selection in GLMs. For Gaussian linear models, the literature on so-called "objective" or "default" priors for model selection is extensive. See Consonni et al. (2018) for a comprehensive review of recent approaches to objective Bayesian analysis, and Bayarri et al. (2012) for a review and discussion of desirable properties. However, in spite of their broad adoption, prior elicitation

for GLMs in the absence of subjective information remains an open problem, particularly in settings where the main goal is variable selection. Chapter 4 develops Laplace power-expected-posterior prior for logistic regression that combines the literature on power priors (Ibrahim and Chen, 2000), expected-posterior priors (Pérez and Berger, 2002) and unit information priors (Kass and Raftery, 1995), provides a systematic way to construct objective priors. The basic idea is to use imaginary training samples to update a (possibly improper) prior into a proper but minimally-informative one. This approach has various advantages over previous proposals for non-informative priors in logistic regression, and can be easily extended to other generalized linear models. We study theoretical properties of the prior and provide a number of empirical studies that demonstrate superior performance both in terms of model selection and of parameter estimation, especially for heavy-tailed versions.

In Chapter 5, we return to model selection in linear models and propose a framework that bridges the gap between the two Bayesian approaches to sparsity, namely g-priors and continuous shrinkage priors. We propose global local g priors that combines strength of (mixture of) g-priors (Liang et al., 2008) with global-local shrinkage priors (Bhadra et al., 2016; Carvalho et al., 2010) allowing for differential shrinkage across coefficients, accounting for prior correlation structure while simultaneously performing variable selection. Additionally, we propose Dirichlet Process (DP) block-g priors that uses Bayesian non-parameterics to cluster shrinkage parameters and in-turn coefficients of similar magnitude and avoids conditional Lindley paradox (CLP) unlike classical g-priors. We compare performances of our proposed priors over real and simulated datasets and show superior performance in terms of model selection, parameter estimation and prediction.

In Chapter 6, we discuss the complete work and propose areas of future research.

Chapter 2

COMPARING METHODS FOR STATISTICAL INFERENCE WITH MODEL UNCERTAINTY

2.1 Introduction

Statistical analysis is often carried out using probability models for the data at hand. In this context, five of the most important statistical tasks are parameter estimation, interval estimation, inference about model parameters, point prediction, and producing prediction intervals. These tasks often have to be carried out in the context of model uncertainty, where several different statistical models are plausible.

In the past thirty years, many satisfactory methods have been proposed. While several previous comparisons of existing methods have been carried out (see for e.g., Fernandez et al., 2001, Eicher et al., 2011, Liang et al., 2008, Celeux et al., 2012, Deckers and Hanck, 2014, Bhadra et al., 2019 and Forte et al., 2018), they have tended to be based on a narrow range of methods, to be based on simulation experiments whose connection to empirical data is less clear, and to base comparisons on a subset of the statistical tasks of interest. In this chapter, we aim to fill this gap by carrying out a broad study that simultaneously compares a wide range of statistical methods for the five statistical tasks described above and for computational efficiency over *real-like* datasets encountered in practical data analysis.

2.2 Methods

We carried out an extensive set of simulation studies based closely on real datasets that span a range of situations encountered in practical data analysis. This is in contrast with many simulation studies in the statistical literature whose design is determined by the investigators without direct reference to data. The simulation design, the metrics and the underlying

datasets are described below.

Figure 2.1 shows the sample size and the number of candidate variables for the different datasets. These include classic statistical situations where the sample size is larger than the number of variables, high-dimensional situations where the number of variables exceeds the sample size, and intermediate situations where the two are of the same order of magnitude.

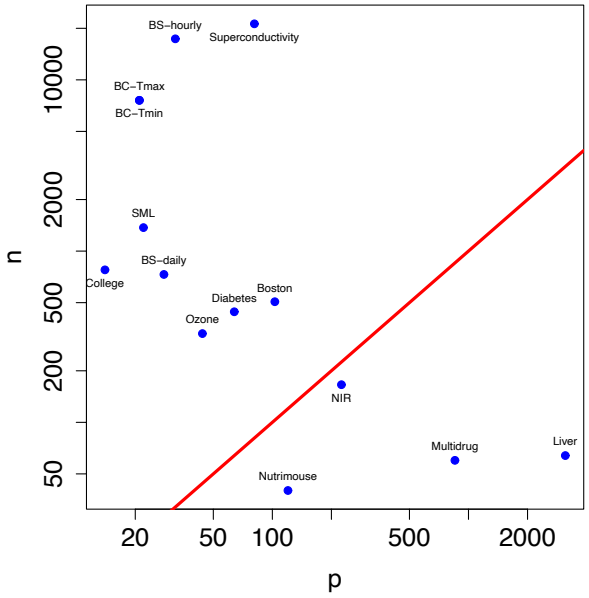


Figure 2.1: Sample size n versus the number of candidate variables p for the 14 datasets on which our simulation studies are based. The $n = p$ line is shown in red.

2.2.1 Statistical methods for comparison

The 21 methods we compare are listed in Table 2.1, along with references, the R package used, and the function call used. All the g-prior methods implemented using the `BAS` package, and the NLP methods implemented using the `mombf` package, use the Beta-Binomial$(1, 1)$ prior as the default model space prior. For high-dimensional datasets with $p > n$, a truncated Beta-Binomial$(1, 1)$ prior is used as the model space prior; this assigns probability zero to

any model with size greater than $n-2$. The BICREG-SIS method assumes a uniform prior over the model space. For methods implemented using the `BAS` package, a combination of the Metropolis-Hastings algorithm, as in the MC^3 algorithm (Raftery et al., 1997), with a random swap between a currently included and a currently excluded variable, is used for model space exploration.

Table 2.1: Variable selection methods compared in this study

Method	Authors	Implementation (R package-Version)	function
$g=\sqrt{n}$	Fernandez et al. (2001)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="g-prior", alpha=sqrt(n))`
Hyper-g	Liang et al. (2008)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="hyper-g")`
EB-local	Hansen and Yu (2003)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="EB-local")`
JZS	Zellner and Siow (1980)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="JZS")`
Horseshoe	Carvalho et al. (2010)	`horseshoe- V0.2.0`(van der Pas et al., 2019)	`horseshoe()`
UIP	Kass and Wasserman (1995)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="g-prior", alpha=n)`
EB-global	Clyde and George (2000); George and Foster (2000)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="EB-global")`
Benchmark	Fernandez et al. (2001)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="g-prior", alpha=max(n,`p^2`))`
NLP	Rossell and Telesca (2017); Johnson and Rossell (2012a)	`mombf-V2.2.9`(Rossell et al., 2019)	`modelSelection()`
LASSO*	Tibshirani (1996)	`glmnet-V3.0.2`(Friedman et al., 2010)	`cv.glmnet()`
SCAD	Fan and Li (2001)	`ncvreg-V3.11.2`(Breheny and Huang, 2011)	`cv.ncvreg(..., penalty="SCAD")`
BIC-BAS	George and Foster (2000)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="BIC")`
BICREG-SIS	Raftery (1995); Fan and Lv (2008)	`BMA - V3.18.12` (Raftery et al., 2020)	`bicreg()`
Spike Slab	George and McCulloch (1997)	`BoomSpikeSlab - V1.2.3`(Scott, 2020)	`lm.spike()`
Elastic Net	Zou and Hastie (2005)	`glmnet-V3.0.2`(Friedman et al., 2010)	`cv.glmnet( , alpha)`
MCP	Zhang (2010)	`ncvreg-V3.11.2`(Breheny and Huang, 2011)	`cv.ncvreg(..., penalty="MCP")`
SS Lasso	Ročková and George (2018)	`SSLASSO-V1.2.2`(Rockova and George, 2018)	`SSLASSO()`
EMVS	Ročková and George (2014)	`EMVS-V1.1`(Rockova and Moran, 2019)	EMVS()
AIC	George and Foster (2000)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="AIC")`
$g=1$	van Zwet (2019)	`BAS - V1.5.5`(Clyde, 2020)	`bas.lm(..., prior="g-prior", alpha=1)`

*LASSO-1se has same reference as LASSO

2.2.2 Datasets

We carried out 14 simulation studies, each one based on a different publicly available real dataset, from a variety of fields including social sciences, healthcare, genome sciences, physical sciences, chemistry and engineering (Table 2.2). We selected several of our datasets by filtering all the datasets in the UCI machine learning repository as follows. We filtered datasets with default task as regression, attribute type as numerical, data type as multivariate/univariate and number of attributes between 10 and 100. We further restricted our

attention to datasets with $p > 20$ and $n < 25,000$. This reduced our list of UCI datasets to four: the Bias Correction, Bike sharing, SML and superconductivity datasets. Note that the Bias correction and Bike sharing datasets each have two versions based on choice of outcome and frequency.

We also included several datasets that have been used as examples in the literature. We included the College dataset (James et al., 2013) as an example dataset where full enumeration of models is feasible. We included the Diabetes (Efron et al., 2004) and Ozone (Liang et al., 2008; Miller, 2002) datasets, and the Boston Housing dataset with squares and interaction terms between its covariates. Finally, we included four high-dimensional datasets from chemometrics and genomics from the `mixOmics` (Rohart et al., 2017) and `chemometrics` R packages (Filzmoser and Varmuza, 2017). For all the datasets, the continuous predictors were standardized to have mean zero and variance 1, and the response variable was centered to have mean zero. The 14 datasets used in the simulation study are listed in Table 2.2. Details of dataset pre-processing are given in the Appendix A.1.

2.2.3 Determining the generating model for the simulation study

For our simulation study, we require a data generating model based on each of our real datasets. For datasets for which $p < 30$, we performed all subsets regression using the `leaps` package in R (Lumley, 2020) and selected the largest model with all variables significant at 0.05 level. For datasets with $p > 30$, all subsets regression can be computationally intensive, and so we performed iterative sure independence screening (ISIS) (Fan and Lv, 2008) to reduce the number of variables. If the filtered list contained more than 30 variables, we further selected the top 30 variables with the highest R^2 values under univariate regression. We then applied all subsets regression to the filtered list of covariates with the above criteria to find the data generating model for our simulation study.

Consider the Boston Housing dataset ($n = 506$, $p = 103$) as an example. This includes 14 geographic housing variables, plus interactions and squares for each continuous variable, leading to 103 possible predictors. All subsets regression is not computationally feasible,

Table 2.2: Datasets used in the study

Dataset Name	Sample size (N)	Covariates (p)	Source
College	777	14	`ISLR` (James et al., 2017)
Bias Correction-Tmax	7590	21	UCI ML repository
Bias Correction-Tmin	7590	21	UCI ML repository
SML2010	1373	22	UCI ML repository
Bike sharing-daily	731	28	UCI ML repository
Bike sharing-hourly	17379	32	UCI ML repository
Superconductivity	21263	81	UCI ML repository
Diabetes	442	64	`spikeslab` (Ishwaran et al., 2013)
Ozone	330	44	`gss`(Gu, 2014)
Boston housing	506	103	`mlbench` (Newman et al., 1998)
NIR	166	225	`chemometrics` (Filzmoser and Varmuza, 2017)
Nutrimouse	40	120	`mixOmics` (Rohart et al., 2017)
Multidrug	60	853	`mixOmics` (Rohart et al., 2017)
Liver toxicity	64	3116	`mixOmics` (Rohart et al., 2017)

so instead we used ISIS to get a filtered list of 81 variables. We then performed univariate regressions for each of the filtered variable to select the top 30 variables with the highest R^2 values. Finally, we performed all subsets regression using the screened variables to get our data generating model with 23 variables and an R^2 of 0.86.

2.2.4 Simulation design

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

To evaluate the predictive performance of the methods, we divided each of the simulated datasets into 100 random 75%-25% train-test splits. We trained the methods on the training data and used the test data to assess the predictive performance using the metrics described below. We calculated point predictions for each of the methods and posterior predictive intervals for Bayesian techniques that allow for uncertainty quantification.

We used the following metrics to compare the methods:

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

$$RMSE = \sqrt{\frac{1}{p}\sum_{i=1}^{p}(\beta_{i,DG} - \hat{\beta}_i)^2}, \tag{2.1}$$

  where $\beta_{i,DG}, i = 1, \ldots, p$ denote the coefficients in the data generating model, and $\hat{\beta}_i, i = 1, \ldots, p$ denote the posterior means of the coefficients for the Bayesian techniques and the estimated optimal coefficients for penalized likelihood based approaches.

- **IntEst**: The Interval Score (IS) (Gneiting and Raftery, 2007) provides a balance between the narrowness of the intervals and the accuracy of the coverage. It is a sum of two components: the first rewards narrow intervals and the second rewards accurate coverage. For a variable z, the IS is given by

$$\begin{aligned} MIS_\alpha(l, u, z) = (u - l) &+ \frac{2}{\alpha}(l - z)\mathbb{1}\{z < l\} \\ &+ \frac{2}{\alpha}(z - u)\mathbb{1}\{u < z\}, \end{aligned} \tag{2.2}$$

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

- **Inference**: To compare the performance of the methods for identifying the appropriate variables, we calculated the area under the precision recall curve (AUPRC) for each of

the methods. This gives an overall assessment of model selection quality and does not require a threshold to be chosen for the posterior inclusion probability of a covariate.

For penalized likelihood based approaches, the AUPRC was obtained by varying the cross-validation parameter λ from close to 0 (no penaliaation) to λ_{max}, defined as the smallest value of λ for which none of the variables is included in the model (Vignes et al., 2011). For the horseshoe, the AUPRC was obtained by varying the credible set levels leading to different number of variables being selected by the method. We report Inference with (1−AUPRC) as our metric, and a lower value is better.

- **Prediction**: To assess the accuracy of point prediction, we calculated R^2_{test} as follows:

$$R^2_{test} = 1 - \frac{\sum_{i \in test}(y_i - \hat{y}_i)^2}{\sum_{i \in test}(y_i - \bar{y}_{train})^2}, \tag{2.3}$$

  where $\{y_i : i \in test\}$ denotes the response variable of the test set, $\hat{y}_i$ denotes the corresponding predictions, and $\bar{y}_{train}$ denotes the mean of the response variable in the training set. Note that this quantity can be less than zero, if the predictions perform worse than the baseline $\bar{y}_{train}$.

- **IntPred**: To assess the quality of the prediction intervals, we calculated the interval score using (2.2) for each of the test set observations. Here, l and u represent the lower and upper bounds of the $(1-\alpha) \times 100\%$ posterior predictive interval for the observation. We calculated the mean interval score (MIS), averaging IS over test set for each of the train-test splits. A lower MIS corresponds to a better interval forecast.

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

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

2.3 Results

The results are shown in Table 2.3. Performance metrics are shown for all 21 methods for each of point estimation, interval estimation, inference, prediction and interval prediction. All metrics are relative to the score for the JZS method, taken as the reference, and averaged across datasets. Detailed results of performance metrics for the simulation studies based on each of the 14 datasets can be found in Appendix A.3. The "Score" column shows the average of the five metrics for each method. For seven of the methods, interval estimation and interval prediction metrics were not available as the methods did not provide uncertainty assessments, and so we calculated the "PartScore," which is the average of the three remaining metrics. In all cases, a lower score is better.

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

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

Table 2.3: Performance of 21 methods for inference in linear regression under model uncertainty: "PointEst" is the RMSE for point estimation, "IntEst" is the Mean Interval Score (MIS) for interval estimation, "Inference" is 1− the area under the precision-recall curve (AUPRC), "Prediction" is the RMSE for point prediction, while "IntPred" is the MIS for interval prediction. "N vars" is the average number of variables used for the task. All metrics are standardized to equal 1 for the JZS method. See the text, and Materials and Methods section for more information about the ranking and coloring of this table, and the definitions of the methods and metrics. Note that BICREG denotes the BICREG-SIS method, in which Sure Independence Screening is used first to reduce the number of variables to 30.

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

LASSO was the top penalized likelihood method, doing particularly well for point prediction, as did the Elastic Net — comparable to the top Bayesian methods for this task, although not for the other tasks. However, they both selected far more variables on average than the Bayesian methods — twice as many or more in most cases without any noticeable increase in predictive performance. Plots of prediction accuracy, given by R^2, versus average model size, denoted by $\hat{p}$, for all datasets are available in Appendix A.4.

A surprise was that two of the top three methods were efficient computationally even though they were Bayesian, comparable to LASSO in spite of the reputation of Bayesian methods for being slow. This is partly because we used a default of 10,000 MCMC iterations, which is far fewer than the default in the `BAS` R package used to implement these methods (Clyde, 2020). This clearly gave adequate performance. Performance might be improved slightly with more iterations, but at the cost of computational efficiency. The hyper-g method is substantially slower, which seems to be due to its greater complexity, but this may be a worthwhile tradeoff given its good performance. Several of the other methods were extremely slow.

One needs to be cautious in interpreting the CPU time results, as they reflect the coding efficiency of the implementations as well as the intrinsic computational efficiency of the methods. For most methods we used the developers' packages with default settings, and these could clearly often be sped up.

One question is whether inferences are sensitive to the choice of model selection/model averaging method. To provide a partial answer, we compared the results for our 14 datasets for the top three methods identified by our study. Scatterplots of parameter estimates and posterior inclusion probabilities are shown in Appendix A.2 for all 14 datasets. We found that the (model-averaged) parameter estimates were very similar between the three methods for the ten tall datasets (with $p < n$), and less similar but still highly correlated for the four wide datasets (with $p > n$). The posterior inclusion probabilities were similar between methods for the tall datasets, but less so for the wide datasets. The $g = \sqrt{n}$ method tended to favor models with slightly more variables than the hyper-g and EB-local methods.

2.3.1 Comparison of BMA with Bayesian model selection:

An alternative to BMA is Bayesian model selection (BMS), in which just one model is selected. When several candidate models are available, a researcher can choose to select one model or perform model averaging. Bayesian model selection refers to selection of one model from a list of candidate models based on the data (Kass and Raftery, 1995; Wasserman, 2000). One choice for BMS is to select the model with the highest posterior probability in model search, also known as the *Maximum a posteriori* (MAP) model. We compared the performance of BMA and BMS for the top three methods identified in the previous section: $g = \sqrt{n}$, Hyper-g and EB-local.

We used the same performance metrics as before. As before, all metrics are relative to BMA under the JZS prior, except for computation time, for which LASSO was used as the reference. The results are shown in Table 2.4. The BMS versions of the top three methods performed worse than the corresponding BMA versions in terms of all the metrics.

Table 2.4: Comparison of BMA and Bayesian model selection (BMS) for top three methods

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

2.4 Discussion

In this chapter, we compared 21 popular model selection methods via an extensive simulation study based on a wide range of real datasets. We found that three adaptive Bayesian model

averaging methods performed best across all the statistical tasks, and that two of these were also among the most computationally efficient. As mentioned earlier, several previous comparisons of existing methods have been carried out. We now discuss how does our study differ from existing comparisons and briefly comment on theoretical properties of our recommended methods.

Fernandez et al. (2001) did a simulation study based on a non-empirical design (Raftery et al., 1997), and compared methods based on their ability to recover the true underlying model as the MAP model, and assessing predictive performance using log-predictive scores. Hence, their comparisons were based on only two statistical tasks, namely Inference and point prediction. They considered only BMA methods. They found a UIP-based method with $g = n$ to work best when $n < p^2$, and an RIC-based method (Foster et al., 1994) with $g = p^2$ to work best otherwise, but they pointed out that the RIC-based method is not model-selection consistent. We have included the resulting combined method in our study under the name "Benchmark prior." The only other method in their study that is also in ours is the $g = \sqrt{n}$ method, which they found to be outperformed by BIC, in contrast with our findings here.

Eicher et al. (2011) considered the same BMA methods as (Fernandez et al., 2001), considered prediction for a well-known economic growth dataset and for several simulations with the same non-empirically-based design, and again found BIC and UIP to do best. Our results here are based on a wider and more empirically-based set of simulations, which may help explain the different results. Liang et al. (2008) also carried out a non-empirically-based simulation study using the design of Cui and George (Cui and George, 2008), and found the hyper-g prior to be competitive with other BMA methods in terms of parameter estimation, including several that we have considered here (but they did not include the $g = \sqrt{n}$ method).

Celeux et al. (2012) carried out another non-empirical simulation study to assess quality of inference, and assessed point prediction using two small real datasets; they assessed 15 methods, of which 7 were in common with ours. They focused on the situation where p is

close to n. Like us, they found Bayesian methods to outperform non-Bayesian ones.

Deckers and Hanck (2014) compared a subset of the Bayesian techniques discussed in our study, specifically the UIP and RIC, or Benchmark prior, and LASSO with multiple testing procedures (MTPs) controlling False discovery rate (FDR). They focused their comparison on the model inference performance of these procedures using 'size-adjusted' power i.e., comparison of power (number of correctly selected variables) in situations where procedures infer similar size models. In their comparison over a non-empirical simulation study, they found that BMA was slightly more powerful given the size than the MTPs and LASSO. Their comparison did not focus on other statistical inference tasks of prediction and estimation.

Bhadra and Polson (Bhadra et al., 2019) compared variants of the horseshoe, LASSO and SCAD in terms of their performance in variable selection, using the non-empirically-based simulation design of Zhao and Yu (Zhao and Yu, 2006). They found the horseshoe to do best, then SCAD, both substantially dominating LASSO. This agrees with our ranking in terms of inference from Table 2.3, but we found LASSO to overtake SCAD when other statistical tasks were also taken into account. Forte et al. (2018) compared different BMA software packages in terms of computational performance, and found the `BAS` package (Clyde, 2020) to dominate others in terms of speed, as we also found. However, they also warned against the use of the "MCMC+BAS" method within BAS, which they reported does not provide reliable estimates of the inclusion probabilities, and instead recommended the method "MCMC." They also commented on the very high memory demands of BAS. Here we used their recommended method "MCMC" and found it to work well.

One can also evaluate methods in terms of theoretical properties. One is model-selection consistency (Fernandez et al., 2001), which says that if the true model is among the candidate models considered, the method will select it with probability approaching 1 as the sample size increases indefinitely. All three of our top-ranked methods satisfy this unless the true model is the null model with no predictors (Fernandez et al., 2001; Liang et al., 2008). However, LASSO does not have this property (Zhao and Yu, 2006).

A second property is whether the method is subject to Bartlett's paradox (Bartlett, 1957),

according to which, if the data are held fixed and the prior variance increases without bound, then BMA will select the null model with probability tending to 1, regardless of the data. None of our top three methods is subject to this, as they do not allow the prior variance to increase without bound.

A third consideration is whether the method is subject to the so-called "information paradox" (Liang et al., 2008). This arises when, for fixed n and p, the data provide maximal support for a larger model, for example when $R^2 \longrightarrow 1$. One could argue that, in this case, the Bayes factor for this model against any submodel should tend to infinity with the sample size. However, g-priors with fixed g do not have this property, and indeed in that case the Bayes factor has a finite (although usually very high) upper bound. It has been argued that this is undesirable, making them subject to the information paradox. The hyper-g and EB-local methods are not subject to this, but the $g = \sqrt{n}$ prior is, which could be argued to be a disadvantage of the latter.

However, one might question the relevance of the "information paradox" to the choice of method (Zellner, 2008). If $R^2 = 1$ when n is small, this will often be because of the inherent discreteness of most data, which are rarely measured or recorded with full precision, but rather to within a certain measurement tolerance (for example a certain number of significant digits). In that case, the fact that the Bayes factor for an additional variable is bounded above could be viewed as an advantage. The linear regression model models the observed response variable as a continuous variable, thus measured with infinite precision. This is actually an approximation, which is usually inconsequential, but is relevant for assessing the relevance of the information paradox. If the discreteness of observed data were accounted for in the model, the "information paradox" would never arise.

For example, the famous data on heights of fathers and sons in England (Pearson and Lee, 1903; Friendly, 2021) are reported to the nearest inch. If one took a sample of size 3 from these data, say (father, son) = (62.5, 64.5), (67.5, 69.5), (70.5, 72.5), and regressed son's height on father's height, one would find that $R^2 = 1$ and the standard F statistic is infinite. In this case, one would not want the Bayes factor for the effect of father's height to

be infinite, but it is infinite for the hyper-g and EB-local priors, while for the $g = \sqrt{n}$ prior it is 1.65. The latter represents positive but weak evidence for an effect, which seems more reasonable than an infinite Bayes factor corresponding to absolute certainty based on three data points.

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

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

Given the good performance of the $g = \sqrt{n}$ prior of (Fernandez et al., 2001), it is of interest how it relates to the popular BIC criterion, which corresponds approximately to $g = n$, and performed less well in our experiments. Let us consider just two models: the null model and a regression model of interest, with d variables. Then if B is the Bayes factor for the regression model against the null model, the exact result is

$$-2\log B = (n-1)\log\{1+\sqrt{n}(1-R^2)\} - (n-1-d)\log\big(1+\sqrt{n}\big).$$

The BIC approximation is

$$-2\log B \approx n\log\big(1-R^2\big) + d\log(n).$$

A similar approximation with the $g = \sqrt{n}$ prior is

$$-2\log B \approx n\log\big(1-R^2\big) + d(\log(n)/2) + \sqrt{n}(1-R^2)R^2.$$

The last term does not involve the number of parameters directly, and so the complexity penalty in the Bayes factor with the $g = \sqrt{n}$ prior is effectively half that in the BIC.

We have focused on one specific type of model uncertainty in one statistical setting, namely uncertainty about which variables to include in a linear regression model. This has been much studied and arises frequently in science, as well as being a canonical example for other statistical models. However, there are many other statistical settings in which the same issue arises, and it would be of interest to carry out similar comparative studies. We expect that our main conclusion, that Bayesian model averaging with an adaptive parameter prior performs well, would carry over to other settings.

Under the BMA framework, while several comparisons exist for default parameter priors, similar comparison for model space priors remain limited. In the next chapter, we provide a brief overview of existing model space priors and evaluate the effect on model priors on the five statistical tasks using a similar simulation study.

Chapter 3

EFFECT OF MODEL SPACE PRIORS ON STATISTICAL INFERENCE WITH MODEL UNCERTAINTY

3.1 Introduction

Chapter 1 discussed the Bayesian framework for model selection that provides a straightforward way to account for model uncertainty by treating the model as a parameter itself, using Bayesian model averaging (BMA). BMA requires the specification of a model space prior and a parameter space prior. However, subjective elicitation of these priors is often not feasible, particularly when p is large.

Several default parameter prior choices have been proposed in the last thirty years (see Chapter 2 for an overview) and several other comparisons of these methods have been carried out (Celeux et al., 2012; Deckers and Hanck, 2014; Eicher et al., 2011; Fernandez et al., 2001; Forte et al., 2018; Liang et al., 2008). However, similar comparisons of default model space priors remain limited. In this chapter, our focus is on understanding the effect of model space priors on the statistical tasks of parameter estimation, interval estimation, statistical inference, point and interval prediction.

To accomplish this, we compare combinations of three default parameter priors with eight choices of model space priors that have been advocated in the literature. These model space priors correspond to different flavors of Bayesian inference with: (i) fixed hyper-parameter choices, (ii) with Bayesian treatment of hyper-parameters, and (iii) estimation of hyper-parameters in an empirical Bayes (EB) manner. Similar to Chapter 2, the comparison is carried out over an extensive simulation study closely based on 14 real datasets that span a wide range of practical data analysis situations.

3.2 Background

3.2.1 Choice of model space priors

Model space priors require specification of the prior probabilities of all models $\mathcal{M}_{\boldsymbol{\gamma}}$, indexed by the binary variable inclusion vector $\boldsymbol{\gamma}$. A common approach is to consider the inclusion of each variable as an independent and exchangeable Bernoulli random variable with a common prior probability of inclusion θ, i.e.

$$p(\mathcal{M}_{\boldsymbol{\gamma}}|\theta) = \prod_{i=1}^{p} \theta^{\gamma_i}(1-\theta)^{1-\gamma_i} = \theta^{p_{\boldsymbol{\gamma}}}(1-\theta)^{p-p_{\boldsymbol{\gamma}}}, \tag{3.1}$$

where θ is the prior expected fraction of the β_j's that are not zero and $p_{\boldsymbol{\gamma}} = \sum_{i=1}^{p} \gamma_i$ is the total number of covariates included in the model $\mathcal{M}_{\boldsymbol{\gamma}}$. For a fixed value of θ, the above prior induces a binomial prior for model size $S = \sum_{i=1}^{p} \gamma_i$ i.e. $S \sim Bin(p, \theta)$, with prior mean $p\theta$ and prior variance $p\theta(1-\theta)$.

In the absence of prior information, a common choice is to set $\theta = 0.5$. This induces a uniform prior over all models with $p(\mathcal{M}_{\boldsymbol{\gamma}}) = 2^{-p}$ for all $\boldsymbol{\gamma} \in \{0,1\}^p$, where p is the total number of covariates considered. The expected prior model size under the uniform model prior is $p/2$. However, choosing $\theta = 0.5$ does not provide any multiplicity control (George and Foster, 2000).

Another way to specify θ is by using the researcher's prior belief about expected model size $E[S]$. Sala-I-Martin et al. (2004) (hereafter SDM) recommended a prior expected model size of 7 based on their growth regression analysis. Similar priors for expected model size have also been proposed elsewhere (Sala-i Martin, 1997; Levine and Renelt, 1992). Hence, we can define an SDM version of the above prior with $\theta_{SDM} = 7/p$.

Any fixed choice of θ can lead to rather informative priors on model size $p_{\boldsymbol{\gamma}}$. One way to address this issue is by estimating θ from the data using an empirical Bayes (EB) approach (George and Foster, 2000). The EB approach involves maximizing the marginal likelihood

of the data given θ:

$$\hat{\theta}_{EB} = \arg\max_{\theta \in [0,1]} \sum_{\gamma} P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}) P(\mathcal{M}_{\boldsymbol{\gamma}}|\theta). \tag{3.2}$$

Maximization of (3.2) can be computationally challenging. This is especially true when p is large, where marginal likelihood evaluation for all models is not feasible. In those cases, the sum can be approximated based on the models explored by MCMC. To optimize the marginal likelihood in (3.2), we implement Algorithm 1, iterating between fitting the BMA approach to find likely models given θ and solving (3.2) using the fitted models to find a new θ:

Algorithm 1 EB optimisation algorithm for $\hat{\theta}_{EB}$

probs $\leftarrow 0.5$

Fit `mod`$\leftarrow$ `bas.lm(..., model.prior=bernoulli(probs))`

Initialize $\theta^{(0)} \leftarrow \frac{\hat{p}}{p}$ ▷ $\hat{p}$ is avg. posterior model size

$i \leftarrow 1$

repeat

probs $\leftarrow \theta^{(i-1)}$

`mod`$\leftarrow$ `bas.lm(..., model.prior=bernoulli(probs))`

Optimise (3.2) over models explored in `mod` to find $\theta^{(i)}$

$i \leftarrow i + 1$

until convergence

An alternative way to reduce the sensitivity of the posterior distribution to prior assumptions is to use hierarchical modeling and specify a weak hyper-prior for θ. One choice of such a hyper-prior is a Beta distribution, $\theta \sim \text{Beta}(a, b)$. Marginalizing out θ in (3.1), gives

$$\begin{aligned} P(\mathcal{M}_{\boldsymbol{\gamma}}|a, b) &= \int_0^1 p(\mathcal{M}_{\boldsymbol{\gamma}}|\theta) p(\theta) d\theta \\ &= \frac{B(p_{\boldsymbol{\gamma}} + a, p - p_{\boldsymbol{\gamma}} + b)}{B(a, b)}, \end{aligned} \tag{3.3}$$

where $B(a', b') = \frac{\Gamma(a')\Gamma(b')}{\Gamma(a'+b')}$ is the Beta function. It thus induces a Beta-Binomial(a, b) prior on the model size S with probability mass function

$$P_S(s) = \binom{p}{s} \frac{B(s+a, p-s+b)}{B(a,b)}.$$

Under a uniform prior on θ, i.e. when $a = b = 1$, (3.3) simplifies to $p(\mathcal{M}_\gamma) = \frac{1}{p+1}\binom{p}{p_\gamma}^{-1}$. This is a combination of a uniform prior over model size with a uniform prior over the models of same size given the model size.

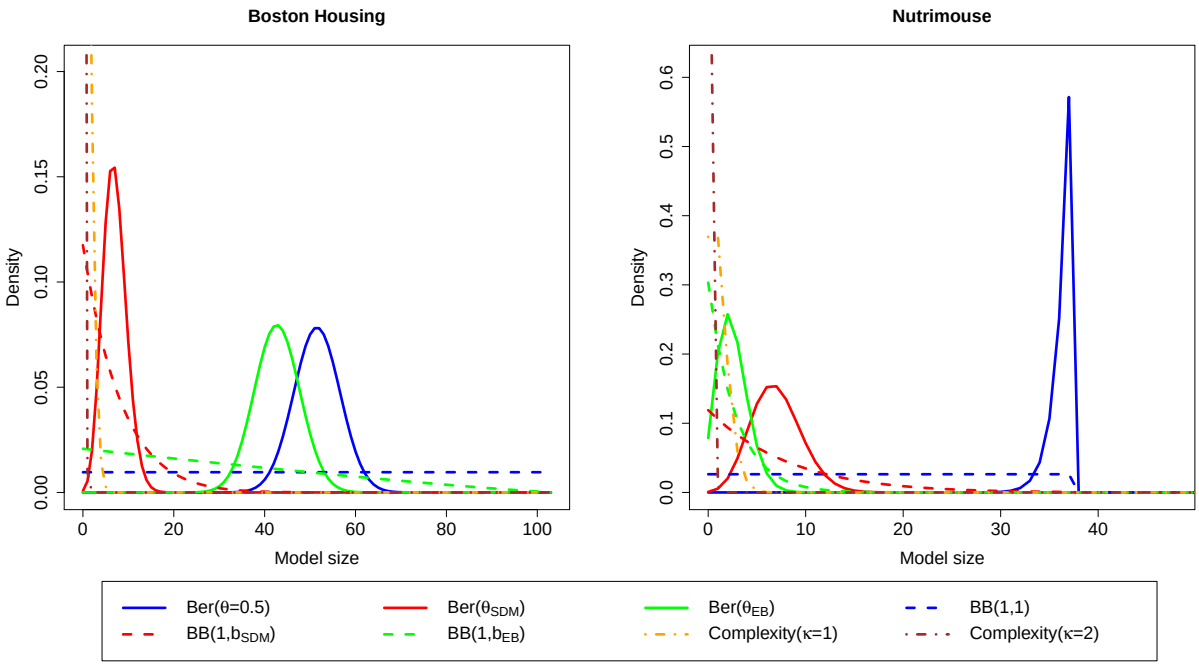


Figure 3.1: Prior model size distribution for the Boston Housing and Nutrimouse datasets

Under a Beta-Binomial (BB) prior, the prior expected model size is $E[S] = \frac{a}{a+b}p$. Similarly to a Bernoulli prior, we can elicit the prior in terms of the prior expected model size $E[S]$. To facilitate prior elicitation, we fix $a = 1$. We can then define an SDM version of the BB prior (BB-SDM) with an expected prior model size, such as $E[S] = 7$, by setting $b_{SDM} = \frac{p}{E[S]} - 1$. Note that SDM themselves (Sala-I-Martin et al., 2004) did not use a Beta-Binomial prior on models, but only a Bernouilli prior.

Alternatively, we can use an EB approach to learn b from the data. This can be done by maximizing the marginal likelihood given b, namely

$$\hat{b}_{EB} = \underset{b\in(0,\infty)}{\arg\max} \sum_{\boldsymbol{\gamma}} P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}})P(\mathcal{M}_{\boldsymbol{\gamma}}|a=1,b). \tag{3.4}$$

We find the optimal value, $\hat{b}_{EB}$, using Algorithm 2.

Algorithm 2 EB optimisation algorithm for $\hat{b}_{EB}$

$b \leftarrow 1$

Fit `mod`$\leftarrow$ `bas.lm(..., model.prior=beta.binomial(1,b))`

Initialize $b^{(0)} \leftarrow \frac{p}{\hat{p}} - 1$ ▷ $\hat{p}$ is avg. posterior model size

$i \leftarrow 1$

repeat

$b \leftarrow b^{(i-1)}$

`mod`$\leftarrow$ `bas.lm(..., model.prior=beta.binomial(1,b))`

Optimise (3.4) over models explored in `mod` to find $b^{(i)}$

$i \leftarrow i + 1$

until convergence

For Zellner's g-prior, we require the model size to be no larger than the number of regression coefficients that can be identified from the data, so that $p_{\boldsymbol{\gamma}} < n - 2$. Thus, for higher dimensional datasets ($p > n$), we require that $P(\mathcal{M}_{\boldsymbol{\gamma}}) = 0$ for all models with model size greater than $n - 2$. Hence, we use truncated versions of the priors defined in (3.1) and (3.3), namely

$$p(\mathcal{M}_{\boldsymbol{\gamma}}|\theta) \propto \theta^{p_{\boldsymbol{\gamma}}}(1-\theta)^{p-p_{\boldsymbol{\gamma}}}\mathbb{1}\{p_{\boldsymbol{\gamma}} < n-2\}, \tag{3.5}$$

$$P(\mathcal{M}_{\boldsymbol{\gamma}}|a,b) \propto \frac{B(p_{\boldsymbol{\gamma}}+a, p-p_{\boldsymbol{\gamma}}+b)}{B(a,b)}\mathbb{1}\{p_{\boldsymbol{\gamma}} < n-2\}. \tag{3.6}$$

Castillo et al. (2015) introduced complexity priors, also known in the literature as diffusing (Narisetty and He, 2014) or power priors (Clyde, 2020). Here, the marginal probability of

inclusion of any variable decays at the rate $p^{-\kappa}$ for some $\kappa > 0$, where p is the total number of possible covariates. This specifies a vanishing prior probability of large models and leads to a faster rate of rejection of spurious parameters, at the cost of slower rates of detection of active parameters (Rossell et al., 2021). Similar priors have also been used elsewhere (Yang et al., 2016; Rossell, 2021).

The complexity prior is defined as

$$p(\mathcal{M}_\gamma) \propto p^{-\kappa|\gamma|}\mathbb{1}\{|\gamma| \leq s_0\},$$

where s_0 is a pre-specified integer specifying the maximum number of important covariates and $|\gamma|$ is the model size. In the absence of external information, we set $s_0 = \min\{n-2, p\}$. This prior is implemented in the BAS package as `tr.power.prior(kappa,trunc)`. We implement Complexity priors with $\kappa = 1$ (Rossell and Rubio, 2018; Rossell et al., 2021), and with $\kappa = 2$ which is the default choice in the `BAS` package.

3.2.2 Model space priors - a graphical illustration

To illustrate the effect of different model space priors, we use two datasets from our analysis: Boston Housing ($n = 506, p = 103$) and Nutrimouse ($n = 40, p = 120$) (Figure 3.1). The solid lines show the independent Bernoulli prior from (3.1), while the dashed lines represent the Beta-Binomial prior in (3.3) and the dash-dotted lines illustrate Complexity priors. For the Nutrimouse dataset ($p > n$), we use the truncated versions as discussed above. The colors group different flavors of methods: (i) Uniform versions with $\theta = 0.5$ or $b = 1$ (blue), (ii) SDM versions with expected prior model size 7 (red), (iii) EB versions with θ or b learned from the data (green), (iv) Complexity prior with $\kappa = 1$ (orange), and (v) Complexity prior with $\kappa = 2$ (brown).

The Bernoulli model space priors are very concentrated around their mean, $p\theta$. The complexity priors are concentrated around smaller model sizes with a mode at 0. The BB priors, on the other hand, are more diffuse, implying more prior uncertainty about model size. For the Nutrimouse dataset, all the model space priors assign zero probability to any

Table 3.1: Summary of prior moments of model size S under different model space priors and `BAS` code to implement them.

Model prior	E[S]	Var(S)	`BAS code`
Bernoulli(θ)	$p\theta$	$p\theta(1-\theta)$	`bas.lm(..., model.prior=bernoulli(probs))`
Beta-Binomial(a, b)	$\frac{pa}{a+b}$	$\frac{pab(a+b+p)}{(a+b)^2(a+b+1)}$	`bas.lm(..., model.prior=beta.binomial(alpha,beta))`
Complexity(κ)	–	–	`bas.lm(..., model.prior=tr.power.prior(kappa,trunc))`

model with size greater than $(n-2) = 38$. Among the Bernoulli versions, $\theta = 0.5$ implies a prior mode around $\min\{p/2, n-2\}$ while θ_{SDM} has a prior model size of 7 (the same as the prior mean). The prior model size distribution induced by the Ber(θ_{EB}) prior adapts based on the data, with a prior mode between $\theta = 0.5$ and θ_{SDM} for the Boston Housing dataset, while having the lowest prior mode among the Bernoulli priors considered for the Nutrimouse dataset. The BB$(1, 1)$ prior corresponds to a uniform prior over model size. The BB$(1, \theta_{SDM})$ and BB$(1, \theta_{EB})$ priors both induce a model size distribution with prior mode at zero.

3.2.3 Choice of parameter priors

Based on an extensive simulation study in Chapter 2, we found that in comparing parameter priors for BMA, three adaptive Zellner's g-priors, given by (5.1), performed the best among many popular choices across the statistical tasks of parameter estimation, interval estimation, model inference, point prediction and interval prediction.

g-priors are popular because of their computational efficiency in evaluating marginal likelihoods and performing model search. It is also attractive because of its intuitive interpretation arising from analysis of a conceptual sample generated using the same design matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$ as in the data at hand. It is a special case of spike-and-lab family with the slab density given by the Normal density above and the spike being a point mass at zero.

In what follows we shall focus only on these three parameter prior choices, briefly reviewed below:

- $\boldsymbol{g = \sqrt{n}}$: First proposed by Fernandez et al. (2001), it corresponds to a prior sample size equal to $\sqrt{n}$ and has been found to work well in high dimensional settings (Young et al., 2014). As shown in Chapter 2, the complexity penalty for a model using this specification is effectively half that in the BIC.

- **EB-local**: An alternative to fixing g to a pre-specified value is to instead estimate it from the data in an empirical Bayes (EB) manner. The local EB approach estimates a different g for each model. Let $P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, g)$ denotes the marginal likelihood of the data under a g-prior. Then

$$\hat{g}_{\boldsymbol{\gamma}} = \arg\max_{g \geq 0} P(\boldsymbol{Y}|\mathcal{M}_{\boldsymbol{\gamma}}, g).$$

  For a linear model, Hansen and Yu (2003) showed that it reduces to $\hat{g}_{\boldsymbol{\gamma}} = \max\{F_{\boldsymbol{\gamma}} - 1, 0\}$ where $F_{\boldsymbol{\gamma}}$ is the F statistic for testing $\boldsymbol{\beta}_{\boldsymbol{\gamma}} = \mathbf{0}$.

- **Hyper-g**: A natural Bayesian way to account for uncertainty about the scale parameter g is to specify a hyper-prior for g and perform fully Bayesian inference. Liang et al. (2008) proposed the hyper-g prior

$$\pi(g) = \frac{a-2}{2}(1+g)^{-a/2},$$

  which is proper for $a > 2$. Liang et al. (2008) recommended $a = 3$ as a default choice for the hyper-g prior. One advantage of using a hyper-g prior is that the posterior distribution of g given the model $\mathcal{M}_{\boldsymbol{\gamma}}$ is available in closed form, simplifying Bayesian inference.

3.3 Numerical Comparison

Similar to Chapter 2, we investigate the performance of different model space priors and parameter prior combinations using an extensive simulation study based closely on real

datasets. We evaluate the effect of prior choices for the statistical tasks of parameter point and interval estimation, inference, point and interval prediction, and computation time using the metrics discussed in Chapter 2.

All the parameter and model space prior combinations were implemented using the `BAS` R package (Clyde, 2020) with skeleton code shown in Table 3.1. A combination of the MC^3 Metropolis-Hastings algorithm for sampling from the posterior distribution of models (Raftery et al., 1997), along with a random swap between a currently included and a currently excluded variable is used for model space exploration. This is implemented by setting the option `method="MCMC"` in the `bas.lm()` function. We used a default of 10,000 MCMC iterations for all methods.

For the EB methods, we used Algorithm 1 (or 2) to find the optimal θ_{EB} (or b_{EB}) before fitting a BAS model with the estimated hyperparameter value. For higher dimensional datasets ($p > n$), a truncated version of the Beta-Binomial prior (3.6) was implemented by setting the option `model.prior="tr.beta.binomial(alpha,beta,trunc=n-2)"` in BAS. Similarly, a truncated version of complexity prior is implemented in the BAS package. A truncated version of the Bernoulli prior (3.5) is not available in BAS. We implemented it by (i) implementing `bas.lm()` with `tr.beta.binomial(1,1,trunc=n-2)`, and then (ii) using importance sampling to calculate updated posterior model probabilities with weights proportional to the ratio of the prior model space densities in (3.5) and (3.6).

3.3.1 Datasets

Similar to Chapter 2, we based our analysis on 14 publicly available datasets listed in Table 2.2. For each dataset, continuous covariates are standardized to have zero mean and variance 1 and the response variable is centered to have zero mean.

3.3.2 Simulation Design

For each dataset, we selected a data generating model that closely approximates the real dataset using the procedure described in Sub-section 2.2.3.

We used the above data generating model and parametric bootstrapping to generate 100 bootstrapped datasets with the same design matrix X and different simulated response vectors $\boldsymbol{Y}$. Each of the resulting simulated datasets had the same design matrix and error distribution as the real dataset on it was based.

3.4 Results

The results are shown in Table 3.2. We used the combination of the g-prior with $g = \sqrt{n}$ as the parameter prior and the Beta-Binomial$(1, 1)$ model space prior as the reference. Note that the g-prior with $g = \sqrt{n}$ was found to the best parameter prior Chapter 2. Metrics for all other combinations were calculated relative to the reference metric, and averaged across datasets. Detailed results of performance metrics for the simulation studies based on each of the 14 datasets can be found in Appendix B.1. The "Score" column contains the average of the scores for PointEst, IntEst, Inference, Prediction and IntPred under each method. We used the Score column to rank the methods.

For each metric, we color the methods based on their performance relative to the reference metric. A method is colored green if it performed similarly or better than the reference method, yellow if it performed somewhat worse, and orange if it performed substantially worse.

For all choices of parameter prior, Beta-Binomial$(1, 1)$ was the top scoring model space prior. The three Beta-Binomial versions with $g = \sqrt{n}$ were the top three methods across statistical tasks. The Complexity priors with $\kappa = 1$ and $\kappa = 2$ were the worst performing model space priors. The uniform prior denoted by Ber$(\theta = 0.5)$ also performed less well than the Beta-Binomial priors. This ranking of methods was consistent across different performance metrics.

Most parameter and model prior combinations selected sparser models than the $g = \sqrt{n}$ and BB$(1, 1)$ prior combination, with the exception of methods involving the uniform model space prior. The complexity priors selected very sparse models compared to our baseline, which may be explained by the strong sparsity induced by the prior. This may also explain

the poor performance of the complexity priors across statistical tasks. Notably, the rankings of the prior combinations are similar for the different task. In particular, the rankings for prediction are consistent with those for point estimation and parameter inference, with a correlation of 0.77 between scores for point estimation and point prediction.

We also note that the EB model space priors tended to outperform the corresponding SDM model space priors when combined with the Hyper-g and EB-local parameter priors. However, the results with the EB model space priors took longer to computer on average because of the optimisation procedure. The hyper-g parameter priors are the slowest due to the integral calculations required in the posterior computation. In general, the Beta-Binomial priors performed better than the Bernoulli and complexity priors.

3.5 Discussion

We have compared BMA techniques with different choices of model space priors and parameter priors using an empirical study based closely on real datasets. We found that the Beta-Binomial$(1, 1)$ model space prior performed the best across various statistical tasks and choices of parameter priors. We found that the hierarchical model space prior with a hyper-prior on the prior inclusion probability θ was more diffuse and led to more efficient exploration of the model space. Fixed choices of θ led to worse performance across statistical tasks and were often quite concentrated. Complexity priors that induce high sparsity on model complexity performed worst among all the methods considered.

We are not the first to compare model space priors in the presence of model uncertainty. Past comparisons have either focused on a subset of the model priors discussed here, or evaluated BMA methods for only a subset of the statistical tasks considered here. In several cases, they also tended to use simulation designs that are at best loosely related to empirical data observed in practice.

Ley and Steel (2009) evaluated the effect of different model priors on model selection performance using three real economic growth regressions datasets. However, they used only two fixed choices of g-priors: the Unit Information prior (UIP) with $g = n$ (Kass

Table 3.2: Performance of different parameter prior and model space prior combinations for inference in linear regression under model uncertainty: "PointEst" is the RMSE for point estimation, "IntEst" is the Mean Interval Score (MIS) for interval estimation, "Inference" is the 1- area under the precision-recall curve (AUPRC), "Prediction" is the RMSE for point prediction, while "IntPred" is the MIS for interval prediction. "N vars" is the average number of variables used for the task. All metrics are standardized to equal 1 for the $g = \sqrt{n}$ with BB$(1, 1)$ prior on model space. For each column, lower value is better.

Rank	Parameter prior	$P(\mathcal{M}_\gamma)$	Score	PointEst	IntEst	Inference	Prediction	IntPred	N vars	CPU time
1	$g = \sqrt{n}$	BB$(1, 1)$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2	$g = \sqrt{n}$	BB$(1, b_{SDM})$	1.005	1.005	1.005	1.013	1.001	1.001	0.918	1.009
3	$g = \sqrt{n}$	BB$(1, b_{EB})$	1.012	1.004	1.025	1.026	1.004	1.001	0.960	3.136
4	Hyper-g	BB$(1, 1)$	1.015	1.024	1.062	0.988	1.006	0.994	0.837	3.271
5	EB-local	BB$(1, 1)$	1.024	1.037	1.071	1.001	1.012	0.999	0.853	0.803
6	EB-local	BB$(1, b_{EB})$	1.025	1.039	1.091	0.992	1.007	0.994	0.811	2.441
7	Hyper-g	BB$(1, b_{EB})$	1.025	1.026	1.091	0.996	1.015	0.996	0.793	9.849
8	EB-local	BB$(1, b_{SDM})$	1.042	1.053	1.124	1.006	1.023	1.002	0.805	0.812
9	Hyper-g	BB$(1, b_{SDM})$	1.042	1.047	1.139	1.009	1.017	0.996	0.786	3.287
10	$g = \sqrt{n}$	Ber(θ_{SDM})	1.052	1.062	1.150	1.036	1.000	1.011	0.837	0.849
11	$g = \sqrt{n}$	Ber(θ_{EB})	1.085	1.091	1.275	1.053	1.004	1.005	1.068	2.595
12	EB-local	Ber(θ_{EB})	1.104	1.115	1.360	1.011	1.026	1.009	0.988	2.163
13	Hyper-g	Ber(θ_{EB})	1.105	1.115	1.358	1.021	1.025	1.010	0.964	9.483
14	Hyper-g	Ber(θ_{SDM})	1.130	1.121	1.472	1.030	1.020	1.007	0.761	3.179
15	EB-local	Ber(θ_{SDM})	1.143	1.132	1.500	1.034	1.037	1.014	0.772	0.730
16	$g = \sqrt{n}$	Uniform	1.200	1.184	1.400	1.165	1.179	1.070	1.343	0.846
17	Hyper-g	Uniform	1.215	1.186	1.482	1.083	1.198	1.125	1.200	3.172
18	EB-local	Uniform	1.256	1.181	1.508	1.092	1.295	1.204	1.211	0.721
19	$g = \sqrt{n}$	Complexity(1)	1.298	1.222	2.045	1.152	1.047	1.023	0.558	0.670
20	EB-local	Complexity(1)	1.409	1.303	2.515	1.162	1.046	1.020	0.546	0.707
21	Hyper-g	Complexity(1)	1.415	1.305	2.526	1.171	1.054	1.017	0.537	3.492
22	$g = \sqrt{n}$	Complexity(2)	1.787	1.490	3.652	1.444	1.255	1.096	0.406	0.680
23	EB-local	Complexity(2)	1.863	1.544	4.032	1.432	1.227	1.081	0.411	0.732
24	Hyper-g	Complexity(2)	1.872	1.553	4.045	1.434	1.245	1.085	0.407	4.023

and Raftery, 1995) and the risk inflation criterion (RIC) with $g = p^2$ (Foster et al., 1994), motivated by the simulation study of Fernandez et al. (2001). In Chapter 2, we found both

of these parameter prior choices to be outperformed by the parameter priors used in this study. Also, their comparison was based only on tall datasets ($n > p$) and their comparison of methods was limited to the statistical tasks of inference and probabilistic prediction using the log-predictive score. They also did not consider EB versions of the Binomial(p, θ) and Beta-Binomial$(1, b)$ model space priors and complexity priors. Like Ley and Steel (2009), we found that random θ versions (or Beta-Binomial versions) performed better since the hierarchical prior is less sensitive to the choice of prior model size $E[S]$. Similarly, they found that priors specified by a fixed θ tended to be quite informative, casting doubt on their appropriateness as default reference priors.

Scott and Berger (2010) discussed the multiplicity correction effect of a subset of the model space priors discussed here, specifically Ber$(\theta = 0.5)$, Ber(θ_{EB}) and BB$(1, 1)$. They used a non-empirical simulation design, and did not compare methods based on the statistical tasks discussed here. Eicher et al. (2011) compared 12 parameter priors (of which $g = \sqrt{n}$ is common with ours) along with two fixed model priors: Uniform model priors with $\theta = 0.5$ and Ber(θ_{SDM}) with a prior expected model size of 7. The comparison was based on non-empirical simulation studies and one real growth regression dataset using predictive performance and inference measures. They found that the UIP with a uniform model prior performed better than Ber(θ_{SDM}) on the three statistical tasks common with ours. In contrast, we found that Ber(θ_{SDM}) was ranked higher than Uniform model priors for all our three preferred parameter priors across the statistical tasks considered.

We found the complexity priors (Castillo et al., 2015) to perform relatively poorly. At first sight, this seems to be in conflict with the theoretical results of Castillo et al. (2015), who showed that under certain assumptions the posterior distribution contracts optimally to recover an unknown sparse parameter vector and gives optimal predictions. However, their theoretical results assume that the data are generated from a spike and slab prior with the Laplace distribution as the slab density, and that the error variance σ^2 is known, which rarely holds in practice. Also, Rossell et al. (2021) argued that complexity priors can introduce very strong sparsity a priori, and showed empirically that when the true model is not sparse,

complexity priors may perform suboptimally for finite n. This is consistent with our results.

We have focused attention on independent model priors, i.e. priors in which the inclusion of each variable is statistically independent of that of the other variables. However, non-independent default priors have been proposed as well. (George, 1999, 2010) proposed dilution priors which dilute the prior model probability within subsets of similar models with highly correlated predictors. There is also research designing dependent model priors based on domain knowledge (Brock et al., 2003; Durlauf et al., 2008). Dellaportas et al. (2012) proposed a joint specification of the prior distribution across models so that the sensitivity of posterior model probabilities to the dispersion of prior distributions for the parameters of individual models (Lindley's paradox) is diminished. Villa and Walker (2015) assigned prior mass to models on the basis of their *worth*, based on the KL-divergence between densities under different models. However, all of these dependent model space priors lead to increased computational complexity and have been shown to work only when p is relatively small. They have also not yet been implemented in publicly available software.

Chapter 4

LAPLACE POWER EXPECTED POSTERIOR PRIOR FOR GENERALIZED LINEAR MODELS

4.1 Introduction

Generalized linear models (GLMs, e.g., see McCullagh and Nelder, 2019) are one of the main workhorses of statistical analysis. They are widely used both to model data directly and as building blocks for more complex hierarchical models. However, in spite of their broad adoption, prior elicitation for GLMs in the absence of subjective information remains an open problem, particularly in settings where the main goal is variable selection. Indeed, because standard non-informative priors for GLMs that work well for parameter estimation are often improper, they cannot be used in model selection problems, as they typically lead to ill-defined Bayes factors (e.g., see Berger et al., 2001).

As discussed in previous chapters, the literature on so-called "objective" or "default" priors for model selection in Gaussian linear models is extensive. Examples include point-mass spike-&-slab priors (Mitchell and Beauchamp, 1988; Geweke, 1996), g-priors (Zellner, 1986b), mixtures of g-priors (Zellner and Siow, 1980; Liang et al., 2008), unit information priors (Kass and Wasserman, 1995), intrinsic Bayes factors (Berger and Pericchi, 1996a), fractional Bayes factors (O'Hagan, 1995; De Santis and Spezzaferri, 2001), non-local priors (Johnson and Rossell, 2010, 2012b), power-expected-posterior priors (Fouskakis et al., 2015) and prior-based Bayesian information criterion (Bayarri et al., 2019), among other approaches. See Consonni et al. (2018) for a comprehensive review of recent approaches to objective Bayesian analysis, and Bayarri et al. (2012) for a review and discussion of desirable properties. The literature on default priors for GLMs is more limited, with three main approaches dominating. These include those introduced by Bové and Held (2011) and

Li and Clyde (2018), both of which consider modifications of mixtures of g-priors that are suitable for GLMs, and Fouskakis et al. (2018), who considers extensions of power-expected-posterior priors that rely on unnormalized power likelihoods. One feature shared by all three approaches is that they can be thought of as being based on the idea of calibrating (possibly improper) priors using either real or imaginary training samples. The idea of using imaginary training samples to calibrate a prior traces back to Good (1950), while the idea of using imaginary samples to find normalizing constants in the Bayes factors when using improper priors goes back to Spiegelhalter and Smith (1982). One popular way to choose imaginary samples is by using the "local" principle, where the imaginary samples fully support the null hypothesis in nested model comparisons. Additionally, the idea of *minimal training samples* and unit information prior (Kass and Raftery, 1995) has been utilised to make information content in the priors as small as possible and to minimize the amount of data utilised for building prior distribution (e.g., see Ibrahim and Chen, 2000 and Pérez and Berger, 2002). Imaginary training samples can be fixed (e.g., see Spiegelhalter and Smith, 1982, Ibrahim and Chen, 2000 and Zellner, 1986a) or can be treated as stochastic components (e.g., see Berger and Pericchi, 1996b, Pérez and Berger, 2002 and Fouskakis et al., 2018). Alternatively, O'Hagan (1995) proposed fractional Bayes factors where instead of using training samples, the improper prior is "trained" to proper prior using a fraction of full sample likelihood.

In this chapter, we introduce a variant of the power-expected-posterior (PEP) prior for GLMs that we call the Laplace PEP, or LPEP. While the formulation is general, this manuscript emphasizes the development of the LPEP for logistic regression models. This is because this subclass of models provides the best illustration of the theoretical and practical advantages of our approach. For example, we note that the prior described in Li and Clyde (2018) is improper when the maximum likelihood estimate (MLE) of the regression coefficients under the observed data does not exist, leading to ill-defined Bayes factors. In the case of logistic regression, this happens when there is separation among the groups (e.g., see Albert and Anderson, 1984, Heinze and Schemper, 2002 and Ghosh, 2019). Separation is reasonably common in practical applications, especially in problems with relatively small

samples and several unbalanced and highly predictive risk factors. A similar issue arises with the PEP priors introduced in Fouskakis et al. (2018) since the imaginary training samples are not restricted to yield finite MLEs. Furthermore, both versions of the PEP prior proposed by Fouskakis et al. (2018) are computationally intractable, requiring the use of reversible Jump Markov Chain Monte Carlo algorithms (Green, 1995; Dellaportas et al., 2002). The LPEP is well defined under separation as long as at least one training sample exists that yields finite MLEs under the full model. Furthermore, because the LPEP can be written as a location-and-scale mixture of Gaussian distribution, it is easy to incorporate into Markov chain Monte Carlo algorithms that rely on data augmentation (e.g., Polson et al., 2013). This feature also facilitates the development of prior-based Bayesian Information Criterion (e.g., see Li and Clyde, 2018 and Bayarri et al., 2019) that avoids data augmentation (at the potential cost of accuracy). Finally, the location-and-scale mixture structure simplifies the theoretical study of the prior, allowing us to show that the procedure is both asymptotically consistent and intrinsically consistent. Like Bové and Held (2011), Li and Clyde (2018) and Fouskakis et al. (2018), LPEPs implicitly assume that q, the size of the largest model under consideration is smaller than n. However, unlike these previous works, we consider the model selection consistency when both q and the number of variables p grow with n. Finally, code implementing our algorithms along with all real and simulated data sets are available here. Code to replicate the results in the paper is available on Github.

The remainder of this chapter is organized as follows: Section 4.2 reviews the general definition and properties of PEP priors. Sections 4.3 and 4.4 discusses in detail the LPEP prior for logistic regression and its theoretical properties such as intrinsic and model selection consistency. Section 4.5 discusses different computational strategies that can be used with the LPEP. Sections 4.6 and 4.7 presents empirical results from simulation studies as well as three different real datasets. Finally, Section 4.8 discusses our results and future directions for research.

4.2 Power-expected-posterior priors: A brief review

Power-expected-posterior (PEP) priors (Fouskakis et al., 2015) extend the expected-posterior (EP) priors introduced by Pérez and Berger (2002) by controlling the amount of information contained in the prior using the power approach originally developed by Ibrahim and Chen (2000) and Chen et al. (2000) in the context of subjective priors.

Briefly, let $\boldsymbol{y}$ denote the n-dimensional vector containing the observed data, $\boldsymbol{\gamma}$ index the model space, and $\boldsymbol{\beta_\gamma}$ represent vector of parameters under model $\boldsymbol{\gamma}$. We start with a (potentially improper) prior $\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta_\gamma})$ under model $\boldsymbol{\gamma}$ and introduce an n^*-dimensional vector of *imaginary* training samples arising from a distribution $m^*(\mathbf{y}^*)$. The EP prior is then constructed as

$$\pi_{\boldsymbol{\gamma}}^{EP}(\boldsymbol{\beta_\gamma}) = \int \frac{f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}\right) \pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta_\gamma}\right)}{\int f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}\right) \pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta_\gamma}\right) d\boldsymbol{\beta_\gamma}} m^*(\boldsymbol{y}^*) d\boldsymbol{y}^*.$$

In words, the EP priors use the imaginary training sample $\mathbf{y}^*$ to update the original prior $\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta_\gamma})$, and addresses the possible effect of using any particular training sample by averaging over the distribution $m^*(\mathbf{y}^*)$. The use of a common $m^*(\boldsymbol{y}^*)$ properly calibrates the priors across the different models, even in situations where $m^*(\boldsymbol{y}^*)$ is improper. Pérez and Berger (2002) discuss various possible choices of $m^*(\boldsymbol{y}^*)$ in both informative and non-informative settings.

Note that an implicit assumption in the formulation of the PEP is that $\mathbf{y}^*$ and n^* are such that the posterior based on it is proper, i.e.,

$$\int f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta_\gamma}\right) \pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta_\gamma}\right) d\boldsymbol{\beta_\gamma} < \infty, \tag{4.1}$$

for any $\boldsymbol{y}^*$ in the support of m^* (see Pérez and Berger, 2002 for details). However, large values of n^* will produce priors that are relatively concentrated. To balance these two goals, it is common to choose n^* as the size of the minimum training sample required to satisfy (4.1) across all models.

Even though the EP prior attempts to ameliorate the effect of the $\boldsymbol{y}^*$ by averaging over m^* and by using training samples that are as small as possible, in some applications the

prior might still be quite concentrated. Power-expected-posterior priors (Fouskakis et al., 2015) address this by scaling the likelihood of the imaginary sample,

$$\pi_{\boldsymbol{\gamma}}^{PEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \int \frac{\tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) \pi_{\boldsymbol{\gamma}}^{N}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)}{\int \tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) \pi_{\boldsymbol{\gamma}}^{N}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}} m^*(\boldsymbol{y}^* \mid \delta) f(\delta \mid \boldsymbol{\gamma}) d\delta d\boldsymbol{y}^*,$$

where $\tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) = \frac{f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^{\frac{1}{\delta}}}{\int f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^{\frac{1}{\delta}} d\boldsymbol{\beta}_{\boldsymbol{\gamma}}}$ is the normalized power likelihood for the training sample $\boldsymbol{y}^*$ based on model $\boldsymbol{\gamma}$, δ is the power parameter, $m^*(\boldsymbol{y}^* \mid \delta)$ is the predictive distribution generating the imaginary samples $\boldsymbol{y}^*$, and $f(\delta \mid \boldsymbol{\gamma})$ is a hyper-prior on δ. Fouskakis et al. (2015) recommended $m^*(\boldsymbol{y}^* \mid \delta) = m_0^N(\boldsymbol{y}^* \mid \delta)$, i.e., the marginal likelihood evaluated using the power likelihood of $\boldsymbol{y}^*$ under the null model and baseline prior $\pi_{\mathbf{0}}^N(\boldsymbol{\beta}_{\mathbf{0}})$. If $\delta = 1$, then PEP prior reduces to the EP prior, while values of $\delta > 1$ yield priors with a larger variance (and therefore, less information) than the EP prior. A particularly appealing choice is $\delta = n^*$ (or, alternatively, a prior on δ that is centered around n^*), which leads to a prior that can be considered as being unit information (Kass and Wasserman, 1995). Note that δ plays a similar role to the g parameter involved in the definition of (mixtures of) g priors. Hence, treating δ as random will typically lead to priors that have heavier tails, and are therefore more robust (in the sense of Dawid, 1973 and Andrade and O'Hagan, 2011). In this chapter, in addition to the unit information setting where $\delta = n^*$, we consider the hyper-g/n prior from Liang et al. (2008) and the robust prior from Bayarri et al. (2012) as two choices for hyper priors for δ. See Section 4.3 for more details. Furthermore, being able to use the parameter δ to control the amount of information contained in the prior means that the choice of the size of the training sample is less critical in the case of PEP priors. In the sequel, we work with $n^* = n$, a choice that is particularly convenient when dealing with regression models. Indeed, taking $n^* = n$ allows us to select $\boldsymbol{X}^*$, the design matrix associated with the training sample $\boldsymbol{y}^*$, as $\boldsymbol{X}^* = \boldsymbol{X}$ (see Section 4.3), the design matrix associated with the observed data.

The PEP prior was originally derived for model selection in Gaussian linear models (Fouskakis et al., 2015). In that case, normalizing constant $\int f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^{\frac{1}{\delta}} d\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ associated

with $\tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right)$ is is usually straightforward to compute. Indeed, for most standard choices of $\pi_{\boldsymbol{\gamma}}^N\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)$, the induced PEP can be written as a location-and-scale mixture of Gaussian distributions, dramatically simplifying computation within a Markov chain Monte Carlo framework. This property, however, does not extend to other GLMs. To address this issue, Fouskakis et al. (2018) introduced two different modifications of the PEP framework that rely on the unnormalized power likelihood $f_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^{\frac{1}{\delta}}$ rather than $\tilde{f}_{\boldsymbol{\gamma}}\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right)$: the concentrated reference PEP (CRPEP) and the diffuse reference PEP (DRPEP). However, while the use of the unnormalized power likelihood avoids some of the computational difficulties associated with the original PEP prior, many of them remain. In particular, neither $\pi_{\boldsymbol{\gamma}}^{CRPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ nor $\pi_{\boldsymbol{\gamma}}^{DRPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ belong to standard families of distributions. This prevents closed-form integration of the regression coefficients and therefore requires the use Reversible Jump Markov chain Monte Carlo algorithms (Green, 1995; Dellaportas et al., 2002). Furthermore, the definition of the CRPEP and the DRPEP and the computational approach introduced by the authors (which relies on Laplace approximations to compute certain normalizing constants needed for the acceptance probabilities of various Metropolis-Hastings steps) implicitly assume that the MLE of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ exists for any $\boldsymbol{y}^*$ and model $\boldsymbol{\gamma}$.

4.3 The Laplace power-expected-posterior prior for logistic regression

Instead of working with the unnormalized power likelihood as in Fouskakis et al. (2018), in this chapter we propose replacing the likelihood of the imaginary samples with its Laplace approximation *before* raising it to the power $1/\delta$. Hence, the name Laplace PEP, or LPEP. More concretely, let the observations $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be generated from a logistic regression likelihood of the form

$$f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \exp\left\{\sum_{i=1}^{n} y_i \boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}} - \log\left(1 + \exp\left\{\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right\}\right)\right\}, \tag{4.2}$$

where $\boldsymbol{x}_i = (1, x_{i,1}, \ldots, x_{i,p})^T$ is the $p+1$ dimensional vector of regressors associated with observation y_i, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ is the $p+1$ dimensional vector of regression coefficients (including the intercept), $\boldsymbol{\gamma}^T = (\gamma_0, \gamma_1, \ldots, \gamma_p)$ is a binary vector of length $p+1$ such that

for all $j \in \{1, \ldots, p\}, \gamma_j = 1$ if the j-th variable is included in the model (i.e., if β_j is different from zero) and $\gamma_j = 0$ otherwise, and $\boldsymbol{x}_{\boldsymbol{\gamma},i}$ and $\boldsymbol{\beta}_\gamma$ denote the sub-vectors of $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ with length $p_{\boldsymbol{\gamma}} + 1$ where $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ that include only those components for which the corresponding γ_j is equal to 1. In the sequel we assume that $\gamma_0 = 1$ (i.e., intercept is always included in the model) and that $n > p$. The Laplace approximation to (4.2) is given by

$$f_\gamma\left(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) \approx f_\gamma^L\left(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma(\boldsymbol{y})\right)^T \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y})\left(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma(\boldsymbol{y})\right)\right\}, \tag{4.3}$$

where $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y})$ denotes the MLE of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ based on sample $\boldsymbol{y}$, and $\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y})$ is the $(p_{\boldsymbol{\gamma}}+1) \times (p_{\boldsymbol{\gamma}}+1)$ observed information matrix

$$\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}) = \sum_{i=1}^{n} \frac{\left(1 + \exp\left\{\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y})\right\}\right)^2}{\exp\left\{\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y})\right\}} \boldsymbol{x}_{\boldsymbol{\gamma},i} \boldsymbol{x}_{\boldsymbol{\gamma},i}^T.$$

In the case of logistic regression models (and of regular exponential families more broadly), it is well known that this approximation is accurate up to an $\mathcal{O}(\frac{1}{n})$ order term (e.g., see Schwarz, 1978b). With this in mind, we define the LPEP as

$$\pi_{\boldsymbol{\gamma}}^{LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \int \frac{\tilde{f}_\gamma^L\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) \pi_\gamma^N\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)}{\int \tilde{f}_\gamma^L\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) \pi_\gamma^N\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}} m^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*) f(\delta \mid \boldsymbol{\gamma}) d\delta d\boldsymbol{y}^*, \tag{4.4}$$

where $\boldsymbol{X}^*$ is the $n^* \times (p+1)$ matrix whose rows correspond to the $\boldsymbol{x}_i^T$ vectors and

$$\tilde{f}_\gamma^L\left(\boldsymbol{y}^* \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta\right) \propto \delta^{-\frac{p_{\boldsymbol{\gamma}}+1}{2}} \exp\left\{-\frac{1}{2\delta}\left(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma(\boldsymbol{y}^*)\right)^T \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\left(\boldsymbol{\beta}_\gamma - \hat{\boldsymbol{\beta}}_\gamma(\boldsymbol{y}^*)\right)\right\}.$$

As discussed in Pérez and Berger (2002), standard choices for the baseline prior include the (improper) flat prior $\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1$ and the Jeffreys prior for GLMs (Ibrahim and Laud, 1991) where $\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto \left|\mathsf{E}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}\left\{\boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y})\right\}\right|^{1/2}$. In this chapter, we focus our attention on the flat prior $\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1$, which was also used in Fouskakis et al. (2018). There are two reasons for this. The first one is greater mathematical tractability, as the Jeffreys prior for GLMs does not lead to tractable expressions for the LPEP. Secondly, as we show in Section 4.4.3, the intrinsic prior associated with the LPEP derived under $\pi_{\boldsymbol{\gamma}}^N(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1$ to the same intrinsic prior associated with Bové and Held (2011).

For the predictive distribution of the imaginary samples we consider

$$m^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*) \propto \tilde{m}^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*)\mathbf{1}\left(\mathbf{y}^* \in A(\boldsymbol{X}^*)\right), \tag{4.5}$$

where $\tilde{m}^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*)$ is an unrestricted predictive distribution of $\boldsymbol{y}^*$ given by

$$\tilde{m}^*(\boldsymbol{y}^*) = \frac{\Gamma\left(\sum_{i=1}^{n^*} y_i^* + \frac{1}{2}\right)\Gamma\left(n^* - \sum_{i=1}^{n^*} y_i^* + \frac{1}{2}\right)}{\Gamma(n^*+1)\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)}, \tag{4.6}$$

(a Beta-Binomial distribution with both parameters equal to $\frac{1}{2}$, which is the predictive distribution under the null model and its reference/Jeffreys prior), and $\mathbf{1}\left(\mathbf{y}^* \in A(\boldsymbol{X}^*)\right)$ is the indicator function on $A(\boldsymbol{X}^*) = \left\{\tilde{\boldsymbol{y}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\left(\tilde{\boldsymbol{y}}, \boldsymbol{X}^*\right) \text{ exists and is finite for all } \boldsymbol{\gamma}\right\}$. Note that our choice for $m^*(\boldsymbol{y}^* \mid \boldsymbol{X}^*)$ differs substantially from that in Fouskakis et al. (2018). For example, (4.5) *does not* depend on the scaling factor δ. This makes intuitive sense (there is no obvious reason why the power factor used to re-scale the information in the training sample should also affect how the training sample is generated), and simplifies both posterior computation and theoretical analysis. Furthermore, (4.5) explicitly depends on $\boldsymbol{X}^*$ in a way that ensures that the LPEP is proper (see Section 4.4.1).

At first sight, the computational implementation of (4.5) might seem daunting, as it potentially requires checking on the existence of the MLE of the parameters for every possible model. However, the following theorem shows that, for a broad class of GLMs that includes logistic regression, it is enough to check its existence for the full model.

Theorem 1. Let $\ell_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \boldsymbol{y}) : S_{\boldsymbol{\gamma}} \to \mathbb{R}$ denote a log-likelihood in the regular exponential family of the form $\ell(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \mathbf{y}) = \sum_{i=1}^{n} T(y_i)\eta\left(\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right) + A\left(\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)$. Assume $S_{\boldsymbol{\gamma}}$ is an open connected subset of $\mathbb{R}^{p+1}$ and let $\boldsymbol{\gamma}_F = (1, 1, \ldots, 1)$ denote the model that includes all potential regressors. If

(i) $\ell_{\boldsymbol{\gamma}_F}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_F}; \boldsymbol{y})$ is continuous and striclty concave on $S_{\boldsymbol{\gamma}_F}$

(ii) $\lim_{\boldsymbol{\beta}_{\boldsymbol{\gamma}_F} \to \boldsymbol{\beta}^*} \ell_{\boldsymbol{\gamma}_F}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_F}; \boldsymbol{y}) = -\infty$ for any $\boldsymbol{\beta}^* \in \partial S_{\boldsymbol{\gamma}_F}$, the closure of $S_{\boldsymbol{\gamma}_F}$.

Then, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y})$, the MLE, exists under any other model $\boldsymbol{\gamma}$.

The proof of Theorem 1 can be seen in Appendix C.1. Note that, in the case of logistic regression, condition (i) is satisfied as long as the design matrix $\mathbf{X}$ is full rank (which, in particular, requires $p < n$), while condition (ii) is satisfied as long as the data does not suffers from complete or quasi-complete separation under the full model (Albert and Anderson, 1984). Separation checks can be carried out using the algorithms implemented in the `R` package `detectseparation` (Kosmidis and Schumacher, 2020). Section 4.6 of Konis (2007) shows that the version of the test based on the dual program has the best empirical worst-case time and that it scales linearly in both sample size n and number of covariates p. Furthermore, the authors empirically showed that the dual program takes approximately the same time as fitting a GLM using iteratively re-weighted least squares (IRLS) algorithm.

Finally, we discuss the specification of the distribution on the power parameter δ. As mentioned in Section 4.2, in this manuscript we consider three alternatives. The first version of the LPEP we investigate is the unit information LPEP (UI-LPEP) obtained by fixing $\delta = n^*$. We also consider a version of the hyper-g/n prior discussed in Liang et al. (2008) and Li and Clyde (2018), $f^{HGN}(\delta) = \left(1 + \frac{\delta}{n^*}\right)^{-2}$. We call this the HGN-LPEP. The median of this hyper-g/n prior is equal to n^*, and the prior places much of its mass around this value. Our third alternative is a version of the robust prior recommended by Bayarri et al. (2012), $f^{R}(\delta \mid \boldsymbol{\gamma}) = \frac{1}{2(p_{\boldsymbol{\gamma}}+1)^{1/2}} \frac{(n^*+1)^{1/2}}{(\delta+1)^{3/2}} \mathbf{1}\left(\delta > \frac{n^*-p_{\boldsymbol{\gamma}}}{p_{\boldsymbol{\gamma}}+1}\right)$, which we call this the R-LPEP. Note that, under this prior, $\mathsf{E}(\delta) = 3\frac{n^*+1}{p^*_{\boldsymbol{\gamma}}+1} = \mathcal{O}(n^*)$.

4.4 Properties of the LPEP prior

4.4.1 Proper prior

The LPEP for logistic regression can be written as a location-and-scale mixture,

$$\pi^{LPEP}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \sum_{\boldsymbol{y}^* \in \{0,1\}^n} \left[\int \phi_{p_{\boldsymbol{\gamma}}+1} \left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*), \delta \boldsymbol{H}^{-1}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \right) f(\delta \mid \boldsymbol{\gamma}) d\delta \right] m^*(\boldsymbol{y}^* \mid \boldsymbol{X}),$$

where $\phi_p(\cdot \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density of the p-variate normal with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Note that, because the training samples are restricted to yield finite MLEs,

$\phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*), \delta \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{y}^*)\right)$ is proper for any model γ. Furthermore, $m^*(\boldsymbol{y}^* \mid \boldsymbol{X})$, by construction, is also proper. Hence, the LPEP prior is also proper for every $\boldsymbol{\gamma}$

4.4.2 Tail behavior

It is straightforward to see that the UIP version of the LPEP (where $\delta = n^*$), $\pi_{\boldsymbol{\gamma}}^{LPEP-UI}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ has Gaussian tails. On the other hand, as the following theorem shows, the hyper-g/n and the robust versions of the LPEP have heavier (polynomial) tails in every direction.

Theorem 2. For any model $\boldsymbol{\gamma}$ and vector $\boldsymbol{v}$ such that $\|\boldsymbol{v}\| = 1$, let $\zeta^{HGN}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}) = \pi_{\boldsymbol{\gamma}}^{HGN-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\big|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}=s\boldsymbol{v}}$ and $\zeta^{R}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}) = \pi_{\boldsymbol{\gamma}}^{R-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\big|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}=s\boldsymbol{v}}$. Then there exist bounded functions $c_{\boldsymbol{\gamma}}^{HGN}(\boldsymbol{v})$ and $c_{\boldsymbol{\gamma}}^{R}(\boldsymbol{v})$ such that

$$\lim_{s\to\infty} \frac{\zeta^{HGN}(s \mid \boldsymbol{v}, \boldsymbol{\gamma})}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = c_{\boldsymbol{\gamma}}^{HGN}(\boldsymbol{v}), \qquad \lim_{s\to\infty} \frac{\zeta^{R}(s \mid \boldsymbol{v}, \boldsymbol{\gamma})}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = c_{\boldsymbol{\gamma}}^{R}(\boldsymbol{v}).$$

The proof is presented in Appendix C.2. One important implication of this result is that, from an estimation (rather than model selection) perspective, $\pi_{\boldsymbol{\gamma}}^{HGN-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ and $\pi_{\boldsymbol{\gamma}}^{R-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ are robust, in the sense of having bounded influence in the case of likelihood-prior conflict (e.g., see Andrade and O'Hagan, 2006 and Andrade and O'Hagan, 2011). A second implication relates to the existence of point estimators such as the posterior mean and the posterior variance. Ghosh et al. (2018) recently showed that, in the presence of separation, priors with Cauchy-like tails might lead to posterior distribution that, while proper, might have infinite means. The results in Theorem 2 guarantee that, even in the presence of separation, the model-averaged posterior means are finite under all of the three versions of the LPEP that we consider in this manuscript.

4.4.3 Intrinsic consistency

The following theorem shows that the LPEP priors for logistic regression are intrinsically consistent, i.e., that they converge to a proper prior as the size of the training sample increases (see criteria 4 of Bayarri et al., 2012). Naturally, the exact form of the intrinsic prior depends

on the asymptotic regime for the covariates associated with new observations, as well as the exact prior used for δ. Theorem 3 below provides a relevant example.

Theorem 3. Assume that, as n^* grows, the covariate vectors $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots$ satisfy either of the following two conditions:

(i) If $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots$ forms a deterministic sequence, then $\frac{1}{n^*}\left[\boldsymbol{X}^*\right]^T \boldsymbol{X}^* \underset{n^*\to\infty}{\longrightarrow} \boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma}$ is positive definite.

(ii) If $\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \ldots$ are random, then they are independent and identically distributed with mean $\mathbf{0}$ and finite, positive definite covariance matrix $\boldsymbol{\Sigma}$.

Then, the unit information ($\delta = n^*$), hyper-g/n and robust versions of the LPEP have proper, non-degenerate intrinsic priors of the form

$$\begin{aligned}
\lim_{n^*\to\infty} \pi_{\boldsymbol{\gamma}}^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) &= \phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, 4\left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right]^{-1}\right), \\
\lim_{n^*\to\infty} \pi_{\boldsymbol{\gamma}}^{HGN-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) &= \int \phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, 4\delta^*\left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right]^{-1}\right)(1+\delta^*)^{-2}\, d\delta^*, \\
\lim_{n^*\to\infty} \pi_{\boldsymbol{\gamma}}^{R-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) &= \int \phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \mathbf{0}, 4\delta^*\left[\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right]^{-1}\right)\frac{(\delta^*)^{-3/2}}{2(p_{\boldsymbol{\gamma}}+1)^{\frac{1}{2}}}\mathbf{1}\left(\delta^* > \tfrac{1}{p_{\boldsymbol{\gamma}}+1}\right) d\delta^*,
\end{aligned}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ is the submatrix of $\boldsymbol{\Sigma}$ that includes only the rows and columns for which $\gamma_j = 1$.

A proof of this result can be seen in Appendix C.3. Interestingly, we note that these are the same intrinsic priors associated with the prior in Bové and Held (2011).

4.4.4 Model selection consistency

Model selection consistency refers to the ability of the procedure to choose the correct model as $n \to \infty$. Intuitively, when p is fixed, because the amount of information in $\pi_{\boldsymbol{\gamma}}^{LPEP}$ is kept approximately constant as n^* increases, the associated Bayes factors would behave asymptotically like those computed from the Bayesian Information Criterion, which are known to be consistent. Our results, which rely on a slight extension of those presented in Barber et al.

(2016) for sparse high-dimensional logistic regression, extend this intuition to situations in which p grows with n as long as n remains larger than p and at most a moderate number of covariates $q < p$ remain active.

We consider a sequence of variable selection problems indexed by the sample size n, where $y(n)$ represents the sample for the n-th problem, p_n is the total number of covariates, $\boldsymbol{\beta}_T(n)$ is the true parameter, which is associated with the true model $\boldsymbol{\gamma}_T(n)$, and $p_{\boldsymbol{\gamma}_T(n)} = \sum_{j=1}^{p_n} \gamma_{T,j}(n)$. Our interest lies in the recovery of $\boldsymbol{\gamma}_T(n)$ over the set $\Gamma = \left\{\boldsymbol{\gamma}(n) : \boldsymbol{\gamma}(n) \in \{0,1\}^{p_n},\, p_{\boldsymbol{\gamma}(n)} \leq q_n\right\}$. An implicit assumption is that the true model is contained in the set of models under consideration. The following theorem, a proof of which can be seen in Appendix C.4, formalizes the result for the LPEP with $\delta = n$.

Theorem 4. Assume that:

(i) $q_n = n^{\psi}$ for $0 \leq \psi < 1/3$

(ii) $p_n = n^{\kappa}$ for $\psi < \kappa < 1$

(iii) $\boldsymbol{\beta}^{min}_{\boldsymbol{\gamma}_T(n)}(n) = \min\limits_{j:\gamma_{T,j}(n)=1} \left|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{T,j}(n)}(n)\right| \geq n^{-\phi/2}$ for some $0 \leq \phi < 1 - \psi$

(iv) $\|\boldsymbol{\beta}_T(n)\|_2 \leq a_0$ for a fixed constant $a_0 \in (0, \infty)$.

(v) For every $i = 1, 2, \ldots$, the vector $\boldsymbol{x}_i$ is such that $\|\boldsymbol{x}_i\|_2$ is bounded by a constant.

(vi) $\forall\ n$, the smallest eigenvalue of $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X}$ is bounded from below by a positive constant.

(vii) $P\left(\boldsymbol{\gamma}(n)\right) \propto \binom{p_n}{p_{\boldsymbol{\gamma}(n)}}^{-1} \mathbb{I}\{p_{\boldsymbol{\gamma}(n)} \leq q_n\}$.

Define $\tilde{\boldsymbol{\gamma}}(n) = \arg\max_{\boldsymbol{\gamma}(n)}\{P\left(\boldsymbol{\gamma}(n)\right) \times m^{UI-LPEP}_{\boldsymbol{\gamma}(n)}(\boldsymbol{y}(n))\}$, where $m^{UI-LPEP}_{\boldsymbol{\gamma}(n)}(\boldsymbol{y}(n))$ is the marginal likelihood of $\boldsymbol{y}(n)$ under UI-LPEP prior. Then $\mathsf{Pr}\left(\tilde{\boldsymbol{\gamma}}(n) = \boldsymbol{\gamma}_T(n) \middle| \boldsymbol{y}(n)\right) \to 1$ as $n \to \infty$.

Conditions (i) and (ii) are statements about the rate of growth of the number of parameters and the maximum size of the model under consideration, while condition (iii) relates them to the rate of *decrease* in the minimum signal size. Note that when $\phi = 0$, condition (iii) holds true for all $0 \leq \psi < 1/3$. However, in the worst case, when q_n grows at the rate that is $\approx n^{1/3}$, we require $\boldsymbol{\beta}^{min}_{\boldsymbol{\gamma}_T(n)}(n) \geq n^{-1/3}$. Condition (iv) is necessary to avoid assymptotic data separation. Conditions (v) and (vi) are standard conditions for the existence of maximum likelihood estimators (e.g., see Wedderburn, 1976). Assumptions (iv), (v) and (vi), or their implications, have appeared previously in the literature (e.g., see Chen and Chen, 2012 and Luo and Chen, 2013). Condition (vii) limits the size of the models under consideration using a truncated Beta-Binomial prior. Assuming priors on models that heavily penalize large models is also common in high-dimensional regression (e.g., see Rossell et al., 2021).

4.4.5 Information consistency

Information consistency refers to the behavior of the model selection criteria for fixed sample n, as the observed sample $\boldsymbol{y}$ becomes increasingly more extreme. In logistic regression, because the sample space for $\boldsymbol{y}$ is finite for any n, information inconsistency do not arise.

4.5 Computation

4.5.1 Markov chain Monte Carlo sampling

The LPEP prior can be easily combined with the Polya-Gamma augmentation of Polson et al. (2013) to generate an efficient Markov chain Monte Carlo algorithm for variable selection in logistic regression. For this purpose, it is convenient to re-express (4.4) as

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*, \delta, \boldsymbol{\gamma} \sim \mathsf{N}\left(\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*), \delta \left\{\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\right\}^{-1}\right), \tag{4.7}$$

with $y^* \sim m^*\left(y^* \mid \boldsymbol{X}^*\right)$ and $\delta \mid \boldsymbol{\gamma} \sim f(\delta \mid \boldsymbol{\gamma})$. We can then use this hierarchical framework to first sample from the conditional posterior distribution of $(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta \mid \boldsymbol{y}^*, \boldsymbol{y})$ followed by sampling from the full conditional posterior distribution of $(\boldsymbol{y}^* \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta, \boldsymbol{y})$.

Consider first sampling $(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta \mid \boldsymbol{y}^*, \boldsymbol{y})$. From Theorem 1 of Polson et al. (2013),

$$f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto \prod_{i=1}^{n} \left(\exp\left\{(y_i - 1/2)\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right\} \int_0^\infty \exp\left\{-\frac{\omega_i}{2}\left(\boldsymbol{x}_{\boldsymbol{\gamma},i}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right)^2\right\} f(\omega_i \mid 1, 0) d\omega_i \right),$$

where $f(\omega \mid a, b)$ denotes the density of a Pòlya-Gamma random variate with parameters a and b. Therefore, after introducing a vector of auxiliary variables $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_n)$,

$$f(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}) \propto f(\boldsymbol{\gamma}) f(\delta \mid \boldsymbol{\gamma}) \\ \phi_{p_{\boldsymbol{\gamma}}+1}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*), \delta \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{y}^*)\right) \phi_n\left(\boldsymbol{z} \mid \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\Omega}^{-1}\right), \quad (4.8)$$

where $\boldsymbol{z} = ((y_1 - 1/2)/\omega_1, \ldots, (y_n - 1/2)/\omega_n)^T$, $\boldsymbol{\Omega} = \text{diag}\{\omega_1, \ldots, \omega_n\}$, $f(\boldsymbol{\gamma})$ is a prior on 2^p dimensional model space and $f(\delta \mid \boldsymbol{\gamma})$ is the prior on scale parameter δ. It is straightforward to see that $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ can be integrated out of (4.8), yielding

$$f(\boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}) = \int f(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}} \propto f(\boldsymbol{\gamma}) f(\delta \mid \boldsymbol{\gamma}) \phi_n(\boldsymbol{z} \mid \boldsymbol{m}_{\boldsymbol{z}}^{\boldsymbol{\gamma}}, \boldsymbol{V}_{\boldsymbol{z}}^{\boldsymbol{\gamma}}), \quad (4.9)$$

where $\boldsymbol{m}_{\boldsymbol{z}}^{\boldsymbol{\gamma}} = \boldsymbol{X}_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)$, $\boldsymbol{V}_{\boldsymbol{z}}^{\boldsymbol{\gamma}} = \boldsymbol{\Omega}^{-1} + \delta \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{H}_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{y}^*) \boldsymbol{X}_{\boldsymbol{\gamma}}^T$. Then, various versions of Metropolis-Hastings algorithms can be implemented to explore the space of models (e.g., see section 4.5 of George and McCulloch, 1997).

Once the model $\boldsymbol{\gamma}$ and the exponent δ have been updated, the regression coefficients can be sampled using the fact that $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta, \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y} \sim \mathsf{N}(\boldsymbol{m}_{\boldsymbol{\gamma},\boldsymbol{\omega}}, \boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\omega}})$, where

$$\boldsymbol{m}_{\boldsymbol{\gamma},\boldsymbol{\omega}} = \boldsymbol{V}_{\boldsymbol{\omega}} \left(\boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\Omega} \boldsymbol{z} + \frac{1}{\delta} \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \right), \qquad \boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\omega}} = \left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{\Omega} \boldsymbol{X}_{\boldsymbol{\gamma}} + \frac{1}{\delta} \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \right)^{-1},$$

and each ω_i can be updated from $f(\omega_i \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta, \mathbf{y}^*, \mathbf{y})$, which corresponds to an updated Pólya-Gamma distribution. Finally, $(\boldsymbol{y}^* \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta, \boldsymbol{y})$ can be easily updated using either Gibbs sampling or random-walk Metropolis-Hastings steps. Further details of the computational algorithm can be seen in Appendix C.5.

4.5.2 Model search using a prior-based Bayesian Information Criteria

In order to accelerate computation, Li and Clyde (2018) propose to use their default prior, $\beta_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta \sim \mathsf{N}(\mathbf{0}, \delta\{\boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y})\}^{-1})$, to construct a prior-based Bayesian Information Criterion

(pBIC). In situations where p is at least moderately large, this pBIC is then embedded into a random walk Metropolis-Hastings on the model space that has many similarities with the algorithm described in the previous section. This approach, can be applied across a wide variety of GLMs. In logistic regression, it sidesteps the need to perform the kind of data augmentation with Pólya Gamma random variables, potentially leading to computational gains (at the potential expense of accuracy).

A similar approach can be developed for the LPEP priors. In particular, we can use the Laplace approximation in (4.3) (applied this time to the likelihood of the observed data) in combination with (4.7) to get

$$f(\mathbf{y} \mid \gamma, \delta, \mathbf{y}^*) \approx f^L(\mathbf{y} \mid \gamma, \delta, \mathbf{y}^*) = f_{\boldsymbol{\gamma}}\left(\boldsymbol{y} \mid \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}\right) \delta^{-\frac{p_{\boldsymbol{\gamma}}+1}{2}} \left(\frac{|\boldsymbol{H}_{\boldsymbol{\gamma}}|}{\left|\tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}}\right|}\right)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\left[\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^T \boldsymbol{H}_{\boldsymbol{\gamma}} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}} + \frac{1}{\delta} \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{*T} \boldsymbol{H}_{\boldsymbol{\gamma}}^* \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^* - \tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}}^T \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}} \tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}}\right]\right\}, \tag{4.10}$$

where

$$\tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}} = \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}}^{-1}\left[\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y})\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}) + \frac{1}{\delta}\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\right], \qquad \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}} = \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}) + \frac{1}{\delta}\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*). \tag{4.11}$$

Equation (4.10) can be used to approximate the acceptance probabilities of a Metropolis-Hastings algorithm that explores the posterior distribution $f(\boldsymbol{\gamma}, \delta, \mathbf{y}^* \mid \mathbf{y})$ by alternating between sampling from $f(\boldsymbol{\gamma}, \delta \mid \mathbf{y}^*, \mathbf{y})$ and $f(\mathbf{y}^* \mid \boldsymbol{\gamma}, \delta, \mathbf{y})$. Then, samples for the coefficients can be obtained from the approximate posterior $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta, \mathbf{y}^*, \mathbf{y} \sim \mathsf{N}\left(\tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}}, \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}}^{-1}\right)$. Further details are provided in Appendix C.6.

4.6 Simulation studies

We conducted two simulation studies to compare the estimation and model selection performance of Laplace PEP priors with other existing model selection techniques. This section discusses the results from the first simulation study. The results for the second one can be seen in Appendix C.8.

The simulation study described in this section uses a sample size of $n = 500$ and a total number of covariates $p = p_{\boldsymbol{\gamma}_F} = 100$, with the vectors of predictors being drawn independently from a zero-mean, unit-scale multivariate normal distribution with pairwise correlations given by $cor(x_{i,j}, x_{i,j'}) = r^{|j-j'|}$ for $1 \leq j < j' \leq p$. It involves eight scenarios, which differ in terms of the sparsity level in the vector of regression coefficients and the correlation structure among predictors. More specifically, we consider all combinations of four different levels of sparsity ($p_{\boldsymbol{\gamma}_T} \in \{0, 5, 10, 20\}$, please see Table 4.1) and two different correlation coefficients ($r \in \{0, 0.9\}$). A total of 100 datasets were generated for each of our

Table 4.1: Value of intercept and coefficients in the true logistic regression model where $\boldsymbol{b} = (2, -1, -1, 0.5, -0.5)^T$

$p_{\boldsymbol{\gamma}_T}$	$\beta_{\boldsymbol{\gamma}_T,0}$	$\beta_{\boldsymbol{\gamma}_T,1:5}$	$\beta_{\boldsymbol{\gamma}_T,6:10}$	$\beta_{\boldsymbol{\gamma}_T,11:15}$	$\beta_{\boldsymbol{\gamma}_T,16:20}$
0	-0.5	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$
5	-0.5	$\boldsymbol{b}$	$\mathbf{0}$	$\mathbf{0}$	$\mathbf{0}$
10	-0.5	$\boldsymbol{b}$	$\mathbf{0}$	$\boldsymbol{b}$	$\mathbf{0}$
20	-0.5	$\boldsymbol{b}$	$0.5\boldsymbol{b}$	$\boldsymbol{b}$	$0.5\boldsymbol{b}$

8 scenarios. We apply both Bayesian procedures and various penalized likelihood approaches to each dataset. In terms of Bayesian procedures, we implement the LPEP prior using (i) the "exact" MCMC procedure using Pólya-Gamma latent variables as discussed in 4.5.1 (denoted as LPEPE in the sequel) and (ii) the "approximate" MCMC using the Laplace approximation of the likelihood discussed in 4.5.2 (denoted as LPEPL). We also consider the methodology of Li and Clyde (2018), which relies on a mixture of g-priors using (i) a Laplace approximation to compute the associated marginal likelihood (denoted LCL in the sequel), as well as (ii) an "exact" version of their procedure based on a latent-variable augmentation similar to the one described in Section 4.5.1 (denoted LCE in the sequel). Comparing LPEPE, LPEPL, LCE and LCL allows us to disentangle the effect of the Laplace

approximation from that of the prior choice on the performance of these techniques. For each of these four approaches, we consider three different settings for the hyperparameter δ: the unit information prior with $\delta = n$, the hyper-g/n prior, and robust prior (recall Section 4.3). We use the `R` package `BAS-V1.5.5` (Clyde, 2020) to implement LCL, and a slight modification of our own code to implement LCE. In all cases, we assume a Beta-Binomial(1,1) prior over the model space, and run the MCMC chain for $2^{17} \approx 131,000$ iterations after a burn-in of $10,000$ iterations. We do not include the CRPEP and DRPEP priors from Fouskakis et al. (2018) in this simulation study for two reasons. First, the computational complexity of the methods makes a simulation study like this prohibitive. Not only is each iteration of the algorithm more expensive than those of the other approaches, but the algorithm mixes more slowly, which means that a much larger number of iterations are required to get accurate results. Secondly, the algorithm provided by the authors broke down for a number of our simulated datasets. We incorporate comparisons with CRPEP and DRPEP in our real data examples in Section 4.7. In terms of penalized likelihood methods, we compare against LASSO (Tibshirani, 1996), smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010). We use the `R` package `glmnet` (Friedman et al., 2010) to implement LASSO, and the package `ncvreg` (Breheny and Huang, 2011) for SCAD and MCP. In all cases, the value of the regularization was selected using the default 10-fold cross validation schemes implemented in the respective packages. Finally, we also include information criteria (IC) based model averaging methods in our comparsion, particularly, Akaike Information criterion (AIC) and Bayesian Information criteria (BIC). They are implemented in `R` package `BAS`.

We evaluate the performance of these various methods in terms of model selection performance using three metrics. First, we report the frequency with which the MAP model matches the true model $\boldsymbol{\gamma}_T$ (see Table 4.2). For the penalized likelihood approaches (for which a single model is reported for each dataset) the equivalent metric is simply the number of datasets for which the technique reported the correct model.

Bayesian methods tend to clearly outperform penalized likelihood approaches, in some

cases quite dramatically. Furthermore, most Bayesian approaches tend to perform very well when the data is generated from the null model, both in the uncorrelated and highly correlated cases. On the other hand, as the number of non-zero coefficients in the true model increases, we observe that all approaches struggle to identify the true model, particularly when the covariates are highly correlated. AIC performs poorly compared to Bayesian methods while BIC performs similar to Unit information versions of Bayesian methods. When $p_{\boldsymbol{\gamma}_T} = 20$, none of the procedures is able to identify the true model. Nonetheless, it appears that, overall, LPEPE (and, specially, the robust and the hyper-g/n versions of LPEPE) perform the best, and that the exact versions of the procedures (LPEPE and LCE) perform better than their approximate counterparts.

While the MAP metric we discussed above provides some insights into model performance, it tends to be less informative when there is substantial uncertainty on the posterior distribution over the model space. Therefore, we also compute for each dataset the F_1 score for the MAP (Bayesian procedures) or selected (penalized likelihood procedures) model (see Figure 4.1). In this setting, the F_1 score is defined as the harmonic mean of the proportion of true positives among selected covariates (the precision) and the proportion of selected covariates among true positive covariates (the recall). We focus on precision and recall rather than false positive and false negatives because of the class imbalance implied by the sparse nature of the true models used in our simulation. Results are not presented for the null model since the F_1 score is not well defined in that case. In all cases, the methods based on LPEP priors tend to have higher F_1 scores, with the robust and hyper-g/n versions performing slightly better than the unit information prior. As before, AIC performs poorly whereas F_1 score under BIC are similar to UIP versions of Bayesian. We also see that, while all Bayesian procedures have very similar performance under the unit information prior, exact versions of LPEP and the Li and Clyde (2018) prior tends to outperform approximate versions under the robust and hyper-g/n priors (in some cases, quite dramatically).

Next, we report the average size of the sampled/selected models for each data set (see Figure 4.2). Under the robust and hyper-g/n priors, LCE and, especially, LCL tend to favor

Table 4.2: Number of times (**over 100 replications**) that the **MAP** model coincides with the true model; **BOLD** represent group maximum; * represent overall maximum.

p		100							
$p(\boldsymbol{\gamma})$		Beta-Binomial(1,1)							
$p_{\boldsymbol{\gamma}_T}$		0		5		10		20	
r		0	0.9	0	0.9	0	0.9	0	0.9
$\delta = n$	LPEPE	99	**100***	47	0	**13**	0	0	0
	LPEPL	**100***	**100***	47	0	12	0	0	0
	LCE	**100***	**100***	**48**	0	9	0	0	0
	LCL	**100***	**100***	44	0	9	0	0	0
$\delta \sim$ robust	LPEPE	99	**100***	**50**	0	**14***	0	0	0
	LPEPL	**100***	**100***	**50**	0	8	0	0	0
	LCE	99	**100***	39	0	0	0	0	0
	LCL	**100***	**100***	45	0	2	0	0	0
$\delta \sim$ hyper g/n	LPEPE	**99**	**98**	**51***	0	**11**	0	0	0
	LPEPL	98	**98**	50	0	6	0	0	0
	LCE	97	**98**	21	0	0	0	0	0
	LCL	66	80	4	0	0	0	0	0
	LASSO	59	66	0	0	0	0	0	0
	SCAD	57	62	0	0	0	0	0	0
	MCP	73	70	**8**	0	**3**	0	0	0
	AIC	54	92	8	**3***	2	0	0	0
	BIC	**99**	**100***	**44**	0	**8**	0	0	0

very large models. Similarly, LPEPL tends to favor larger models compared to LPEPE under robust and hyper-g/n priors. On the other hand, all Bayesian procedures under the unit information prior seem to underestimate the model size when $p_{\boldsymbol{\gamma}_T} = 20$. The best performing approaches with average model size very close to the true model size are again the robust and hyper-g/n versions of LPEPE.

Next, we compare the procedures using the average mean squared error (AMSE) of the estimated coefficients, $AMSE(\boldsymbol{\beta}) = \frac{1}{p}\sum_{j=1}^{p}(\hat{\beta}_j - \beta_{j,\boldsymbol{\gamma}_T})^2$, where $\hat{\beta}_j$ and $\beta_{j,\boldsymbol{\gamma}_T}$ are the estimated and true values of j^{th} covariate, respectively. For the Bayesian procedures, model-averaged posterior mean estimates are used while for penalized likelihood methods, the sparse point estimates of the coefficients are used. The results in Table 4.3 indicate that, as the true model size $p_{\boldsymbol{\gamma}_T}$ and the true correlation between covariates increases, the AMSE increases for all techniques. However, heavy tailed versions of LPEP significantly outperforms all other techniques in terms of estimation performance under non-null true model scenarios.

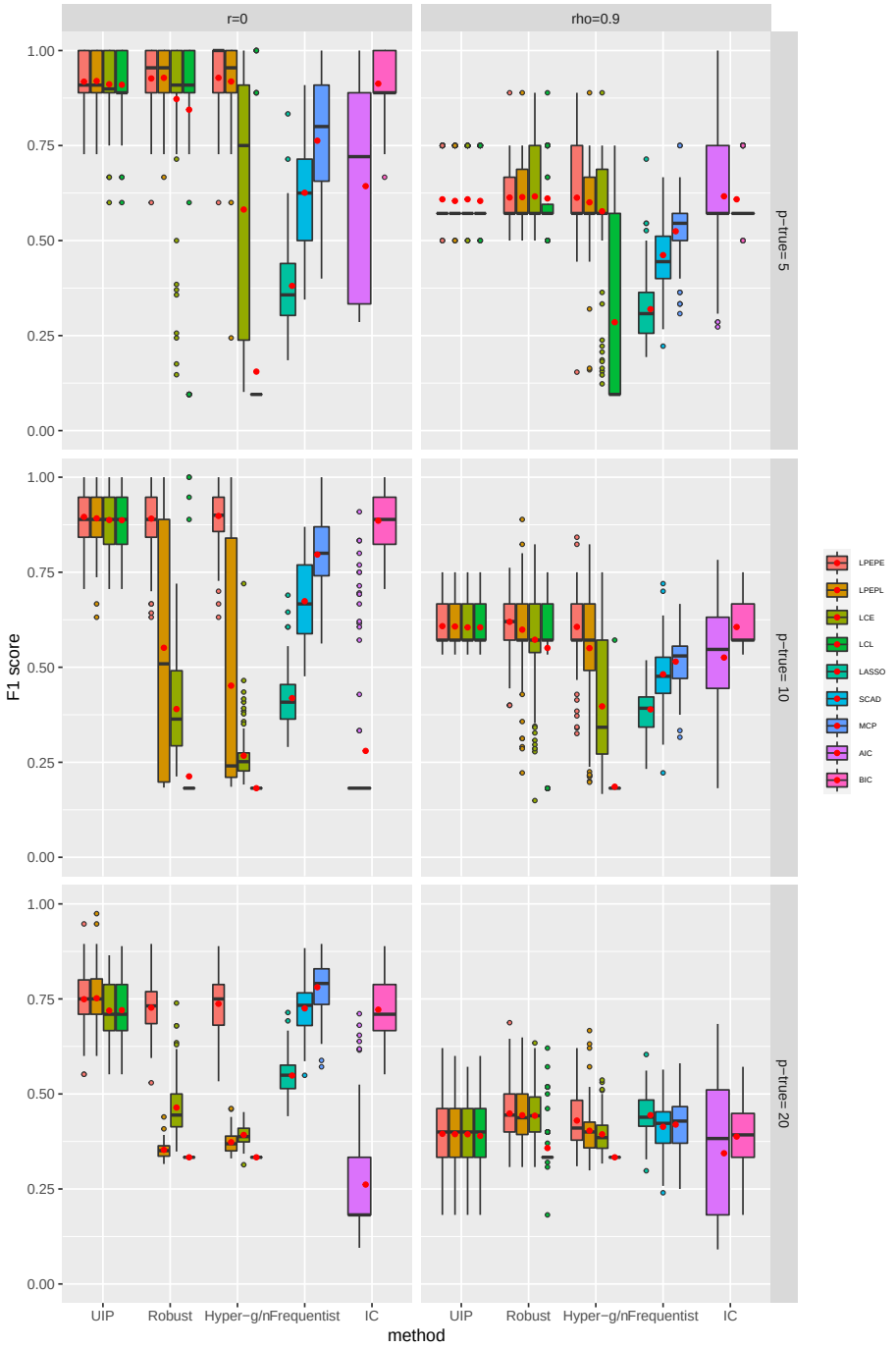


Figure 4.1: F1 score for the MAP model estimated by various methods and prior combinations for 100 simulated datasets ($n = 500, p = 100$) under different scenarios of correlation ($r = 0$: left; $r = 0.9$: right) and true number of non-zero coefficients specified in rows ($p - true = p_{\boldsymbol{\gamma}_T}$); Red dots represent the average F1 score across 100 simulated datasets.

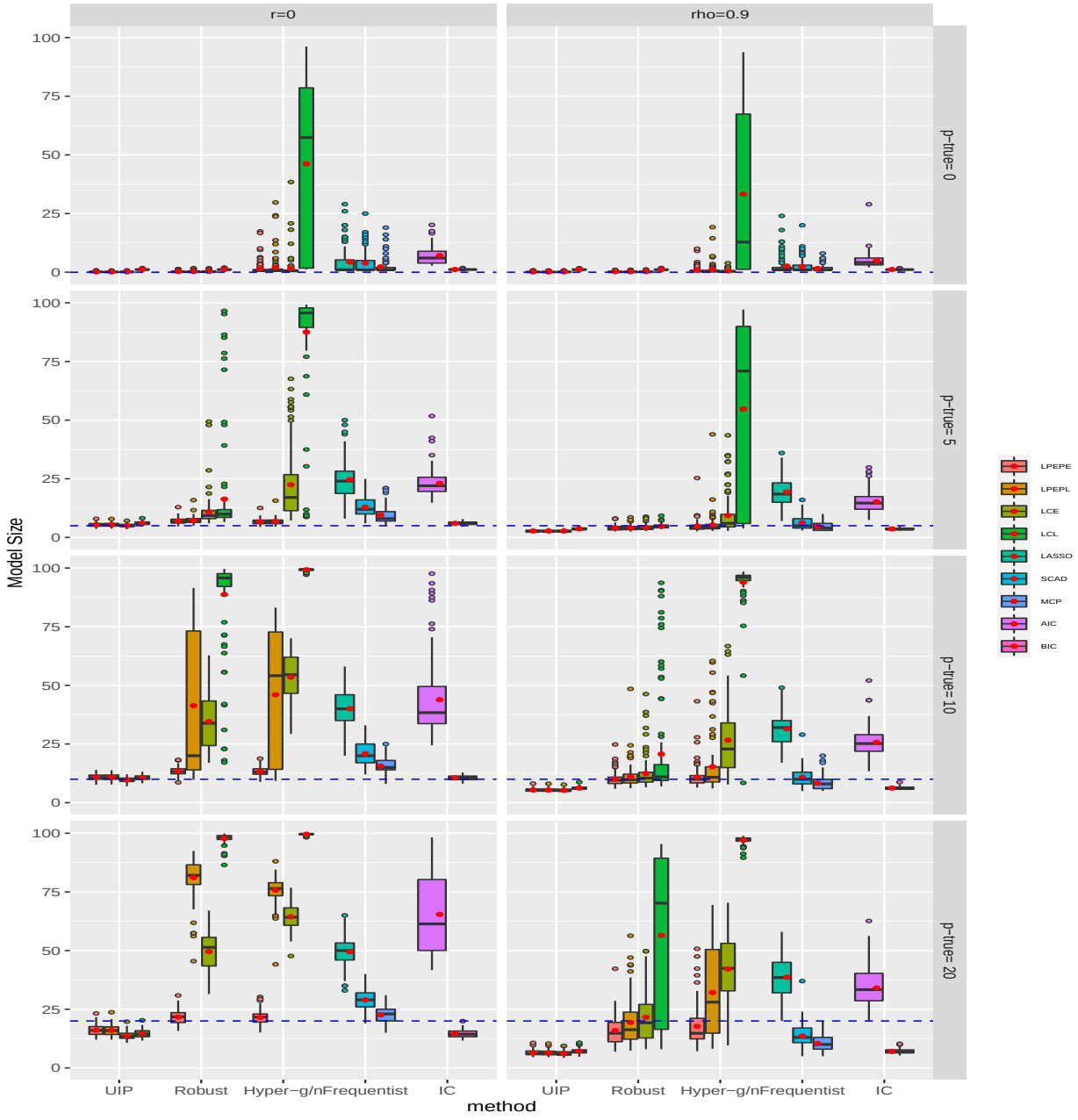


Figure 4.2: Average size of models selected by various methods and prior combinations for 100 simulated datasets ($n = 500, p = 100$) under different scenarios of correlation ($r = 0$: left; $r = 0.9$: right) and true number of non-zero coefficients specified in rows ($p - true = p_{\boldsymbol{\gamma}_T}$); Dotted blue line indicates the true model size $p - true = p_{\boldsymbol{\gamma}_T}$ and red dots represent the average model size over 100 simulated datasets.

Table 4.3: 1000 times the AMSE for estimated coefficients over 100 replications; **BOLD** represent group minimum; * represent overall minimum.

p		100							
$p(\boldsymbol{\gamma})$		Beta-Binomial(1,1)							
$p_{\boldsymbol{\gamma}_T}$		0		5		10		20	
r		0	0.9	0	0.9	0	0.9	0	0.9
$\delta = n$	LPEPE	0.11	**0.10***	2.93	16.06	7.84	31.56	15.89	**59.19**
	LPEPL	**0.10***	**0.10***	**2.73**	**15.84**	**6.64**	**31.46**	**14.65**	59.24
	LCE	0.11	**0.10***	3.05	16.17	8.30	31.92	16.96	59.94
	LCL	**0.10***	**0.10***	2.87	16.05	7.09	31.57	16.31	60.09
$\delta \sim$ robust	LPEPE	0.12	**0.10***	2.64	14.58	**6.35***	26.61	**13.47**	48.96
	LPEPL	0.11	**0.10***	**2.51***	**14.37**	42.64	**26.34***	125.96	**47.21***
	LCE	0.12	**0.10***	5.57	14.47	68.70	29.67	147.02	58.43
	LCL	**0.10***	**0.10***	8.80	14.58	3067.22	41.31	1744.03	112.74
$\delta \sim$ hyper g/n	LPEPE	**0.15**	0.14	2.71	**14.11***	**6.59**	**26.52**	**13.29***	**48.16**
	LPEPL	0.18	0.15	**2.58**	14.14	62.95	27.61	146.06	55.29
	LCE	0.22	**0.12**	8.02	14.45	53.17	33.20	76.12	55.21
	LCL	0.30	0.43	35.33	27.52	146.80	73.60	164.25	94.32
	LASSO	0.25	**0.21**	7.08	19.62	16.96	**33.66**	28.99	**54.68**
	SCAD	**0.21**	0.79	3.07	**19.31**	6.85	41.61	14.99	70.28
	MCP	0.22	0.23	**2.82**	19.96	**6.63**	45.80	**14.78**	69.89
	AIC	0.77	0.83	9.13	**14.84**	2810.47	35.62	1856.46	61.73
	BIC	**0.11**	**0.10***	**2.87**	15.99	**7.13**	**31.59**	**16.34**	**59.85**

Further discussions of the tradeoff between computational complexity and accuracy associated with the use of LPEPE, LPEPL, LCE and LCL can be found in Appendix C.7

4.7 Real data applications

This section discusses the performance of the LPEP in two real datasets. Three additional datasets are considered in Appendixes C.9, C.10 and C.11. Based on simulations studies, AIC is excluded based on its poor performance while BIC is excluded because its performance is similar to UIP versions of Bayesian priors.

4.7.1 `URINARY`: Determinants of urinary incontinence

The `URINARY` data set describes the results from a small drug study with 21 subjects. The response corresponds to whether the subject developed urinary incontinence after receiving the drug. The explanatory variables capture drug-induced physiological changes, which were in the same direction for most subjects. This data set was first presented in Potter (2005), and is further discussed in Mansournia et al. (2018). While very small, the data set is challenging to analyze because it exhibits full separation.

Table 4.4 presents estimates for the regression coefficients for various Bayesian and penalized likelihood methods. The results for LPEPE, CRPEP and DRPEP are based on 10,000 iterations of the MCMC algorithm obtained after a burn-in period of 10,000 iterations. On the other hand, for LCL we use the full model enumeration procedure in the `R` package `BAS`. In the case of Bayesian procedures, Table 4.4 presents model-averaged posterior means, as well as 95% credible intervals for the coefficients. Confidence intervals for the penalized likelihood procedures are not presented since they are not straightforward to obtain and the `R` packages we used to fit these models do not readily provide them. Furthermore, results for CRPEP and DRPEP are not included under the robust hyper-prior because such a procedure is not implemented in Fouskakis et al. (2018).

Note that LCL produces large point estimates and very wide credible for the model coefficients under all hyperpriors. This is no surprise; the prior proposed by Li and Clyde (2018) is proper only for models for which the maximum likelihood estimates (MLEs) are finite. Hence, for a data set like `URINARY`, some of the Bayes factors associated with LCL

are ill-defined. This is also why we do not show results for LCE and LPEPL; the posterior distribution for the associated Markov chain Monte Carlo algorithm is improper if the full model is included in the analysis. Furthermore, note that CRPEP yields point estimates that appear to be different from those generated by LPEPE, DRPEP, and the penalized likelihood methods. This is clearer when looking at the intercept of the model, which is negative with high probability under CRPEP but positive with high probability under LPEPE and DRPEP under all hyperpriors.

Next, we present in Table 4.5 the posterior inclusion probabilities (PIPs) associated with each of the three variables under each one of the Bayesian approaches, along with the model selected by each of the penalized likelihood methods. In all cases, LCL consistently places probability one on all variables.This is consistent with the results generated by the penalized likelihood methods. On the other hand, LPEPE places very high probability of inclusion of the second and third variables, but only moderately high probability of inclusion on the first one. These results are consistent for all specification of δ. In contrast, the results for the original PEP procedures in Fouskakis et al. (2018) are completely different and, more importantly, inconsistent across the CRPEP and DRPEP, and across various choices of δ. For example, while the CRPEP consistently favors excluding the third covariate, the DRPEP favors dropping either the second, or both the first and the third covariates depending on which hyperprior is used for δ.

4.7.2 `GUSTO-I`: Survival to treatments for occluded coronary arteries

Next, we consider data from the Global Utilization of Streptokinase and TPA for Occluded Coronary Arteries (GUSTO-I) trial (Califf et al., 1996), which has been previously analyzed in Held et al. (2015) and Li and Clyde (2018), and is publicly available at `http://www.clinicalpredictionmodels.org/` (Steyerberg, 2019). We aim to model the binary endpoint of 30-day survival for a subgroup of $n = 2188$ patients using 17 clinical covariates described in Appendix C.12.

Figure 4.3 displays the marginal posterior inclusion probabilities (PIPs) for Bayesian

Table 4.4: Estimated BMA coefficients and 95% credible intervals for different Bayesian techniques for `urinary` dataset; For frequentist techniques, estimated coefficient is displayed.

		β_0	β_1	β_2	β_3
$\delta = n$	LPEPE	0.56 (-1.66 , 2.85)	-0.70 (-2.32 , 0.10)	-0.39 (-0.81 , -0.13)	0.15 (0.00 , 0.37)
	LCL	-83.84 (-6009.26 , 5897.13)	-2333.88 (-161488.87 , 158312.76)	-1578.58 (-109266.01 , 107118.14)	296.17 (-19896.81 , 20678.41)
	CRPEP	-1.15 (-3.21 , 0.52)	-0.70 (-1.88 , 0.30)	-0.34 (-0.63 , -0.10)	0.00 (-0.00 , 0.00)
	DRPEP	0.69 (-0.55 , 2.13)	-1.00 (-2.09 , -0.19)	0.00 (0.00 , 0.00)	0.06 (-0.03 , 0.16)
$\delta \sim$ robust	LPEPE	0.71 (-1.74 , 3.55)	-0.98 (-3.82 , 0.05)	-0.52 (-1.89 , -0.12)	0.19 (0.00 , 0.54)
	LCL	-83.84 (-6250.14 , 5980.74)	-2148.67 (-161065.38 , 154146.29)	-1453.30 (-108979.51 , 104298.98)	272.66 (-19890.08 , 20102.78)
$\delta \sim$ hyper g/n	LPEPE	0.61 (-1.65 , 3.07)	-0.75 (-2.74 , 0.10)	-0.41 (-1.02 , -0.09)	0.15 (0.00 , 0.39)
	LCL	-83.84 (-6252.44 , 5650.45)	-1288.82 (-124412.67 , 113166.12)	-871.72 (-84179.77 , 76570.77)	163.55 (-15457.94 , 14685.16)
	CRPEP	-1.04 (-3.04 , 0.62)	-0.66 (-1.78 , 0.30)	-0.33 (-0.65 , -0.08)	0.00 (-0.00 , 0.00)
	DRPEP	-0.89 (-3.11 , 0.69)	0.00 (0.00 , -0.00)	-0.36 (-0.76 , -0.11)	0.00 (-0.00 , 0.00)
	LASSO	0.36	-0.70	-0.31	0.11
	SCAD	0.41	-0.23	-0.20	0.07
	MCP	0.40	-0.17	-0.20	0.07

methods and the inferred model under the penalized likelihood techniques. For techniques related to LCL, we again rely on full model enumeration. For all other Bayesian techniques, we use 131,000 iterations with a burn-in of 10,000 iterations. As in our simulation studies,

Table 4.5: Marginal posterior inclusion probabilities (PIPs) for `urinary` dataset (Bayesian procedures) and variables included in the model (penalized likelihood methods)

		Posterior inclusion probabilities		
		$P(\gamma_1 \neq 0 \mid \boldsymbol{y})$	$P(\gamma_2 \neq 0 \mid \boldsymbol{y})$	$P(\gamma_3 \neq 0 \mid \boldsymbol{y})$
$\delta = n$	LPEPE	0.725	0.996	0.908
	LCL	1.000	1.000	1.000
	CRPEP	1.000	1.000	0.000
	DRPEP	1.000	0.000	1.000
$\delta \sim$ robust	LPEPE	0.743	0.996	0.901
	LCL	1.000	1.000	1.000
$\delta \sim$ hyper g/n	LPEPE	0.732	0.992	0.891
	LCL	1.000	1.000	1.000
	CRPEP	1.000	1.000	0.000
	DRPEP	0.000	1.000	0.000
	LASSO	1.000	1.000	1.000
	SCAD	1.000	1.000	1.000
	MCP	1.000	1.000	1.000

all penalized likelihood techniques select denser models than the Bayesian procedures. In line with Li and Clyde (2018) and Held et al. (2015), we observe that `AGE`, `KILLIP`, `HYP`, `HRT` and `STE` have high PIPs under all methods. However, the different versions of LCL perform quite differently. In particular, the version of LCL that relies on a hyper-g/n hyperprior tends to explore very dense models leading, to PIPs close to 0.5 for all variables. Similarly, the hyper-g/n versions of CRPEP and DRPEP seem to differ from their $\delta = n$ versions with respect to `PMI` and `SEX` variables. On the other hand, the different versions of the LPEP

prior are roughly in agreement for all variables

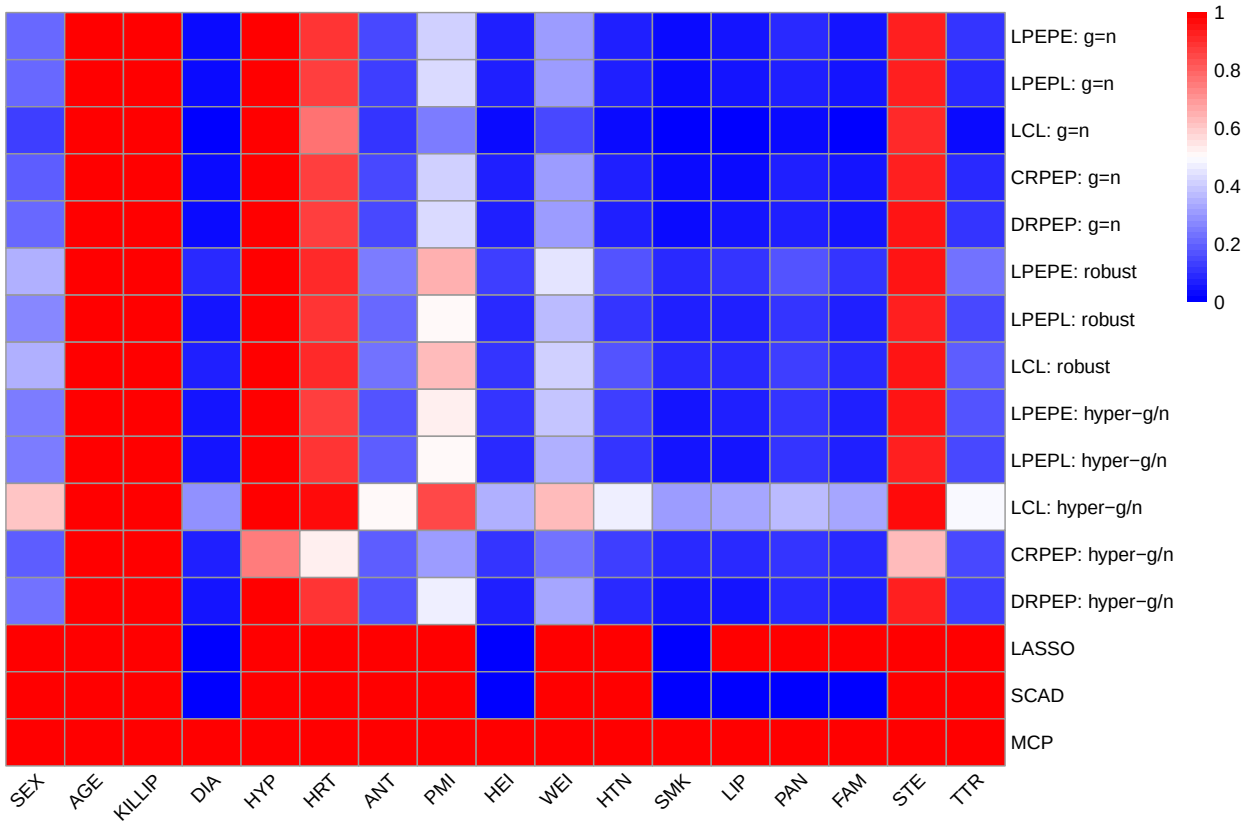


Figure 4.3: Marginal posterior inclusion probabilities (PIPs) for `GUSTO-I` dataset (Bayesian procedures) and variables included in the model (penalized likelihood methods).

We also compare the different procedures in terms of their out-of-sample predictive performance. For this purpose, we performed a 10-fold cross-validation study. Table 4.6 presents the average value of four different metrics across all 10 folds. The metrics we rely on are the same ones employed in Li and Clyde (2018): the area under the ROC curve (AUC), the Calibration Slope (CS), the Logarithmic Score (LS) and the Brier score (BRIER). AUC and CS allow us to evaluate the methods in terms of discrimination and calibration. In both cases, scores closer to 1 indicate better performance. On the other hand, LS and BRIER measure the predictive accuracy of methods; in both cases lower scores indicate better performance.

Most methods perform similarly under these metrics. The main exceptions are both versions of CRPEP and DRPEP, which seem to substantially underperform across all metrics. LPEP procedures slightly outperform other methods in terms of AUC and CS. On the other hand, LASSO seems to slightly outperform all Bayesian procedures in terms of LS and Brier score, but at the cost of selecting much denser models.

Table 4.6: Average prediction accuracy measures in a 10-fold cross validation study for `GUSTO-I` dataset; Bold represents group maximum for AUC, for CS closest to one, and group minimum for LS and Brier score; * represents the corresponding best score among all methods.

		AUC	CS	LS	BRIER
$\delta = n$	LPEPE	**0.8324***	**0.9971***	**0.1824**	0.0496
	LPEPL	**0.8324***	1.0082	**0.1824**	**0.0495**
	LCL	0.8300	0.9931	0.1831	0.0497
	CRPEP	0.7789	1.0578	0.1965	0.0521
	DRPEP	0.7790	1.0569	0.1963	0.0521
$\delta \sim$ robust	LPEPE	**0.8322**	**1.0129**	0.1822	**0.0495**
	LPEPL	0.8320	1.0239	**0.1820**	**0.0495**
	LCL	0.8316	0.9804	0.1822	**0.0495**
$\delta \sim$ hyper g/n	LPEPE	0.8319	**1.0074**	0.1823	0.0495
	LPEPL	**0.8322**	1.0197	0.1821	0.0495
	LCL	0.8311	1.0109	**0.1818**	**0.0493**
	CRPEP	0.7956	1.1677	0.1951	0.0522
	DRPEP	0.7800	1.0571	0.1961	0.0520
	LASSO	**0.8305**	**1.0369**	**0.1816***	**0.0492***
	SCAD	0.8243	0.9135	0.1838	0.0496
	MCP	0.8250	0.9196	0.1838	0.0496

4.8 Discussion

The results from our theoretical and empirical studies show that the LPEP approach to creating non-informative priors for logistic regression is superior to existing techniques, both in terms of model selection and of parameter estimation. The differences are particularly striking when comparing the LPEP with the original CRPEP and DRPEP approaches proposed in Fouskakis et al. (2018), and for heavy-tailed versions of the priors of the mixtures of g-priors introduced in Li and Clyde (2018). When compared against the CRPEP and DRPEP priors (both of which require the use of Reversible Jump MCMC schemes), LPEP priors substantially reduce the computational burden associated with the uses of imaginary samples. Furthermore, our empirical analyses show that the results generated by the CRPEP and DRPEP can differ substantially from each other and from the consensus of other Bayesian and non-Bayesian methods, and that they can can be greatly affected by the choice of hyperpriors.

We were surprised by the poor behavior of the heavy-tailed versions of LCL and LCE procedures in some of our simulation studies. One point to note is that the setup of the simulations in Section 4.6 ($n = 500$, $p = 100$) was only briefly studied in Li and Clyde (2018). Indeed, most of the simulation studies in Li and Clyde (2018) focus on settings involving fewer covariates ($p = 20$). In this lower dimensional setting, LCL and LCE behave quite well (see Appendix C.8). A comparison of the results for LCL against its exact version (LCE), as well as those under the unit information ($\delta = n$) with those under the robust and hyper-g/n versions of the procedure, suggests these results are driven by a combination of sensitivity to the choice of hyperprior for δ and issues with the way `BAS` integrates over δ. Interestingly, the sensitivity to the hyperprior does not seem to be present for the LPEP procedures. We believe that this stability represents a key advantage of our method.

In addition to providing a mechanism to improve computational efficiency, the prior-based BIC developed in Section 4.5.2 also provides useful intuition about the working of "trained" priors such as the EP and PEP priors. As Equations (4.10) illustrates, approaches

based on training sample average information in the observed data with that coming from the imaginary training samples, with the exponent δ controlling the relative weight of the information coming from training samples.

Chapter 5

DIFFERENTIAL SHRINKAGE BLOCK-g PRIORS FOR LINEAR REGRESSION

5.1 Introduction

As seen in Chapter 2, Bayesian model selection methods based on (mixture of) g-priors (Zellner, 1986a; Liang et al., 2008) of the form,

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\alpha, \sigma^2, g \sim \mathcal{N}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid 0, g\sigma^2(X_{\boldsymbol{\gamma}}^T X_{\boldsymbol{\gamma}})^{-1}), \tag{5.1}$$

with $\pi_{\boldsymbol{\gamma}}(\alpha, \sigma^2) \propto \sigma^{-2}$, $g|\boldsymbol{\gamma}, \tau \sim p(g \mid \boldsymbol{\gamma}, \tau)$ and $\tau \sim p(\tau)$ performs well across statistical tasks when compared to other methods for variable selection and have lots of desirable properties like information consistency and asymptotic model consistency. However, Som et al. (2016) showed that use of a single shrinkage parameter makes g-priors susceptible to the conditional Lindley's paradox (CLP). The CLP states that when comparing nested models, if at least one of the regression coefficients common to both models is large relative to other coefficients in the bigger model, the Bayes factor will place too much weights on the smaller model irrespective of the data generating model. Additionally, using single shrinkage parameter g, prevents differential shrinkage of weak signals compared to strong signals.

To be more concrete, consider the two models,

$$\mathcal{M}_1 : \boldsymbol{Y} = \mathbf{1}_n\alpha + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}, \quad \mathcal{M}_2 : \boldsymbol{Y} = \mathbf{1}_n\alpha + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{5.2}$$

with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $n \times p_1$ and $n \times p_2$ dimensional matrices, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathbb{R}^p$ and $p = p_1 + p_2$. Consider a sequence of linear models with a fixed n and p defined by

$$\Psi_N = (\boldsymbol{X}_1, \boldsymbol{X}_2, \alpha, \boldsymbol{\beta}_1(N), \boldsymbol{\beta}_2, \epsilon) \tag{5.3}$$

where $\|\boldsymbol{\beta}_1(N)\| \to \infty$ as $N \to \infty$. Then, Som et al. (2016) showed that under the asymptotic regime discussed in (5.3) and hyper-g prior, log Bayes factor of $\mathcal{M}_2$ to $\mathcal{M}_1$, denoted by $log(BF_{2:1}) = log(BF(\mathcal{M}_2 : \mathcal{M}_1)) \to 0$, irrespective of the data. To illustrate the paradox empirically, let $p_1 = p_2 = 1$, $\alpha = 0.5, \beta_2 = 1$, $\sigma^2 = 1$ and $\beta_1 \in \{1, 3, 5, 10, 20, 40, 60, 100\}$. X_i are drawn independently from $\mathcal{N}(0, 1)$ and normalized to have mean 0 and standard deviation 1. We generate 100 datasets from $\mathcal{M}_2$ for each of the β_1 values and fit BMA framework with g-prior on coefficients given by (5.1) and a hyper-g prior on parameter g. Figure 5.1 shows that as β_1 increases, $log(BF_{2:1})$ averaged over 100 datasets decreases (left) and the posterior mean of β_2, denoted by $\hat{\beta}_2$, shrinks towards zero (right).

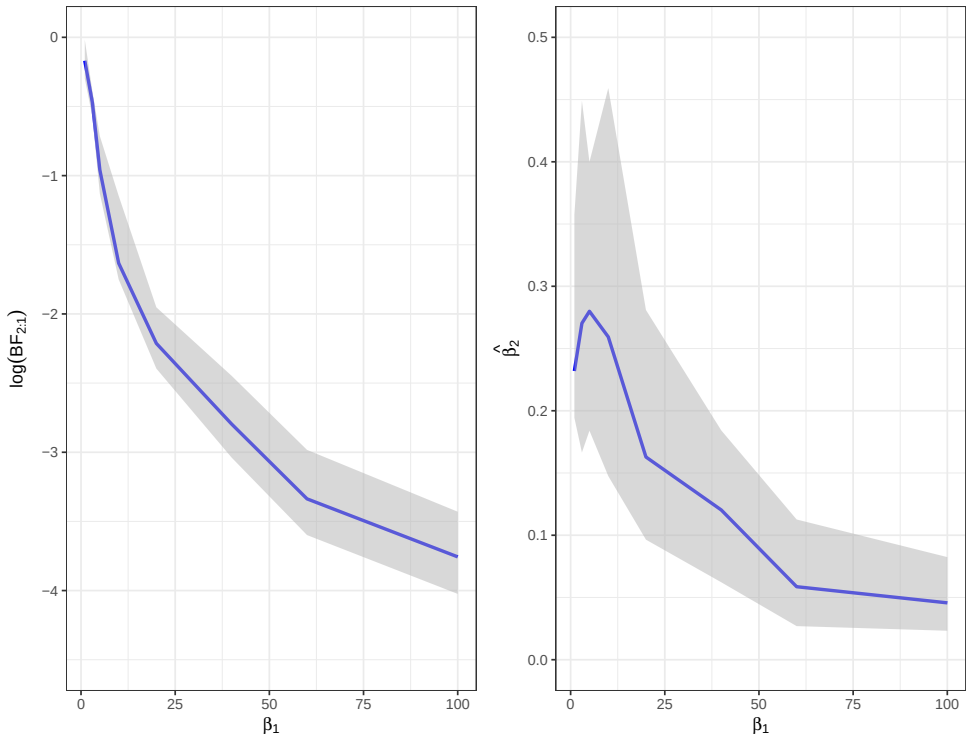


Figure 5.1: Empirical illustration of the conditional Lindley paradox under hyper-g prior; Left displays $log(BF_{2:1})$ and right figure displays $\hat{\beta}_2$ as β_1 increases under the asymptotic regime described in (5.3).

Mixtures of g-priors induce sparsity in the model coefficients by zeroing out some of the model coefficients and estimating a single parameter to shrink signal coefficients. An alternative approach, as discussed in Chapter 1, is to specify heavy-tailed global-local shrinkage (GLS) priors for coefficients defined in (1.3). Unlike g-priors, GLS priors assume that coefficients are independent apriori. It specifies a global shrinkage parameter τ that controls the global sparsity in a data adaptive way and a local shrinkage parameter per coefficient that allows for differential shrinkage by significantly shrinking noise components while leaving large signals unshrunk. However, since they are absolutely continuous distributions on parameter space, the posterior samples drawn under these priors are never identical to zero. Variable selection under these models can only be obtained by ad-hoc thresholding procedures.

To bridge the gap between the two Bayesian paradigms, we propose a unified framework for parameter priors in Section 5.2 that discusses connections between several existing parameter priors and showcases several existing continuous shrinkage priors, classical g-priors and block hyper-g priors as special cases of our framework. In Section 5.3, we develop Dirichlet process (DP) block-g priors for variable selection that allows for simultaneous partitioning of covariates for differential shrinkage and perform variable selection. Section 5.4 discusses theoretical properties of the prior and how this data-adaptive blocking procedure avoids CLP. We then discuss an efficient MCMC procedure that implements DP block-g procedure under a wide variety of priors on g. Sections 5.6 and 5.7 presents empirical results from simulation studies as well as two real datasets. Finally, Section 5.8 discusses our results and future directions for research.

5.2 Unifying continuous shrinkage priors and g-priors

Consider the model selection problem described in (1.1). Given the covariate matrix $\boldsymbol{X}$ and response vector y, our goal is to perform model selection while accounting for model uncertainty in presence of p possible predictors.

To account for the correlation structure of covariates in the prior while still allowing for

differential shrinkage of coefficients, we propose the global local (GL) g-priors of the form

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \boldsymbol{G}, \boldsymbol{\gamma}, \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{G}_{\boldsymbol{\gamma}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{1/2}), \quad \pi_{\boldsymbol{\gamma}}(\alpha, \sigma^2) \quad \propto \sigma^{-2}, \tag{5.4}$$

where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ is a binary vector of length p such that for all $j \in \{1, \ldots, p\}, \gamma_j = 1$ if the j-th variable is included in the model (i.e., if β_j is different from zero) and $\gamma_j = 0$ otherwise. Let $S_{\boldsymbol{\gamma}} = \{i : \gamma_i = 1; i = 1, \ldots, p\}$, then $\boldsymbol{G}_{\boldsymbol{\gamma}} = \text{diag}\{g_i : i \in S_{\boldsymbol{\gamma}}\}$ is a $p_{\boldsymbol{\gamma}} \times p_{\boldsymbol{\gamma}}$ diagonal matrix where $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ and $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite matrix. $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denote the $p_{\boldsymbol{\gamma}} \times 1$ sub-vector of $\boldsymbol{\beta}$, that include only those rows and columns for which the corresponding γ_j is equal to 1. Finally, we specify a Jeffreys' prior for the shared parameters (α, σ^2) and $g_i, i \in S$ are independently and identically distributed, i.e. $g_i \sim^{i.i.d} p(g|\tau, \boldsymbol{\gamma})$ where $p(g_i|\tau, \boldsymbol{\gamma})$ is the hyper-prior on g_i and τ is a hyperparameter.

Similar to Zellner's g-prior, GL g-priors are computationally efficient since all marginal likelihoods given $\boldsymbol{G}$ are available in closed form. Under model specified in (1.1) and prior specified in (5.4), the marginal likelihood is given by

$$p(\boldsymbol{Y}|\boldsymbol{G}, \boldsymbol{\gamma}) = \frac{\Gamma\left(\frac{n-1}{2}\right)}{\sqrt{n}(\sqrt{\pi})^{n-1}} \frac{[\boldsymbol{Y}^T(\boldsymbol{I} + \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{1/2} \boldsymbol{X}_{\boldsymbol{\gamma}}^T)^{-1} \boldsymbol{Y} - n\bar{\boldsymbol{Y}}^2]^{-\frac{n-1}{2}}}{|\boldsymbol{I} + \boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{1/2} \boldsymbol{X}_{\boldsymbol{\gamma}}^T|^{1/2}}. \tag{5.5}$$

Several existing methods can be considered special cases of our framework. When $\boldsymbol{G}_{\boldsymbol{\gamma}} = g\boldsymbol{I}_{p_{\boldsymbol{\gamma}}}$, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = (\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$ and $\boldsymbol{\gamma} \sim p(\boldsymbol{\gamma})$ where $p(\boldsymbol{\gamma})$ denotes a non-degenerate prior on model space (for e.g. uniform or beta-binomial prior), then we recover various versions of classical (mixtures of) g-priors defined in Zellner and Siow (1980) and Liang et al. (2008). Note that under this setting, (5.5) reduces to marginal likelihood under Zellner's g-prior (see Equation (5) of Liang et al., 2008). On the other hand, when $\boldsymbol{\gamma} = \boldsymbol{\gamma}_F = (1, 1, \ldots, 1)$, i.e., when we fix our model to the full model that includes all potential regressors, and $\boldsymbol{\Sigma} = \boldsymbol{I}_p$ a $p \times p$ dimensional identity matrix, the prior on coefficients in (5.4), along with priors on g_i and τ, reduces to an independent scale mixture of normals given by

$$\beta_i \sim \mathcal{N}(0, \sigma^2 g_i) \quad g_i \sim p(g_i|\tau) \quad \tau \sim p(\tau)$$

for $i = 1, \ldots, p$. This is the continuous shrinkage prior framework discussed in (1.3) where g_i (under appropriate choices of mixing density $p(g_i|\tau)$) plays the same role as local shrinkage

parameter λ_i^2. For example, when the mixing density is exponential, we recover the Bayesian LASSO (Park and Casella, 2008; Hans, 2009). Similarly, when the mixing density is half-Cauchy, then we recover the Horseshoe prior (Carvalho et al., 2010). A more extensive but non-exhaustive list of examples includes the Student t-prior (Tipping, 2001), Normal-Exponential-Gamma prior (Griffin and Brown, 2005), Normal-Gamma prior (Brown and Griffin, 2010), the Dirichlet-Laplace prior (Bhattacharya et al., 2015), global-local shrinkage priors (Polson and Scott, 2012), Horseshoe+ prior, (Bhadra et al., 2017), Beta-prime prior (Bai and Ghosh, 2018), and tail-adaptive shrinkage prior (Lee et al., 2020).

The previous discussion suggests that GL g-priors provide a bridge between continuous shrinkage priors and mixtures of g-priors by allowing differential shrinkage for different covariates while incorporating the correlation between predictors into the prior and performing model selection. However, adding p new parameters can lead to a computationally intensive procedure, specially when p is large (e.g., see Section 5.6). A natural remedy is to block covariates in a given model so that members of same block share the same local shrinkage parameter. This approach has been recently discussed in Som et al. (2014) and Boss et al. (2023) , who suggest ad-hoc ways to block covariates based on correlation between covariates or domain specific knowledge. In practice, however, these guidance are difficult to implement and the number of blocks as well as the block membership of covariates are rarely known. Additionally, incorrect grouping of covariates can lead to poor model selection performace (again, see Section 5.6). Hence, there is a need to adaptively learn the blocking structure of covariates based on the data.

5.3 Dirichlet Process block-g priors

One way to group similar coefficients is to specify a Bayesian non-parametric prior for g_i using a Dirichlet Process (DP) mixture model as follows

$$
\begin{aligned}
\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \sigma^2, \boldsymbol{G}, \boldsymbol{\gamma}, \boldsymbol{X} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{G}_{\boldsymbol{\gamma}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} \boldsymbol{G}_{\boldsymbol{\gamma}}^{1/2}), \\
\pi_{\boldsymbol{\gamma}}(\alpha, \sigma^2) &\propto \sigma^{-2}, \\
g_i | H &\sim H, \quad i \in S_{\boldsymbol{\gamma}}, \\
H &\sim DP(a_0, H_0),
\end{aligned}
\tag{5.6}
$$

where $DP(a_0, H_0)$ denote a Dirichlet Process with concentration parameter a_0. As $a_0 \to 0$, the prior implies a shrinkage factor for all covariates like classical g-priors while as $a_0 \to \infty$, the prior implies a different shrinkage factor per cofficient similar to GL-g priors discussed in Section 5.2. Hence, this approach avoids fixing the number of clusters, allowing us to interpolate between classical g-priors and the GL g-priors defined above. This means that we can share statistical strength across similar covariates through the common block shrinkage parameters and adaptively learn the level of differential shrinkage required based on data.

Note that above prior offers two levels of classification. $\boldsymbol{\gamma}$ allows us to disentangle signals from noise components, while clustering induced by shared shrinkage parameter allows us to separate signals of different magnitude. In the sequel, we let $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = (\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$ and refer our prior specification in (5.6) as DP block-g priors.

It is important to note that Bayesian non-parametric tools have been previously employed in sparse Bayesian regression problems and robust modeling. Shahbaba and Johnson (2013) proposed a Bayesian non-parametric prior for discovering differentially expressed genes with structure similar to (5.6) with model fixed at full model, $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ fixed to be identity matrix and base measure chosen to log-normal distribution. Recently proposed discrete mixture of continuous mixtures (Denti et al., 2021) can similarly be considered as a special case of our method with $\boldsymbol{\gamma} = \boldsymbol{\gamma}_F, \boldsymbol{\Sigma} = \boldsymbol{I}_p$ and H_0 being the Half-Cauchy density. However, unlike our specification, neither of the above priors do model selection. Furthermore, theoretical investigation of both these DP mixture based approaches have been limited. In the context of robust graphical modeling, Finegold and Drton (2014) proposed Dirichlet t-distributions to identify batches of variables taking on extreme values in a high dimensional random variable. Dirichlet t prior has similar structure to (5.6) when the base density H_0 is Gamma$(\nu/2, \nu/2)$

however its application for variable selection in linear regression has not been explored.

The centering measure H_0 can be any existing hyper priors specified for g (see for e.g., Liang et al., 2008) or priors on scale under continuous shrinkage priors (see for e.g., Bhadra et al., 2019). In what follows, we will adopt a beta-prime density, denoted by $BP(a, b, \tau^2)$ for the centering measure H_0:

$$f(g|\tau^2) = \frac{1}{(\tau^2)^{b+1}B(a+1, b+1)} g^b \left(1 + \frac{g}{\tau^2}\right)^{-a-b-2} \mathbb{1}_{(0,\infty)}(g),$$

with $a > -1, b > -1$ and $\tau^2 > 0$. Several existing priors on scale, both continuous shrinkage and g-priors can be considered special cases of the beta-prime priors. For example, when $a = b = 0$, we retrieve the hyper-g prior when $\tau^2 = 1$ and the hyper-g/n prior with $\tau^2 = n$ (see Liang et al., 2008 and Cui and George, 2008). When $a = b = -0.5$, we recover the half-Cauchy density that is used in the Horseshoe prior (Carvalho et al., 2010) when $\tau \sim C^+(0, 1)$.

To complete the model specification, we assume a Beta-binomial$(1, 1)$ prior on model space that was recommended through the simulation study in Chapter 3.

5.4 Properties of the DP block-g prior

5.4.1 Tail behavior

To understand tail behavior, we introduce the notion of regularly varying tails. A positive, measurable function is said to be regularly varying at infinity with index $\omega \in \mathbb{R}$ if $\lim_{x\to\infty} f(tx)/f(x) = t^\omega$, for all $t > 0$.

Under the DP block-g priors defined in (5.6) with $\sigma^2 = \tau^2 = 1$ and centering measure $H_0 = BP(a, b, 1)$, the marginal prior on jth coefficient included in the model, denoted by $\beta_{\boldsymbol{\gamma},j}$, is given by

$$f(\beta_{\boldsymbol{\gamma},j}|\boldsymbol{\gamma}, \boldsymbol{X}) = \int \mathcal{N}(0, g\boldsymbol{\Sigma}_{\boldsymbol{\gamma},jj}) BP(g; a, b, 1) dg$$

where $BP(g; a, b, 1)$ denote the Beta-prime density on scale parameter g. This has same structure as the marginal prior under Group Inverse-Gamma Gamma (GIGG) prior proposed by Boss et al. (2023). Then, the index of regular variation of the marginal prior is $\omega = -2b-3$

(see Theorem 2.1 of Boss et al. (2023)). This implies that our implied marginal prior has heavy (polynomial) tails. As discussed in 4, this implies that marginal priors implied by DP block-g priors are robust in the sense of having bounded influence in the case of likelihood-prior conflict.

5.4.2 Conditional Lindley paradox

We return to the conditional Lindley paradox discussed in Section 5.1 and show that DP block-g priors avoid the CLP. Note that Som (2014) showed that classical g-priors suffer from CLP because of use of single shrinkage factor. Subsequently, Som et al. (2014) showed that if the blocking structure of covariates is known in linear regression, we can avoid CLP by using a different shrinkage parameter for each block. We extend these results, under certain assumptions on design, to cases when blocking structure is unknown.

Let $\mathcal{P}$ denote the set of all partitions of $[p] = \{1, \ldots, p\}$ where p is the total number of covariates. A partition is denoted by $\rho = \{S_1, \ldots, S_K\}$ where K, in our context, denotes the number of blocks of covariates.

Definition 5.4.1 (Refinement of a partition)**.** Let $\rho = \{S_1, \ldots, S_K\}$ and $\rho' = \{S'_1, \ldots, S'_{K'}\}$ denote two unique partition of the set $[p]$ with K and K' unique blocks respectively such that $1 \leq K' \leq K \leq p$. Then, ρ is a refinement of ρ', denoted by $\rho \prec \rho'$, if for every $S_k \in \rho, k \in \{1, \ldots, K\}$, there exist a $S'_j \in \rho', j \in \{1, \ldots, K'\}$ such that $S_k \subseteq S'_j$.

Based on $\mathcal{M}_2$ in (5.2), the true partition of p covariates is given by $\rho^{(0)} = \{S_1^{(0)}, S_2^{(0)}\}$ where $S_1^{(0)} = \{1, \ldots, p_1\}$ and $S_2^{(0)} = \{p_1 + 1, \ldots, p_2\}$.

Lemma 5. Consider the model

$$\mathcal{M} : \boldsymbol{Y} = \mathbf{1}_n \alpha + \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $n \times p_1$ and $n \times p_2$ dimensional matrices, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T \in \mathbb{R}^p$ and $p = p_1 + p_2$. Assume that $\boldsymbol{X}$ is an orthogonal matrix such that $\boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{I}_p$ and sample variance σ^2 is known. Let $\rho = \{S_1, \ldots, S_K\}$ denote an

arbitrary partition of set $[p]$. Under the DP block-g prior discussed in (5.6) with BP(a,b,1) prior as the base measure H_0 and asymptotic regime discussed in (5.3),

1. For $\rho \not\asymp \rho^{(0)}$, $\frac{P(\boldsymbol{Y}|\rho,\mathcal{M},\sigma^2)}{P(\boldsymbol{Y}|\rho^{(0)},\mathcal{M},\sigma^2)} \to 0$, and

2. For $\rho \prec \rho^{(0)}$, $\frac{P(\boldsymbol{Y}|\rho,\mathcal{M},\sigma^2)}{P(\boldsymbol{Y}|\rho^{(0)},\mathcal{M},\sigma^2)} \to c > 0$.

Proof. For all $k \in \{1,\ldots,K\}$, let $|S_k \cap S_1^{(0)}| = m_{1,k}$ and $|S_k \cap S_2^{(0)}| = m_{2,k}$ respectively denote the number of covariates from true partition sets $S_1^{(0)}$ and $S_2^{(0)}$ in a block k of partition $\rho \in \mathcal{P}$. Note that $0 \leq m_{1,k}, m_{2,k} \leq |S_k|$ and $m_{1,k} + m_{2,k} = |S_k|$ and $\sum_k |S_k| = p_1 + p_2 = p$ where $|.|$ denotes the cardinality of a set. Note that for true partition $\rho^{(0)}$, $m_{1,1} = p_1, m_{1,2} = 0, m_{2,1} = 0$ and $m_{2,2} = p_2$.

We start by deriving the marginal likelihood of the data under model $\mathcal{M}$ and DP block-g prior with BP$(a,b,1)$ density for $g_k, k = 1,\ldots,K$ and the assumptions, i) $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$ and ii) sample variance σ^2 is known as follows

$$
\begin{aligned}
P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M}) &= \int\cdots\int f(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}},\alpha,\sigma^2,\boldsymbol{X},\mathcal{M}) f(\boldsymbol{\beta} \mid \sigma^2,\boldsymbol{G},\rho,,\mathcal{M},\boldsymbol{X})\pi(\alpha)\prod_{k=1}^{K} p(g_k)d\boldsymbol{\beta}dg_k d\alpha \\
&\propto \prod_{k=1}^{K}\left(\int_0^{\infty} g_k^b(1+g_k)^{-a-b-2-\frac{|S_k|}{2}} \exp\left(\frac{1}{2\sigma^2}\frac{g_k}{1+g_k}\left\|\hat{\boldsymbol{\beta}}_{MLE,k}\right\|^2\right) dg_k\right),
\end{aligned}
$$

where a and b are hyper-parameters associated with BP prior, $|S_k|$ is the size of block k and $\hat{\boldsymbol{\beta}}_{MLE,k} = \boldsymbol{X}_k^T\boldsymbol{Y}$ denote the MLE for block k, since $\boldsymbol{X}_k^T\boldsymbol{X}_k = \boldsymbol{I}$. Using the transformation $u_k = \frac{g_k}{1+g_k}$, we can write above expression as

$$
\begin{aligned}
P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M}) &\propto \prod_{k=1}^{K}\left(\int_0^1 u_k^b(1-u_k)^{a+\frac{|S_k|}{2}} \exp\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}\right\|^2 u_k}{2\sigma^2}\right) du_k\right) \\
&\propto \prod_{k=1}^{K} M\left(b+1, a+b+\frac{|S_k|}{2}+2, \frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}\right\|^2}{2\sigma^2}\right) B\left(b+1, a+\frac{|S_k|}{2}+1\right).
\end{aligned}
$$

Therefore, we can write

$$\lim_{N\to\infty}\frac{P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M})}{P(\boldsymbol{Y}|\sigma^2,\rho^{(0)},\mathcal{M})}\propto\lim_{N\to\infty}\frac{\prod_{k=1}^{K}M\left(b+1,a+b+\frac{|S_k|}{2}+2,\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|^2}{2\sigma^2}\right)}{\prod_{k=1}^{2}M\left(b+1,a+b+\frac{|S_k^{(0)}|}{2}+2,\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|^2}{2\sigma^2}\right)}$$
$$\propto\lim_{N\to\infty}\frac{\prod_{k:m_{1,k}\neq 0}M\left(b+1,a+b+\frac{|S_k|}{2}+2,\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|^2}{2\sigma^2}\right)}{M\left(b+1,a+b+\frac{p_1}{2}+2,\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,1}(N)\right\|^2}{2\sigma^2}\right)}.$$

The second proportionality follows from the fact that the blocks with $m_{1,k}=0$, i.e. no element from $S_1^{(0)}$, do not depend on N and hence can be dropped as constant. For large values of N, we can then approximate $M(a_0,b_0,z)\approx\Gamma(b_0)\frac{\exp(z)z^{a_0-b_0}}{\Gamma(a_0)}$ (see 13.5.1 on page 508 of Abramowitz et al., 1988). This approximation is true for large values of z when $a_0>0$ which are both true in our setting. We can then re-write the above ratio as

$$\lim_{N\to\infty}\frac{P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M})}{P(\boldsymbol{Y}|\sigma^2,\rho^{(0)},\mathcal{M})}\propto\lim_{N\to\infty}\frac{\prod_{k:m_{1,k}\neq 0}\exp\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|^2}{2\sigma^2}\right)\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|^2}{2\sigma^2}\right)^{-\frac{a+|S_k|+1}{2}}}{\exp\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,1}(N)\right\|^2}{2\sigma^2}\right)\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,1}(N)\right\|^2}{2\sigma^2}\right)^{-\frac{a+p_1+1}{2}}}.$$

Since the exponential terms in both numerator and denominator are of same order with respect to N, we can simplify the expression as

$$\lim_{N\to\infty}\frac{P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M})}{P(\boldsymbol{Y}|\sigma^2,\rho^{(0)},\mathcal{M})}\propto\lim_{N\to\infty}\frac{\prod_{k:m_{1,k}\neq 0}\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|^2}{2\sigma^2}\right)^{-\frac{a+|S_k|+1}{2}}}{\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,1}(N)\right\|^2}{2\sigma^2}\right)^{-\frac{a+p_1+1}{2}}}.$$

Now, we can analyze the above ratio for refinements and non-refinements of $\rho^{(0)}$ separately. For k such that $m_{1,k}\neq 0$, we can approximate $\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|=\mathcal{O}(N)$. For a non-refinement, $\rho\not\prec\rho^{(0)}$, there exists at least one set S_k such that $m_{1,k}\neq 0$ and $m_{2,k}\neq 0$. Since $S_k=m_{1,k}+m_{2,k}$, this implies that

$$p_1=\sum_k m_{1,k}=\sum_{k:m_{1,k}\neq 0}m_{1,k}<\sum_{k:m_{1,k}\neq 0}|S_k|.$$

We can then show that

$$\lim_{N\to\infty}\frac{P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M})}{P(\boldsymbol{Y}|\sigma^2,\rho^{(0)},\mathcal{M})}\propto\lim_{N\to\infty}\frac{\prod_{k:m_{1,k}\neq 0}\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|^2}{2\sigma^2}\right)^{-\frac{a+|S_k|+1}{2}}}{\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,1}(N)\right\|^2}{2\sigma^2}\right)^{-\frac{a+p_1+1}{2}}}$$
$$\propto\lim_{N\to\infty}\frac{\mathcal{O}(N^{-\sum_{k:m_{1,k}\neq 0}|S_k|})}{\mathcal{O}(N^{-p_1})}\to 0.$$

The limit follows by the fact that $p_1\leq\sum_{k:m_{1,k}\neq 0}|S_k|$ and therefore the numerator shrinks to 0 faster than denominator. However, for a refinement, since there is no block with entries both from $S_1^{(0)}$ and $S_2^{(0)}$,

$$p_1=\sum_k m_{1,k}=\sum_{k:m_{1,k}\neq 0}m_{1,k}=\sum_{k:m_{1,k}\neq 0}|S_k|.$$

Therefore,

$$\lim_{N\to\infty}\frac{P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M})}{P(\boldsymbol{Y}|\sigma^2,\rho^{(0)},\mathcal{M})}\propto\lim_{N\to\infty}\frac{\prod_{k:m_{1,k}\neq 0}\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,k}(N)\right\|^2}{2\sigma^2}\right)^{-\frac{a+|S_k|+1}{2}}}{\left(\frac{\left\|\hat{\boldsymbol{\beta}}_{MLE,1}(N)\right\|^2}{2\sigma^2}\right)^{-\frac{a+p_1+1}{2}}}$$
$$\propto\lim_{N\to\infty}\frac{\mathcal{O}(N^{-\sum_{k:m_{1,k}\neq 0}\frac{|S_k|}{2}})}{\mathcal{O}(N^{-\frac{p_1}{2}})}\to c>0.$$

□

Theorem 6. Consider the models $\mathcal{M}_1$ and $\mathcal{M}_2$ in (5.2). Assume that $\boldsymbol{X}$ is an orthogonal matrix such that $\boldsymbol{X}^T\boldsymbol{X}=\boldsymbol{I}_p$ and sample variance σ^2 is known. Under the DP block-g prior discussed in (5.6) with $BP(a,b,1)$ prior as the base measure H_0 and asymptotic regime discussed in (5.3), $BF(\mathcal{M}_2:\mathcal{M}_1|\sigma^2)>0$.

Proof. Note that

$$BF(\mathcal{M}_2:\mathcal{M}_1|\sigma^2)=\frac{\sum_{\rho\in\mathcal{P}[p_1+p_2]}P(\rho)P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M}_2)}{\sum_{\rho\in\mathcal{P}[p_1]}P(\rho)P(\boldsymbol{Y}|\sigma^2,\rho,\mathcal{M}_1)},$$

where $\mathcal{P}[p_1+p_2]$ denotes the set of all partitions of $[p_1+p_2]$ and $P(\rho) \propto a_0^K \prod_{k=1}^K (|S_k|-1)!$ under a DP prior with concentration parameter a_0. Since $\rho^{(0)}$ is a refinement of itself, $P(\boldsymbol{y}|\sigma^2, \rho^{(0)}, \mathcal{M}_i) > 0$ for $i = 1, 2$. Dividing and multiplying numerator and denominator by $P(\boldsymbol{y}|\sigma^2, \rho^{(0)}, \mathcal{M}_2)$ and $P(\boldsymbol{y}|\sigma^2, \rho^{(0)}, \mathcal{M}_1)$ respectively, we can re-write $BF(\mathcal{M}_2 : \mathcal{M}_1|\sigma^2)$ as

$$BF(\mathcal{M}_2 : \mathcal{M}_1|\sigma^2) = \frac{P(\boldsymbol{y}|\sigma^2, \rho^{(0)}, \mathcal{M}_2)}{P(\boldsymbol{y}|\sigma^2, \rho^{(0)}, \mathcal{M}_1)} \frac{\sum_{\rho \in \mathcal{P}[p_1+p_2]} P(\rho) \frac{P(\boldsymbol{Y}|\sigma^2, \rho, \mathcal{M}_2)}{P(\boldsymbol{Y}|\sigma^2, \rho^{(0)}, \mathcal{M}_2)}}{\sum_{\rho \in \mathcal{P}[p_1]} P(\rho) \frac{P(\boldsymbol{Y}|\sigma^2, \rho, \mathcal{M}_1)}{P(\boldsymbol{Y}|\sigma^2, \rho^{(0)}, \mathcal{M}_1)}}.$$

Note that as $N \to \infty$, $\frac{P(\boldsymbol{y}|\sigma^2, \rho^{(0)}, \mathcal{M}_2)}{P(\boldsymbol{y}|\sigma^2, \rho^{(0)}, \mathcal{M}_1)} \to c_1 > 0$. Since, $\frac{P(\boldsymbol{Y}|\sigma^2, \rho, \mathcal{M}_i)}{P(\boldsymbol{y}|\sigma^2, \rho^{(0)}, \mathcal{M}_i)} \to c_{2,i} > 0$ for $i = 1, 2$ and $\rho \prec \rho^{(0)}$ and $P(\rho) > 0$ for all partition ρ,

$$\lim_{N \to \infty} BF(\mathcal{M}_2 : \mathcal{M}_1|\sigma^2) > 0.$$

To illustrate how DP block-g priors avoid CLP, we return to our empirical setting discussed in Section 5.1 and fit BMA framework with DP block-g prior with centering measure H_0 to be hyper-g prior and $\alpha_0 \sim Gamma(1, 1)$. Figure 5.2 shows that, unlike classical g-priors, $log(BF_{2:1})$ and $\hat{\beta}_2$ stabilizes as β_1 increases under DP block-g prior. Similar empirical illustration of GL-g priors avoiding CLP is shown in Appendix D.1.

□

5.5 Computation

5.5.1 Markov chain Monte Carlo sampling

An efficient MCMC sampler for variable selection in linear regression under DP block-g priors can be easily derived by resorting to a truncation of DP prior (Ishwaran and James, 2001; Ishwaran and Zarepour, 2002). This approximation allows us to resort to methods for computation of finite mixture models for posterior inference. We augment our prior in (5.6) with latent class membership labels $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^p$ with $\xi_i \in \{1, \ldots, K\}$ where K denotes the maximum number of clusters. Thus, $\xi_i = k$ implies that i-th coefficient is assigned cluster k and has a shrinkage factor g_k^*. When DP concentration measure H_0 is $BP(a, b, \tau^2)$, we can

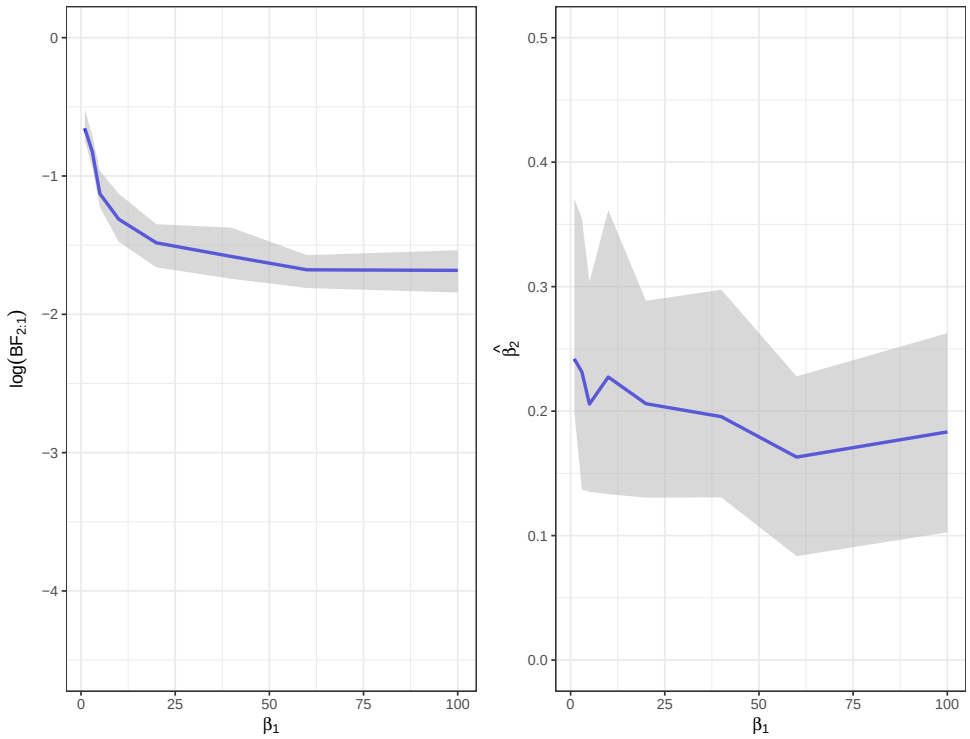


Figure 5.2: Empirical illustration of DP block-g prior with hyper-g prior as centering measure avoiding CLP; Left displays $log(BF_{2:1})$ and right figure displays $\hat{\beta}_2$ as β_1 increases under the asymptotic regime described in (5.3).

re-parameterize model in (5.6) as follows

$$\begin{aligned}
\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \boldsymbol{G}^*, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{X}, \tau^2 &\sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2 {\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} {\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{1/2}), \\
\pi_{\boldsymbol{\gamma}}(\alpha, \sigma^2) &\propto \sigma^{-2}, \\
g^*_k|\boldsymbol{\xi} &\sim BP(a, b, 1), \\
\xi_i|\pi &\sim \sum_{k=1}^{K} \pi_k \delta_k(.), \\
\boldsymbol{\pi} &\sim \text{Dir}\left(\frac{a_0}{K}, \ldots, \frac{a_0}{K}\right),
\end{aligned} \tag{5.7}$$

where $\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}} = \text{diag}\{g^*_{\xi_i} : i \in S_{\boldsymbol{\gamma}}\}$ and $S_{\boldsymbol{\gamma}} = \{i : \gamma_i = 1; i = 1, \ldots, p\}$. We can then use this hierarchical framework to first sample from conditional posterior distribution of $(\boldsymbol{\gamma}|\boldsymbol{G}, \boldsymbol{\xi})$,

using Metropolis-Hastings algorithms implemented to explore the space of models (see for e.g. Section 4.5 of George and McCulloch, 1997). This is followed by sampling other model specific parameters. Using the latent membership labels, the conditional posterior distribution of all parameters except $g_k^*, k = 1, \ldots, K$ are available and correspond to standard family. We design a novel slice sampler using rejection sampler for extended Gamma distribution (Liu et al., 2012) to sample from conditional posterior distribution of shrinkage parameters $g_k^*, k = 1, \ldots, K$. When τ has a standard half-Cauchy prior, we can efficiently sample τ using an augmentation strategy introduced by Makalic and Schmidt (2015). If the concentration parameter a_0 is not fixed, we employ Gamma$(1, 1)$ prior for a_0 and use the parameter augmentation sampling strategy discussed in Section 6 of Escobar and West (1995) to efficiently sample a_0. Further details of the computational algorithm can be seen in Appendix D.2.

5.6 Simulation study

We conduct the following simulation study to compare the model selection, parameter estimation and prediction performance with other existing model selection techniques. The setup of our simulation study is motivated by that in Denti et al. (2021). The simulation study uses a sample size of $n = 500$, with three choices of $p = \{250, 500, 750\}$ corresponding to low, medium and high dimensional datasets observed in practice. For each scenario, the n vectors of predictors are drawn independently from a zero-mean, unit-scale multivariate normal distribution with pairwise correlations given by $cor(x_{i,j}, x_{i,j'}) = r$ for $1 \leq j < j' \leq p$. The corresponding regression coefficients are sampled as follows: $\beta_j \sim \mathcal{N}(0, 100)$ for $j = 1, \ldots, 100$, $\beta_j \sim \mathcal{N}(0, 1)$ for $j = 101, \ldots, 200$ and $\beta_j = 0$ otherwise. We then set $\boldsymbol{Y} = 5 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. We consider two different correlation coefficients $r \in \{0, 0.5\}$ among predictors. More specifically, our simulation study considers six scenarios which vary in terms of sparsity, dimensionality and correlation between predictors.

A total of 100 datasets were generated for each of the six scenarios. We compare multiple existing Bayesian procedures along with several special cases of our prior incorporating

different features. Existing competing methods include g-priors (Liang et al., 2008), Horseshoe (Carvalho et al., 2010) and the Horseshoe mixture (HSM) (Denti et al., 2021). We implement two versions of global-local (GL) g-priors discussed in Section 5.2 with $\tau = 1$ and $\tau \sim \mathcal{C}^+(0,1)$ which allow a different shrinkage parameter per component. Additionally, we also implement two modified versions of block g-priors by Som et al. (2014) using (i) $K = 2$ blocks where block label $\xi_i = 1$ for $i = 1, \ldots, 100$, $\xi_i = 2$ for $i = 101, \ldots, 200$ and other covariates are randomly assigned one of the two blocks and (ii) $K = 3$ blocks where $\xi_i = 1$ for $i = 1, \ldots, 100$, $\xi_i = 2$ for $i = 101, \ldots, 200$ and $\xi_i = 3$ otherwise. Both of the above procedures assumes blocking structure to be known but doesn't require block orthogonality assumptions. Finally, we implement two versions of DP block-g priors discussed in Section 5.1 with $\tau = 1$ and $\tau \sim \mathcal{C}^+(0,1)$ that allow for simultaneous blocking of covariates, model selection and parameter estimation unlike any of the above procedures. For all the different versions of g-priors, we assume the base measure for shrinkage parameter to be Hyper-g prior i.e. BP$(-0.5, 0, 1)$ given by $p(g) \propto (1+g)^{-3/2}$ based on the recommendation in Chapter 2. Additionally, based on the recommendation in Chapter 3, we assume a Beta-Binomial$(1,1)$ prior on model space when $p < n$ and a truncated Beta-Binomial$(1,1)$ truncated to models of size less than $n-2$ for $p > n$ settings. We use the `R` package `BAS-V1.6.1` (Clyde, 2020) to implement g-prior approach, `bayesreg-V1.2` (Makalic and Schmidt, 2016) and implement HSM with code available on Github. For all the methods, we require 15,000 burn-in samples and followed by retaining 5000 posterior draws with a thinning of 5. A brief summary of different features of the above method can be found in Table 5.1.

Comparing g priors with versions of GL g-prior, modified versions of Som's block-g priors and DP block-g-priors allows us to disentangle the effect of using single versus multiple shrinkage parameters and effect of known vs learning blocking structure. Comparing Horseshoe and HSM with versions of g-priors allows us to understand the effect of fixed model versus model averaged solutions and effect of accounting correlation into prior, if any.

We first evaluate the performance of these methods in terms of model selection using F_1 score for the median probability model (MPM), see Figure 5.3. The F_1 score is defined as the

Table 5.1: Summary of methods compared in the study; Methods where $\boldsymbol{\gamma}$ is assigned a non-degenerate prior perform model selection; Note that all the above methods can be considered specials cases of general framework discussed in Section 5.2.

Method	$\boldsymbol{\xi}$ (Cluster labels)	$\boldsymbol{\gamma}$ (Model space)	$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}$ (prior correlation)
Horseshoe	Fixed($\xi_i = i; i = 1, \ldots, p$)	$\boldsymbol{\gamma}_F$	$\boldsymbol{I}_p$
HSM	Adaptive	$\boldsymbol{\gamma}_F$	$\boldsymbol{I}_p$
g-prior	Fixed($\xi_i = 1; i = 1, \ldots, p$)	Beta-Binomial$(1, 1)$	$(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$
Modified Som's block-g	Fixed (domain knowledge)	Beta-Binomial$(1, 1)$	$(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$
GL g-prior	Fixed($\xi_i = i; i = 1, \ldots, p$)	Beta-Binomial$(1, 1)$	$(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$
DP block-g prior	Adaptive	Beta-Binomial$(1, 1)$	$(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{X}_{\boldsymbol{\gamma}})^{-1}$

harmonic mean of proportion of true positives among "selected" covariates (the precision) and the proportion of "selected" covariates among true positive covariates (the recall). The F_1 score ranges between 0 and 1, with a higher value indicating better model selection performance. Note that we have excluded Horseshoe and HSM from this analysis since they don't perform model selection without post-processing posterior samples using ad-hoc thresholding.

In all the scenarios, GL-g priors with fixed $\tau = 1$, classical g-prior and modified Som's Block-g prior with $K = 3$ perform worse compared to other methods. The drastically different performance of the two versions of Som's block-g prior shows evidence on how (potentially) incorrect domain specific blocking structure can lead to concerning results. The performance of GL-g priors improves significantly when τ is random, signifying the importance of adaptive global shrinkage parameter when using a different shrinkage parameter per covariate. In all the scenarios, two versions of DP block-g performs well in terms of model selection performance. The effect of τ being fixed versus random is negligible since DP allows to borrow information across coefficients of similar size.

Next, we compare the block wise mean squared error (MSE) for estimated coefficients of

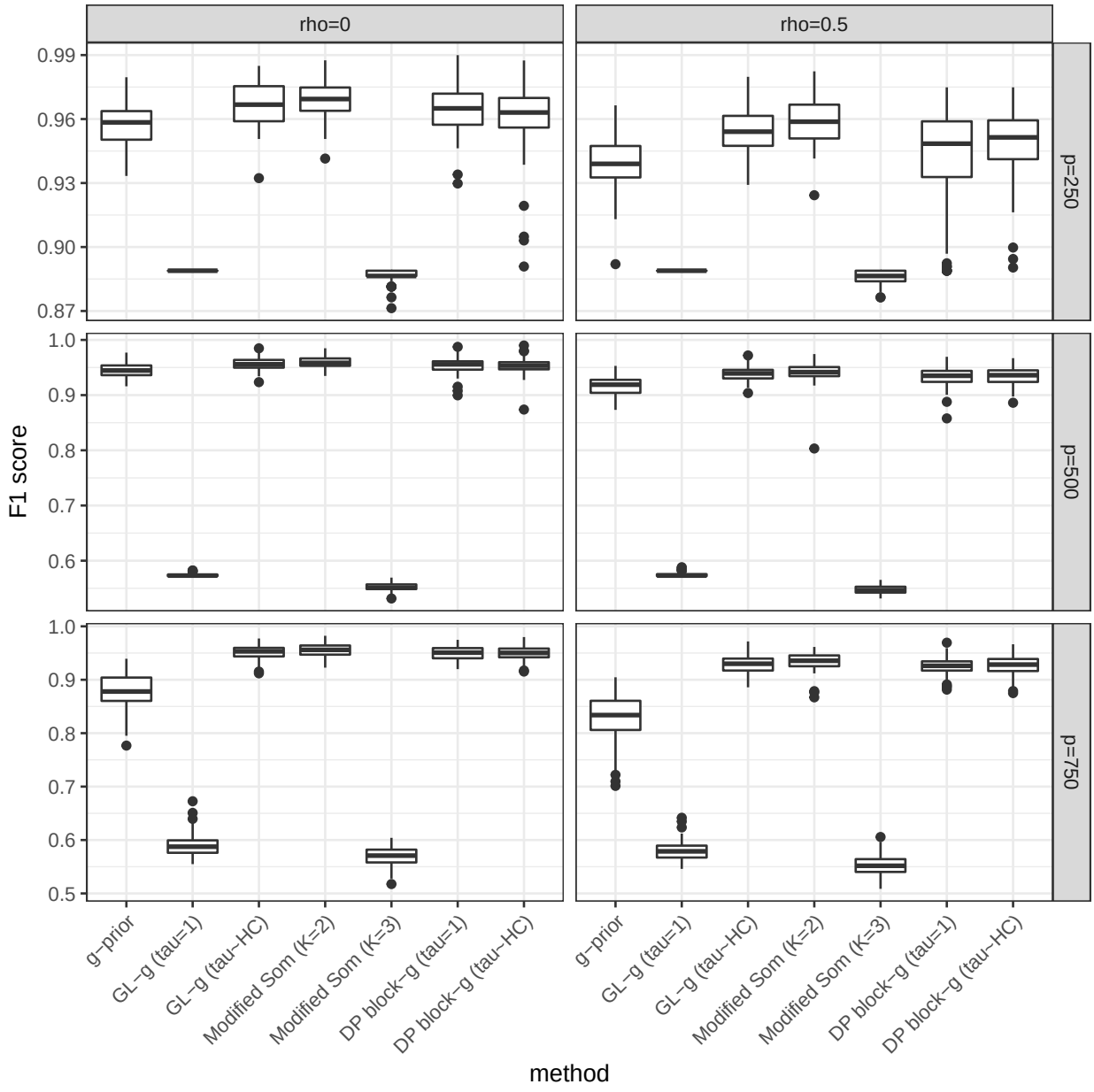


Figure 5.3: F1 score for the MPM estimated by various methods for 100 simulated datasets ($n = 500$) under different scenarios of correlation ($r = 0$: left; $r = 0.5$: right) and different number of possible covariates p in rows ($p \in \{250, 500, 750\}$).

different magnitude across different scenarios. The results can be seen in Table 5.2 and Table 5.3. Note that most methods do well in terms of parameter estimation with the exception of g-prior, Horseshoe and GL-g with fixed τ which perform significantly worse compared to other methods particularly for higher dimensional datasets. We highlight the top four methods with lowest MSE for each block and scenario. Note that Som's block-g prior with $K = 3$ performs the best in terms of MSE across the board but at the expense of poor model selection performance despite knowing the blocking structure as seen in Figure 5.3 potentially due to over-fitting. DP block-g and HSM also perform well in terms of parameter estimation for all three blocks of coeffcients across the scenarios. However, DP block-g is the only method that performs well in parameter estimation while simulataneously blocking the coefficients and performing model selection.

Table 5.2: 1000 times the AMSE for estimated coefficient blocks of different magnitude over 100 replications when $r = 0$; Bold represents top four methods with lowest AMSE.

$n = 500, r = 0$									
p	250			500			750		
Method	Signal 1	Signal 10	Noise	Signal 1	Signal 10	Noise	Signal 1	Signal 10	Noise
g-prior	9.092	43.654	**0.293**	36.540	227.747	0.458	97.249	583.893	0.663
GL-g ($\tau = 1$)	5.188	4.378	1.865	9.418	8.445	3.897	23.623	23.389	8.840
GL-g ($\tau \sim \mathcal{C}^+(0,1)$)	4.054	**3.737**	**0.649**	5.019	4.009	**0.167**	5.737	4.335	**0.128**
Modified Som ($K = 2$)	**3.862**	3.747	**0.543**	**4.654**	**4.006**	**0.143**	**5.072**	**4.244**	**0.091**
Modified Som ($K = 3$)	**3.422**	**3.502**	**0.137**	**3.766**	**3.577**	**0.156**	**4.565**	**4.200**	0.328
DP block-g ($\tau = 1$)	4.016	**3.737**	0.722	**4.912**	**4.005**	**0.208**	**5.536**	**4.283**	**0.130**
DP block-g ($\tau \sim \mathcal{C}^+(0,1)$)	4.063	3.768	0.733	4.980	4.018	0.210	5.815	4.356	**0.132**
HSM	**3.873**	**3.711**	0.992	**4.686**	**3.939**	0.238	**5.102**	**4.113**	0.139
Horseshoe	**3.845**	3.885	2.770	6.810	6.449	2.975	9.167	8.083	2.254

Finally, we compare methods using out-of-sample prediction performance using 100 test datasets for each of the six scenarios in Table 5.4. The test datasets were simulated same as

Table 5.3: 1000 times the AMSE for estimated coefficient blocks of different magnitude over 100 replications when $r = 0.5$; Bold represents top four methods with lowest AMSE.

$n = 500, r = 0.5$									
p	250			500			750		
Method	Signal 1	Signal 10	Noise	Signal 1	Signal 10	Noise	Signal 1	Signal 10	Noise
g-prior	15.625	46.921	**0.485**	49.637	222.990	**0.490**	139.121	600.685	0.811
GL-g ($\tau = 1$)	10.691	8.911	3.636	19.510	16.921	7.372	46.415	44.308	15.123
GL-g ($\tau \sim \mathcal{C}^+(0,1)$)	8.298	7.595	**1.504**	11.161	8.377	**0.473**	14.190	9.702	**0.316**
Modified Som ($K = 2$)	8.532	8.047	**1.134**	13.522	12.087	0.899	**13.743**	11.257	**0.408**
Modified Som ($K = 3$)	**6.825**	**7.007**	**0.260**	**7.553**	**7.175**	**0.300**	**9.556**	**8.650**	0.701
DP block-g ($\tau = 1$)	**8.113**	**7.555**	1.779	**10.813**	**8.337**	0.549	**13.565**	**9.616**	**0.378**
DP block-g ($\tau \sim \mathcal{C}^+(0,1)$)	8.152	**7.545**	1.736	**10.782**	**8.307**	**0.544**	13.768	**9.674**	**0.384**
HSM	**7.783**	**7.495**	2.411	**10.065**	**8.140**	0.634	**11.707**	**8.959**	0.419
Horseshoe	**7.646**	7.754	5.307	13.164	12.347	5.278	18.650	16.321	4.326

training datasets using the parametric model described above for each of the scenarios. We compare the prediction mean square error averaged over 100 test sets.

Similar to parameter estimation and model selection performance, we observe that g-priors, GL-g with fixed τ and Horseshoe performs poorly. Modified Som's method with $K = 3$ performs well in terms of Prediction MSE however at the expense of poor model selection performance. HSM and DP block-g with $\tau = 1$ are consistently among the top four methods. However, unlike HSM, only DP block-g priors performs model selection while performing superior or competitively with other existing methods.

5.7 Real data applications

This section discusses the performance of GL-g priors and DP block-g priors on two real datasets.

Table 5.4: Prediction MSE averaged over 100 test datasets for $p \in \{250, 500, 750\}$ and $r \in \{0, 0.5\}$; Bold represents top four methods with lowest Prediction MSE.

r	0			0.5		
p	250	500	750	250	500	750
g-prior	6.323	27.355	68.988	4.206	14.687	38.091
GL-g ($\tau = 1$)	2.062	3.959	10.591	2.097	3.930	9.733
GL-g ($\tau \sim \mathcal{C}^+(0,1)$)	**1.823**	1.945	2.089	1.853	2.042	2.294
Modified Som ($K = 2$)	**1.800**	**1.900**	**1.994**	1.884	2.355	2.384
Modified Som ($K = 3$)	**1.712**	**1.777**	**2.065**	**1.720**	**1.778**	**2.112**
DP block-g ($\tau = 1$)	**1.823**	**1.949**	**2.059**	**1.849**	**2.037**	**2.264**
DP block-g ($\tau \sim \mathcal{C}^+(0,1)$)	1.831	1.955	2.100	**1.850**	**2.028**	**2.285**
HSM	**1.819**	**1.925**	**2.005**	**1.845**	**1.995**	**2.154**
Horseshoe	1.927	3.194	3.949	1.928	3.043	3.925

5.7.1 `NIR`: Determination of Glucose by near infrared spectroscopy

The `NIR` dataset consists of 235 variables containing first derivatives of near infrared spectroscopy (NIR) absorbance values for 166 alcoholic fermentation mashes of different feedstock. The outcome variables is the concentration of glucose (in g/L) in the feedstock. The dataset was first studied by Liebmann et al. (2009) and is available in package `chemometrics` (Filzmoser and Varmuza, 2017).

We compare different procedures in terms of their out-of-sample predictive performance using 25 random $80 - 20\%$ train-test splits. We exclude Modified Som's procedures since the blocking structure in not known. For GL-g priors and DP block-g priors, we assume two versions: $\tau \sim \mathcal{C}^+(0,1)$ and $\tau = 1$. For all the methods, 5,000 posterior samples were drawn with a thinning of 5 and a burn-in of 15,000 samples. Table 5.5 shows the median value of three different metrics across train-test splits: Prediction mean squared error (PMSE), prediction mean interval score (PMIS) and median size of the median probability model (MPM). For both PMSE and PMIS, a lower score is better. Since HSM and Horseshoe does

not perform model selection, MPM model size is not available for these two methods.

Table 5.5: Results for out-of-sample prediction analysis for `NIR` dataset; Bold represents the top three methods

Method	PMSE	PMIS	MPM size
g-prior	0.229	**2.80**	1
GL-g ($\tau = 1$)	**0.201**	2.98	3
GL-g ($\tau \sim \mathcal{C}^+(0,1)$)	0.233	3.11	5
DP block-g ($\tau = 1$)	0.265	3.17	2
DP block-g ($\tau \sim \mathcal{C}^+(0,1)$)	0.213	**2.83**	3
HSM	**0.187**	2.87	-
Horseshoe	**0.193**	**2.78**	-

We observe that HSM, GL-g priors with fixed tau and Horseshoe performs the best in terms of PMSE. However, as noted above, HSM and Horseshoe does not perform model selection unlike GL-g priors. Horseshoe performs the best in terms of PMIS followed by g-priors and DP block-g priors. Average MPM size and PMSE performance of DP block-g lies between the two extremes: GL-g priors with a different g per covariate and g-priors with a single g for all covariates.

5.7.2 `GGT`: Contaminant exposures associated with liver functionality

Next, we consider the data from National Health and Nutrition Examination Survey (NHANES) conducted by National Center for Health Statistics and previous considered by Boss et al. (2023). The goal here is model Gamma glutamyl transferese (GGT), an enzymatic marker of liver functionality, as a function of 35 measured contaminants for a subset of 990 individuals from NHANES 2003-2004 dataset. The contaminants can be grouped into five exposure

classes: metals (3 exposures), phthalates (7 exposures), organochlorine pesticides (8 exposures), polybrominanted diphenyl ethers (7 exposures) and polycyclic aromatic hydrocarbons (PAHs) (10 exposures). Boss et al. (2023) showed that the contaminants are block correlated with high correlation among contaminants with in the same exposure classes (see Figure 5.4). GGT and all the exposures were log transformed and subsequently standardized. Additionally, GGT was adjusted by discretized age, sex, gender, body mass index, poverty-to-income ratio, ethnicity and urine creatanine. Subsequently, adjusted GGT was used as the outcome variable of interest.

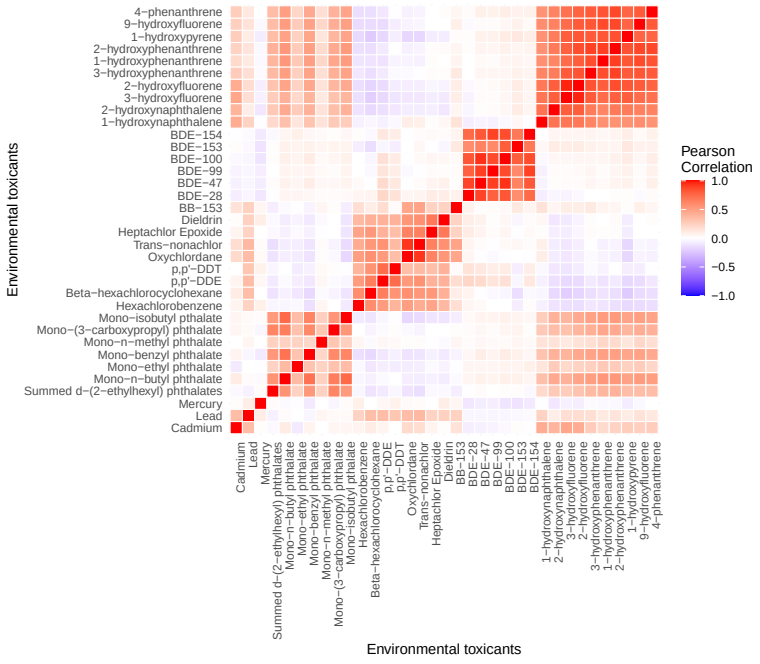


Figure 5.4: Pearson correlation among covariates of `GGT` dataset

Figure 5.5 displays estimated percentage change in GGT for a two fold change in an exposure and their 95% credible intervals for the four Bayesian methods. The dashed line represents that posterior inclusion probability (PIP) for that exposure was greater than

0.5. All the methods use 5,000 MCMC iterations with a thinning of 5 and a burn-in of 15,000 iterations. We display five methods: g-priors (orange), Modified Som (green) with five blocks given by the five exposure classes discussed above and separated by the dotted line in the figure, GL-g prior (blue) with a random τ, DP block-g prior (red) with $\tau = 1$ and HSM (cyan). Note that all methods except HSM can perform variable selection. All four remaining methods select Cadmium, Lead and Mono-n-methyl phthalate in their Median probability model. Additionally, Mono-(3-carboxypropyl) phthalate was selected by g-priors and GL-g with a PIP$>$ 0.5. GL-g priors picks three other variables: 2 from PAHs and one from pesticides. It is worth noting that GL-g priors have and average credible interval length 3.88 which is 25.5% shorter than the average credible interval length of 5.21 for g-priors. GL-g priors is followed by HSM with a average credible interval length of 4.21, followed by DP block-g with average length of 5.15 and Modified Som's prior with an average length of 5.04. GL-g priors and DP block-g priors seem to be more efficient with narrower credible intervals while simultaneously performing model selection.

5.8 Discussion

The principal methodological contribution of this paper is to develop global local g-priors, a unifying framework that bridges the gap between g-priors and continuous shrinkage priors. It allows a different shrinkage component per predictor like shrinkage priors, while still allowing for incorporation of correlation structure into prior and performing model selection. We additionally developed Dirichlet process block-g priors as an extension to block-g priors developed by Som et al. (2014) that avoids conditional Lindley paradox and provides differential shrinkage across coefficients of different magnitude. Unlike Som's block-g priors, our proposal does not require block orthogonal structure among predictors and blocking structure to be known apriori. DP block-g priors allows clustering the shrinkage parameters associated with coefficients while simultaneously perform model selection. Our simulation studies showed that GL-g priors with a random global shrinkage parameter τ and DP block-g priors with a fixed τ perform competitively or superior to other existing methods, both in

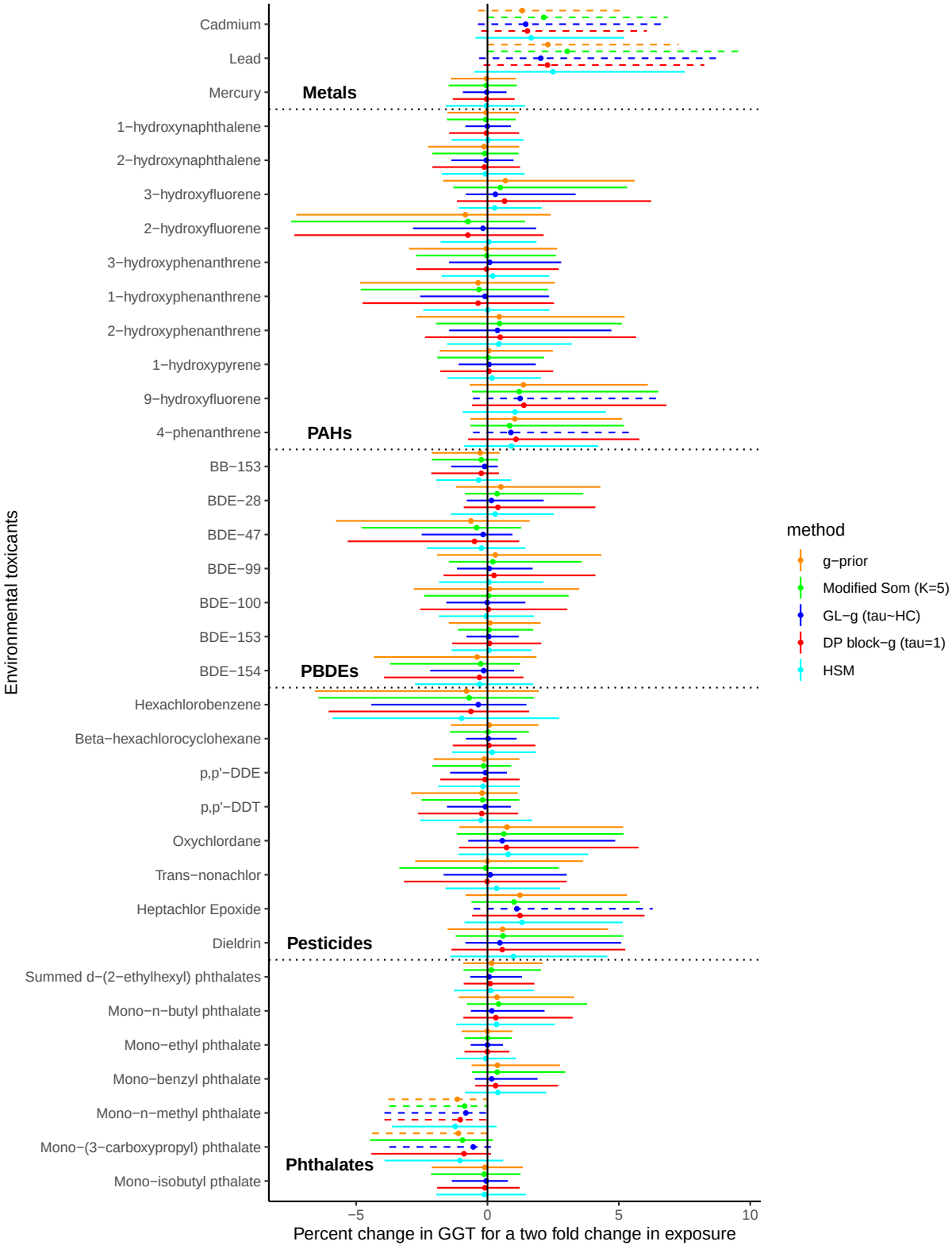


Figure 5.5: Estimated change in Gamma glutamyl transferese (GGT) based on a two fold change in environmental toxicants for NHANES 2003-04 dataset ($n = 990$).

terms of model selection, parameter estimation and out-of-sample prediction performance. This difference is particularly striking when compared to mixture of g-priors and Horseshoe prior.

One motivation for the priors proposed in this chapter was to address CLP. However, one may question the relevance of CLP in practice. Under the asymptotic regime discussed (5.3), $R_1^2 \to 1$ where R_1^2 is R^2 under $\mathcal{M}_1$. In that case, some may argue that $BF_{2:1} \to 0$ may rather be desirable. However, while asymptotic regime in CLP may not arise in practice, our simulation and real dataset studies show that g-prior do lose power to identify weaker but important signals. This is specially undesirable in settings where there are large number of potentially weaker signals and some strong signals, for example, a neuroscience study where the goal is to detect regional levels of brain activation associated with an activity at cellular resolution (Denti et al., 2021). Allowing differential shrinkage using GL-g and DP block-g priors remedies this problem and identifies both strong and weak signals more efficiently than classical g-priors.

There are several possible directions of future research. It is straightforward to extend this framework to generalized linear models using Laplace approximation of likelihood similar to Li and Clyde (2018) or using data augmentation tricks similar to Polson and Scott (2012). Additionally, we will like provide additional theoretical support to our proposed priors by studying their intrinsic and model selection consistency properties. It will also be interesting to understand why random global shrinkage parameter τ does not provide additional performance gains under DP block-g priors in contrast to GL-g priors.

Chapter 6

CONCLUSION AND FUTURE DIRECTIONS

In this thesis, we provided an overview and comparison of statistical techniques developed for inference and variable selection in presence of model uncertainty. Additionally, we proposed a novel Bayesian approach for model selection in GLMs and two approaches for model selection in linear regression. We discuss our main contributions and future directions below.

In Chapter 2, we focused on canonical task of variable selection in linear regression and evaluated 21 existing approaches over statistical tasks of inference, estimation and prediction. While several such comparisons exist, they are often based on a subset of methods included in our study, evaluates methods on a smaller number of tasks and is often based on synthetic simulation experiments. We proposed a parametric bootstrapping framework to compare Bayesian and frequentist approaches based on 14 real datasets on statistical tasks of parameter estimation, interval estimation, point prediction, interval prediction and model inference. We found three Bayesian model averaging (BMA) based approaches based on (mixture of) g-priors to consistently perform well across statistical taks and real datasets. In addition to parameter priors, BMA framework requires specification of prior over discrete model space. While parameter priors have been extensively studied (e.g., see George and McCulloch, 1993, George and McCulloch, 1997, Liang et al., 2008, Johnson and Rossell, 2012a), model space priors have received little attention. In Chapter 3, we extend our parametric bootstrapping framework to compare 8 prominent model space priors over the above mentioned statistical tasks. We found that the Beta-Binomial$(1, 1)$ prior, that induces a uniform prior over model sizes, performs the best across statistical tasks, datasets and choice of three parameter priors recommended by Chapter 2.

In Chapter 4, we proposed Laplace power expected-posterior (LPEP) priors, a novel default Bayesian prior for model selection in generalized linear models. This prior borrows ideas from power priors, expected-posterior priors and unit information priors to transform (potentially) improper priors to minimally informative priors for model selection. Since LPEP can be expressed as a location and scale mixture of normals, we device an efficient MCMC approach for inference based on Pólya-Gamma augmentation strategy by Polson et al. (2013). Additionally, we showed that the proposed prior is intrinsically and asymptotically consistent and performs superior to existing methods on simulated and several real datasets.

Finally, we proposed differential shrinkage g-priors in Chapter 5 for linear regression. Our proposed framework unifies that two Bayesian paradigms for modeling sparsity: continuous shrinkage priors and discrete mixture priors. Unlike classical g-priors that uses a single shrinkage parameter for all coefficients, our proposed global-local (GL) g-priors models shrinkage for each coefficient separately in similar spirit as global-local shrinkage priors while still performing model selection. We also proposed Dirichlet process (DP) block-g that allows to cluster coefficients and in turn predictors requiring similar shrinkage level. We theoretically showed how above proposed prior avoids undesirable conditional Lindley paradox (CLP) and demonstrated empirircal performance of our approach over real and simulated datasets.

6.1 Future directions

While we focused on one example of statistical inference in presence of model uncertainty in Chapter 2, there are many other statistical settings in which the same issue arises, and it would be of interest to carry out similar comparative studies. In linear regression itself, there are the choices of error distribution and functional form of the variables. The same issues arise in generalized linear models such as logistic regression and Poisson regression, in addition to the choice of link function and mean-variance relationship. Similar model choice issues arise with Bayesian hierarchical models and many other model classes. It will be useful

to carry out similar comparison studies for the above problems to allow practitioners to decide which methods works well to account for model uncertainty in their scientific process.

Chapter 4 focuses on developing the LPEP for logistic regression. However, the formulation is very general and can be extended to many other generalized linear models. Many of the computational advantages of our procedure extend to binomial, negative binomial and multinomial logistic models where the data augmentation approach of Polson et al. (2013) can be readily applied. This is also true for probit models in which computation can rely on the data augmentation approach of Albert and Chib (1993), as well as for loglinear regression using the approach of Frühwirth-Schnatter et al. (2009). It will be interesting to explore theoretical and empirical properties of LPEP prior in the above settings. Additionally, a key drawback of LPEP prior is that it assumes $n > p$, i.e. the approach is developed for low dimensional settings. Tzoumerkas and Fouskakis (2021) recently developed PEP priors for high-dimensional settings in linear regression where the default baseline prior is a shrinkage prior rather than a non-informative improper prior. Another way to extend PEPs to high-dimensional settings could be to use penalized likelihood approaches like LASSO, SCAD and MCP to filter out irrelevant predictors before using LPEP on filtered list of p^* relevant covariates where $p^* < n$. It will be interesting to develop extensions to LPEP methodology for high-dimensional settings in linear and generalized linear models.

The focus of Chapter 5 was to design default Bayesian prior for variable selection in linear regression setting. It is straightforward to extend our proposals to GLMs using Laplace approximations similar to Li and Clyde (2018). While we theoretically showed that DP block-g and GL-g priors avoid CLP, it will be of interest to study asymptotic model selection and intrinsic consistency of the proposed priors.

Ideas similar to those developed in Chapter 5 can also be extended for graphical modeling. Gaussian graphical models (GGM) provide an appealing way to uncover latent network structures by encoding conditional independence between variables using a graph (Lauritzen, 1996). Consider a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ and an undirected graph $G = (V, E)$ with vertex set $V = \{1, \ldots, p\}$. Under GGM given by graph G, we assume that $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

and $\boldsymbol{\Sigma}^{-1}_{j,k} = 0 \iff X_j$ and X_k are conditionally independent given remain variables $\boldsymbol{X}_{\setminus\{j,k\}}$. That is, the pattern of zero entries in the precision matrix corresponds to conditional independence between the p variables. Thus, model selection priors finds a natural application in sparse precision matrix estimation for GGMs. Model selection priors can drastically reduce the number of parameters, prevent overfitting and help uncover hidden structures in a network (Dobra et al., 2011). A natural Bayesian way to infer graph G is by specifying a prior on covariance matrix $\boldsymbol{\Sigma}$ corresponding to the graph. A common prior choice for $\boldsymbol{\Sigma}$ is given by hyper-inverse Wishart distribution, denoted by $\boldsymbol{\Sigma} \sim HIW_G(b, D)$ where $b \in \mathbb{R}^+$ denotes the degrees of freedom D is a $p \times p$-dimensional symmetric positive definite scale matrix (Dawid and Lauritzen, 1993). Carvalho and Scott (2009) developed hyper-inverse Wishart-g (HIW-g) prior for model selection in GGM where $D = g\boldsymbol{X}^T\boldsymbol{X}$ and name derived from it's analogy to Zellner's g-priors. It will be interesting to extend this framework to develop *hyper-inverse Wishart-GL-g prior* by replacing D above with $\boldsymbol{G}^{-1/2}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{G}^{-1/2}$ where $\boldsymbol{G}$ is a p-dimensional diagonal matrix allowing for differential shrinkage as opposed to single shrinkage parameter under HIW-g prior. An alternative way to infer G, particularly in high dimensional settings, is by defining and inferring p separate sparse conditional linear regression where X_i is regressed on $\boldsymbol{X}_{-i}$ (Dobra et al., 2004). While these regressions may not cohere in the sense of being consistent with the proper joint distribution of $\boldsymbol{X}$, it can still be attrative to infer higher dimensional relationships by fitting p sparse regression models using GL-g priors discussed in Chapter 5.

Another possible future direction can be to utilize DP block-g distribution for robust graphical modeling. Despite the appeal, GGM can often be susceptible to measurement errors. To remedy this, Finegold and Drton (2014) replaced multivariate normal distribution for $\boldsymbol{Y}$ with Dirichlet-t distributions for robust graphical modeling. If a random latent vector $\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}, \Psi)$ and $\tau_j \sim H$ with $H \sim DP(\alpha, H_0)$ where H_0 is Gamma$(\nu/2, \nu/2)$ distribution then $\boldsymbol{Y} \in \mathbb{R}^p$ with coordinates $X_j = \mu_j + Z_j/\tau_j$ follows a Dirichlet-t distribution. Dirichlet-t distribution allows to identify pockets of similarly contaminated observations out of the p observations rather than weighing an entire observations using a single divisor, as in the

classical t-model. However, Dirichlet-t distribution can be seen as a special case of DP block-g distribution. It will be interesting to see straightforward extensions of robust Dirichlet-t distributions by replacing with other heavy tailed prior choices like hyper-g prior and Beta-prime density as the base measure H_0.

Model averaging provides principled way to simultaneously account for uncertainty in parameter space and model space for a statistical model. This thesis was focused on comparing existing model selection techniques in practice, provide recommendation to practitioners. Another goal of this thesis was to develop new methodology to address theoretical and empirical concerns with existing methods. To this end, we proposed a new framework to compare existing methods grounded in real datasets observed in practice and proposed three principled ways to perform model selection in linear and generalized linear models. We hope that this work helps paves way for new ground-breaking developments in field of modeling sparsity and performing variable selection.

BIBLIOGRAPHY

Abramowitz, M., Stegun, I. A., and Romer, R. H. (1988). Handbook of mathematical functions with formulas, graphs, and mathematical tables.

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.

Akaike, H. (1983). Information measures and model selection. *Bulletin of the International Statistical Institute*, 44:277–291.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, 88(422):669–679.

Andrade, J. A. A. and O'Hagan, A. (2006). Bayesian robustness modeling using regularly varying distributions. *Bayesian Analysis*, 1(1):169–188.

Andrade, J. A. A. and O'Hagan, A. (2011). Bayesian robustness modelling of location and scale parameters. *Scandinavian Journal of Statistics*, 38(4):691–711.

Bai, R. and Ghosh, M. (2018). On the beta prime prior for scale parameters in high-dimensional Bayesian regression models. *arXiv preprint arXiv:1807.06539*.

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.

Bartlett, M. S. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44:533–534.

Bayarri, M., Berger, J. O., Jang, W., Ray, S., Pericchi, L. R., and Visser, I. (2019). Prior-based Bayesian information criterion. *Statistical Theory and Related Fields*, 3(1):2–13.

Bayarri, M. J., Berger, J. O., Forte, A., García-Donato, G., et al. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics*, 40(3):1550–1577.

Berger, J. O. (2021). Four types of frequentism and their interplay with Bayesianism. De Finetti Lecture, Meeting of the International Society for Bayesian Analysis.

Berger, J. O. and Pericchi, L. R. (1996a). The intrinsic Bayes factor for linear models. In J. M. Bernardo, J. O. Berger, A. P. D. and Smith, A. F. M., editors, *Bayesian Statistics 5*, pages 25–44. Oxford Univ. Press.

Berger, J. O. and Pericchi, L. R. (1996b). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91(433):109–122.

Berger, J. O., Pericchi, L. R., Ghosh, J., Samanta, T., De Santis, F., Berger, J., and Pericchi, L. (2001). Objective Bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series*, pages 135–207.

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika*, 103(4):955–969.

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

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

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.

Boss, J., Datta, J., Wang, X., Park, S. K., Kang, J., and Mukherjee, B. (2023). Group inverse-gamma gamma shrinkage for sparse linear models with block-correlated regressors. *Bayesian Analysis*, 1(1):1–30.

Bové, D. S. and Held, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis*, 6(3):387–410.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, 5(1):232–253.

Brock, W., Durlauf, S. N., and West, K. D. (2003). Policy evaluation in uncertain economic environments.

Brown, P. J. and Griffin, J. E. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Burnham, K. P. and Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretical approach*. Springer-Verlag, New York, 2nd edition.

Califf, R. M., White, H. D., Van de Werf, F., Sadowski, Z., Armstrong, P. W., Vahanian, A., Simoons, M. L., Simes, R. J., Lee, K. L., and Topol, E. J. (1996). One-year results from the Global Utilization of Streptokinase and TPA for Occluded Coronary Arteries (GUSTO-I) trial. *Circulation*, 94(6):1233–1238.

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

Carvalho, C. M. and Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika*, 96(3):497–512.

Casella, G., Ghosh, M., Gill, J., and Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411.

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.

Celeux, G., El Anbari, M., Marin, J.-M., and Robert, C. P. (2012). Regularization in regression: Comparing Bayesian and frequentist methods in a poorly informative situation. *Bayesian Analysis*, 7:477–502.

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-p sparse glm. *Statistica Sinica*, pages 555–574.

Chen, M.-H., Ibrahim, J. G., and Shao, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference*, 84(1-2):121–137.

Clyde, M. (2020). *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. R package version 1.5.5.

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B, Statistical Methodology*, 62(4):681–698.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Statistical Science*, 19(1):81–94.

Consonni, G., Fouskakis, D., Liseo, B., Ntzoufras, I., et al. (2018). Prior distributions for objective Bayesian analysis. *Bayesian Analysis*, 13(2):627–679.

Cragg, M. I., Zhou, Y., Gurney, K., and Kahn, M. E. (2013). Carbon geography: the political economy of congressional support for legislation intended to mitigate greenhouse gas production. *Economic Inquiry*, 51(2):1640–1650.

Cui, W. and George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138(4):888–900.

Dawid, A. P. (1973). Posterior expectations for large observations. *Biometrika*, 60(3):664–667.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, pages 1272–1317.

De Santis, F. and Spezzaferri, F. (2001). Consistent fractional Bayes factor for nested normal linear models. *Journal of statistical planning and inference*, 97(2):305–321.

Deckers, T. and Hanck, C. (2014). Variable selection in cross-section regressions: Comparisons and extensions. *Oxford Bulletin of Economics and Statistics*, 76(6):841–873.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing*, 12(1):27–36.

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2012). Joint specification of model space and parameter space prior distributions. *Statistical Science*, 27(2):232–246.

Denti, F., Azevedo, R., Lo, C., Wheeler, D., Gandhi, S. P., Guindani, M., and Shahbaba, B. (2021). A horseshoe pit mixture model for Bayesian screening with an application to light sheet fluorescence microscopy in brain imaging. *arXiv preprint arXiv:2106.08281*.

Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G., and West, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, 90(1):196–212.

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, 106(496):1418–1433.

Durlauf, S. N., Kourtellos, A., and Tan, C. M. (2008). Are any growth theories robust? *The Economic Journal*, 118(527):329–346.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Eicher, T. S., Papageorgiou, C., and Raftery, A. E. (2011). Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics*, 26(1):30–55.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588.

Facchini, G. and Steinhardt, M. F. (2011). What drives us immigration policy? Evidence from congressional roll call votes. *Journal of Public Economics*, 95(7-8):734–743.

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 Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911.

Fernandez, C., Ley, E., and Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100(2):381–427.

Filzmoser, P. and Varmuza, K. (2017). *chemometrics: Multivariate Statistical Analysis in Chemometrics*. R package version 1.4.2.

Finegold, M. and Drton, M. (2014). Robust Bayesian graphical modeling using Dirichlet t-distributions. *Bayesian Analysis*, 9(3):521–550.

Forte, A., Garcia-Donato, G., and Steel, M. F. J. (2018). Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *International Statistical Review*, 86(2):237–258.

Foster, D. P., George, E. I., et al. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, 22(4):1947–1975.

Fouskakis, D., Ntzoufras, I., and Draper, D. (2015). Power-expected-posterior priors for variable selection in Gaussian linear models. *Bayesian Analysis*, 10(1):75–107.

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2018). Power-expected-posterior priors for generalized linear models. *Bayesian Analysis*, 13(3):721–748.

Freedman, D. A. (1983). A note on screening regression equations. *American Statistician*, 37:152–155.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, 1(2):302–332.

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

Friendly, M. (2021). *HistData: Data Sets from the History of Statistics and Data Visualization*. R package version 0.8-7.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L., and Rue, H. (2009). Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statistics and Computing*, 19(4):479–492.

George, E. (1999). Discussion of "model averaging and model search strategies" by M. Clyde. In *Bayesian Statistics 6–Proceedings of the Sixth Valencia International Meeting*.

George, E. I. (2010). Dilution priors: Compensating for model space redundancy. In *Borrowing Strength: Theory Powering Applications–A Festschrift for Lawrence D. Brown*, pages 158–165. Institute of Mathematical Statistics.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7(2):339–373.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5*.

Ghosh, J. (2019). Cauchy and other shrinkage priors for logistic regression in the presence of separation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(6):e1478.

Ghosh, J., Li, Y., and Mitra, R. (2018). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*, 13(2):359–383.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Good, I. J. (1950). Probability and the weighing of evidence.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732.

Griffin, J. and Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. *University of Kent Technical Report*.

Gu, C. (2014). Smoothing spline ANOVA models: R package gss. *Journal of Statistical Software*, 58(5):1–25.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4):835–845.

Hansen, M. H. and Yu, B. (2003). Minimum description length model selection criteria for generalized linear models. *Lecture Notes-Monograph Series*, pages 145–163.

Heinze, G. and Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statistics in Medicine*, 21(16):2409–2419.

Held, L., Bové, D. S., and Gravestock, I. (2015). Approximate Bayesian model selection with the deviance statistic. *Statistical Science*, pages 242–257.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14:382–417.

Holmes, C. C. (2011). Discussion of 'Regression shrinkage and selection via the lasso: a retrospective'. *Journal of the Royal Statistical Society, Series B*, 73:279–280.

Ibrahim, J. G. and Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

Ibrahim, J. G. and Laud, P. W. (1991). On Bayesian analysis of generalized linear models using Jeffreys's prior. *Journal of the American Statistical Association*, 86(416):981–986.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American statistical Association*, 96(453):161–173.

Ishwaran, H., Rao, J., and Kogalur, U. (2013). *spikeslab : Prediction and variable selection using spike and slab regression.* R package version 1.1.5.

Ishwaran, H. and Zarepour, M. (2002). Dirichlet prior sieves in finite normal mixtures. *Statistica Sinica*, pages 941–963.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2017). *ISLR: Data for an Introduction to Statistical Learning with Applications in R.* R package version 1.2.

Jeffreys, H. (1961). *Theory of Probability.* Oxford University Press, Oxford, U.K., 3rd edition.

Jessee, S. A. and Theriault, S. M. (2012). The two faces of congressional roll-call voting. *Party Politics*, page 1354068812458612.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170.

Johnson, V. E. and Rossell, D. (2012a). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Johnson, V. E. and Rossell, D. (2012b). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, 107(498):649–660.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90(431):928–934.

Konis, K. (2007). *Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models*. PhD thesis, University of Oxford.

Kosmidis, I. and Schumacher, D. (2020). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates*. R package version 0.1.

Lauritzen, S. L. (1996). *Graphical models*, volume 17. Clarendon Press.

Leamer, E. E. (1978). *Specification Searches: Ad hoc Inference with Nonexperimental Data*, volume 53. Wiley.

Lee, S. Y., Pati, D., and Mallick, B. K. (2020). Continuous shrinkage prior revisited: a collapsing behavior and remedy. *arXiv preprint arXiv:2007.02192*.

Levine, R. and Renelt, D. (1992). A sensitivity analysis of cross-country growth regressions. *The American economic review*, pages 942–963.

Ley, E. and Steel, M. F. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics*, 24(4):651–674.

Li, H. and Pati, D. (2017). Variable selection using shrinkage priors. *Computational Statistics & Data Analysis*, 107:107–119.

Li, Y. and Clyde, M. A. (2018). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association*, 113(524):1828–1845.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423.

Liebmann, B., Friedl, A., and Varmuza, K. (2009). Determination of glucose and ethanol in bioethanol production by near infrared spectroscopy and chemometrics. *Analytica Chimica Acta*, 642(1-2):171–178.

Liu, Y., Wichura, M. J., and Drton, M. (2012). Rejection sampling for an extended gamma distribution. *Unpublished manuscript*.

Lofland, C. L., Rodríguez, A., and Moser, S. (2017). Assessing differences in legislators' revealed preferences: A case study on the 107th U.S. Senate. *The Annals of Applied Statistics*, 11(1):456–479.

Lumley, T. (2020). *leaps: Regression Subset Selection*. R package version 3.1.

Luo, S. and Chen, Z. (2013). Selection consistency of EBIC for GLIM with non-canonical links and diverging number of parameters. *Statistics and its Interface*, pages 275–284.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Makalic, E. and Schmidt, D. F. (2016). High-dimensional Bayesian regularised regression with the bayesreg package. *arXiv preprint arXiv:1611.06649*.

Mäkeläinen, T., Schmidt, K., and Styan, G. P. (1981). On the existence and uniqueness of the maximum likelihood estimate of a vector-valued parameter in fixed-size samples. *The Annals of Statistics*, pages 758–767.

Mansournia, M. A., Geroldinger, A., Greenland, S., and Heinze, G. (2018). Separation in logistic regression: causes, consequences, and control. *American Journal of Epidemiology*, 187(4):864–870.

Mattei, P. A. (2020). A parsimonious tour of Bayesian model uncertainty. https://arxiv.org/abs/1902.05539.

McCullagh, P. and Nelder, J. A. (2019). *Generalized linear models*. Routledge.

Miller, A. (2002). *Subset Selection in Regression*. CRC Press.

Miquel, G. P. I. and Snyder Jr, J. M. (2006). Legislative effectiveness and legislative careers. *Legislative Studies Quarterly*, 31(3):347–381.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Moser, S., Rodríguez, A., and Lofland, C. L. (2021). Multiple ideal points: Revealed preferences in different domains. *Political Analysis*, 29(2):139–166.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, 42(2):789–817.

Newman, D., Hettich, S., Blake, C., and Merz, C. (1998). UCI Repository of machine learning databases.

O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):99–118.

O'Roark, J. B. and Wood, W. C. (2011). Determinants of congressional minimum wage support: the role of economic education. *Public Choice*, 147(1-2):209–225.

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., and Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics*, 42:570–575.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Pearson, K. and Lee, A. (1903). On the laws of inheritance in man: I. inheritance of physical characters. *Biometrika*, 2:357–462.

Pérez, J. M. and Berger, J. O. (2002). Expected-posterior prior distributions for model selection. *Biometrika*, 89(3):491–512.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for mcmc. *R News*, 6(1):7–11.

Polson, N. G. and Scott, J. G. (2012). Local shrinkage rules, lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):287–311.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349.

Potter, D. M. (2005). A permutation test for inference in logistic regression with small-and moderate-sized data sets. *Statistics in Medicine*, 24(5):693–708.

Raftery, A., Hoeting, J., Volinsky, C., Painter, I., and Yeung, K. Y. (2020). *BMA: Bayesian Model Averaging*. R package version 3.18.12.

Raftery, A. E. (1988). Approximate Bayes factors for generalized linear models. Technical Report 121, Department of Statistics, University of Washington. https://stat.uw.edu/sites/default/files/files/reports/1988/tr121.pdf.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179–191.

Raftery, A. E. and Zheng, Y. (2003). Discussion: Performance of Bayesian model averaging. *Journal of the American Statistical Association*, 98(464):931–938.

Rockova, V. and George, E. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Ročková, V. and George, E. I. (2014). EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association*, 109(506):828–846.

Ročková, V. and George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, 113(521):431–444.

Rockova, V. and Moran, G. (2019). *EMVS: The Expectation-Maximization Approach to Bayesian Variable Selection*. R package version 1.1.

Rohart, F., Gautier, B., Singh, A., and Le Cao, K.-A. (2017). mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752.

Rossell, D. (2021). Concentration of posterior model probabilities and normalized l0 criteria. *Bayesian Analysis*, 1(1):1–27.

Rossell, D., Abril, O., and Bhattacharya, A. (2021). Approximate Laplace approximations for scalable model selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):853–879.

Rossell, D., Cook, J. D., Telesca, D., Roebuck, P., and Abril, O. (2019). *mombf: Bayesian Model Selection and Averaging for Non-Local and Local Priors*. R package version 2.2.9.

Rossell, D. and Rubio, F. J. (2018). Tractable Bayesian Variable Selection: Beyond Normality. *Journal of the American Statistical Association*.

Rossell, D. and Telesca, D. (2017). Nonlocal priors for high-dimensional estimation. *Journal of the American Statistical Association*, 112(517):254–265.

Rubin, D. B. and Schenker, N. (1986). Efficiently simulating the coverage properties of interval estimates. *Journal of the Royal Statistical Society: Series C, Applied Statistics*, 35(2):159–167.

Sala-i Martin, X. (1997). I just ran four million regressions.

Sala-I-Martin, X., Doppelhofer, G., and Miller, R. I. (2004). Determinants of long-term growth: A Bayesian averaging of classical estimates (BACE) approach. *American economic review*, 94(4):813–835.

Schwarz, G. (1978a). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Schwarz, G. (1978b). Estimating the dimension of a model. *The Annals of Statistics*, pages 461–464.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Scott, S. L. (2020). *BoomSpikeSlab: MCMC for Spike and Slab Regression*. R package version 1.2.3.

Shahbaba, B. and Johnson, W. O. (2013). Bayesian nonparametric variable selection as an exploratory tool for discovering differentially expressed genes. *Statistics in Medicine*, 32(12):2114–2126.

Smith, J. W., Everhart, J. E., Dickson, W., Knowler, W. C., and Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care*, page 261. American Medical Informatics Association.

Som, A. (2014). *Paradoxes and Priors in Bayesian Regression*. PhD thesis, The Ohio State University.

Som, A., Hans, C. M., and MacEachern, S. N. (2014). Block hyper-g priors in Bayesian regression. *arXiv preprint arXiv:1406.6419*.

Som, A., Hans, C. M., and MacEachern, S. N. (2016). A conditional Lindley paradox in Bayesian linear models. *Biometrika*, 103(4):993–999.

Spiegelhalter, D. J. and Smith, A. F. (1982). Bayes factors for linear and log-linear models with vague prior information. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(3):377–387.

Steyerberg, E. W. (2019). *Clinical prediction models*. Springer.

Thawornwattana, Y., Dalquen, D., and Yang, Z. (2018). Designing simple and efficient Markov chain Monte Carlo proposal kernels. *Bayesian Analysis*, 13(4):1037–1063.

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244.

Tzoumerkas, G. and Fouskakis, D. (2021). Power-expected-posterior methodology with baseline shrinkage priors. In *Interational Conference on Bayesian Young Statistician Meeting*, pages 35–44. Springer.

van der Pas, S., Scott, J., Chakraborty, A., and Bhattacharya, A. (2019). *horseshoe: Implementation of the Horseshoe Prior*. R package version 0.2.0.

van Zwet, E. (2019). A default prior for regression coefficients. *Statistical Methods in Medical Research*, 28(12):3799–3807.

Vignes, M., Vandel, J., Allouche, D., Ramadan-Alban, N., Cierco-Ayrolles, C., Schiex, T., Mangin, B., and De Givry, S. (2011). Gene regulatory network reconstruction using Bayesian networks, the Dantzig selector, the lasso and their meta-analysis. *PloS one*, 6(12):e29165.

Villa, C. and Walker, S. (2015). An objective Bayesian criterion to determine model prior probabilities. *Scandinavian Journal of Statistics*, 42(4):947–966.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, 44(1):92–107.

Wedderburn, R. W. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, 63(1):27–32.

Womack, A. J., León-Novelo, L., and Casella, G. (2014). Inference from intrinsic Bayes'

procedures under model selection and uncertainty. *Journal of the American Statistical Association*, 109(507):1040–1053.

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). On the computational complexity of high-dimensional bayesian variable selection. *The Annals of Statistics*, 44(6):2497–2532.

Young, W. C., Raftery, A. E., and Yeung, K. Y. (2014). Fast Bayesian inference for gene regulatory networks using scanbma. *BMC Systems Biology*, 8(1):47.

Zellner, A. (1986a). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian Inference and Decision Techniques*, 6.

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

Zellner, A. (2008). Comments on 'Mixtures of g-priors for Bayesian variable selection'.

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadística y de Investigaciów Operativa*, 31(1):585–603.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Appendix A

APPENDIX A

A.1 Datasets Description

Here we present a brief description of the data and the transformations used in the analyses:

- `College`: We model log of applications received on statistics collected for 777 US colleges from the 1995 issue of US news and world report. We remove `Enroll, Accept` due to potential causal relationship with number of applications. We perform a log transformation on number of full time undergraduates (`F.undergrad`) and part-time graduates (`P.undergrad`).

- `Bias Correction`: This data is for the purpose of bias correction of next-day maximum and minimum air temperatures forecast of the LDAPS model operated by the Korea Meteorological Administration over Seoul, South Korea. The input data is largely composed of the LDAPS model's next-day forecast data, in-site maximum and minimum temperatures of present-day, and geographic auxiliary variables. There are two outputs (i.e. next-day maximum and minimum air temperatures) in this data. We excluded station and date variable from our modelling. We model square root of maximum and minimum air temperature as a function of available covariates. Maximum temperature was not used in modeling of minimum temperature and vice versa.

- `SML2010`: The following dataset contains forecasts of several weather parameters and actuators state collected at a solar-powered house, known as Small Medium Large System (SMLsystem). The data was sampled every minute, computing and uploading it smoothed with 15 minute means. The goal is to model indoor temperature as a

function of the above parameters. We excluded time and data variables. We also excluded Enthalpic motor and Enthalpic motor turbo since there was no variability in these variables across observations. We treated day of week as a factor variable.

- `Bike sharing`: We consider both the hourly and daily version of Bike sharing dataset that contains data for count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. We remove `instant, dteday` variables as these variable are indexing variables. We also remove `casual, registered` variables since `count` is defined as sum of these two variables. We rescale the normalised versions of `temp, atemp, windspeed` and `hum` using normalization constants available in the read me file available on UCI. We treat `yr, weathersit, holiday, mnth, weekday` and `season` as a factor variable. We remove `workingday` as it is determined by `weekday` deterministically. The outcome variable for daily data is square root of `count`.

  For the hourly data, we group `hr` into 5 categories: `Latenight, EarlyMorning, Morning, Evening` and `Night`. The outcome variable for daily data is cube root of `count`.

- `Superconductivity`: We model the cube root of critical temperature based on based on the 81 features extracted from the superconductor's chemical formula. The sample size is 21,263.

- `Diabetes`: The data consists of 442 patients in which the response of interest is a quantitative measure of disease progression. The data includes 10 baseline measurements for each patient, in addition to 45 interactions and 9 quadratic terms, for a total of 64 variables for each patient.

- `Ozone`: The dataset consists of daily measurements of the maximum ozone concentration near Los Angeles and eight meteorological variables. We model log of ozone

concentration using the eight meteorological variables, plus interactions and squares, leading to 44 possible predictors.

- `Boston Housing`: This dataset consists of housing data for 506 census tracts of Boston from the 1970 census. We examine median value of owner-occupied homes (in USD 1000's) using 13 base variables, their interactions and quadratic terms, for a total of 103 variables per observation.

- `NIR`: This dataset consists of 235 variables containing first derivatives of near infrared spectroscopy (NIR) absorbance values for 166 alcoholic fermentation mashes of different feedstock. The outcome variables is concentration of glucose (in g/L) in the feedstock.

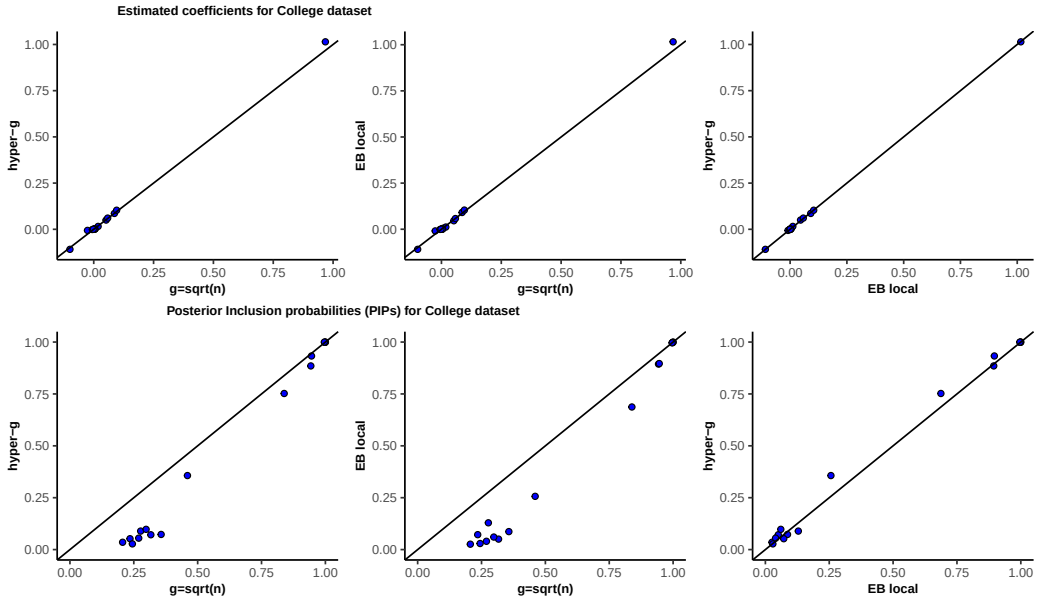
- `nutrimouse`: The data set come from a nutrigenomic study in the mouse containing observations of 40 mice where hepatic fatty-acid concentrations are regressed upon the expression of 120 potentially relevant genes measured in liver cells. The response variable is `C16.0`.

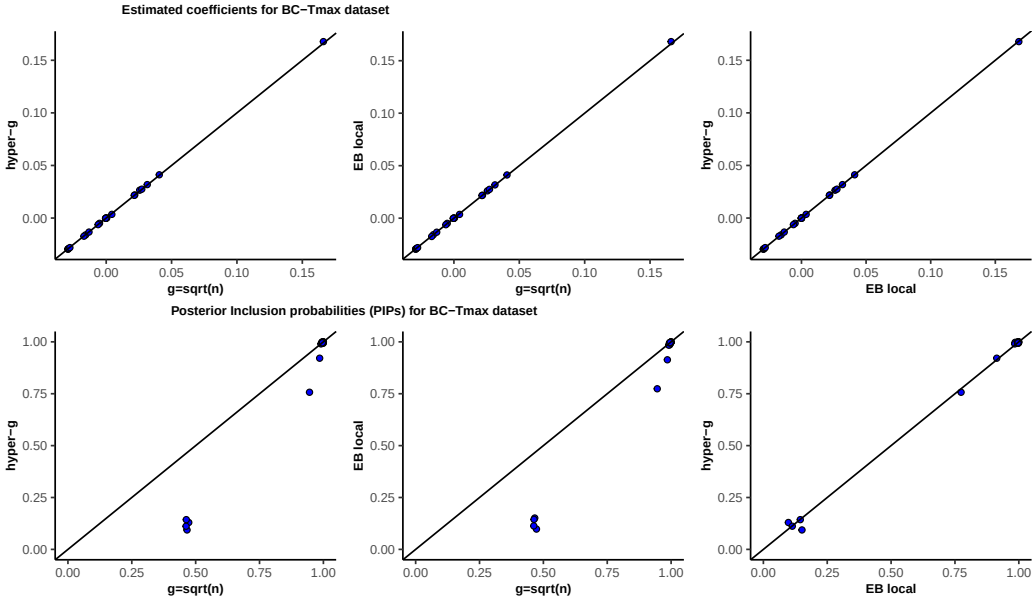
- `multidrug`: The Multidrug data are from a pharmacogenomic study investigating the relationship between the drug concentration and expression of the adenosine triphosphate binding cassette transporter (ABC3A). The X matrix comprises observations of the activity of 853 drugs on 60 different human cell lines, expressed as the concentration at which each drug leads to a 50% inhibition of growth for each cell line. The y variable is the measured expression of ABC3A in each cell line.

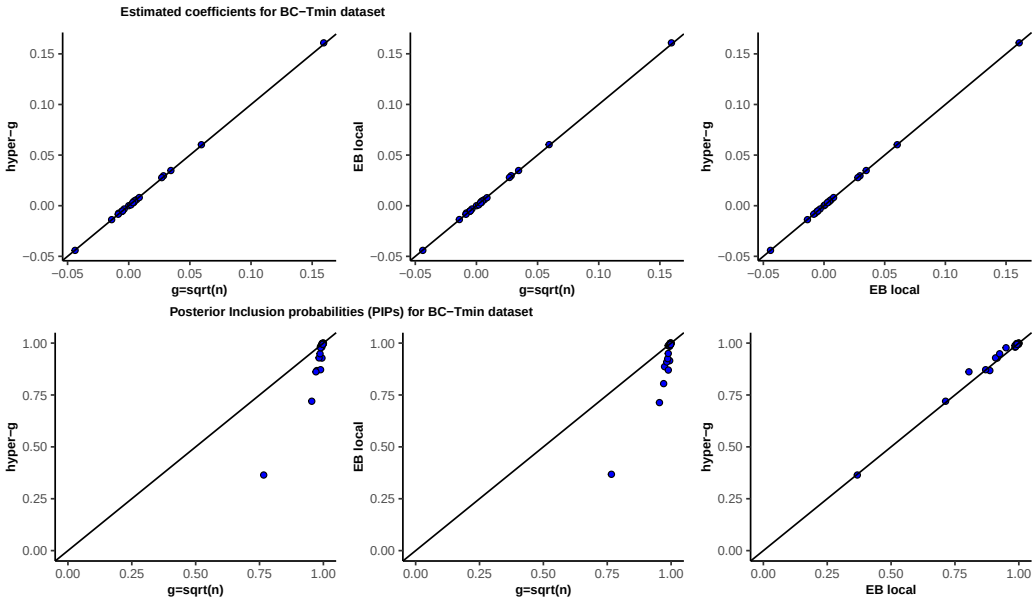
- `liver.toxicity`: This dataset comes from a liver toxicity study containing the expression scores for 3116 genes in 64 rat subjects. We model the cholesterol concentration in the liver (in mg/dL) as a function of the expression scores of genes.

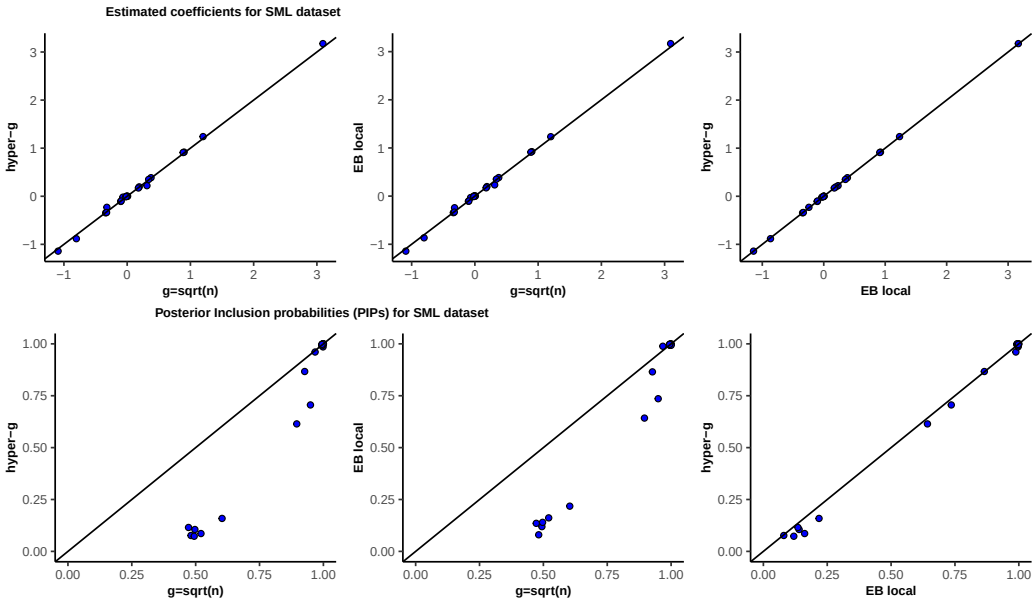
A.2 Scatter plots for estimated posterior inclusion probabilities (PIPs) and coefficients of top 3 methods for all datasets

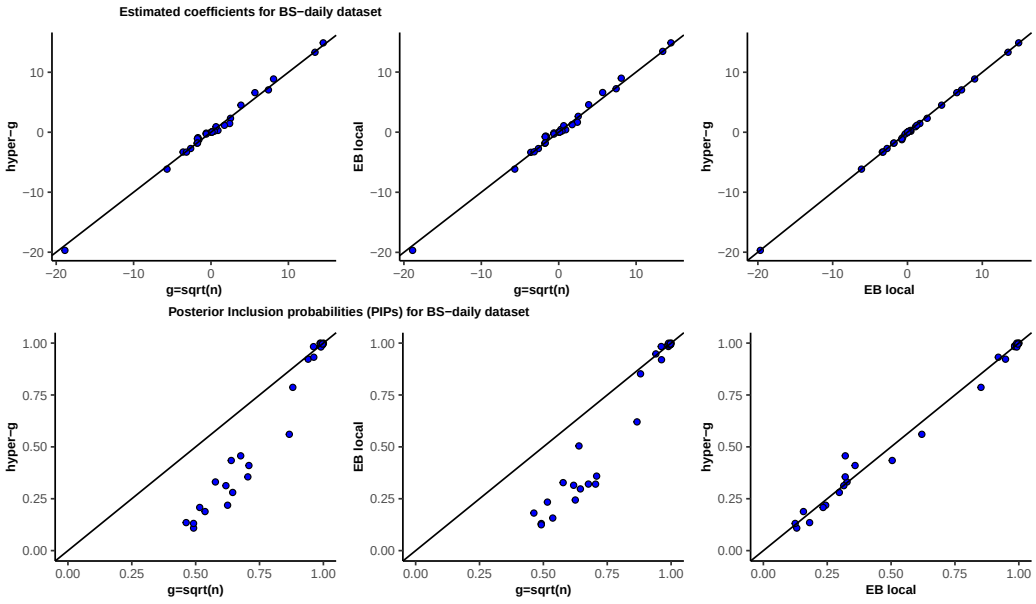
Here are the scatter plots for PIPs and coefficients of top three methods. We observe that for tall datasets, Hyper-g and EBlocal agree and tend to select sparser models compared to g=sqrt(n). For wide datasets, there is no agreement between methods on estimated coefficients and PIPs possibly due to presence of several highly correlated variables.

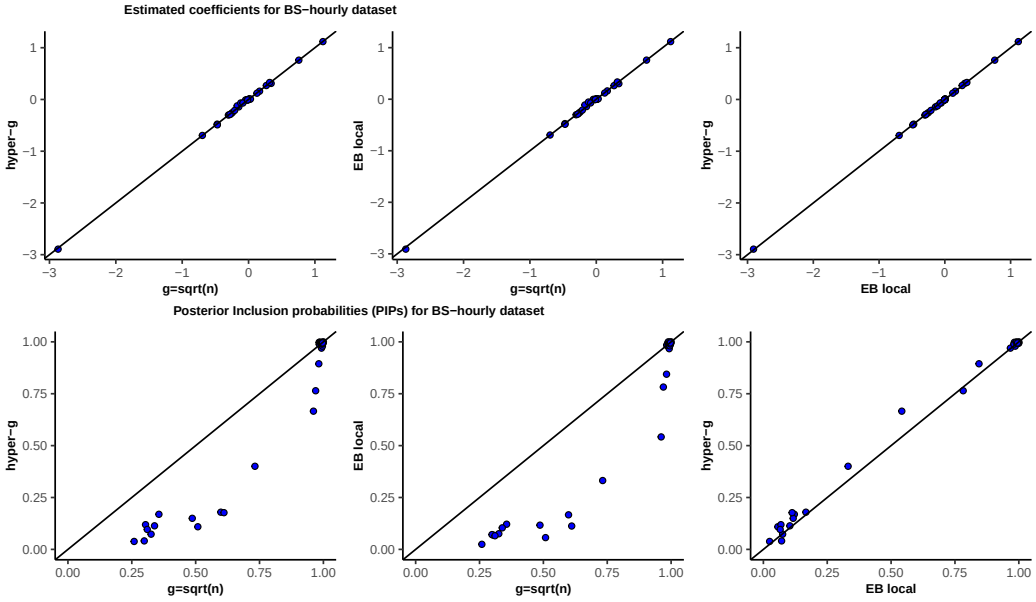


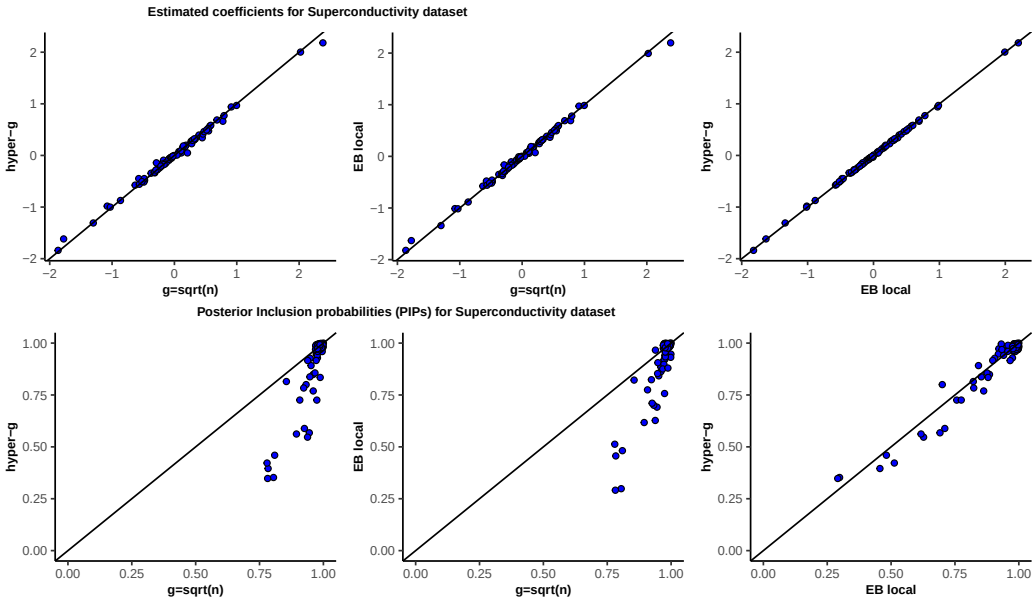

Estimated coefficients for BC−Tmax dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for BC−Tmax dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local

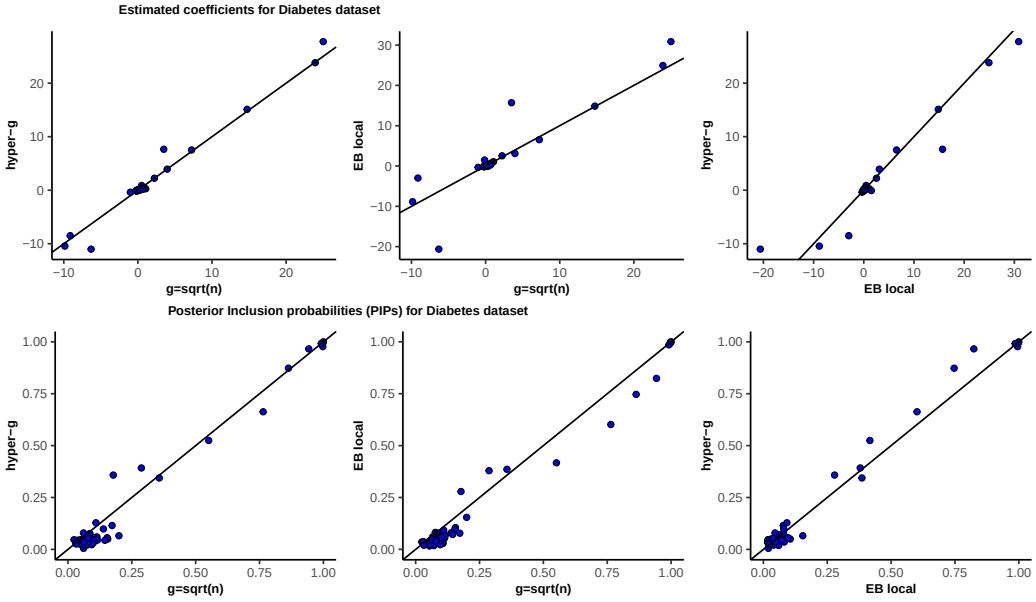

Estimated coefficients for BC−Tmin dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for BC−Tmin dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local

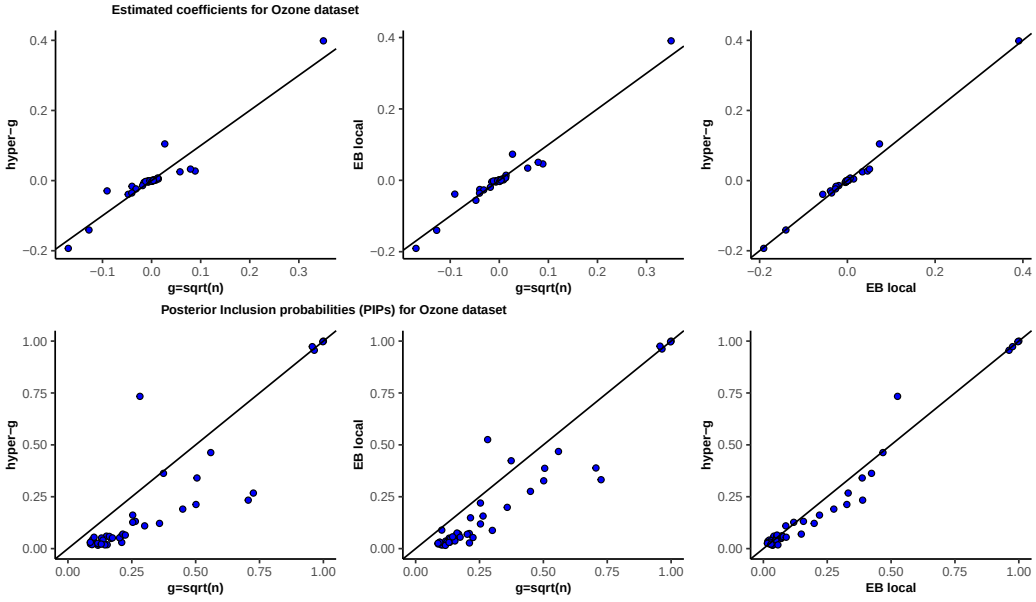

Estimated coefficients for SML dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for SML dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local

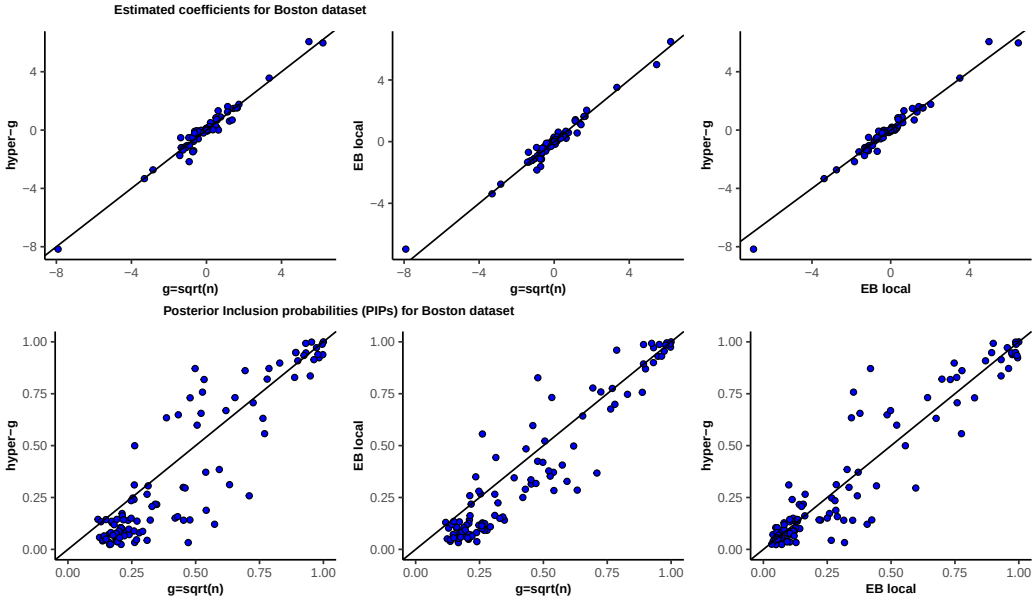

Estimated coefficients for BS−daily dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for BS−daily dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local

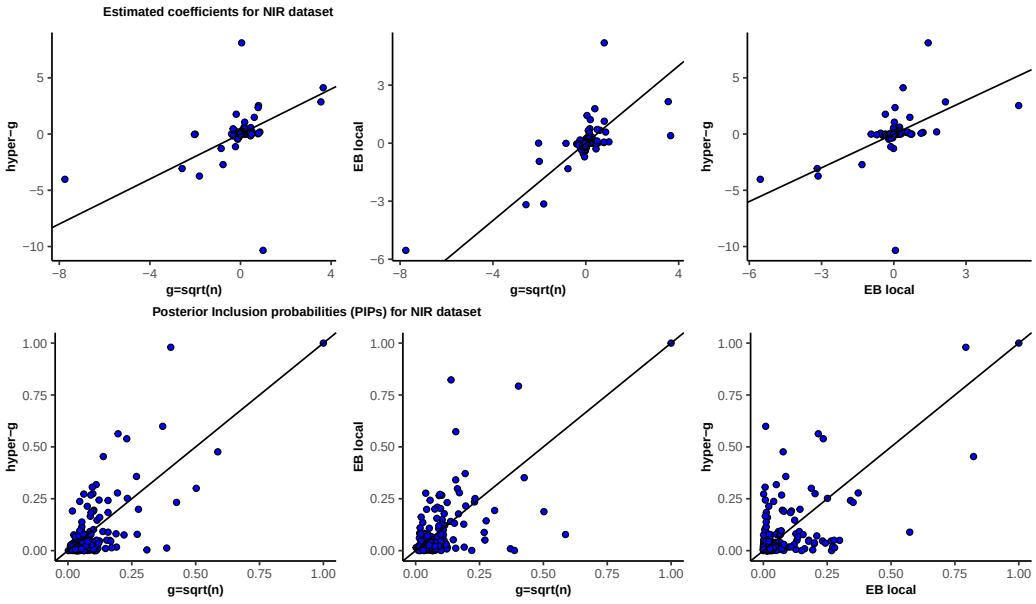

Estimated coefficients for BS−hourly dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for BS−hourly dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local

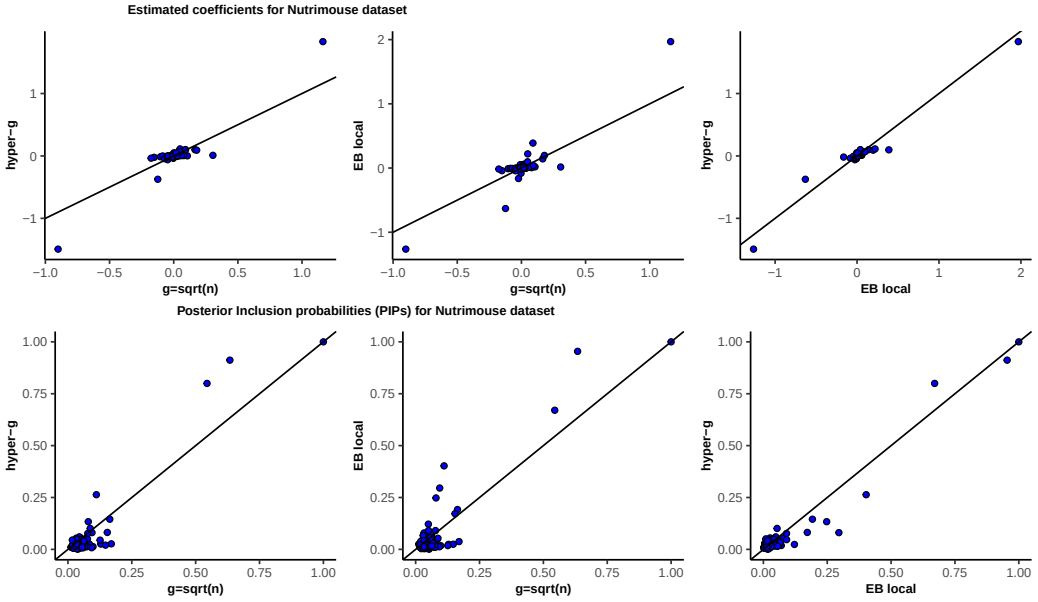

Estimated coefficients for Superconductivity dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for Superconductivity dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local

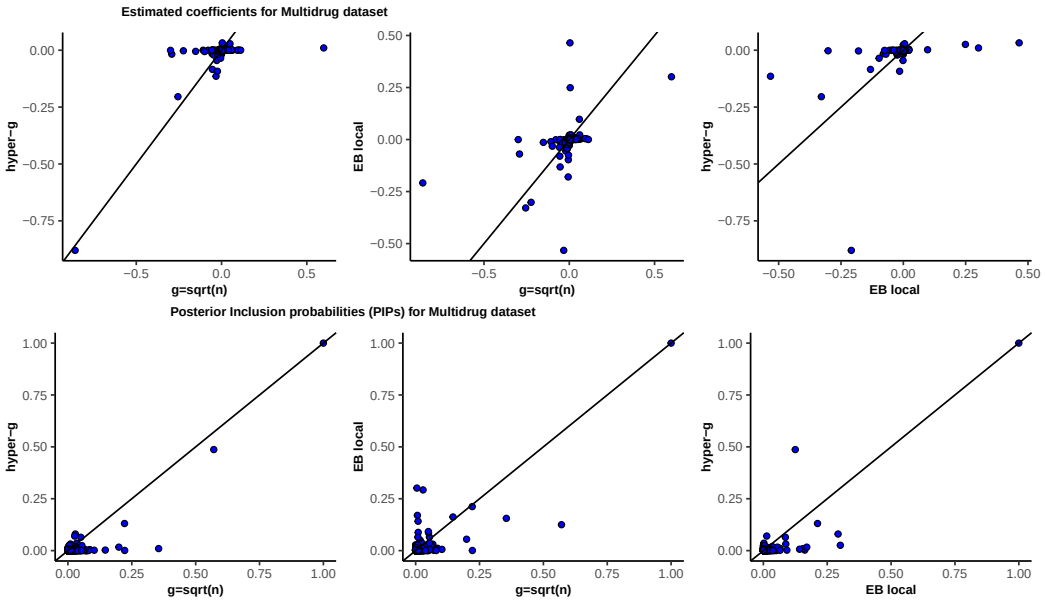

Estimated coefficients for Diabetes dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for Diabetes dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local

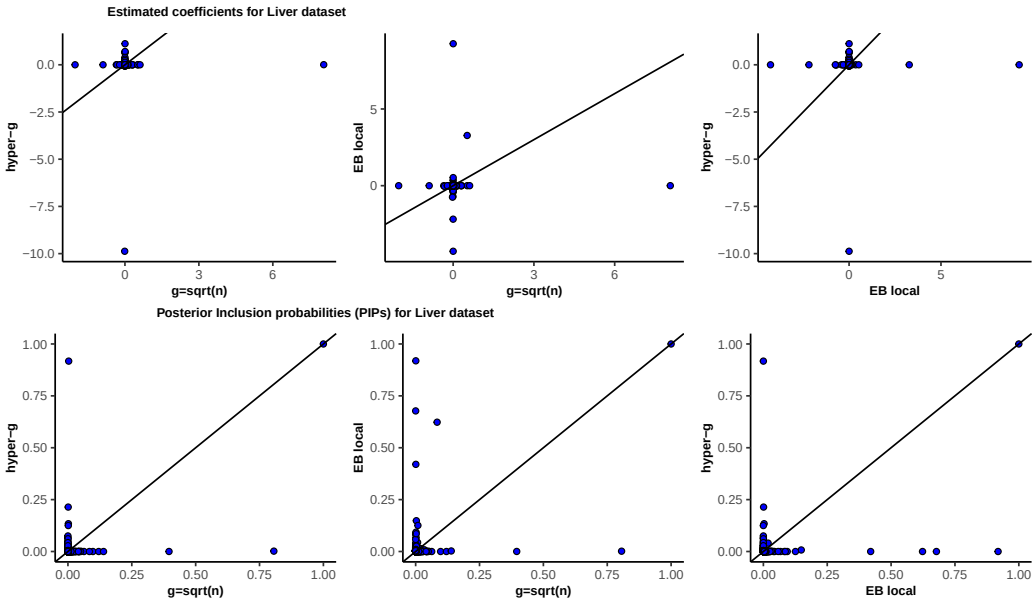

Estimated coefficients for Ozone dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for Ozone dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local


Estimated coefficients for Boston dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for Boston dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local


Estimated coefficients for NIR dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for NIR dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local


Estimated coefficients for Nutrimouse dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for Nutrimouse dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local


Estimated coefficients for Multidrug dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for Multidrug dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local


Estimated coefficients for Liver dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local
Posterior Inclusion probabilities (PIPs) for Liver dataset
hyper−g
g=sqrt(n)
EB local
g=sqrt(n)
hyper−g
EB local

A.3 Dataset specific results for all metrics from Table 1 of paper

	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	22.227 (7.245)	2.537 (0.705)	2.226 (0.46)	1034.446 (298.497)	34.805 (9.633)	77.082 (21.875)	1687.135 (451.607)	24.079 (6.611)	20.003 (9.14)	616.984 (232.253)	273.596 (16.804)	114.398 (8.609)	529.441 (288.16)	580.114 (23.619)
Hyper-g	22.586 (7.721)	2.551 (0.768)	2.625 (0.58)	956.709 (267.792)	36.598 (9.394)	85.687 (27.862)	1759.479 (583.806)	24.356 (7.092)	18.695 (9.885)	632.621 (238.504)	236.12 (35.997)	116.132 (11.514)	618.573 (385.086)	592.795 (33.797)
EB-local	22.545 (7.621)	2.557 (0.776)	2.629 (0.578)	957.555 (272.537)	36.649 (9.496)	85.558 (28)	1747.042 (575.786)	23.984 (5.97)	18.871 (9.964)	636.934 (241.049)	236.609 (35.172)	115.641 (13.319)	589.34 (299.583)	594.547 (28.683)
JZS	22.794 (7.656)	2.569 (0.783)	2.667 (0.59)	957.291 (270.3)	37.391 (9.869)	86.346 (27.435)	1764.418 (498.718)	23.698 (5.913)	19.202 (9.753)	635.859 (230.23)	244.999 (43.769)	119.442 (12.467)	553.072 (339.772)	592.573 (32.734)
Horseshoe	18.981 (5.038)	2.558 (0.722)	2.205 (0.527)	1077.88 (269.534)	39.9 (10.175)	77.353 (22.17)	1606.644 (306.906)	31.088 (4.986)	20.421 (7.797)	645.507 (112.005)	227.22 (31.935)	114.84 (13.471)	292.609 (73.925)	609.424 (51.451)
UIP	22.476 (7.659)	2.651 (0.818)	2.884 (0.617)	978.286 (274.767)	40.422 (9.93)	86.912 (27.717)	1788.751 (381.509)	27.207 (9.039)	18.973 (9.816)	728.504 (230.12)	237.289 (32.48)	118.724 (13.716)	557.655 (314.593)	591.81 (31.976)
EB-global	22.726 (7.71)	2.549 (0.771)	2.633 (0.591)	951.792 (279.081)	36.33 (9.442)	86.094 (28.139)	1819.712 (728.26)	22.314 (5.894)	18.806 (10.462)	714.041 (316.618)	260.116 (41.044)	122.481 (11.471)	674.471 (500.263)	603.453 (37.125)
NLP	27.621 (8.958)	3.289 (0.964)	4.143 (0.528)	1478.751 (411.05)	43.159 (8.815)	80.118 (27.627)	2226.935 (993.586)	27.833 (16.504)	20.422 (10.555)	594.975 (148.281)	240.53 (33.305)	131.037 (23.351)	324.223 (24.929)	569.775 (36.2)
Benchmark	22.598 (7.623)	2.655 (0.814)	2.878 (0.609)	979.953 (272.206)	40.904 (10.893)	87.01 (27.811)	1926.384 (429.822)	25.354 (9.86)	19.68 (9.939)	844.045 (224.59)	298.62 (44.423)	119.506 (6.163)	586.319 (383.109)	586.448 (26.409)
LASSO	21.012 (7.077)	2.636 (0.741)	2.214 (0.593)	1278.31 (385.763)	45.321 (13.169)	77.527 (22.319)	1553.861 (389.273)	43.975 (6.013)	21.347 (7.567)	737.581 (77.445)	256.163 (31.322)	107.926 (7.468)	253.289 (45.72)	579.167 (6.186)
SCAD	24.506 (10.092)	2.517 (0.711)	2.121 (0.49)	1529.596 (568.747)	41.786 (16.041)	83.61 (30.896)	1739.529 (494.73)	40.783 (6.677)	23.485 (12.932)	878.922 (109.591)	299.729 (68.661)	113.935 (11.363)	384.369 (41.925)	590.617 (19.728)
BIC-BAS	22.494 (7.604)	2.646 (0.8)	2.885 (0.609)	972.276 (271.268)	40.549 (11.248)	86.579 (27.942)	1787.937 (375.556)	28.104 (12.947)	19.224 (10.07)	707.524 (263.783)	558.517 (298.812)	127.602 (20.128)	634.026 (405.168)	599.152 (30.036)
BICREG-SIS	22.161 (7.761)	2.822 (0.883)	3.311 (0.562)	980.928 (272.486)	41.712 (10.311)	96.856 (27.552)	1700.149 (483.911)	56.813 (10.349)	30.084 (14.05)	758.83 (196.946)	302.505 (93.614)	140.1 (24.626)	472.652 (161.101)	513.219 (133.737)
Spikeslab	31.239 (7.861)	3.836 (1.253)	4.189 (0.46)	1623.362 (375.367)	62.433 (20.77)	138.361 (23.25)	2182.946 (364.524)	37.74 (13.209)	22.325 (10.339)	975.739 (220.227)	291.055 (34.78)	120.872 (9.248)	467.337 (189.504)	588.711 (45.202)
Elastic net	22.366 (7.616)	2.646 (0.713)	2.215 (0.557)	1338.567 (433.078)	43.148 (12.819)	74.776 (20.983)	1585.461 (395.18)	57.759 (7.87)	23.52 (8.681)	726.937 (72.811)	260.475 (29.084)	107.066 (6.03)	266.535 (23.26)	576.978 (3.694)
MCP	24.511 (10.457)	2.522 (0.739)	2.132 (0.508)	1204.207 (488.497)	40.207 (13.934)	86.5 (32.141)	1768.974 (503.851)	42.068 (8.602)	29.733 (12.781)	900.889 (118.698)	341.086 (76.503)	120.677 (13.14)	390.171 (43.232)	610.649 (10.71)
SS Lasso	33.331 (8.237)	3.909 (1.425)	4.459 (0.39)	1531.5 (377.188)	50.695 (9.345)	123.306 (20.843)	2309.326 (165.478)	44.571 (12.755)	21.425 (11.731)	756.073 (106.423)	279.246 (24.555)	137.111 (19.451)	383.014 (93.883)	594.029 (16.983)
LASSO-1se	44.082 (5.856)	8.022 (1.141)	6.951 (0.967)	2605.602 (305.118)	111.284 (7.785)	158.858 (20.793)	1766.498 (267.723)	100.979 (11.861)	27.973 (6.491)	870.685 (44.102)	278.472 (31.503)	110.447 (6.027)	280.294 (42.131)	577.01 (4.227)
EMVS	26.166 (4.174)	2.643 (0.593)	2.418 (0.619)	1212.878 (482.105)	62.759 (14.277)	130.979 (7.863)	13454.014 (2456.259)	48.824 (4.566)	28.987 (6.895)	751.219 (211.286)	291.633 (13.067)	120.745 (11.997)	397.282 (32.126)	595.787 (5.419)
AIC	23.98 (8.102)	2.558 (0.69)	2.104 (0.457)	1293.306 (431.048)	42.407 (12.324)	73.978 (21.287)	11652.227 (66671.38)	59.213 (26.662)	30.316 (16.117)	2161.99 (4044.25)	831.263 (471.679)	225.118 (128.153)	33136.847 (61879.495)	628.459 (91.614)
$g = 1$	132.911 (3.471)	21.23 (0.44)	20.481 (0.256)	3277.732 (174.725)	309.368 (2.856)	437.009 (8.451)	7873.668 (13099.449)	133.435 (5.263)	44.624 (4.009)	842.916 (38.247)	319.35 (3.959)	115.955 (1.276)	386.166 (17.578)	574.697 (1.539)

Table A.1: Average RMSE($\times$1000) for parameter estimates using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	0.103 (0.041)	0.012 (0.003)	0.01 (0.002)	4.973 (2.017)	0.159 (0.042)	0.332 (0.082)	6.371 (2.472)	0.089 (0.018)	0.07 (0.036)	2.337 (0.88)	1.36 (0.389)	0.397 (0.063)	1.141 (0.371)	0.956 (0.099)
Hyper-g	0.117 (0.081)	0.012 (0.004)	0.014 (0.006)	4.447 (2.588)	0.179 (0.09)	0.415 (0.291)	6.785 (3.589)	0.095 (0.049)	0.06 (0.068)	3.04 (1.843)	1.195 (0.518)	0.409 (0.075)	1.401 (1.172)	0.985 (0.111)
EB-local	0.113 (0.079)	0.012 (0.004)	0.014 (0.006)	4.503 (2.728)	0.182 (0.097)	0.421 (0.296)	6.805 (3.747)	0.096 (0.046)	0.057 (0.069)	2.959 (1.737)	1.151 (0.49)	0.397 (0.078)	1.269 (0.655)	0.987 (0.117)
JZS	0.116 (0.081)	0.012 (0.005)	0.014 (0.007)	4.501 (2.62)	0.186 (0.1)	0.411 (0.294)	7.13 (4.058)	0.087 (0.033)	0.062 (0.071)	2.937 (1.671)	1.258 (0.522)	0.432 (0.082)	1.323 (0.784)	0.976 (0.114)
Horseshoe	0.092 (0.021)	0.012 (0.004)	0.01 (0.003)	4.9 (1.902)	0.179 (0.056)	0.342 (0.127)	8.721 (1.279)	0.14 (0.021)	0.101 (0.027)	2.717 (0.892)	1.098 (0.143)	0.416 (0.088)	0.868 (0.197)	1.141 (0.241)
UIP	0.115 (0.079)	0.013 (0.006)	0.019 (0.01)	4.819 (3.211)	0.242 (0.129)	0.427 (0.306)	8.547 (4.719)	0.129 (0.099)	0.065 (0.082)	4.375 (2.143)	1.121 (0.431)	0.431 (0.086)	1.426 (1.057)	0.976 (0.114)
EB-global	0.12 (0.087)	0.012 (0.005)	0.014 (0.007)	4.559 (3.083)	0.186 (0.106)	0.433 (0.321)	8.762 (4.827)	0.107 (0.056)	0.075 (0.101)	5.677 (2.963)	1.87 (0.658)	0.497 (0.069)	2.543 (2.011)	1.057 (0.133)
NLP	0.195 (0.136)	0.025 (0.011)	0.073 (0.019)	7.449 (4.721)	0.312 (0.152)	0.334 (0.201)	12.268 (5.033)	0.12 (0.102)	0.076 (0.09)	3.175 (1.776)	1.271 (0.495)	0.464 (0.174)	1.429 (0.355)	0.821 (0.188)
Benchmark	0.119 (0.083)	0.013 (0.005)	0.019 (0.01)	4.867 (3.137)	0.243 (0.129)	0.43 (0.311)	10.836 (5.247)	0.127 (0.127)	0.081 (0.102)	7.755 (2.901)	2.546 (0.436)	0.533 (0.029)	1.957 (1.388)	0.976 (0.099)
BIC-BAS	0.113 (0.079)	0.013 (0.005)	0.019 (0.01)	4.797 (3.143)	0.241 (0.123)	0.433 (0.309)	8.531 (4.934)	0.146 (0.148)	0.064 (0.087)	4.117 (2.431)	1.918 (0.843)	0.45 (0.076)	1.565 (1.148)	1.009 (0.104)
BMA -bicreg	0.103 (0.064)	0.014 (0.006)	0.029 (0.013)	4.587 (3.164)	0.239 (0.129)	0.55 (0.431)	5.756 (3.519)	0.798 (0.25)	0.282 (0.195)	6.894 (2.591)	2.477 (1.225)	0.507 (0.136)	2.859 (1.183)	0.614 (0.328)
Spikeslab	0.273 (0.16)	0.034 (0.02)	0.082 (0.018)	16.851 (6.23)	0.7 (0.43)	1.679 (0.773)	15.265 (4.798)	0.285 (0.219)	0.1 (0.094)	10.525 (2.886)	2.445 (0.214)	0.484 (0.056)	1.379 (1.024)	0.939 (0.198)
AIC	0.111 (0.039)	0.012 (0.003)	0.009 (0.002)	5.916 (2.208)	0.194 (0.058)	0.329 (0.087)	26.489 (66.571)	0.259 (0.045)	0.16 (0.06)	5.848 (3.77)	4.906 (3.133)	0.524 (0.117)	60.811 (101.497)	1.04 (0.133)
$g = 1$	1.452 (0.07)	0.38 (0.009)	0.327 (0.007)	44.855 (4.177)	4.667 (0.141)	7.163 (0.262)	55.205 (58.045)	1.491 (0.059)	0.32 (0.047)	5.46 (0.527)	1.959 (0.299)	0.497 (0.043)	1.959 (0.315)	0.979 (0.038)

Table A.2: Average parameter Mean Interval Score using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	0.033 (0.034)	0.002 (0.003)	0.002 (0.003)	0.017 (0.02)	0.008 (0.009)	0.004 (0.006)	0.162 (0.106)	0.026 (0.015)	0.086 (0.113)	0.152 (0.075)	0.517 (0.091)	0.802 (0.097)	0.644 (0.163)	0.932 (0.094)
Hyper-g	0.034 (0.038)	0.001 (0.003)	0.003 (0.004)	0.016 (0.025)	0.01 (0.01)	0.005 (0.009)	0.166 (0.105)	0.029 (0.023)	0.094 (0.133)	0.182 (0.097)	0.397 (0.142)	0.806 (0.114)	0.653 (0.191)	0.939 (0.097)
EB-local	0.034 (0.038)	0.002 (0.003)	0.003 (0.004)	0.015 (0.016)	0.011 (0.012)	0.005 (0.009)	0.165 (0.104)	0.03 (0.023)	0.09 (0.121)	0.182 (0.088)	0.4 (0.13)	0.792 (0.132)	0.693 (0.167)	0.953 (0.089)
JZS	0.033 (0.038)	0.002 (0.003)	0.003 (0.004)	0.015 (0.017)	0.011 (0.01)	0.005 (0.008)	0.162 (0.106)	0.026 (0.016)	0.092 (0.115)	0.177 (0.084)	0.421 (0.153)	0.829 (0.111)	0.665 (0.187)	0.951 (0.091)
Horseshoe	0.03 (0.035)	0.002 (0.009)	0.002 (0.003)	0.023 (0.021)	0.016 (0.018)	0.006 (0.014)	0.137 (0.093)	0.052 (0.021)	0.088 (0.122)	0.185 (0.069)	0.398 (0.112)	0.764 (0.15)	0.495 (0.113)	0.953 (0.071)
UIP	0.035 (0.04)	0.002 (0.003)	0.003 (0.005)	0.015 (0.016)	0.013 (0.011)	0.005 (0.008)	0.172 (0.109)	0.043 (0.039)	0.096 (0.123)	0.236 (0.098)	0.397 (0.128)	0.809 (0.122)	0.68 (0.201)	0.947 (0.08)
EB-global	0.034 (0.037)	0.001 (0.003)	0.003 (0.004)	0.014 (0.018)	0.009 (0.009)	0.005 (0.008)	0.163 (0.105)	0.028 (0.022)	0.09 (0.125)	0.208 (0.092)	0.446 (0.145)	0.86 (0.089)	0.704 (0.186)	0.945 (0.081)
NLP	0.033 (0.038)	0.002 (0.003)	0.006 (0.007)	0.034 (0.025)	0.017 (0.009)	0.004 (0.007)	0.196 (0.092)	0.026 (0.021)	0.102 (0.109)	0.173 (0.082)	0.458 (0.111)	0.693 (0.104)	0.476 (0.059)	0.869 (0.117)
Benchmark	0.034 (0.037)	0.001 (0.003)	0.003 (0.005)	0.015 (0.018)	0.011 (0.01)	0.005 (0.008)	0.185 (0.105)	0.039 (0.044)	0.111 (0.138)	0.356 (0.112)	0.643 (0.125)	0.95 (0.059)	0.737 (0.191)	0.939 (0.091)
LASSO	0.132 (0.099)	0.031 (0.02)	0.006 (0.008)	0.224 (0.087)	0.082 (0.016)	0.137 (0.053)	0.148 (0.092)	0.411 (0.024)	0.154 (0.109)	0.321 (0.042)	0.703 (0.06)	0.756 (0.084)	0.75 (0.066)	0.938 (0.066)
SCAD	0.04 (0.05)	0.001 (0.003)	0.003 (0.007)	0.027 (0.021)	0.025 (0.02)	0.076 (0.045)	0.151 (0.105)	0.32 (0.042)	0.257 (0.168)	0.349 (0.05)	0.764 (0.078)	0.82 (0.093)	0.814 (0.055)	0.969 (0.035)
BIC-BAS	0.033 (0.037)	0.002 (0.003)	0.003 (0.005)	0.016 (0.019)	0.013 (0.011)	0.005 (0.009)	0.172 (0.108)	0.049 (0.053)	0.093 (0.12)	0.217 (0.109)	0.567 (0.189)	0.845 (0.104)	0.707 (0.179)	0.967 (0.068)
BICREG-SIS	0.033 (0.036)	0.001 (0.002)	0.004 (0.006)	0.014 (0.017)	0.014 (0.016)	0.006 (0.009)	0.165 (0.122)	0.27 (0.08)	0.396 (0.256)	0.29 (0.106)	0.649 (0.2)	0.806 (0.127)	0.772 (0.16)	0.625 (0.247)
Spikeslab	0.037 (0.043)	0.002 (0.004)	0.008 (0.008)	0.034 (0.021)	0.03 (0.024)	0.021 (0.025)	0.203 (0.092)	0.075 (0.063)	0.125 (0.115)	0.45 (0.091)	0.584 (0.1)	0.787 (0.067)	0.678 (0.165)	0.883 (0.144)
Elastic net	0.16 (0.098)	0.035 (0.019)	0.007 (0.009)	0.263 (0.066)	0.091 (0.023)	0.247 (0.074)	0.159 (0.09)	0.423 (0.024)	0.224 (0.133)	0.323 (0.04)	0.762 (0.061)	0.754 (0.077)	0.799 (0.06)	0.914 (0.088)
MCP	0.036 (0.047)	0.001 (0.002)	0.002 (0.003)	0.028 (0.023)	0.025 (0.019)	0.108 (0.056)	0.156 (0.114)	0.317 (0.042)	0.303 (0.184)	0.363 (0.051)	0.801 (0.078)	0.883 (0.081)	0.79 (0.049)	0.993 (0.019)
SS Lasso	0.054 (0.07)	0.001 (0.002)	0.005 (0.007)	0.05 (0.05)	0.018 (0.019)	0.02 (0.013)	0.347 (0.097)	0.292 (0.087)	0.124 (0.187)	0.322 (0.099)	0.456 (0.103)	0.79 (0.113)	0.678 (0.147)	0.965 (0.043)
EMVS	0.021 (0.025)	0.003 (0.004)	0.002 (0.004)	0.184 (0.091)	0.042 (0.021)	0.022 (0.01)	0.88 (0.01)	0.103 (0.021)	0.13 (0.085)	0.427 (0.098)	0.708 (0.058)	0.813 (0.095)	0.79 (0.063)	0.99 (0.006)
AIC	0.038 (0.039)	0.002 (0.007)	0.002 (0.003)	0.025 (0.029)	0.012 (0.013)	0.004 (0.008)	0.179 (0.114)	0.062 (0.025)	0.122 (0.13)	0.183 (0.076)	0.718 (0.166)	0.894 (0.105)	0.788 (0.16)	0.962 (0.063)
$g = 1$	0.034 (0.037)	0.002 (0.004)	0.003 (0.004)	0.039 (0.024)	0.015 (0.013)	0.011 (0.011)	0.2 (0.114)	0.09 (0.026)	0.13 (0.1)	0.277 (0.056)	0.738 (0.076)	0.827 (0.095)	0.742 (0.096)	0.943 (0.057)

Table A.3: Average 1-AUPRC for variable selection using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	0.14 (0.022)	0.221 (0.009)	0.163 (0.006)	0.177 (0.031)	0.277 (0.006)	0.073 (0.008)	0.506 (0.054)	0.226 (0.006)	0.282 (0.051)	0.16 (0.044)	0.38 (0.121)	0.889 (0.17)	0.149 (0.045)	0.518 (0.181)
Hyper-g	0.14 (0.024)	0.221 (0.009)	0.164 (0.006)	0.177 (0.031)	0.277 (0.006)	0.072 (0.008)	0.51 (0.057)	0.226 (0.006)	0.282 (0.051)	0.168 (0.048)	0.42 (0.199)	0.921 (0.2)	0.155 (0.059)	0.532 (0.213)
EB-local	0.14 (0.023)	0.221 (0.009)	0.164 (0.006)	0.177 (0.031)	0.277 (0.006)	0.072 (0.008)	0.509 (0.056)	0.226 (0.006)	0.283 (0.054)	0.167 (0.047)	0.438 (0.246)	0.926 (0.215)	0.154 (0.053)	0.535 (0.235)
JZS	0.14 (0.024)	0.221 (0.009)	0.164 (0.006)	0.178 (0.031)	0.277 (0.006)	0.073 (0.008)	0.514 (0.058)	0.226 (0.006)	0.283 (0.052)	0.167 (0.048)	0.411 (0.203)	0.948 (0.196)	0.154 (0.056)	0.546 (0.242)
Horseshoe	0.139 (0.023)	0.221 (0.009)	0.163 (0.006)	0.176 (0.031)	0.277 (0.006)	0.072 (0.008)	0.51 (0.057)	0.226 (0.006)	0.276 (0.045)	0.155 (0.046)	0.419 (0.195)	1.048 (0.366)	0.148 (0.04)	0.765 (0.383)
UIP	0.14 (0.023)	0.221 (0.009)	0.164 (0.006)	0.178 (0.031)	0.277 (0.006)	0.073 (0.008)	0.52 (0.059)	0.227 (0.006)	0.282 (0.049)	0.174 (0.05)	0.407 (0.209)	0.969 (0.291)	0.153 (0.045)	0.531 (0.214)
EB-global	0.14 (0.024)	0.221 (0.009)	0.164 (0.006)	0.178 (0.031)	0.277 (0.006)	0.073 (0.008)	0.516 (0.059)	0.226 (0.006)	0.288 (0.056)	0.187 (0.056)	0.415 (0.228)	0.947 (0.171)	0.175 (0.053)	0.607 (0.248)
NLP	0.141 (0.024)	0.221 (0.009)	0.165 (0.006)	0.177 (0.03)	0.277 (0.006)	0.073 (0.008)	0.536 (0.062)	0.227 (0.006)	0.288 (0.051)	0.185 (0.055)	0.373 (0.174)	1.193 (0.416)	0.195 (0.049)	0.574 (0.183)
Benchmark	0.14 (0.024)	0.221 (0.009)	0.164 (0.006)	0.179 (0.03)	0.277 (0.006)	0.073 (0.008)	0.533 (0.06)	0.227 (0.006)	0.282 (0.045)	0.193 (0.055)	0.445 (0.209)	1.006 (0.094)	0.198 (0.071)	0.673 (0.225)
LASSO	0.139 (0.022)	0.221 (0.009)	0.163 (0.006)	0.176 (0.031)	0.277 (0.006)	0.072 (0.008)	0.508 (0.053)	0.228 (0.006)	0.286 (0.049)	0.159 (0.046)	0.431 (0.208)	0.86 (0.195)	0.174 (0.04)	0.578 (0.171)
SCAD	0.14 (0.024)	0.221 (0.009)	0.163 (0.006)	0.177 (0.031)	0.277 (0.006)	0.072 (0.008)	0.508 (0.059)	0.229 (0.006)	0.302 (0.065)	0.185 (0.056)	0.428 (0.233)	0.893 (0.221)	0.253 (0.055)	0.624 (0.224)
BIC-BAS	0.14 (0.024)	0.221 (0.009)	0.164 (0.006)	0.178 (0.031)	0.277 (0.006)	0.073 (0.008)	0.52 (0.059)	0.227 (0.006)	0.284 (0.052)	0.174 (0.049)	2.06 (3.051)	1.052 (0.407)	0.155 (0.06)	0.538 (0.217)
BICREG-SIS	0.139 (0.024)	0.221 (0.009)	0.164 (0.006)	0.179 (0.031)	0.277 (0.006)	0.073 (0.008)	0.503 (0.055)	0.243 (0.006)	0.361 (0.07)	0.192 (0.053)	0.716 (0.53)	1.511 (0.539)	0.198 (0.058)	0.642 (0.262)
Spikeslab	0.143 (0.024)	0.222 (0.009)	0.165 (0.006)	0.194 (0.032)	0.278 (0.006)	0.074 (0.008)	0.543 (0.06)	0.23 (0.006)	0.286 (0.046)	0.211 (0.067)	0.451 (0.179)	0.978 (0.139)	0.172 (0.049)	0.596 (0.232)
Elastic Net	0.14 (0.023)	0.221 (0.009)	0.163 (0.006)	0.176 (0.03)	0.277 (0.006)	0.072 (0.008)	0.512 (0.056)	0.227 (0.006)	0.285 (0.051)	0.156 (0.045)	0.457 (0.229)	0.838 (0.223)	0.175 (0.04)	0.559 (0.174)
MCP	0.14 (0.024)	0.221 (0.009)	0.163 (0.006)	0.176 (0.031)	0.277 (0.006)	0.072 (0.008)	0.509 (0.062)	0.229 (0.006)	0.3 (0.067)	0.187 (0.061)	0.577 (0.29)	0.951 (0.231)	0.244 (0.063)	0.683 (0.284)
SS Lasso	0.143 (0.024)	0.222 (0.009)	0.166 (0.006)	0.189 (0.031)	0.278 (0.006)	0.073 (0.008)	0.57 (0.043)	0.228 (0.006)	0.295 (0.051)	0.197 (0.056)	0.46 (0.399)	1.413 (0.48)	0.193 (0.057)	0.772 (0.261)
LASSO-1se	0.151 (0.019)	0.225 (0.009)	0.167 (0.006)	0.197 (0.029)	0.28 (0.006)	0.076 (0.008)	0.535 (0.048)	0.231 (0.006)	0.303 (0.04)	0.186 (0.067)	0.424 (0.149)	0.946 (0.099)	0.21 (0.057)	0.629 (0.128)
EMVS	0.141 (0.023)	0.221 (0.009)	0.163 (0.006)	0.873 (0.136)	0.324 (0.009)	0.083 (0.009)	0.622 (0.086)	0.23 (0.006)	0.292 (0.04)	0.177 (0.053)	0.484 (0.308)	0.941 (0.227)	0.366 (0.077)	0.636 (0.201)
AIC	0.14 (0.023)	0.221 (0.009)	0.163 (0.006)	0.177 (0.032)	0.277 (0.006)	0.072 (0.008)	0.515 (0.063)	0.226 (0.006)	0.304 (0.07)	0.169 (0.048)	5.911 (16.127)	2.223 (1.856)	6.825 (14.002)	0.647 (0.343)
$g = 1$	0.355 (0.014)	0.415 (0.006)	0.372 (0.006)	0.379 (0.024)	0.458 (0.004)	0.303 (0.009)	0.624 (0.035)	0.42 (0.004)	0.448 (0.031)	0.361 (0.037)	0.54 (0.082)	0.93 (0.055)	0.371 (0.042)	0.723 (0.077)

Table A.4: Average $1 - R^2_{test}$ for predictions using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	2.179 (0.234)	0.713 (0.019)	0.546 (0.015)	38.002 (5.12)	5.231 (0.08)	3.602 (0.186)	247.793 (15.455)	2.97 (0.061)	1.993 (0.221)	20.316 (4.29)	11.705 (0.993)	16.757 (3.279)	3.75 (0.286)	50.295 (14.652)
Hyper-g	2.14 (0.268)	0.715 (0.02)	0.546 (0.016)	38.426 (5.62)	5.24 (0.082)	3.593 (0.286)	248.108 (16.003)	2.974 (0.063)	2.011 (0.25)	20.57 (4.897)	10.968 (2.041)	16.79 (3.159)	3.655 (0.804)	51.785 (17.85)
EB-local	2.141 (0.269)	0.715 (0.019)	0.546 (0.016)	38.45 (5.65)	5.24 (0.084)	3.598 (0.285)	248.693 (15.881)	2.975 (0.063)	2.01 (0.251)	20.535 (4.929)	11.628 (4.026)	17.474 (4.342)	3.687 (0.792)	51.399 (15.28)
JZS	2.143 (0.269)	0.715 (0.019)	0.546 (0.016)	38.457 (5.684)	5.24 (0.083)	3.601 (0.291)	248.934 (16.13)	2.977 (0.062)	2.016 (0.257)	20.451 (4.948)	11.388 (3.056)	17.315 (4.267)	3.648 (0.775)	53.562 (16.963)
Horseshoe	2.132 (0.268)	0.716 (0.02)	0.546 (0.016)	38.532 (5.75)	5.229 (0.082)	3.582 (0.289)	248.065 (16.767)	2.976 (0.063)	1.986 (0.242)	20.042 (4.891)	12.461 (4.638)	22.87 (14.105)	3.588 (0.708)	155.243 (91.875)
UIP	2.139 (0.269)	0.715 (0.019)	0.546 (0.016)	38.498 (5.739)	5.243 (0.083)	3.6 (0.283)	248.952 (15.804)	2.978 (0.063)	2.023 (0.26)	21.355 (5.177)	10.918 (2.632)	17.722 (4.814)	3.645 (0.684)	51.671 (17.637)
EB-global	2.139 (0.271)	0.715 (0.019)	0.546 (0.016)	38.681 (5.775)	5.24 (0.083)	3.603 (0.292)	248.977 (16.472)	2.978 (0.063)	2.045 (0.262)	23.215 (5.58)	13.262 (8.906)	17.065 (4.434)	4.185 (1.167)	60.925 (24.562)
NLP	2.154 (0.267)	0.715 (0.019)	0.548 (0.016)	38.467 (5.719)	5.241 (0.085)	3.579 (0.283)	251.78 (16.576)	2.978 (0.063)	2.033 (0.267)	22.942 (5.337)	10.865 (3.875)	29.404 (18.712)	4.101 (0.935)	60.889 (22.19)
Benchmark	2.143 (0.272)	0.715 (0.019)	0.546 (0.016)	38.592 (5.719)	5.241 (0.082)	3.6 (0.288)	250.164 (14.532)	2.977 (0.063)	2.037 (0.268)	22.871 (5.435)	11.278 (2.755)	16.377 (3.082)	4.095 (0.715)	58.899 (20.197)
BIC-BAS	2.139 (0.268)	0.715 (0.019)	0.546 (0.016)	38.572 (5.616)	5.242 (0.084)	3.6 (0.286)	249.039 (15.323)	2.977 (0.062)	2.023 (0.253)	21.249 (4.973)	42.039 (34.339)	18.389 (6.603)	3.706 (0.831)	53.948 (21.88)
BICREG-SIS	2.139 (0.268)	0.714 (0.019)	0.546 (0.016)	38.724 (5.711)	5.235 (0.082)	3.625 (0.301)	248.026 (17.193)	3.041 (0.058)	2.271 (0.253)	24.199 (5.976)	34.378 (22.277)	54.369 (24.741)	4.323 (1.081)	90.871 (53.324)
Spikeslab	2.16 (0.27)	0.715 (0.019)	0.549 (0.016)	39.647 (5.737)	5.231 (0.087)	3.697 (0.294)	251.747 (15.192)	2.984 (0.062)	2.063 (0.282)	24.519 (6.453)	11.586 (3.199)	16.638 (3.788)	3.82 (0.895)	57.489 (23.42)
AIC	2.144 (0.268)	0.716 (0.019)	0.545 (0.016)	38.481 (5.804)	5.23 (0.081)	3.565 (0.292)	256.391 (21.61)	2.976 (0.062)	2.037 (0.261)	18.909 (3.655)	89.41 (48.949)	26.811 (14.668)	7.875 (2.61)	60.016 (25.634)
$g = 1$	3.278 (0.07)	0.923 (0.009)	0.861 (0.015)	51.261 (2.987)	6.435 (0.016)	7.356 (0.063)	273.295 (6.171)	3.781 (0.014)	2.574 (0.147)	32.661 (2.993)	14.934 (1.201)	15.953 (1.415)	5.897 (0.108)	60.632 (11.203)

Table A.5: Average Mean Interval Score for predictions using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	9.07 (0.42)	19.78 (0.31)	21.06 (0.35)	21.8 (0.97)	25.34 (1.02)	19.56 (0.5)	11.26 (1.56)	77.37 (0.85)	12.13 (1.32)	38.55 (3.18)	8.12 (1.53)	6.04 (3.05)	16.2 (1.74)	5.13 (1.46)
Hyper-g	6.79 (0.34)	17.79 (0.42)	17.3 (1.55)	17.64 (1)	20.66 (1.03)	15.93 (0.68)	8.96 (1.69)	71.27 (1.9)	8.48 (1.12)	31.47 (3.22)	7.19 (1.92)	5.68 (3.48)	13.88 (1.6)	4.52 (1.28)
EB-local	6.8 (0.32)	17.79 (0.42)	17.3 (1.48)	17.66 (1.02)	20.67 (1.02)	15.92 (0.66)	9.18 (1.72)	71.23 (1.78)	8.47 (1.14)	31.67 (3.4)	7.63 (2.04)	6.37 (2.9)	13.72 (1.64)	4.51 (1.22)
JZS	6.72 (0.32)	17.69 (0.44)	16.94 (1.57)	17.31 (1.03)	20.28 (0.99)	15.87 (0.66)	7.37 (1.2)	70.05 (2.07)	7.94 (1.06)	30.85 (3.28)	6.62 (2.01)	3.65 (2.39)	13.64 (1.71)	4.33 (1.2)
Horseshoe	8.64 (1.05)	17.82 (0.77)	20.8 (0.65)	21.4 (1.33)	27.16 (1.38)	18.93 (0.97)	8.02 (1.59)	73.52 (1.62)	12.3 (2)	41.84 (3.85)	5.29 (3.24)	2.24 (2.65)	3.2 (1.24)	8.59 (4.81)
UIP	6.81 (0.32)	17.11 (0.46)	15.25 (1.29)	15.97 (0.95)	18.92 (0.73)	15.78 (0.62)	5.88 (0.86)	61.39 (1.39)	7.29 (0.82)	26.33 (2.74)	6.88 (1.85)	3.6 (1.85)	13.15 (1.66)	4.69 (1.35)
EB-global	6.71 (0.33)	17.85 (0.43)	17.53 (1.67)	16.81 (1.08)	20.41 (1.09)	15.9 (0.69)	6.48 (1.1)	74.72 (3.42)	6.75 (0.87)	27.62 (3.96)	3.72 (0.83)	1.76 (0.38)	10.57 (1.61)	3.13 (0.73)
NLP	4.57 (0.32)	14.06 (0.43)	9.57 (0.59)	18.28 (1.1)	17.89 (0.5)	16.05 (0.85)	3 (0.62)	66.23 (1.42)	5.1 (0.82)	28.3 (4.08)	20.01 (2.86)	32.74 (3.25)	35.04 (2.39)	28.19 (5.09)
Benchmark	6.82 (0.32)	17.11 (0.47)	15.31 (1.3)	15.58 (0.89)	18.93 (0.73)	15.79 (0.63)	4.21 (0.53)	61.24 (1.42)	5.85 (0.52)	15.71 (2.38)	2.62 (0.39)	1.09 (0.2)	6.66 (0.99)	2.06 (0.32)
LASSO	12.2 (1.62)	20.83 (1.01)	21.83 (0.38)	26.77 (1.43)	31.38 (1.15)	21.31 (1.57)	18.23 (4.82)	79.51 (1.48)	17.52 (4.29)	67.29 (11.57)	15.75 (7.41)	19.48 (12.21)	36.46 (3.31)	24.4 (12.81)
SCAD	10.15 (2.19)	18.51 (0.73)	21.73 (0.45)	23.24 (3.31)	29.43 (1.41)	17.63 (1.87)	15.35 (3.04)	62.58 (1.58)	12.04 (3.26)	48.89 (17.74)	7.02 (2.44)	9.64 (4.45)	11.59 (1.9)	11.66 (3.87)
BIC-BAS	6.84 (0.32)	17.12 (0.47)	15.31 (1.3)	16.07 (0.95)	18.91 (0.71)	15.86 (0.66)	5.95 (0.89)	61.25 (1.32)	7.32 (0.86)	27.08 (3.01)	25.76 (2.57)	6.64 (6.09)	14.05 (1.72)	5.2 (1.41)
BICREG-SIS	6.99 (0.27)	16.48 (0.46)	13.73 (0.72)	15.44 (0.78)	18.56 (0.59)	14.95 (0.48)	8.33 (0.61)	29.9 (0.32)	3.29 (1.24)	23.94 (2.24)	7.68 (0.91)	10.4 (1.09)	9 (1.18)	8.27 (1.64)
Spikeslab	4.03 (0.34)	13.4 (0.31)	9.06 (0.49)	10.14 (0.6)	15.81 (0.9)	12.35 (0.62)	2.22 (0.19)	45.35 (2.04)	3.94 (0.18)	8.91 (1.38)	1.55 (0.35)	0.35 (0.24)	5.66 (0.39)	1.84 (0.47)
Elastic Net	13.49 (1.57)	20.94 (1.04)	21.84 (0.37)	26.88 (1.48)	31.62 (1.05)	22.06 (0.98)	20.24 (6.26)	80.27 (0.95)	20.53 (4.75)	76.05 (9.84)	26.94 (14.26)	51.66 (35.9)	47.06 (10)	60.37 (39.08)
MCP	9.31 (2.39)	18.26 (0.61)	21.74 (0.44)	23.58 (3.6)	29.17 (1.58)	17.54 (1.94)	11.37 (2.36)	62.64 (1.46)	10.49 (3.1)	47.92 (18.13)	5.3 (2.15)	4.27 (2.17)	11.97 (2.16)	6.08 (2.41)
SS Lasso	5.25 (0.48)	14.27 (0.45)	9.4 (0.64)	12.1 (1.09)	16.48 (0.69)	13.48 (0.67)	3 (0)	66.43 (1.78)	5.04 (0.55)	17.9 (2.49)	3.55 (1.01)	5.89 (2.81)	7.6 (0.79)	2.56 (0.83)
LASSO-1se	8.9 (0.81)	18.66 (0.76)	16.81 (1.32)	17.8 (2.04)	18.48 (1.84)	17.6 (1.01)	8.62 (2.35)	67.56 (1.8)	10.39 (1.75)	41.52 (11.98)	9.12 (4.61)	5.43 (6.62)	29.65 (5.13)	10.27 (6.73)
EMVS	1 (0)	1 (0)	1 (0)	18.63 (1.44)	7.53 (0.72)	8.25 (0.69)	56.58 (2.37)	36.45 (1.43)	1.69 (0.51)	40.11 (3.95)	5.56 (1.48)	0.34 (0.59)	3.09 (1.13)	5.46 (2.95)
AIC	12.72 (0.46)	21.31 (0.1)	21.72 (0.07)	26.46 (0.45)	31.55 (0.28)	21.9 (0.18)	22.85 (3.69)	80.15 (0.21)	22.35 (3.27)	77.2 (5.82)	27.85 (0.49)	26.86 (8.92)	95.85 (12.23)	14.79 (8.66)
$g = 1$	11.56 (0.4)	21.34 (0.08)	20.82 (0.19)	25.7 (0.52)	31.58 (0.27)	20.74 (0.33)	30.97 (5.26)	79.69 (0.28)	21.07 (2.06)	41.5 (3.99)	10.15 (2.14)	5.11 (2.92)	20.29 (2.81)	4.24 (1.59)

Table A.6: Average model size for predictions using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	0.055 (0.003)	0.166 (0.01)	0.15 (0.008)	0.201 (0.009)	0.288 (0.051)	0.15 (0.007)	0.191 (0.027)	0.626 (0.133)	0.144 (0.016)	0.363 (0.107)	0.185 (0.018)	0.389 (0.076)	0.277 (0.024)	2.306 (0.663)
Hyper-g	0.481 (0.092)	1.214 (0.073)	1.117 (0.045)	1.29 (0.115)	2.434 (0.139)	0.466 (0.019)	0.307 (0.033)	3.135 (0.207)	0.34 (0.036)	0.982 (0.138)	0.36 (0.099)	0.436 (0.096)	0.538 (0.048)	2.34 (0.647)
EB-local	0.045 (0.003)	0.156 (0.01)	0.148 (0.008)	0.154 (0.013)	0.242 (0.061)	0.134 (0.009)	0.167 (0.024)	0.567 (0.112)	0.111 (0.011)	0.261 (0.075)	0.174 (0.023)	0.385 (0.079)	0.23 (0.016)	2.189 (0.608)
JZS	1.833 (0.154)	3.24 (0.195)	2.881 (0.116)	1.478 (0.065)	3.374 (0.86)	2.251 (0.102)	1.892 (0.131)	5.207 (1.763)	1.637 (0.104)	1.415 (0.308)	1.822 (0.098)	1.409 (0.231)	1.881 (0.065)	5.781 (3.093)
Horseshoe	3 (0.089)	6.857 (0.453)	6.558 (0.161)	3.492 (0.059)	16.174 (1.275)	3.716 (0.092)	7.627 (0.376)	32.715 (8.086)	5.351 (0.183)	8.558 (1.741)	3.966 (0.082)	10.246 (1.57)	18.304 (0.518)	187.127 (130.548)
UIP	0.044 (0.003)	0.155 (0.053)	0.144 (0.031)	0.141 (0.056)	0.22 (0.029)	0.134 (0.025)	0.139 (0.035)	0.413 (0.087)	0.109 (0.03)	0.219 (0.093)	0.179 (0.035)	0.348 (0.089)	0.224 (0.031)	2.204 (0.641)
EB-global	0.05 (0.004)	0.161 (0.01)	0.154 (0.008)	0.168 (0.015)	0.243 (0.018)	0.141 (0.008)	0.188 (0.026)	0.569 (0.11)	0.122 (0.013)	0.283 (0.079)	0.205 (0.029)	0.413 (0.101)	0.246 (0.019)	2.348 (0.705)
NLP	0.037 (0.005)	0.671 (0.08)	0.343 (0.067)	1.01 (0.174)	1.325 (0.186)	0.734 (0.105)	1.02 (0.66)	7.537 (2.069)	0.706 (0.187)	6.963 (2.985)	27.281 (11.043)	407.185 (181.362)	75.628 (19.532)	1044.758 (828.306)
Benchmark	0.044 (0.003)	0.155 (0.03)	0.146 (0.033)	0.139 (0.032)	0.224 (0.021)	0.136 (0.027)	0.131 (0.042)	0.561 (0.295)	0.102 (0.035)	0.16 (0.042)	0.122 (0.006)	0.3 (0.06)	0.194 (0.01)	2.132 (0.611)
LASSO	0.11 (0.009)	0.237 (0.019)	0.232 (0.009)	0.12 (0.008)	0.49 (0.026)	0.14 (0.007)	0.485 (0.13)	1.434 (0.16)	0.23 (0.036)	0.478 (0.157)	0.168 (0.013)	0.178 (0.032)	0.438 (0.081)	1.172 (1.062)
SCAD	0.155 (0.011)	1.525 (0.108)	1.458 (0.081)	0.383 (0.029)	6.449 (0.464)	0.589 (0.054)	2.936 (0.708)	61.655 (7.135)	0.798 (0.191)	4.109 (0.884)	0.098 (0.006)	0.361 (0.042)	0.221 (0.023)	3.821 (3.5)
BIC-BAS	0.044 (0.003)	0.152 (0.01)	0.144 (0.01)	0.141 (0.033)	0.227 (0.033)	0.133 (0.009)	0.136 (0.016)	0.539 (0.103)	0.104 (0.009)	0.221 (0.078)	0.474 (0.04)	0.373 (0.105)	0.223 (0.015)	2.155 (0.536)
BICREG-SIS	0.5 (0.05)	1.434 (0.102)	1.958 (0.489)	2.996 (0.181)	5.904 (1.39)	1.498 (0.059)	12.444 (2.161)	34.744 (4.794)	0.332 (0.075)	63.468 (32.842)	0.729 (0.138)	1.991 (0.551)	2.704 (0.56)	5.651 (2.524)
Spikeslab	0.359 (0.028)	1.398 (0.099)	0.937 (0.104)	1.482 (0.175)	2.561 (0.186)	1.387 (0.101)	1.496 (0.175)	9.037 (2.267)	1.211 (0.088)	4.055 (1.063)	5.487 (0.207)	21.35 (5.115)	12.186 (0.846)	113.032 (35.903)
Elastic net	6.775 (0.273)	14.366 (0.696)	17.329 (1.196)	10.89 (0.61)	30.132 (6.08)	9.925 (0.384)	20.376 (2.659)	68.895 (5.133)	10.484 (0.85)	19.929 (4.058)	10.138 (0.612)	14.569 (1.167)	24.1 (1.785)	59.483 (25.401)
MCP	0.15 (0.008)	1.435 (0.125)	1.447 (0.088)	0.318 (0.029)	5.443 (0.349)	0.545 (0.044)	2.964 (0.672)	59.008 (6.804)	0.792 (0.196)	4.279 (0.875)	0.09 (0.004)	0.273 (0.029)	0.176 (0.011)	0.817 (0.145)
SS Lasso	0.015 (0.006)	0.14 (0.028)	0.211 (0.05)	0.015 (0.004)	0.442 (0.105)	0.039 (0.007)	0.013 (0.002)	4.179 (0.917)	0.082 (0.015)	0.07 (0.013)	0.082 (0.023)	0.539 (0.129)	0.225 (0.041)	0.86 (0.866)
LASSO-1se	0.11 (0.009)	0.237 (0.019)	0.232 (0.009)	0.12 (0.008)	0.49 (0.026)	0.14 (0.007)	0.485 (0.13)	1.434 (0.16)	0.23 (0.036)	0.478 (0.157)	0.168 (0.013)	0.178 (0.032)	0.438 (0.081)	1.172 (1.062)
EMVS	0.046 (0.003)	0.134 (0.008)	0.131 (0.004)	0.439 (0.058)	0.348 (0.021)	0.071 (0.004)	8.293 (5.131)	2.731 (0.566)	0.348 (0.043)	3.016 (0.982)	0.465 (0.01)	0.771 (0.081)	9.176 (0.41)	4.204 (1.54)
AIC	0.056 (0.003)	0.168 (0.011)	0.153 (0.012)	0.19 (0.009)	0.298 (0.051)	0.151 (0.008)	0.356 (0.113)	1.365 (0.437)	0.246 (0.057)	1.055 (0.357)	0.525 (0.036)	0.638 (0.144)	2.021 (2.429)	2.344 (0.665)
$g = 1$	0.056 (0.004)	0.141 (0.009)	0.114 (0.007)	0.171 (0.012)	0.253 (0.053)	0.114 (0.007)	0.608 (0.147)	1.657 (0.364)	0.321 (0.047)	0.664 (0.236)	0.23 (0.028)	0.435 (0.095)	0.456 (0.069)	2.716 (0.837)

Table A.7: Average computation using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

A.4 R^2_{test} *vs.* $\hat{p}$ *plots for all datasets*

For tall datasets ($n > p$), we observe that Bayesian techniques select sparser models with similar/superior prediction accuracy. Figure A.1 illustrates our claim for tall datasets. Note that $g = 1$ is excluded from the graphs since it had significantly lower R^2_{test} compared to other techniques and EMVS is excluded since it does not provide a sparse estimate in its default settings. In particular, we note that AIC, LASSO-λ_{min} and elastic net consistently tend to select denser models for each dataset without any increase in the prediction accuracy. We also note that while $LASSO = \lambda_{1se}$ selects sparser model, it tends to have lower accuracy than other Bayesian techniques at the same sparsity levels. We also note that Spike-Slab and SSLASSO tent to select very sparse models with slightly lower accuracy compared to other methods. Most other methods are clustered together in terms of both average model size and point predictive accuracy.

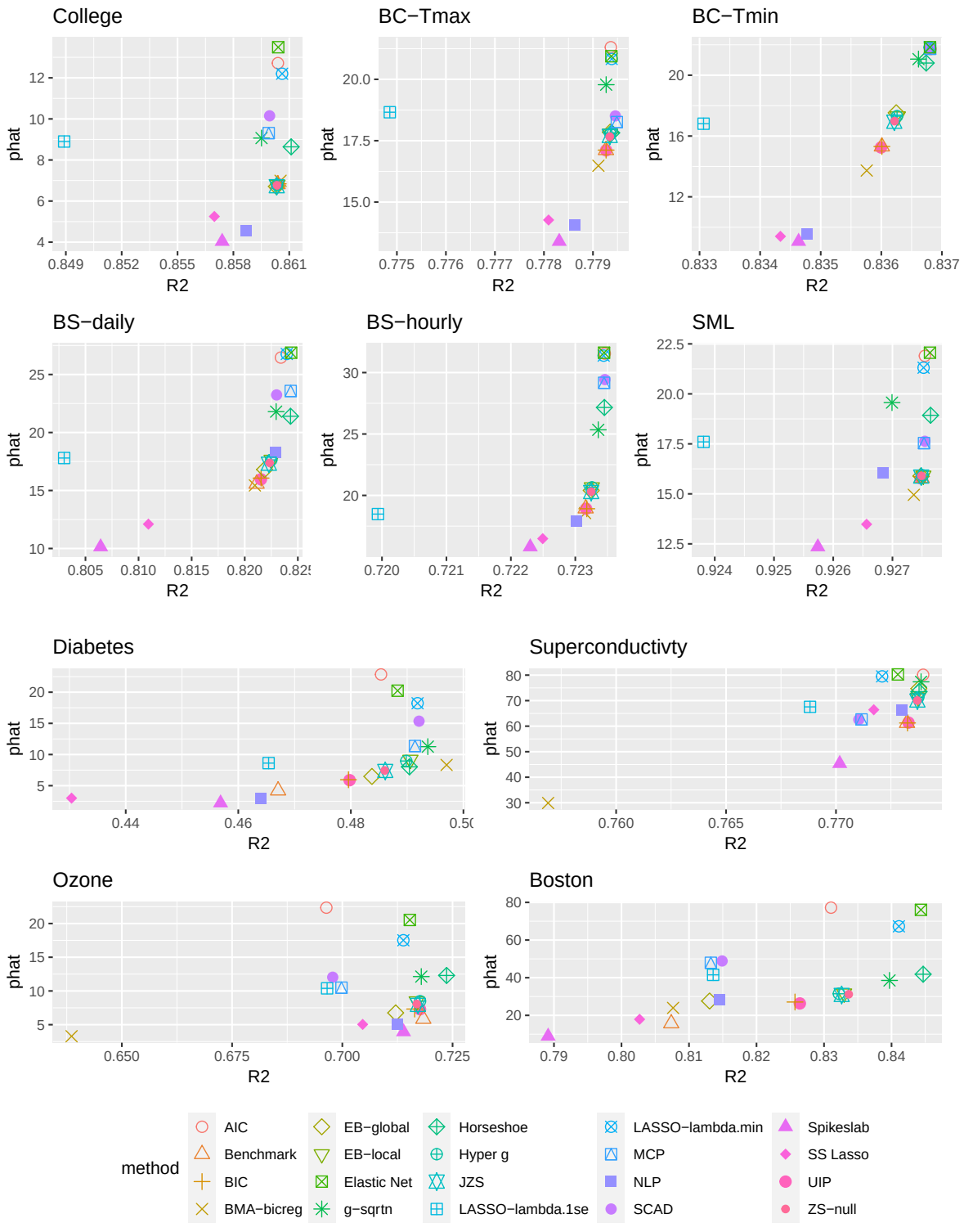


Figure A.1: R^2_{test} vs. $\hat{p}$ plotted for all the tall datasets; $g = 1$ is excluded because it had much lower R^2 than other techniques in all the datasets

For wide datasets ($p > n$), we observe much more variability in the predictive accuracy along with average model size. We note that AIC, BIC and bicreg, tend to overfit models by selecting models with $p \approx n_{train}$ and often perform worse than baseline thus having negative R^2_{test}. We excluded $g = 1$ and EMVS for the same reason as tall datasets. We observe Elastic Net and LASSO-λ_{min} still tend to have much higher average model size without significant increase in accuracy with an exception in Multidrug dataset. Among Bayesian techniques, NLP tends to select denser models. We note that $g = \sqrt{n}$ tend to consistently outperform the other techniques in terms of superior accuracy with sparse models.

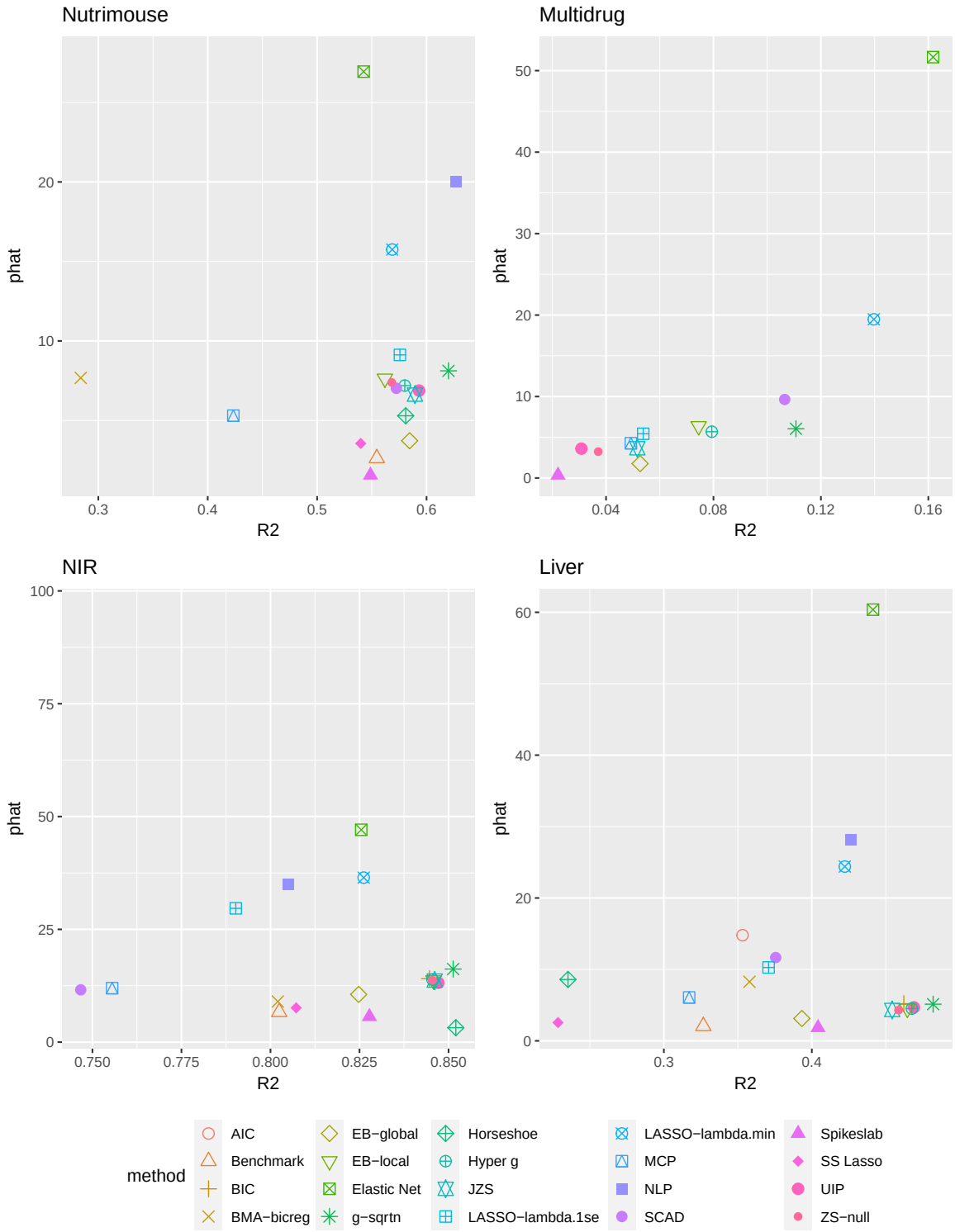


Figure A.2: R^2_{test} vs. $\hat{p}$ plotted for all the wide datasets; Methods with negative R^2_{test} and $g = 1$ are excluded from the plot since they had significantly lower R^2 compared to other techniques in the study for all the wide datasets

Appendix B

APPENDIX B

B.1 Dataset specific results for all metrics from Table 2 of paper

The following document contains results for each datasets under various combinations of parameter and model space priors discussed in the paper. We report the mean and standard deviation of metrics over 100 bootstrapped datasets (or train-test splits for prediction metrics) under different methods.

Table B.1: Average RMSE($\times$1000) for parameter estimates using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

Parameter prior	$P(\mathcal{M}_\gamma)$	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g=\sqrt{n}$	BB(1, 1)	22.207 (5.761)	2.598 (0.59)	2.18 (0.457)	1043.521 (358.873)	31.68 (8.793)	77.741 (18.987)	1668.608 (379.832)	23.521 (5.407)	18.606 (8.831)	591.299 (155.459)	274.28 (21.292)	113.247 (8.97)	487.037 (196.942)	582.706 (20.282)
$g=\sqrt{n}$	BB(1, b_{SDM})	22.259 (5.923)	2.596 (0.594)	2.233 (0.476)	1013.931 (347.818)	31.823 (9.784)	79.172 (20.186)	1653.906 (354.742)	22.834 (5.507)	18.259 (8.638)	605.003 (167.447)	278.589 (20.737)	114.133 (9.059)	519.534 (249.779)	579.186 (20.63)
$g=\sqrt{n}$	BB(1, b_{EB})	22.291 (5.931)	2.658 (0.564)	2.097 (0.429)	1106.556 (402.817)	31.758 (8.48)	79.042 (17.516)	1669.14 (343.569)	23.048 (5.423)	18.484 (9.048)	600.231 (177.353)	285.239 (20.997)	113.864 (9.783)	462.058 (158.909)	582.47 (20.867)
Hyper-g	BB(1, 1)	22.973 (6.909)	2.599 (0.686)	2.56 (0.602)	985.092 (353.002)	32.814 (10.165)	87.417 (27.023)	1673.154 (368.317)	22.604 (5.763)	18.182 (9.556)	616.533 (202.572)	240.157 (41.378)	115.706 (9.372)	554.777 (329.007)	587.943 (27.507)
EB-local	BB(1, 1)	22.894 (6.792)	2.6 (0.671)	2.562 (0.604)	976.349 (339.751)	33.419 (11.641)	87.299 (27.342)	1693.532 (400.038)	23.479 (6.302)	18.623 (10.603)	621.8 (197.197)	241.141 (51.57)	118.252 (11.948)	585.861 (357.749)	584.972 (33.328)
EB-local	BB(1, b_{EB})	23.219 (6.962)	2.587 (0.676)	2.466 (0.629)	980.613 (340.176)	33.136 (10.873)	87.118 (27.055)	1714.391 (382.961)	23.277 (6.184)	18.483 (10.136)	658.388 (210.368)	249.552 (44.873)	113.78 (10.206)	590.075 (426.387)	589.585 (25.596)
Hyper-g	BB(1, b_{EB})	23.191 (6.867)	2.6 (0.677)	2.451 (0.655)	980.642 (347.538)	33.807 (12.319)	86.686 (27.094)	1685.186 (377.698)	22.849 (5.199)	18.418 (11.17)	619.326 (201.491)	259.113 (47.915)	114.823 (10.135)	529.459 (294.107)	587.238 (27.956)
EB-local	BB(1, b_{SDM})	22.953 (6.842)	2.616 (0.689)	2.66 (0.595)	981.754 (345.046)	33.782 (11.2)	89.817 (27.069)	1664.905 (370.299)	23.26 (5.927)	18.832 (10.785)	685.187 (404.012)	247.745 (46.778)	115.979 (11.977)	595.952 (299.174)	587.21 (26.846)
Hyper-g	BB(1, b_{SDM})	22.887 (6.83)	2.619 (0.689)	2.678 (0.603)	981.395 (333.401)	33.28 (10.418)	90.045 (27.104)	1688.816 (371.326)	23.897 (6.156)	19.011 (11.437)	656.405 (213.028)	253.786 (43.173)	116.242 (9.703)	543.659 (293.605)	590.748 (28.218)
$g=\sqrt{n}$	Ber(θ_{SDM})	22.006 (5.811)	2.684 (0.672)	2.803 (0.529)	1032.178 (311.581)	34.193 (10.257)	99.292 (22.799)	1640.878 (341.53)	23.553 (5.512)	18.297 (9.328)	728.053 (209.298)	274.731 (16.995)	113.882 (11.289)	491.486 (198.711)	577.676 (33.901)
$g=\sqrt{n}$	Ber(θ_{EB})	21.939 (5.761)	2.627 (0.638)	2.581 (0.526)	992.723 (334.607)	31.659 (9.388)	88.082 (23.45)	2392.042 (729.524)	24.052 (6.078)	23.864 (8.086)	656.221 (242.858)	290.414 (22.604)	116.446 (9.979)	519.892 (281.369)	588.495 (32.884)
EB-local	Ber(θ_{EB})	22.176 (6.585)	2.727 (0.721)	3.093 (0.523)	979.474 (327.011)	33.448 (10.046)	98.627 (26.455)	2338.274 (623.536)	22.522 (5.834)	20.498 (9.18)	686.677 (551.51)	260.832 (42.39)	118.179 (12.938)	601.838 (438.893)	595.501 (38.973)
Hyper-g	Ber(θ_{EB})	22.12 (6.722)	2.731 (0.721)	3.077 (0.534)	977.353 (326.038)	33.81 (10.356)	98.781 (25.752)	2300.879 (567.991)	22.907 (5.748)	20.668 (9.366)	628.917 (232.159)	269.944 (46.379)	117.796 (12.389)	629.506 (426.174)	598.202 (36.251)
Hyper-g	Ber(θ_{SDM})	22.302 (6.692)	2.83 (0.747)	3.326 (0.501)	1042.542 (311.955)	38.716 (12.02)	107.584 (23.778)	1676.267 (400.587)	26.251 (9.113)	18.091 (10.374)	796.317 (276.339)	245.653 (45.012)	117.331 (10.424)	524.908 (290.718)	595.861 (29.346)
EB-local	Ber(θ_{SDM})	22.377 (6.763)	2.831 (0.759)	3.328 (0.49)	1049.635 (315.02)	38.59 (12.126)	108.076 (23.499)	1670.476 (414.775)	26.83 (9.475)	18.445 (10.998)	752.631 (224.586)	241.437 (43.73)	117.516 (13.266)	618.783 (376.452)	589.002 (36.801)
$g=\sqrt{n}$	Uniform	21.96 (5.807)	2.616 (0.641)	2.587 (0.528)	994.988 (337.794)	32.325 (9.969)	88.505 (23.417)	2382.497 (629.129)	24.523 (6.044)	24.056 (8.09)	645.285 (255.093)	336.183 (74.749)	151.053 (26.343)	897.529 (580.793)	600.036 (38.725)
Hyper-g	Uniform	22.173 (6.597)	2.717 (0.71)	3.084 (0.54)	979.619 (326.742)	34.301 (10.824)	99.142 (25.94)	2348.285 (669.215)	22.888 (5.604)	20.642 (8.924)	618.653 (218.722)	309.117 (88.412)	148.506 (30.53)	889.166 (578.477)	609.027 (44.628)
EB-local	Uniform	22.138 (6.676)	2.726 (0.73)	3.091 (0.524)	979.713 (330.133)	33.808 (10.574)	98.361 (26.194)	2363.269 (666.496)	22.223 (5.8)	20.397 (9.205)	628.714 (278.923)	305.528 (93.645)	155.001 (30.58)	849.621 (484.487)	614.925 (39.509)
$g=\sqrt{n}$	Complexity(1)	25.482 (6.802)	3.074 (0.766)	3.603 (0.467)	1307.143 (356.039)	41.933 (11.081)	123.797 (14.064)	1775.838 (325.383)	28.86 (9.303)	19.073 (9.772)	873.637 (155.562)	294.929 (24.626)	116.806 (8.923)	516.183 (334.78)	582.38 (21.028)
EB-local	Complexity(1)	27.497 (7.155)	3.38 (0.879)	3.985 (0.453)	1317.758 (373.046)	47.482 (12.872)	125.009 (15.645)	1864.424 (347.932)	34.743 (13.949)	20.431 (11.544)	891.94 (212.928)	283.775 (31.757)	116.708 (9.363)	601.749 (385.919)	588.244 (30.382)
Hyper-g	Complexity(1)	27.447 (7.145)	3.383 (0.89)	3.987 (0.43)	1327.962 (376.787)	47.684 (13.027)	124.562 (14.129)	1856.406 (349.399)	34.638 (13.01)	20.157 (11.617)	924.667 (228.576)	281.134 (20.111)	116.22 (8.649)	597.397 (412.165)	592.085 (26.063)
$g=\sqrt{n}$	Complexity(2)	31.134 (9.316)	3.934 (0.955)	4.185 (0.457)	1924.936 (411.71)	57.613 (13.49)	134.089 (14.555)	2191.659 (291.97)	49.882 (15.316)	22.828 (15.087)	941.107 (58.866)	361.813 (25.643)	117.937 (0.697)	496.401 (220.271)	581.388 (13.825)
EB-local	Complexity(2)	33.128 (9.434)	4.239 (1.067)	4.334 (0.442)	1866.438 (437.31)	64.63 (15.908)	132.629 (15.154)	2264.994 (328.33)	56.986 (27.617)	22.202 (13.413)	950.344 (69.237)	356.894 (57.982)	118.076 (1.999)	512.882 (226.216)	581.809 (18.456)
Hyper-g	Complexity(2)	33.138 (9.245)	4.248 (1.079)	4.371 (0.464)	1863.606 (440.693)	63.52 (14.751)	132.217 (14.762)	2263.618 (326.846)	59.805 (28.784)	22.034 (13.212)	953.545 (74.598)	355.904 (51.698)	118.021 (0.357)	531.29 (318.641)	579.524 (21.755)

Table B.2: Average Mean interval score for parameter estimates using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

Parameter prior	$P(\mathcal{M}_\gamma)$	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	BB(1, 1)	0.104 (0.048)	0.012 (0.003)	0.01 (0.002)	4.895 (2.406)	0.141 (0.037)	0.33 (0.09)	6.684 (2.82)	0.091 (0.027)	0.066 (0.04)	2.271 (0.686)	1.395 (0.411)	0.385 (0.071)	1.154 (0.39)	0.969 (0.091)
$g = \sqrt{n}$	BB(1, b_{SDM})	0.105 (0.05)	0.012 (0.003)	0.01 (0.002)	4.685 (2.354)	0.137 (0.039)	0.333 (0.099)	6.43 (2.98)	0.087 (0.027)	0.059 (0.038)	2.467 (1.092)	1.579 (0.442)	0.403 (0.066)	1.192 (0.443)	0.964 (0.104)
$g = \sqrt{n}$	BB(1, b_{EB})	0.104 (0.048)	0.012 (0.003)	0.01 (0.002)	5.102 (2.505)	0.14 (0.034)	0.343 (0.097)	6.528 (2.931)	0.088 (0.021)	0.056 (0.039)	2.382 (0.845)	2.001 (0.348)	0.402 (0.072)	1.1 (0.431)	0.969 (0.103)
Hyper-g	BB(1, 1)	0.12 (0.079)	0.012 (0.005)	0.014 (0.008)	4.455 (2.501)	0.135 (0.076)	0.412 (0.283)	6.652 (3.369)	0.087 (0.038)	0.057 (0.068)	2.847 (1.67)	1.191 (0.546)	0.407 (0.073)	1.388 (0.998)	0.978 (0.099)
EB-local	BB(1, 1)	0.121 (0.08)	0.012 (0.004)	0.014 (0.008)	4.451 (2.497)	0.134 (0.07)	0.409 (0.275)	6.483 (3.389)	0.094 (0.039)	0.06 (0.072)	2.896 (1.695)	1.201 (0.558)	0.409 (0.063)	1.421 (0.867)	0.956 (0.13)
EB-local	BB(1, b_{EB})	0.125 (0.087)	0.012 (0.004)	0.013 (0.007)	4.528 (2.582)	0.14 (0.083)	0.413 (0.275)	6.917 (3.478)	0.095 (0.047)	0.06 (0.074)	3.091 (1.801)	1.414 (0.595)	0.387 (0.066)	1.366 (0.769)	0.981 (0.094)
Hyper-g	BB(1, b_{EB})	0.125 (0.08)	0.012 (0.004)	0.013 (0.008)	4.465 (2.561)	0.151 (0.145)	0.413 (0.277)	6.806 (3.651)	0.088 (0.036)	0.063 (0.081)	2.759 (1.659)	1.556 (0.61)	0.409 (0.069)	1.313 (0.635)	0.983 (0.107)
EB-local	BB(1, b_{SDM})	0.121 (0.078)	0.012 (0.005)	0.016 (0.01)	4.611 (2.757)	0.143 (0.089)	0.451 (0.329)	6.636 (3.364)	0.096 (0.044)	0.062 (0.076)	3.259 (2.004)	1.249 (0.547)	0.403 (0.083)	1.482 (0.726)	0.987 (0.093)
Hyper-g	BB(1, b_{SDM})	0.121 (0.083)	0.012 (0.005)	0.016 (0.01)	4.52 (2.66)	0.149 (0.097)	0.456 (0.321)	6.955 (3.831)	0.101 (0.054)	0.063 (0.083)	3.365 (2.02)	1.378 (0.56)	0.414 (0.07)	1.365 (0.711)	0.977 (0.1)
$g = \sqrt{n}$	Ber(θ_{SDM})	0.101 (0.052)	0.012 (0.005)	0.016 (0.008)	4.968 (2.96)	0.156 (0.109)	0.473 (0.277)	6.319 (2.977)	0.098 (0.041)	0.055 (0.038)	4.457 (2.219)	1.478 (0.326)	0.398 (0.076)	1.148 (0.385)	0.95 (0.132)
$g = \sqrt{n}$	Ber(θ_{EB})	0.1 (0.05)	0.012 (0.004)	0.013 (0.005)	4.452 (2.354)	0.132 (0.047)	0.366 (0.171)	16.673 (2.906)	0.097 (0.02)	0.141 (0.02)	2.848 (0.536)	2.228 (0.304)	0.409 (0.075)	1.169 (0.424)	0.993 (0.118)
EB-local	Ber(θ_{EB})	0.109 (0.067)	0.013 (0.006)	0.024 (0.013)	4.529 (2.717)	0.143 (0.089)	0.584 (0.431)	16.056 (2.817)	0.088 (0.037)	0.099 (0.035)	2.526 (0.89)	1.691 (0.605)	0.413 (0.078)	1.612 (1.376)	1.009 (0.133)
Hyper-g	Ber(θ_{EB})	0.107 (0.068)	0.013 (0.006)	0.024 (0.013)	4.519 (2.661)	0.149 (0.094)	0.593 (0.457)	15.826 (2.734)	0.088 (0.035)	0.099 (0.031)	2.414 (0.67)	1.735 (0.615)	0.414 (0.072)	1.61 (1.21)	1.017 (0.116)
Hyper-g	Ber(θ_{SDM})	0.112 (0.072)	0.015 (0.008)	0.032 (0.015)	5.754 (3.856)	0.208 (0.141)	0.837 (0.556)	6.493 (3.297)	0.131 (0.088)	0.05 (0.061)	5.543 (2.753)	1.176 (0.521)	0.419 (0.07)	1.495 (1.124)	0.98 (0.107)
EB-local	Ber(θ_{SDM})	0.11 (0.069)	0.015 (0.008)	0.032 (0.015)	5.928 (4.017)	0.22 (0.191)	0.854 (0.555)	6.398 (3.3)	0.143 (0.117)	0.059 (0.079)	5.213 (2.557)	1.181 (0.475)	0.417 (0.081)	1.616 (1.066)	0.975 (0.114)
$g = \sqrt{n}$	Uniform	0.101 (0.05)	0.012 (0.004)	0.013 (0.005)	4.457 (2.371)	0.134 (0.049)	0.37 (0.171)	16.71 (2.927)	0.098 (0.021)	0.141 (0.023)	2.84 (0.637)	2.168 (0.708)	0.584 (0.12)	2.568 (1.257)	1.056 (0.127)
Hyper-g	Uniform	0.107 (0.065)	0.013 (0.007)	0.024 (0.013)	4.542 (2.653)	0.153 (0.108)	0.604 (0.443)	16.164 (2.953)	0.086 (0.032)	0.098 (0.031)	2.407 (0.65)	1.97 (0.966)	0.59 (0.15)	2.698 (1.865)	1.089 (0.152)
EB-local	Uniform	0.109 (0.068)	0.013 (0.007)	0.024 (0.013)	4.618 (2.857)	0.147 (0.094)	0.601 (0.454)	16.227 (2.953)	0.085 (0.035)	0.099 (0.034)	2.404 (0.673)	2.041 (1.074)	0.631 (0.199)	2.885 (1.612)	1.101 (0.144)
$g = \sqrt{n}$	Complexity(1)	0.163 (0.11)	0.018 (0.01)	0.044 (0.015)	9.408 (5.227)	0.284 (0.18)	1.251 (0.5)	8.882 (3.732)	0.165 (0.14)	0.067 (0.071)	8.729 (2.307)	2.479 (0.151)	0.462 (0.058)	1.574 (1.028)	0.973 (0.09)
EB-local	Complexity(1)	0.232 (0.155)	0.023 (0.012)	0.066 (0.016)	10.865 (6.235)	0.411 (0.257)	1.469 (0.53)	10.553 (4.287)	0.229 (0.187)	0.089 (0.089)	8.863 (2.608)	2.283 (0.362)	0.436 (0.065)	1.964 (1.387)	0.981 (0.109)
Hyper-g	Complexity(1)	0.235 (0.158)	0.023 (0.012)	0.065 (0.017)	11.128 (6.442)	0.405 (0.228)	1.417 (0.514)	10.099 (4.139)	0.244 (0.197)	0.085 (0.093)	9.427 (2.664)	2.277 (0.285)	0.452 (0.059)	1.928 (1.516)	1 (0.104)
$g = \sqrt{n}$	Complexity(2)	0.305 (0.186)	0.032 (0.015)	0.084 (0.015)	24.468 (9.68)	0.652 (0.29)	1.716 (0.505)	18.395 (4.999)	0.488 (0.26)	0.131 (0.165)	12.649 (1.382)	3.122 (0.446)	0.543 (0.009)	2.223 (0.892)	0.973 (0.079)
EB-local	Complexity(2)	0.377 (0.208)	0.041 (0.019)	0.094 (0.016)	25.186 (10.624)	0.781 (0.342)	1.808 (0.519)	19.203 (5.069)	0.601 (0.394)	0.139 (0.15)	12.881 (1.515)	3.169 (0.642)	0.54 (0.018)	2.14 (0.64)	0.951 (0.095)
Hyper-g	Complexity(2)	0.371 (0.213)	0.041 (0.019)	0.095 (0.014)	24.418 (10.474)	0.754 (0.32)	1.737 (0.514)	19.453 (5.247)	0.652 (0.388)	0.136 (0.144)	13.024 (1.585)	3.112 (0.619)	0.544 (0.007)	2.228 (1.202)	0.95 (0.096)

Table B.3: Average 1-AUPRC for variable selection using various techniques for all datasets averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

Parameter prior	$P(\mathcal{M}_\gamma)$	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	BB(1, 1)	0.036 (0.034)	0.001 (0.003)	0.001 (0.002)	0.017 (0.018)	0.004 (0.008)	0.004 (0.007)	0.168 (0.105)	0.025 (0.02)	0.078 (0.113)	0.144 (0.068)	0.515 (0.11)	0.781 (0.112)	0.671 (0.164)	0.947 (0.083)
$g = \sqrt{n}$	BB(1, b_{SDM})	0.035 (0.033)	0.001 (0.003)	0.001 (0.002)	0.016 (0.016)	0.004 (0.009)	0.004 (0.007)	0.17 (0.103)	0.025 (0.019)	0.075 (0.103)	0.159 (0.075)	0.537 (0.106)	0.787 (0.1)	0.685 (0.161)	0.935 (0.1)
$g = \sqrt{n}$	BB(1, b_{EB})	0.035 (0.032)	0.006 (0.015)	0.003 (0.006)	0.018 (0.023)	0.004 (0.008)	0.01 (0.018)	0.171 (0.1)	0.025 (0.017)	0.082 (0.107)	0.15 (0.068)	0.575 (0.109)	0.79 (0.106)	0.666 (0.164)	0.942 (0.095)
Hyper-g	BB(1, 1)	0.034 (0.034)	0.001 (0.003)	0.002 (0.004)	0.016 (0.019)	0.004 (0.007)	0.006 (0.01)	0.168 (0.102)	0.027 (0.019)	0.084 (0.112)	0.17 (0.088)	0.406 (0.149)	0.797 (0.103)	0.693 (0.173)	0.94 (0.098)
EB-local	BB(1, 1)	0.036 (0.036)	0.001 (0.003)	0.002 (0.004)	0.016 (0.019)	0.005 (0.009)	0.006 (0.009)	0.167 (0.101)	0.028 (0.018)	0.094 (0.128)	0.175 (0.087)	0.413 (0.167)	0.814 (0.097)	0.702 (0.162)	0.933 (0.096)
EB-local	BB(1, b_{EB})	0.037 (0.038)	0.001 (0.003)	0.002 (0.003)	0.016 (0.018)	0.005 (0.008)	0.006 (0.009)	0.167 (0.098)	0.027 (0.019)	0.08 (0.106)	0.187 (0.085)	0.434 (0.156)	0.769 (0.11)	0.68 (0.169)	0.951 (0.075)
Hyper-g	BB(1, b_{EB})	0.034 (0.035)	0.001 (0.003)	0.002 (0.003)	0.016 (0.017)	0.006 (0.012)	0.006 (0.009)	0.162 (0.102)	0.026 (0.016)	0.077 (0.112)	0.176 (0.092)	0.455 (0.15)	0.788 (0.105)	0.675 (0.185)	0.952 (0.081)
EB-local	BB(1, b_{SDM})	0.036 (0.034)	0.001 (0.003)	0.002 (0.004)	0.017 (0.024)	0.005 (0.009)	0.007 (0.01)	0.164 (0.101)	0.027 (0.017)	0.086 (0.112)	0.19 (0.095)	0.424 (0.151)	0.788 (0.131)	0.714 (0.173)	0.949 (0.076)
Hyper-g	BB(1, b_{SDM})	0.035 (0.036)	0.001 (0.003)	0.002 (0.004)	0.016 (0.023)	0.005 (0.008)	0.007 (0.01)	0.17 (0.105)	0.031 (0.023)	0.09 (0.113)	0.194 (0.098)	0.453 (0.147)	0.802 (0.099)	0.672 (0.182)	0.941 (0.095)
$g = \sqrt{n}$	Ber(θ_{SDM})	0.035 (0.034)	0.001 (0.003)	0.002 (0.004)	0.017 (0.018)	0.005 (0.009)	0.007 (0.009)	0.167 (0.099)	0.029 (0.018)	0.078 (0.101)	0.247 (0.1)	0.532 (0.098)	0.788 (0.124)	0.672 (0.156)	0.931 (0.112)
$g = \sqrt{n}$	Ber(θ_{EB})	0.034 (0.035)	0.001 (0.003)	0.002 (0.003)	0.016 (0.021)	0.003 (0.007)	0.005 (0.009)	0.201 (0.123)	0.026 (0.018)	0.092 (0.11)	0.141 (0.059)	0.616 (0.103)	0.805 (0.1)	0.678 (0.163)	0.946 (0.09)
EB-local	Ber(θ_{EB})	0.034 (0.034)	0.001 (0.003)	0.003 (0.004)	0.015 (0.02)	0.005 (0.008)	0.008 (0.01)	0.194 (0.12)	0.027 (0.024)	0.082 (0.107)	0.143 (0.066)	0.486 (0.14)	0.799 (0.117)	0.69 (0.18)	0.94 (0.094)
Hyper-g	Ber(θ_{EB})	0.034 (0.035)	0.001 (0.003)	0.003 (0.005)	0.015 (0.017)	0.005 (0.009)	0.008 (0.011)	0.195 (0.121)	0.025 (0.017)	0.088 (0.121)	0.139 (0.063)	0.492 (0.163)	0.815 (0.112)	0.697 (0.167)	0.942 (0.084)
Hyper-g	Ber(θ_{SDM})	0.032 (0.031)	0.001 (0.002)	0.004 (0.005)	0.018 (0.024)	0.007 (0.009)	0.01 (0.011)	0.167 (0.097)	0.043 (0.039)	0.079 (0.113)	0.267 (0.112)	0.425 (0.158)	0.816 (0.1)	0.677 (0.172)	0.944 (0.086)
EB-local	Ber(θ_{SDM})	0.034 (0.033)	0.001 (0.003)	0.004 (0.005)	0.019 (0.025)	0.008 (0.014)	0.01 (0.011)	0.169 (0.101)	0.044 (0.039)	0.085 (0.122)	0.264 (0.103)	0.422 (0.145)	0.806 (0.126)	0.699 (0.181)	0.942 (0.088)
$g = \sqrt{n}$	Uniform	0.033 (0.034)	0.001 (0.003)	0.002 (0.004)	0.016 (0.018)	0.004 (0.008)	0.006 (0.009)	0.202 (0.117)	0.027 (0.018)	0.096 (0.115)	0.139 (0.06)	0.765 (0.129)	0.892 (0.087)	0.814 (0.143)	0.952 (0.081)
Hyper-g	Uniform	0.034 (0.033)	0.001 (0.003)	0.003 (0.004)	0.015 (0.016)	0.006 (0.01)	0.008 (0.01)	0.199 (0.121)	0.026 (0.02)	0.077 (0.106)	0.14 (0.063)	0.574 (0.2)	0.858 (0.12)	0.785 (0.156)	0.949 (0.084)
EB-local	Uniform	0.034 (0.035)	0.001 (0.002)	0.003 (0.004)	0.016 (0.018)	0.005 (0.008)	0.008 (0.011)	0.2 (0.124)	0.025 (0.018)	0.079 (0.107)	0.138 (0.066)	0.56 (0.185)	0.868 (0.105)	0.807 (0.14)	0.958 (0.074)
$g = \sqrt{n}$	Complexity(1)	0.038 (0.038)	0.001 (0.003)	0.005 (0.006)	0.025 (0.026)	0.01 (0.011)	0.013 (0.01)	0.183 (0.097)	0.051 (0.049)	0.089 (0.1)	0.403 (0.09)	0.639 (0.092)	0.813 (0.1)	0.683 (0.171)	0.95 (0.073)
EB-local	Complexity(1)	0.042 (0.039)	0.003 (0.005)	0.007 (0.007)	0.026 (0.027)	0.016 (0.017)	0.017 (0.012)	0.191 (0.1)	0.067 (0.056)	0.111 (0.123)	0.394 (0.11)	0.567 (0.118)	0.801 (0.099)	0.743 (0.159)	0.954 (0.085)
Hyper-g	Complexity(1)	0.039 (0.04)	0.002 (0.004)	0.006 (0.007)	0.026 (0.025)	0.018 (0.016)	0.017 (0.012)	0.187 (0.095)	0.075 (0.06)	0.106 (0.122)	0.416 (0.105)	0.566 (0.099)	0.811 (0.088)	0.745 (0.178)	0.956 (0.084)
$g = \sqrt{n}$	Complexity(2)	0.04 (0.039)	0.005 (0.006)	0.008 (0.008)	0.048 (0.029)	0.027 (0.02)	0.021 (0.013)	0.297 (0.103)	0.145 (0.082)	0.155 (0.132)	0.507 (0.058)	0.863 (0.055)	0.98 (0.023)	0.854 (0.121)	0.947 (0.072)
EB-local	Complexity(2)	0.048 (0.043)	0.006 (0.007)	0.01 (0.009)	0.053 (0.026)	0.036 (0.02)	0.023 (0.013)	0.289 (0.112)	0.161 (0.081)	0.161 (0.157)	0.511 (0.074)	0.825 (0.1)	0.97 (0.04)	0.848 (0.134)	0.915 (0.098)
Hyper-g	Complexity(2)	0.043 (0.041)	0.007 (0.007)	0.01 (0.01)	0.052 (0.027)	0.034 (0.019)	0.024 (0.014)	0.29 (0.101)	0.174 (0.086)	0.162 (0.139)	0.516 (0.072)	0.824 (0.096)	0.985 (0.009)	0.819 (0.152)	0.923 (0.091)

Table B.4: Average $1 - R^2_{test}$ for predictions using various techniques for all datasets averaged over 100 train-test splits; Numbers in brackets represent standard deviation over 100 train-test splits

Parameter prior	$P(\mathcal{M}_\gamma)$	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	BB(1, 1)	0.142 (0.021)	0.222 (0.009)	0.165 (0.006)	0.174 (0.032)	0.277 (0.006)	0.073 (0.007)	0.505 (0.054)	0.227 (0.005)	0.269 (0.041)	0.159 (0.05)	0.367 (0.126)	0.897 (0.205)	0.165 (0.061)	0.505 (0.261)
$g = \sqrt{n}$	BB(1, b_{SDM})	0.142 (0.021)	0.222 (0.009)	0.165 (0.006)	0.175 (0.032)	0.277 (0.006)	0.073 (0.007)	0.506 (0.053)	0.227 (0.005)	0.269 (0.041)	0.162 (0.05)	0.367 (0.121)	0.892 (0.173)	0.168 (0.064)	0.509 (0.24)
$g = \sqrt{n}$	BB(1, b_{EB})	0.142 (0.021)	0.222 (0.009)	0.164 (0.006)	0.175 (0.032)	0.277 (0.006)	0.073 (0.007)	0.507 (0.053)	0.227 (0.005)	0.269 (0.04)	0.159 (0.05)	0.384 (0.122)	0.881 (0.18)	0.169 (0.065)	0.516 (0.275)
Hyper-g	BB(1, 1)	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.174 (0.031)	0.277 (0.006)	0.072 (0.007)	0.508 (0.055)	0.227 (0.005)	0.27 (0.044)	0.165 (0.054)	0.371 (0.168)	0.901 (0.171)	0.172 (0.069)	0.507 (0.269)
EB-local	BB(1, 1)	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.174 (0.032)	0.277 (0.006)	0.073 (0.007)	0.508 (0.056)	0.227 (0.005)	0.27 (0.044)	0.165 (0.052)	0.377 (0.184)	0.895 (0.193)	0.175 (0.081)	0.53 (0.343)
EB-local	BB(1, b_{EB})	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.175 (0.032)	0.277 (0.006)	0.072 (0.007)	0.511 (0.055)	0.227 (0.005)	0.27 (0.044)	0.166 (0.054)	0.373 (0.186)	0.911 (0.191)	0.174 (0.069)	0.493 (0.352)
Hyper-g	BB(1, b_{EB})	0.142 (0.023)	0.222 (0.009)	0.165 (0.006)	0.175 (0.032)	0.277 (0.006)	0.072 (0.007)	0.512 (0.055)	0.227 (0.005)	0.27 (0.044)	0.165 (0.055)	0.364 (0.157)	0.924 (0.173)	0.173 (0.072)	0.523 (0.332)
EB-local	BB(1, b_{SDM})	0.142 (0.023)	0.222 (0.009)	0.165 (0.006)	0.175 (0.031)	0.277 (0.006)	0.073 (0.007)	0.51 (0.055)	0.227 (0.005)	0.27 (0.044)	0.166 (0.051)	0.38 (0.228)	0.927 (0.248)	0.173 (0.067)	0.535 (0.357)
Hyper-g	BB(1, b_{SDM})	0.142 (0.023)	0.222 (0.009)	0.165 (0.006)	0.175 (0.031)	0.277 (0.006)	0.073 (0.007)	0.51 (0.055)	0.227 (0.005)	0.27 (0.044)	0.167 (0.053)	0.365 (0.164)	0.921 (0.19)	0.171 (0.066)	0.535 (0.431)
$g = \sqrt{n}$	Ber(θ_{SDM})	0.142 (0.021)	0.222 (0.009)	0.165 (0.006)	0.177 (0.031)	0.278 (0.006)	0.074 (0.007)	0.506 (0.053)	0.228 (0.005)	0.269 (0.04)	0.181 (0.057)	0.364 (0.121)	0.872 (0.245)	0.181 (0.064)	0.492 (0.365)
$g = \sqrt{n}$	Ber(θ_{EB})	0.142 (0.021)	0.222 (0.009)	0.165 (0.006)	0.175 (0.031)	0.277 (0.006)	0.073 (0.007)	0.509 (0.06)	0.227 (0.005)	0.271 (0.043)	0.154 (0.047)	0.402 (0.132)	0.887 (0.237)	0.175 (0.064)	0.483 (0.25)
EB-local	Ber(θ_{EB})	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.175 (0.032)	0.277 (0.006)	0.073 (0.007)	0.509 (0.059)	0.228 (0.005)	0.271 (0.046)	0.158 (0.05)	0.368 (0.174)	0.923 (0.278)	0.174 (0.064)	0.573 (0.46)
Hyper-g	Ber(θ_{EB})	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.175 (0.031)	0.277 (0.006)	0.073 (0.007)	0.509 (0.059)	0.228 (0.005)	0.27 (0.045)	0.157 (0.05)	0.368 (0.158)	0.928 (0.319)	0.173 (0.068)	0.563 (0.419)
Hyper-g	Ber(θ_{SDM})	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.177 (0.031)	0.278 (0.006)	0.073 (0.007)	0.507 (0.055)	0.229 (0.005)	0.27 (0.044)	0.184 (0.061)	0.373 (0.164)	0.924 (0.268)	0.176 (0.068)	0.513 (0.314)
EB-local	Ber(θ_{SDM})	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.177 (0.031)	0.278 (0.006)	0.073 (0.007)	0.508 (0.055)	0.229 (0.005)	0.27 (0.044)	0.183 (0.06)	0.396 (0.222)	0.919 (0.246)	0.182 (0.08)	0.558 (0.452)
$g = \sqrt{n}$	Uniform	0.142 (0.021)	0.222 (0.009)	0.165 (0.006)	0.175 (0.031)	0.277 (0.006)	0.073 (0.007)	0.509 (0.06)	0.227 (0.005)	0.271 (0.043)	0.153 (0.045)	0.614 (0.363)	1.344 (0.732)	0.177 (0.065)	0.543 (0.297)
Hyper-g	Uniform	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.175 (0.031)	0.277 (0.006)	0.073 (0.007)	0.509 (0.059)	0.228 (0.005)	0.27 (0.046)	0.156 (0.049)	0.728 (0.528)	1.253 (0.684)	0.177 (0.066)	0.595 (0.381)
EB-local	Uniform	0.142 (0.022)	0.222 (0.009)	0.165 (0.006)	0.175 (0.032)	0.277 (0.006)	0.073 (0.007)	0.51 (0.059)	0.228 (0.005)	0.27 (0.045)	0.157 (0.049)	0.884 (0.818)	1.452 (1.011)	0.19 (0.078)	0.626 (0.478)
$g = \sqrt{n}$	Complexity(1)	0.144 (0.021)	0.223 (0.009)	0.166 (0.006)	0.185 (0.029)	0.278 (0.006)	0.074 (0.007)	0.526 (0.052)	0.229 (0.005)	0.274 (0.042)	0.198 (0.06)	0.418 (0.126)	0.903 (0.145)	0.192 (0.068)	0.535 (0.27)
EB-local	Complexity(1)	0.143 (0.023)	0.223 (0.009)	0.166 (0.006)	0.185 (0.03)	0.278 (0.006)	0.073 (0.007)	0.53 (0.055)	0.23 (0.005)	0.273 (0.046)	0.197 (0.064)	0.372 (0.176)	0.933 (0.184)	0.195 (0.089)	0.542 (0.262)
Hyper-g	Complexity(1)	0.143 (0.023)	0.223 (0.009)	0.166 (0.006)	0.185 (0.03)	0.278 (0.006)	0.073 (0.007)	0.53 (0.054)	0.23 (0.005)	0.273 (0.046)	0.197 (0.067)	0.38 (0.165)	0.931 (0.131)	0.192 (0.081)	0.572 (0.462)
$g = \sqrt{n}$	Complexity(2)	0.147 (0.022)	0.223 (0.009)	0.166 (0.006)	0.196 (0.031)	0.278 (0.006)	0.076 (0.007)	0.545 (0.052)	0.232 (0.005)	0.286 (0.046)	0.218 (0.063)	0.642 (0.119)	1 (0.006)	0.352 (0.106)	0.845 (0.198)
EB-local	Complexity(2)	0.146 (0.023)	0.223 (0.009)	0.167 (0.006)	0.195 (0.031)	0.279 (0.006)	0.075 (0.007)	0.544 (0.055)	0.232 (0.005)	0.282 (0.05)	0.218 (0.067)	0.596 (0.225)	0.998 (0.013)	0.333 (0.118)	0.805 (0.429)
Hyper-g	Complexity(2)	0.146 (0.023)	0.224 (0.009)	0.166 (0.006)	0.195 (0.031)	0.279 (0.006)	0.075 (0.007)	0.544 (0.055)	0.232 (0.005)	0.283 (0.051)	0.22 (0.07)	0.644 (0.179)	0.999 (0.008)	0.333 (0.106)	0.823 (0.216)

Table B.5: Average Mean Interval Score for predictions using various techniques for all datasets averaged over 100 train-test splits; Numbers in brackets represent standard deviation over 100 train-test splits

Parameter prior	$P(\mathcal{M}_\gamma)$	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	BB$(1,1)$	2.2 (0.227)	0.711 (0.019)	0.546 (0.015)	38.038 (5.529)	5.247 (0.072)	3.62 (0.154)	243.262 (12.529)	2.985 (0.054)	1.985 (0.213)	19.845 (3.711)	11.699 (0.872)	17.063 (3.378)	3.897 (0.565)	49.897 (14.894)
$g = \sqrt{n}$	BB$(1, b_{SDM})$	2.201 (0.226)	0.711 (0.019)	0.546 (0.015)	38.048 (5.554)	5.248 (0.072)	3.62 (0.153)	243.761 (12.482)	2.984 (0.054)	1.987 (0.216)	20.195 (3.906)	11.515 (1.087)	16.9 (3.207)	3.896 (0.555)	50.512 (12.012)
$g = \sqrt{n}$	BB$(1, b_{EB})$	2.202 (0.228)	0.711 (0.019)	0.546 (0.015)	38.056 (5.592)	5.247 (0.072)	3.611 (0.158)	243.549 (12.913)	2.985 (0.054)	1.984 (0.215)	19.89 (3.864)	11.726 (1.552)	16.754 (2.866)	3.933 (0.644)	50.933 (12.999)
Hyper-g	BB$(1,1)$	2.171 (0.27)	0.713 (0.02)	0.545 (0.016)	38.126 (6.085)	5.255 (0.073)	3.642 (0.232)	244.007 (14.093)	2.989 (0.056)	1.992 (0.247)	19.889 (4.43)	10.693 (2.161)	16.913 (2.972)	3.889 (1.081)	49.971 (14.028)
EB-local	BB$(1,1)$	2.169 (0.268)	0.713 (0.02)	0.545 (0.016)	38.135 (6.14)	5.256 (0.074)	3.643 (0.235)	243.333 (13.824)	2.989 (0.056)	1.998 (0.253)	19.976 (4.575)	10.669 (2.351)	17.203 (4.003)	3.97 (1.302)	51.391 (18.202)
EB-local	BB$(1, b_{EB})$	2.172 (0.27)	0.713 (0.02)	0.545 (0.016)	38.13 (6.094)	5.255 (0.074)	3.644 (0.235)	244.144 (13.509)	2.99 (0.056)	2.003 (0.252)	19.88 (4.583)	10.386 (2.628)	17.143 (4.019)	3.939 (1.066)	49.633 (16.974)
Hyper-g	BB$(1, b_{EB})$	2.17 (0.268)	0.713 (0.02)	0.545 (0.016)	38.134 (6.035)	5.254 (0.075)	3.644 (0.233)	244.431 (12.88)	2.99 (0.057)	2 (0.251)	19.921 (4.54)	10.359 (1.778)	16.939 (3.483)	3.894 (1.147)	52.25 (18.827)
EB-local	BB$(1, b_{SDM})$	2.17 (0.268)	0.713 (0.02)	0.545 (0.016)	38.195 (6.118)	5.254 (0.073)	3.646 (0.234)	244.231 (13.387)	2.99 (0.056)	1.993 (0.25)	20.076 (4.36)	10.623 (2.616)	17.593 (4.303)	3.963 (1.111)	52.309 (16.788)
Hyper-g	BB$(1, b_{SDM})$	2.172 (0.269)	0.713 (0.02)	0.545 (0.016)	38.101 (6.08)	5.255 (0.074)	3.645 (0.231)	243.9 (12.664)	2.99 (0.055)	1.998 (0.253)	19.861 (4.184)	10.575 (2.155)	16.95 (3.286)	3.9 (1.084)	51.86 (19.041)
$g = \sqrt{n}$	Ber(θ_{SDM})	2.197 (0.226)	0.71 (0.019)	0.546 (0.015)	38.209 (5.422)	5.252 (0.07)	3.646 (0.152)	243.056 (12.186)	2.99 (0.055)	1.991 (0.219)	21.728 (4.308)	11.488 (1.159)	17.288 (4.321)	4.141 (0.545)	49.402 (13.041)
$g = \sqrt{n}$	Ber(θ_{EB})	2.2 (0.225)	0.711 (0.019)	0.546 (0.015)	38.072 (5.462)	5.25 (0.071)	3.637 (0.15)	248.373 (17.237)	2.986 (0.054)	1.979 (0.213)	19.273 (3.583)	11.793 (1.727)	17.413 (3.972)	4.023 (0.622)	50.369 (17.719)
EB-local	Ber(θ_{EB})	2.172 (0.27)	0.713 (0.02)	0.545 (0.016)	38.212 (6.172)	5.256 (0.073)	3.653 (0.231)	248.966 (17.38)	2.991 (0.056)	1.981 (0.252)	19.109 (4.208)	10.4 (2.539)	17.79 (5.306)	3.992 (1.073)	58.569 (30.279)
Hyper-g	Ber(θ_{EB})	2.17 (0.269)	0.713 (0.02)	0.545 (0.016)	38.118 (6.087)	5.256 (0.074)	3.656 (0.234)	249.086 (16.593)	2.991 (0.057)	1.983 (0.249)	19.087 (4.206)	10.491 (1.931)	18.021 (6.735)	3.916 (1.011)	59.275 (29.751)
Hyper-g	Ber(θ_{SDM})	2.173 (0.268)	0.712 (0.02)	0.545 (0.016)	38.349 (6.031)	5.257 (0.073)	3.666 (0.233)	243.399 (13.206)	2.995 (0.056)	1.998 (0.251)	21.639 (4.928)	10.588 (2.641)	17.514 (4.795)	3.911 (0.968)	52.241 (17.626)
EB-local	Ber(θ_{SDM})	2.173 (0.27)	0.712 (0.02)	0.545 (0.016)	38.25 (6.036)	5.257 (0.075)	3.667 (0.238)	243.314 (13.299)	2.994 (0.056)	1.996 (0.254)	21.489 (5.083)	10.801 (3.528)	17.78 (5.278)	4.011 (1.108)	54.697 (23.818)
$g = \sqrt{n}$	Uniform	2.198 (0.225)	0.71 (0.019)	0.546 (0.015)	38.098 (5.48)	5.25 (0.071)	3.638 (0.148)	248.262 (17.871)	2.985 (0.055)	1.982 (0.217)	19.273 (3.621)	15.614 (9.982)	26.009 (13.58)	3.922 (0.766)	55.582 (24.191)
Hyper-g	Uniform	2.171 (0.267)	0.713 (0.02)	0.545 (0.016)	38.102 (6.109)	5.255 (0.074)	3.653 (0.237)	248.876 (16.895)	2.992 (0.055)	1.984 (0.251)	18.994 (4.171)	19.913 (14.005)	28.311 (26.019)	4.085 (1.241)	67.612 (36.888)
EB-local	Uniform	2.17 (0.268)	0.713 (0.02)	0.545 (0.016)	38.132 (6.151)	5.255 (0.074)	3.657 (0.233)	248.799 (17.463)	2.991 (0.056)	1.978 (0.249)	18.959 (4.055)	28.6 (24.734)	33.044 (25.819)	4.338 (1.421)	68.942 (34.534)
$g = \sqrt{n}$	Complexity(1)	2.211 (0.227)	0.71 (0.019)	0.547 (0.015)	38.883 (5.485)	5.252 (0.072)	3.675 (0.147)	245.616 (10.13)	2.992 (0.054)	2.02 (0.229)	22.938 (4.954)	11.885 (1.709)	16.698 (3.125)	4.122 (0.625)	52.243 (14.754)
EB-local	Complexity(1)	2.179 (0.268)	0.713 (0.02)	0.547 (0.016)	38.867 (6.028)	5.254 (0.072)	3.712 (0.244)	246.685 (12.127)	2.998 (0.056)	2.02 (0.267)	22.602 (5.211)	10.504 (2.447)	17.178 (4.214)	4.078 (1.353)	55.215 (22.201)
Hyper-g	Complexity(1)	2.179 (0.266)	0.712 (0.02)	0.547 (0.016)	38.893 (6.094)	5.254 (0.073)	3.708 (0.241)	246.735 (12.381)	2.999 (0.054)	2.019 (0.256)	22.682 (5.472)	10.584 (1.68)	16.5 (2.658)	4.032 (1.116)	55.065 (21.51)
$g = \sqrt{n}$	Complexity(2)	2.237 (0.22)	0.71 (0.018)	0.549 (0.015)	39.838 (5.64)	5.248 (0.071)	3.715 (0.151)	248.406 (11.031)	3.001 (0.054)	2.084 (0.251)	24.059 (4.981)	14.167 (2.297)	16.065 (1.505)	5.261 (0.945)	72.857 (31.807)
EB-local	Complexity(2)	2.204 (0.264)	0.712 (0.02)	0.548 (0.016)	39.764 (6.316)	5.251 (0.072)	3.765 (0.243)	249.215 (12.787)	3.006 (0.055)	2.064 (0.285)	23.815 (5.746)	12.933 (2.636)	16.055 (1.521)	5.197 (1.571)	69.18 (31.614)
Hyper-g	Complexity(2)	2.2 (0.263)	0.712 (0.02)	0.548 (0.016)	39.837 (6.323)	5.251 (0.072)	3.77 (0.249)	249.19 (12.751)	3.006 (0.057)	2.065 (0.288)	24 (5.734)	13.851 (2.731)	16.156 (1.652)	5 (1.135)	69.718 (30.77)

Table B.6: Average model size for predictions using various techniques for all datasets averaged over 100 train-test splits; Numbers in brackets represent standard deviation over 100 train-test splits

Parameter prior	$P(\mathcal{M}_\gamma)$	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	BB(1, 1)	9.03 (0.58)	19.78 (0.28)	21.1 (0.29)	21.85 (0.96)	25.33 (0.91)	19.55 (0.53)	11.5 (1.66)	77.32 (0.93)	12.13 (1.43)	38.61 (3.38)	7.96 (1.24)	5.72 (2.74)	16.27 (1.71)	4.76 (1.49)
$g = \sqrt{n}$	BB(1, b_{SDM})	8.97 (0.53)	19.21 (0.27)	20.57 (0.41)	20.33 (0.88)	24.02 (0.82)	18.6 (0.54)	10.31 (1.54)	72.09 (1.39)	11.12 (1.14)	34.34 (2.88)	6.53 (0.91)	4.5 (2.11)	14.57 (1.54)	4.48 (1.29)
$g = \sqrt{n}$	BB(1, b_{EB})	8.93 (0.63)	21.67 (0.31)	21.86 (0.04)	23.51 (1.93)	25.63 (1.02)	21.78 (0.94)	10.07 (1.59)	79.71 (0.44)	10.81 (1.32)	37.33 (3.37)	4.92 (0.66)	4.8 (2.29)	14.76 (1.42)	4.72 (1.57)
Hyper-g	BB(1, 1)	6.79 (0.43)	17.78 (0.36)	17.52 (1.39)	17.69 (1.01)	20.58 (0.77)	15.93 (0.72)	9.08 (1.51)	71.14 (1.96)	8.63 (1.2)	31.39 (3.28)	6.8 (1.7)	5.24 (2.98)	13.8 (1.62)	4.49 (1.3)
EB-local	BB(1, 1)	6.81 (0.44)	17.77 (0.35)	17.57 (1.4)	17.67 (1.02)	20.64 (0.8)	15.92 (0.71)	9.27 (1.55)	71.41 (2.16)	8.58 (1.24)	31.28 (2.98)	7.26 (1.79)	5.78 (2.76)	13.88 (1.54)	4.75 (1.27)
EB-local	BB(1, b_{EB})	6.68 (0.44)	17.89 (0.39)	18.74 (2.12)	17.67 (1.08)	20.64 (0.83)	15.96 (0.71)	8.11 (1.45)	72.82 (3.32)	7.96 (1.12)	31.02 (3.32)	4.93 (1)	5.11 (2.61)	13.04 (1.76)	4.59 (1.25)
Hyper-g	BB(1, b_{EB})	6.69 (0.43)	17.92 (0.4)	19.25 (2.24)	17.62 (1.1)	20.59 (0.79)	15.98 (0.74)	7.97 (1.48)	73.05 (3.28)	7.92 (1.06)	30.81 (2.88)	4.81 (0.95)	3.99 (2.18)	12.91 (1.56)	4.4 (1.29)
EB-local	BB(1, b_{SDM})	6.78 (0.44)	17.56 (0.36)	16.73 (1.27)	17.1 (0.97)	20.17 (0.73)	15.66 (0.68)	8.35 (1.42)	66.68 (1.62)	8.08 (1.07)	28.95 (2.77)	6.31 (1.68)	5.15 (2.78)	12.82 (1.59)	4.62 (1.27)
Hyper-g	BB(1, b_{SDM})	6.78 (0.46)	17.57 (0.36)	16.71 (1.31)	17.1 (0.96)	20.21 (0.71)	15.69 (0.68)	8.35 (1.27)	66.77 (1.63)	8.15 (1.04)	29.35 (3.2)	5.91 (1.31)	4.45 (2.64)	12.67 (1.58)	4.14 (1.3)
$g = \sqrt{n}$	Ber(θ_{SDM})	8.42 (0.31)	17.19 (0.32)	16.24 (0.76)	15.73 (0.62)	19.79 (0.54)	15.23 (0.4)	9.3 (0.76)	57.84 (1.22)	9.67 (0.62)	20.12 (1.66)	7.43 (0.55)	6.84 (1.42)	11.26 (0.72)	5.56 (1.41)
$g = \sqrt{n}$	Ber(θ_{EB})	8.63 (0.29)	17.7 (0.27)	17.76 (0.66)	18.51 (0.59)	22.28 (0.6)	16.42 (0.41)	23.56 (1.03)	66.23 (1.06)	18.12 (0.7)	45.95 (1.55)	3.96 (0.51)	7.38 (2.06)	12.43 (1.34)	7.4 (2.48)
EB-local	Ber(θ_{EB})	7.11 (0.36)	16.83 (0.36)	14.75 (0.79)	16.65 (0.74)	19.74 (0.52)	15.02 (0.5)	23.88 (1.11)	62.68 (1.11)	15.04 (0.81)	40.71 (2.21)	4.49 (1.23)	7.6 (1.99)	12.22 (1.41)	6.48 (2.04)
Hyper-g	Ber(θ_{EB})	7.09 (0.37)	16.84 (0.37)	14.74 (0.79)	16.65 (0.71)	19.76 (0.54)	15.03 (0.51)	23.88 (1.2)	62.64 (1.15)	15.01 (0.9)	40.76 (2.05)	4.22 (0.71)	6.56 (2.31)	12.21 (1.34)	5.9 (1.94)
Hyper-g	Ber(θ_{SDM})	6.98 (0.36)	16.43 (0.41)	13.58 (0.64)	14.63 (0.75)	18.33 (0.41)	14.39 (0.42)	8.27 (0.79)	54.9 (1.88)	8.05 (0.75)	19.79 (2.28)	6.94 (0.88)	5.82 (1.79)	11.14 (0.86)	5.25 (1.5)
EB-local	Ber(θ_{SDM})	6.98 (0.36)	16.43 (0.41)	13.62 (0.66)	14.67 (0.72)	18.3 (0.39)	14.39 (0.42)	8.27 (0.79)	54.87 (1.8)	8 (0.72)	19.54 (2.09)	7.24 (0.94)	6.43 (1.75)	11.22 (0.98)	5.26 (1.45)
$g = \sqrt{n}$	Uniform	8.66 (0.32)	17.71 (0.27)	17.75 (0.65)	18.47 (0.58)	22.24 (0.64)	16.42 (0.43)	23.55 (0.94)	66.25 (1.14)	18.11 (0.74)	46.08 (1.49)	18.63 (4.02)	14 (7.18)	25.26 (3.15)	7.67 (2.79)
Hyper-g	Uniform	7.1 (0.37)	16.84 (0.36)	14.74 (0.8)	16.67 (0.7)	19.78 (0.55)	15.03 (0.49)	23.88 (1.07)	62.65 (1.12)	15.03 (0.84)	40.78 (1.91)	16.01 (4.09)	13.17 (7.16)	20.17 (2.35)	6.75 (2.25)
EB-local	Uniform	7.09 (0.38)	16.83 (0.38)	14.76 (0.8)	16.71 (0.71)	19.74 (0.55)	15 (0.49)	23.98 (1.15)	62.61 (1.05)	15.01 (0.84)	40.91 (1.99)	16.25 (3.91)	13.98 (6.92)	20.21 (3.11)	6.57 (2.4)
$g = \sqrt{n}$	Complexity(1)	5.86 (0.3)	15.78 (0.45)	12.52 (0.5)	12.01 (0.74)	17.65 (0.39)	13.23 (0.34)	4.96 (0.54)	50.74 (1.89)	5.77 (0.33)	12.17 (1.34)	3.25 (0.24)	2.4 (0.53)	7.37 (0.5)	3.25 (0.65)
EB-local	Complexity(1)	5.54 (0.34)	15.21 (0.44)	11.49 (0.51)	11.8 (0.84)	17.02 (0.43)	13.27 (0.38)	4.49 (0.56)	48.19 (2.41)	5.47 (0.27)	12.76 (1.58)	3.38 (0.29)	2.47 (0.56)	7.9 (0.91)	3.29 (0.63)
Hyper-g	Complexity(1)	5.53 (0.35)	15.22 (0.44)	11.49 (0.52)	11.79 (0.86)	17.02 (0.44)	13.26 (0.37)	4.49 (0.58)	47.62 (2.39)	5.46 (0.32)	12.81 (1.58)	3.27 (0.24)	2.19 (0.49)	7.61 (0.8)	3.11 (0.6)
$g = \sqrt{n}$	Complexity(2)	4.67 (0.42)	14.45 (0.34)	10.37 (0.5)	9.28 (0.35)	15.86 (0.72)	11.56 (0.6)	3.11 (0.15)	41.6 (1.91)	4.61 (0.3)	6.57 (0.51)	1.68 (0.19)	1.01 (0.03)	4.21 (0.61)	1.44 (0.39)
EB-local	Complexity(2)	4.68 (0.47)	14.06 (0.49)	9.77 (0.48)	9.4 (0.44)	14.88 (1.02)	12.17 (0.54)	3.09 (0.14)	40.81 (1.8)	4.69 (0.32)	7.03 (0.85)	1.93 (0.2)	1.01 (0.06)	4.56 (0.88)	1.66 (0.36)
Hyper-g	Complexity(2)	4.68 (0.47)	13.99 (0.54)	9.76 (0.48)	9.41 (0.46)	14.9 (0.98)	12.16 (0.52)	3.08 (0.15)	41.08 (1.89)	4.68 (0.33)	7.08 (0.87)	1.75 (0.26)	1.01 (0.04)	4.56 (0.98)	1.53 (0.39)

Table B.7: Average computation time (in secs) for various techniques averaged over 100 bootstrapped samples; Numbers in brackets represent standard deviation over 100 bootstrapped samples

Parameter prior	$P(\mathcal{M}_\gamma)$	College	Tmax-Bias Correction	Tmin-Bias Correction	Bike Sharing Daily	Bike Sharing Hourly	SML 2010	Diabetes	Superconducitivity	Ozone	Boston Housing	Nutrimouse	multidrug	NIR	Liver
$g = \sqrt{n}$	BB(1, 1)	0.032 (0.002)	0.107 (0.004)	0.186 (0.049)	0.291 (0.087)	0.244 (0.074)	0.172 (0.016)	0.123 (0.012)	0.736 (0.144)	0.079 (0.006)	0.513 (0.151)	0.193 (0.018)	0.492 (0.172)	0.322 (0.029)	1.834 (0.481)
$g = \sqrt{n}$	BB(1, b_{SDM})	0.033 (0.002)	0.106 (0.004)	0.183 (0.048)	0.283 (0.072)	0.232 (0.071)	0.17 (0.014)	0.116 (0.01)	0.689 (0.13)	0.076 (0.006)	0.448 (0.117)	0.176 (0.015)	0.484 (0.176)	0.303 (0.024)	1.805 (0.445)
$g = \sqrt{n}$	BB(1, b_{EB})	0.102 (0.006)	0.345 (0.014)	0.646 (0.171)	0.915 (0.355)	0.744 (0.222)	0.564 (0.049)	0.372 (0.034)	2.15 (0.403)	0.227 (0.018)	1.512 (0.397)	0.498 (0.029)	1.426 (0.49)	0.906 (0.051)	5.429 (1.397)
Hyper-g	BB(1, 1)	0.106 (0.02)	0.747 (0.025)	1.036 (0.111)	0.963 (0.13)	2.091 (0.171)	0.413 (0.02)	0.294 (0.027)	3.446 (0.231)	0.24 (0.024)	1.18 (0.186)	0.397 (0.136)	0.522 (0.176)	0.59 (0.044)	1.835 (0.444)
EB-local	BB(1, 1)	0.026 (0.002)	0.103 (0.004)	0.186 (0.05)	0.226 (0.066)	0.206 (0.063)	0.159 (0.016)	0.113 (0.011)	0.613 (0.117)	0.055 (0.004)	0.377 (0.099)	0.18 (0.03)	0.469 (0.172)	0.25 (0.018)	1.744 (0.425)
EB-local	BB(1, b_{EB})	0.079 (0.005)	0.309 (0.013)	0.573 (0.156)	0.685 (0.177)	0.607 (0.181)	0.478 (0.039)	0.332 (0.033)	1.857 (0.344)	0.18 (0.013)	1.116 (0.29)	0.513 (0.06)	1.428 (0.493)	0.769 (0.043)	5.166 (1.258)
Hyper-g	BB(1, b_{EB})	0.32 (0.061)	2.234 (0.066)	3.23 (0.372)	2.937 (0.438)	6.232 (0.49)	1.248 (0.049)	0.857 (0.084)	10.312 (0.676)	0.717 (0.069)	3.5 (0.513)	0.942 (0.214)	1.608 (0.548)	1.742 (0.116)	5.433 (1.266)
EB-local	BB(1, b_{SDM})	0.026 (0.002)	0.102 (0.005)	0.186 (0.048)	0.228 (0.167)	0.203 (0.06)	0.158 (0.018)	0.107 (0.009)	0.603 (0.112)	0.054 (0.004)	0.349 (0.093)	0.164 (0.025)	0.477 (0.177)	0.245 (0.02)	1.735 (0.423)
Hyper-g	BB(1, b_{SDM})	0.107 (0.021)	0.745 (0.021)	1.043 (0.107)	0.95 (0.137)	2.082 (0.154)	0.411 (0.021)	0.288 (0.027)	3.436 (0.229)	0.236 (0.024)	1.139 (0.176)	0.351 (0.151)	0.515 (0.193)	0.574 (0.047)	1.832 (0.441)
$g = \sqrt{n}$	Ber(θ_{SDM})	0.028 (0.001)	0.097 (0.004)	0.176 (0.045)	0.21 (0.094)	0.197 (0.059)	0.145 (0.014)	0.109 (0.005)	0.587 (0.11)	0.065 (0.004)	0.288 (0.071)	0.217 (0.019)	0.527 (0.197)	0.347 (0.031)	1.859 (0.469)
$g = \sqrt{n}$	Ber(θ_{EB})	0.084 (0.004)	0.303 (0.01)	0.552 (0.138)	0.792 (0.35)	0.669 (0.204)	0.489 (0.042)	0.607 (0.038)	2.245 (0.42)	0.325 (0.02)	2.183 (0.536)	0.697 (0.045)	1.604 (0.555)	1.057 (0.067)	5.524 (1.42)
EB-local	Ber(θ_{EB})	0.07 (0.007)	0.288 (0.015)	0.51 (0.134)	0.607 (0.157)	0.587 (0.183)	0.441 (0.04)	0.579 (0.045)	1.895 (0.35)	0.244 (0.022)	1.601 (0.412)	0.611 (0.07)	1.552 (0.535)	0.823 (0.048)	5.25 (1.257)
Hyper-g	Ber(θ_{EB})	0.308 (0.06)	2.212 (0.061)	3.116 (0.303)	2.875 (0.78)	6.194 (0.494)	1.207 (0.054)	1.192 (0.07)	10.44 (0.713)	0.923 (0.065)	4.126 (0.572)	1.298 (0.327)	1.742 (0.556)	1.815 (0.123)	5.52 (1.292)
Hyper-g	Ber(θ_{SDM})	0.103 (0.02)	0.747 (0.023)	1.046 (0.099)	0.865 (0.127)	2.088 (0.167)	0.391 (0.019)	0.277 (0.022)	3.39 (0.23)	0.248 (0.022)	1 (0.149)	0.415 (0.11)	0.551 (0.17)	0.609 (0.05)	1.857 (0.449)
EB-local	Ber(θ_{SDM})	0.024 (0.001)	0.091 (0.004)	0.154 (0.042)	0.201 (0.177)	0.184 (0.056)	0.133 (0.014)	0.094 (0.006)	0.571 (0.105)	0.058 (0.004)	0.259 (0.067)	0.202 (0.029)	0.491 (0.175)	0.273 (0.022)	1.765 (0.42)
$g = \sqrt{n}$	Uniform	0.027 (0.001)	0.101 (0.024)	0.183 (0.06)	0.244 (0.164)	0.221 (0.076)	0.158 (0.034)	0.241 (0.036)	0.764 (0.157)	0.122 (0.026)	0.837 (0.218)	0.221 (0.02)	0.518 (0.182)	0.346 (0.031)	1.825 (0.46)
Hyper-g	Uniform	0.103 (0.021)	0.738 (0.022)	1.045 (0.098)	0.96 (0.328)	2.066 (0.166)	0.392 (0.021)	0.437 (0.024)	3.512 (0.232)	0.346 (0.035)	1.461 (0.212)	0.437 (0.125)	0.568 (0.201)	0.605 (0.045)	1.821 (0.416)
EB-local	Uniform	0.023 (0.002)	0.094 (0.025)	0.164 (0.057)	0.223 (0.115)	0.192 (0.07)	0.14 (0.035)	0.233 (0.085)	0.638 (0.134)	0.097 (0.027)	0.618 (0.211)	0.203 (0.046)	0.507 (0.199)	0.278 (0.071)	1.771 (0.49)
$g = \sqrt{n}$	Complexity(1)	0.022 (0.001)	0.093 (0.004)	0.152 (0.039)	0.178 (0.124)	0.187 (0.056)	0.13 (0.012)	0.082 (0.004)	0.549 (0.108)	0.046 (0.002)	0.193 (0.048)	0.111 (0.007)	0.404 (0.143)	0.209 (0.014)	1.692 (0.416)
EB-local	Complexity(1)	0.023 (0.002)	0.09 (0.004)	0.136 (0.037)	0.17 (0.067)	0.178 (0.054)	0.128 (0.013)	0.074 (0.004)	0.541 (0.106)	0.051 (0.002)	0.19 (0.048)	0.116 (0.006)	0.384 (0.137)	0.21 (0.012)	1.679 (0.405)
Hyper-g	Complexity(1)	0.113 (0.022)	0.776 (0.026)	1.066 (0.097)	0.846 (0.125)	2.191 (0.154)	0.407 (0.024)	0.23 (0.022)	3.394 (0.225)	0.225 (0.02)	0.838 (0.125)	0.207 (0.03)	0.439 (0.16)	0.498 (0.043)	1.735 (0.404)
$g = \sqrt{n}$	Complexity(2)	0.022 (0.002)	0.086 (0.004)	0.129 (0.033)	0.142 (0.048)	0.17 (0.052)	0.121 (0.012)	0.069 (0.003)	0.507 (0.099)	0.052 (0.002)	0.131 (0.031)	0.106 (0.006)	0.34 (0.123)	0.187 (0.012)	1.621 (0.388)
EB-local	Complexity(2)	0.024 (0.002)	0.088 (0.004)	0.135 (0.033)	0.666 (5.234)	0.171 (0.051)	0.133 (0.013)	0.074 (0.003)	0.493 (0.096)	0.047 (0.002)	0.141 (0.032)	0.102 (0.006)	0.348 (0.127)	0.187 (0.015)	1.631 (0.398)
Hyper-g	Complexity(2)	0.131 (0.026)	0.786 (0.023)	1.124 (0.105)	0.811 (0.123)	2.239 (0.171)	0.421 (0.025)	0.223 (0.021)	3.347 (0.216)	0.254 (0.022)	0.667 (0.101)	0.159 (0.027)	0.35 (0.128)	0.372 (0.069)	1.665 (0.402)

Appendix C

APPENDIX C

C.1 Proof of Theorem 1 (Existence of MLEs)

The two conditions in the theorem together imply that the maximum likelihood estimator for the full model,

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}_F} = \arg\max_{\boldsymbol{\beta}_{\boldsymbol{\gamma}_F}} \ell_{\boldsymbol{\gamma}_F}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_F}; \boldsymbol{y}),$$

exists, is finite, and, furthermore, is unique (e.g., see Mäkeläinen et al., 1981).

Now, note that for any other model $\boldsymbol{\gamma}$, maximizing $\ell_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \boldsymbol{y})$ is equivalent to maximizing $\ell_{\boldsymbol{\gamma}_F}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_F}; \boldsymbol{y})$ subject to the constraint $\boldsymbol{\beta}_{\boldsymbol{\gamma}_F} \in S_{\boldsymbol{\gamma}}$, where $S_{\boldsymbol{\gamma}} = \{\boldsymbol{\theta} \in S_{\boldsymbol{\gamma}_F} : \theta_j = 0 \text{ if } \gamma_j = 0\}$. Furthermore, $S_{\boldsymbol{\gamma}}$ is also an open connected set for all $\boldsymbol{\gamma}$. Because $\ell_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_F}; \boldsymbol{y})$ is continuous and strongly concave, then its restriction to $S_{\boldsymbol{\gamma}}$ is also continuous and strongly concave for any $\boldsymbol{\gamma}$. Furthermore, we also have $\lim_{\boldsymbol{\beta}_{\boldsymbol{\gamma}} \to \boldsymbol{\beta}^*} \ell_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \boldsymbol{y}) = -\infty$ when $\boldsymbol{\beta}^* \in \partial S_{\boldsymbol{\gamma}}$ since $\partial S_{\boldsymbol{\gamma}} \subset \partial S_{\boldsymbol{\gamma}_F}$. Therefore, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ also exists and is finite and unique for any $\boldsymbol{\gamma}$. □

C.2 Proof of Theorem 2 (Tail behavior)

We develop the argument only for $\pi_{\boldsymbol{\gamma}}^{R-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$. The proof for $\pi_{\boldsymbol{\gamma}}^{HGN-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ follows along almost identical lines. Note that

$$\lim_{s\to\infty} \frac{\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma})}{(1 + s^2/(p_{\boldsymbol{\gamma}} + 1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = \sum_{\boldsymbol{y}^*} m^*(\mathbf{y}^*) \lim_{s\to\infty} \frac{\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{y}^*)}{(1 + s^2/(p_{\boldsymbol{\gamma}} + 1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}},$$

where

$$\begin{aligned}
&\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, y^*) = \pi_{\boldsymbol{\gamma}}^{R-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*)\big|_{\boldsymbol{\beta}_{\boldsymbol{\gamma}} = s\boldsymbol{v}} \\
&= \int_0^\infty \left(\frac{1}{2\pi\delta}\right)^{\frac{p_{\boldsymbol{\gamma}}+1}{2}} |\boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y}^*)|^{1/2} \exp\left\{-\frac{1}{2\delta}\left(s\boldsymbol{v} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\right)^T \boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y}^*)\left(s\boldsymbol{v} - \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\right)\right\} f^R(\delta|\boldsymbol{\gamma}) d\delta
\end{aligned}$$

is conditional on a given training sample $\boldsymbol{y}^*$. (We can exchange the summation and the limit in this case because, for any n^*, the number of potential training samples is finite.) In the sequel, it will also be important to remember that $m^*(\boldsymbol{y}^*)$ is defined so that the maximum likelihood estimators exist for any sample $\boldsymbol{y}^*$.

To simplify notation, define

$$\zeta^{R-C}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, y^*) = \int_0^\infty \left(\frac{1}{2\pi\delta}\right)^{\frac{p_{\boldsymbol{\gamma}}+1}{2}} |\boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y}^*)|^{1/2} \exp\left\{-\frac{1}{2\delta} s^2 \boldsymbol{v}^T \boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y}^*) \boldsymbol{v}\right\} f^R(\delta|\boldsymbol{\gamma}) d\delta.$$

Note that

$$\lim_{s\to\infty} \frac{\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{y}^*)}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} =$$

$$\lim_{s\to\infty} \frac{\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{y}^*)}{\zeta^{R-C}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, y^*)} \times \lim_{s\to\infty} \frac{(s^2)^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} \times \lim_{s\to\infty} \frac{\zeta^{R-C}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, y^*)}{(s^2)^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}}.$$

Clearly, the first two limits converge to finite functions that depend on $\boldsymbol{v}$, the training sample $\boldsymbol{y}^*$ and/or the model $\boldsymbol{\gamma}$. Hence,

$$\lim_{s\to\infty} \frac{\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{y}^*)}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = c_{\boldsymbol{\gamma},1}(\boldsymbol{v}, y^*) \lim_{s\to\infty} \left(s^2\right)^{\frac{p_{\boldsymbol{\gamma}}+2}{2}} \zeta^{R-C}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, y^*),$$

where $0 < c_{\boldsymbol{\gamma},1}(\boldsymbol{v}, y^*) < \infty$. Plugging in $\zeta^{R-C}(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, y^*)$ and $f^R(\delta \mid \boldsymbol{\gamma})$, we can write

$$\lim_{s\to\infty} \frac{\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{y}^*)}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = c_{\boldsymbol{\gamma},1}(\boldsymbol{v}, y^*) \lim_{s\to\infty} \left(s^2\right)^{\frac{p_{\boldsymbol{\gamma}}+2}{2}} \times$$

$$\int_0^\infty \left(\frac{1}{2\pi\delta}\right)^{\frac{p_{\boldsymbol{\gamma}}+1}{2}} |\boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y}^*)|^{1/2} \exp\left\{-\frac{1}{2\delta} s^2 \boldsymbol{v}^T \boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y}^*) \boldsymbol{v}\right\} \times$$

$$\frac{1}{2(p_{\boldsymbol{\gamma}}+1)^{1/2}} \frac{(n^*+1)^{1/2}}{(\delta+1)^{3/2}} \mathbf{1}\left(\delta > \frac{n^* - p_{\boldsymbol{\gamma}}}{p_{\boldsymbol{\gamma}}+1}\right) d\delta.$$

Substituting $\lambda = \left(\frac{n^*+1}{p_{\boldsymbol{\gamma}}+1}\right) \frac{1}{\delta+1}$, we can write above equation as

$$\lim_{s\to\infty} \frac{\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{y}^*)}{(1 + s^2/(p_{\boldsymbol{\gamma}}+1))^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = c_{\boldsymbol{\gamma},1}(\boldsymbol{v}, y^*) \left(\frac{1}{2\pi}\right)^{\frac{p_{\boldsymbol{\gamma}}+1}{2}} |\boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y}^*)|^{1/2} \times$$

$$\lim_{s\to\infty} \left(s^2\right)^{\frac{p_{\boldsymbol{\gamma}}+2}{2}} \int_0^1 \left(\frac{\lambda}{m-\lambda}\right)^{\frac{p_{\boldsymbol{\gamma}}+1}{2}} \lambda^{-\frac{1}{2}} \exp\left\{-\left(\frac{\lambda}{m-\lambda}\right) q s^2\right\} d\lambda$$

where $q = \frac{1}{2}\boldsymbol{v}^T \boldsymbol{H}_{\boldsymbol{\gamma}}(\mathbf{y}^*)\boldsymbol{v}$, $m = \frac{n^*+1}{p_{\boldsymbol{\gamma}}+1} > 1$ since $n^* > p_{\boldsymbol{\gamma}}$. Now, from Lemma 2 in Bayarri et al. (2012),

$$\lim_{s\to\infty} \left(s^2\right)^{\frac{p_{\boldsymbol{\gamma}}+2}{2}} \int_0^1 \left(\frac{\lambda}{m-\lambda}\right)^{\frac{p_{\boldsymbol{\gamma}}+1}{2}} \lambda^{-\frac{1}{2}} \exp\left\{-\left(\frac{\lambda}{m-\lambda}\right) qs^2\right\} d\lambda = c_{\boldsymbol{\gamma},2}(\boldsymbol{v}, y^*),$$

where $0 < c_{\boldsymbol{\gamma},2}(\boldsymbol{v}, y^*) < \infty$, and therefore

$$\lim_{s\to\infty} \frac{\zeta^R(s \mid \boldsymbol{v}, \boldsymbol{\gamma}, \boldsymbol{y}^*)}{\left(1 + s^2/(p_{\boldsymbol{\gamma}}+1)\right)^{-\frac{p_{\boldsymbol{\gamma}}+2}{2}}} = c_{\boldsymbol{\gamma},1}(\boldsymbol{v}, y^*) c_{\boldsymbol{\gamma},2}(\boldsymbol{v}, y^*) = c_{\boldsymbol{\gamma},3}(\boldsymbol{v}, y^*).$$

To complete the proof simply define $c_{\boldsymbol{\gamma}}(\boldsymbol{v}) = \sum_{\boldsymbol{y}^*} m^*(\mathbf{y}^*) c_{\boldsymbol{\gamma},3}(\boldsymbol{v}, y^*)$. Since we have a finite number of terms in the sum and each term is both positive and finite, so is $c_{\boldsymbol{\gamma}}(\boldsymbol{v})$.

C.3 Proof of Theorem 3 (Intrinsic consistency)

Note that if $\boldsymbol{y}^*$ leads to separability, then $\boldsymbol{y}^{**} = \mathbf{1} - \boldsymbol{y}^*$ does as well. Hence, our choice of $m^*\left(\boldsymbol{y}^* \mid \boldsymbol{X}\right)$ is symmetric and, as $n^* \to \infty$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \xrightarrow{p} \mathbf{0}$ under m^*. In turn, this implies that

$$\hat{\theta}_{\boldsymbol{\gamma},i}(\boldsymbol{y}^*) = \frac{1}{1 + \exp\left\{\boldsymbol{x}^*_{\boldsymbol{\gamma},i}{}^T \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*)\right\}} \xrightarrow{p} 1/2$$

for all i, and therefore

$$\frac{1}{n^*}\boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) = \frac{1}{n^*}\sum_{i=1}^{n^*} \frac{1}{\hat{\theta}_{\boldsymbol{\gamma},i}(\boldsymbol{y}^*)(1 - \hat{\theta}_{\boldsymbol{\gamma},i}(\boldsymbol{y}^*))} \boldsymbol{x}^*_{\boldsymbol{\gamma},i} \boldsymbol{x}^*_{\boldsymbol{\gamma},i}{}^T \xrightarrow{p} \frac{1}{4}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}.$$

This completes the proof when $\delta = n^*$.

In the case where δ is given a prior distribution define $\delta = n^*\delta^*$. Then, under $f^{HGN}(\delta)$, δ^* has density

$$f^{HGN}(\delta^*) = (1 + \delta^*)^{-2},$$

which is a proper, non-degenerate prior. The argument for the robust prior follows along similar lines.

□

C.4 Proof of Theorem 4 (Model selection consistency)

The proof builds on the results of Barber et al. (2016). As a first step, we must show that the Laplace approximation to the marginal likelihood for all sparse models $\boldsymbol{\gamma}(n)$ with $p_{\boldsymbol{\gamma}(n)} \leq q_n$,

$$m_{\boldsymbol{\gamma}(n)}^{UI-LPEP}(\boldsymbol{y}(n)) = \int f_{\boldsymbol{\gamma}(n)}(\boldsymbol{y}(n) \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}) \pi^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}$$

leads to an approximation error which is uniformly bounded across all models with high probability. This fact, along with appropriate conditions on the prior distribution over models and the rate of growth on the number of covariates with n, can be used to proof the desired consistency result. In the sequel, we drop (n) for ease of notation.

Completing the first step of the proof requires that we verify conditions on both the likelihood and the prior. For the likelihood function, we need to show:

(A1) The Euclidean norm of the true signal is bounded, that is, $\|\boldsymbol{\beta}_{\boldsymbol{\gamma}_T}\|_2 \leq a_0$ for a fixed constant $a_0 \in (0, \infty)$.

(A2) Let $S_{\boldsymbol{\gamma}}$ denote an open connected subset of $\mathbb{R}^{p_{\boldsymbol{\gamma}}+1}$. There is a decreasing function $c_{\text{lower}} : [0, \infty) \to (0, \infty)$ and an increasing function $c_{\text{upper}} : [0, \infty) \to (0, \infty)$ such that for all models $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| \leq 2q$ and all $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \in S_{\boldsymbol{\gamma}}$, the Hessian of the negative log-likelihood function is bounded as

$$c_{\text{lower}}\left(\left\|\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right\|_2\right) \mathbf{I}_{\boldsymbol{\gamma}} \preceq \frac{1}{n} H_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \preceq c_{\text{upper}}\left(\left\|\boldsymbol{\beta}_{\boldsymbol{\gamma}}\right\|_2\right) \mathbf{I}_{\boldsymbol{\gamma}}.$$

(A3) There is a constant $c_{\text{change}} \in (0, \infty)$ such that for all $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| \leq 2q$ and all $\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}'_{\boldsymbol{\gamma}} \in S_{\boldsymbol{\gamma}}$,

$$\frac{1}{n}\|H_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) - H_{\boldsymbol{\gamma}}(\boldsymbol{\beta}'_{\boldsymbol{\gamma}})\|_{\text{sp}} \leq c_{\text{change}} \cdot \|\boldsymbol{\beta}_{\boldsymbol{\gamma}} - \boldsymbol{\beta}'_{\boldsymbol{\gamma}}\|_2,$$

where $\|\cdot\|_{\text{sp}}$ is the spectral norm of a matrix.

Condition (A1) is satisfied by assumption (iv) in Theorem 4. On the other hand, for logistic regression models, (A2) and (A3) are direct consequences of assumptions (v) and (vi)

(which also happen to be standard assumptions for the consistency of maximum likelihood estimators in Generalized Linear Models).

Next we need to show that the LPEP satisfies:

(A4) $\log \pi^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ is F_1^*-Lipschitz on the ball $B_R(0) = \{\boldsymbol{\beta}_{\boldsymbol{\gamma}} \in S_{\boldsymbol{\gamma}} : \|\boldsymbol{\beta}_{\boldsymbol{\gamma}}\|_2 \leq R\}$ for some constant F_1^*.

(A5) For all $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \in S_{\boldsymbol{\gamma}}$ and some constant F_2,

$$\log \pi^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) - \log \pi^{UI-LPEP}(\mathbf{0}) \leq F_2$$

To show that (A4) is satisfied, note that $\pi^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ is a finite mixture of strictly positive (and bounded) C^1 functions. Therefore, $\log \pi^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ is also C^1. This observation, together with the fact that $B_R(0)$ is a bounded convex set, implies that, for each n, $\log \pi^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})$ is locally Lipschitz on $B_R(0)$.

To show that (A5) is satisfied, note first that

$$\min_{\boldsymbol{y}^*} \left\{\log \tilde{m}^*(\boldsymbol{y}^*) + \log \phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*)\right\} + n \log 2 \leq \log \pi^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \leq \max_{\boldsymbol{y}^*} \left\{\log \tilde{m}^*(\boldsymbol{y}^*) + \log \phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*)\right\} + n \log 2. \quad \text{(C.1)}$$

we can show that

$$\begin{aligned}
&\log \pi^{UI-LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) - \log \pi^{UI-LPEP}(0) \\
&\leq \max_{\boldsymbol{y}^*}\{\log \tilde{m}^*(\boldsymbol{y}^*) + \log \phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*)\} - \min_{\boldsymbol{y}^*}\{\log \tilde{m}^*(\boldsymbol{y}^*) + \log \phi(0 \mid \boldsymbol{y}^*)\} \\
&\leq \log \tilde{m}^*(\boldsymbol{y}^*_{max}) + \log \phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*_{max}) - \log \tilde{m}^*(\boldsymbol{y}^*_{min}) - \log \phi(0 \mid \boldsymbol{y}^*_{min}) \\
&\leq \log \phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*_{max}) + \underbrace{\log \tilde{m}^*(\boldsymbol{y}^*_{max}) - \log \tilde{m}^*(\boldsymbol{y}^*_{min}) - \log \phi(0 \mid \boldsymbol{y}^*_{min})}_{C_1} \\
&\leq C_1 + \max_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} \log \phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*_{max}) = C_2
\end{aligned}$$

where $\boldsymbol{y}^*_{max} = \arg\max_{\boldsymbol{y}^*}\{\log \tilde{m}^*(\boldsymbol{y}^*) + \log \phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*)\}$ and $\boldsymbol{y}^*_{min} = \arg\min_{\boldsymbol{y}^*}\{\log \tilde{m}^*(\boldsymbol{y}^*) + \log \phi(0 \mid \boldsymbol{y}^*)\}$. Note that $C_2 < \infty$ because $\tilde{m}^*(\boldsymbol{y}^*) < \infty$ for all $\boldsymbol{y}^*$ and $\phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*) < \infty$ for all $\boldsymbol{y}^*, \boldsymbol{\beta}_{\boldsymbol{\gamma}}$ since $\phi(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*_{max})$ is a Gaussian density, and it is bounded above by a constant.

Using Theorem 1 of Barber et al. (2016), we can now establish that the Laplace approximation has an approximation error which is uniformly bounded across all models with high probability. The remainder of the proof follows from a direct application of Theorem 2 in Barber et al. (2016). The conditions required for this Theorem to apply either match those in the statement of our Theorem 4, follow from the discussion above, or are trivial to check.

C.5 Details of the Markov chain Monte Carlo algorithm for logistic regression

Using the hierarchical representation of the LPEP prior discussed at the start of Section 4.5.1, the posterior distribution for the augmented model can be written as

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\omega}, \boldsymbol{y}^*, \delta \mid \boldsymbol{y}, \boldsymbol{X}) \propto f_{\boldsymbol{\gamma}}(\boldsymbol{y} \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\omega}) \pi_{\boldsymbol{\gamma}}^{LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta, \boldsymbol{y}^*) f(\delta \mid \boldsymbol{\gamma}) f(\boldsymbol{\gamma}) f(\boldsymbol{\omega}) m^*(\boldsymbol{y}^*).$$

From this, it is easy to devise samplers for the full conditional posterior distributions of various blocks of parameters. We focus below on the more general setting where δ has been assigned a hyperprior. The simplifications for the case where δ is fixed are straightforward and we do not discuss them explicitly.

1. Since the prior support of δ maybe dependent on model indicator $\boldsymbol{\gamma}$, the parameters $(\boldsymbol{\gamma}, \delta, \boldsymbol{\beta}_{\boldsymbol{\gamma}})$ are updated jointly by sampling from $f(\boldsymbol{\gamma}, \delta, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y})$ given in (4.8). To do this, we write

   $$f(\boldsymbol{\gamma}, \delta, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}) = f(\boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta, \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y}),$$

   where $f(\boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}) \propto \int f(\boldsymbol{\gamma}, \delta, \boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{y}^*, \boldsymbol{\omega}) d\boldsymbol{\beta}_{\boldsymbol{\gamma}}$.

   The expression for the conditional posterior $f(\boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y})$, up to a proportionality constant, is given by (4.9). To generate samples from it, we generate proposal by combining a random walk Metropolis-Hastings algorithm for $\boldsymbol{\gamma}$ (George and McCulloch, 1997) and a reflective Gaussian random walk for δ (similar to section 2.1 of Thawornwattana et al., 2018). More specifically, we factorize the joint proposal for $(\boldsymbol{\gamma}, \delta)$ as:

   $$q\left(\delta^{(prop)}, \boldsymbol{\gamma}^{(prop)} \mid \delta, \boldsymbol{\gamma}\right) = q\left(\boldsymbol{\gamma}^{(prop)} \mid \boldsymbol{\gamma}\right) q\left(\delta^{(prop)} \mid \delta, \boldsymbol{\gamma}^{(prop)}\right).$$

For $q(\boldsymbol{\gamma}^{(prop)} \mid \boldsymbol{\gamma})$, we use a symmetric random walk proposal similar to equation (46) of George and McCulloch (1997) as follows:

- We define two probability vectors $p_1 = (0.9, 0.1)$ and $p_2 = (0.6, 0.2, 0.15, 0.05)$.
- Each time, we decide on one of two types of moves according to the probability vector p_1.
  - If a move type 1 is selected, then the proposed new model $\boldsymbol{\gamma}^{(prop)}$ is generated by randomly flipping $d \in \{1, 2, 3, 4\}$ components of $\boldsymbol{\gamma}$ with probability $p_{2,d}$. The components of $\boldsymbol{\gamma}$ to be flipped are selected uniformly at random given d.
  - If a move type 2 is selected, then the proposed model $\boldsymbol{\gamma}^{(prop)}$ is generated by removing one variable currently included in the model and replacing it with a variable that is currently excluded, leaving the dimensionality of the model unchanged. The variables to add and remove are chosen uniformly at random within each set.

Next, given $\boldsymbol{\gamma}^{(prop)}$, we propose δ using a reflective Gaussian random walk with a left reflection boundary $a_{\boldsymbol{\gamma}^{(prop)}}$. More specifically, we define $\delta^{(prop)} = a_{\boldsymbol{\gamma}^{(prop)}} + |\epsilon - a_{\boldsymbol{\gamma}^{(prop)}}|$ where $\epsilon \sim N(\delta, \tau^2)$ and $a_{\boldsymbol{\gamma}} = 0$ under the hyper-g/n prior and $a_{\boldsymbol{\gamma}} = \frac{n - p_{\boldsymbol{\gamma}}}{p_{\boldsymbol{\gamma}} + 1}$ under the robust prior. In both cases we found $\tau = n/2$ to be an efficient tuning parameter in our studies.

Since the proposal distribution of $\boldsymbol{\gamma}$, given by $q(\boldsymbol{\gamma}^{(prop)} \mid \boldsymbol{\gamma})$ is symmetric, the proposed model $(\boldsymbol{\gamma}^{(prop)}, \delta^{(prop)})$ is then accepted with probability

$$\min\left\{\frac{f(\boldsymbol{\gamma}^{(prop)}, \delta^{(prop)} \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y})}{f(\boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y})} \frac{q(\delta \mid \delta^{(prop)}, \boldsymbol{\gamma})}{q(\delta^{(prop)} \mid \delta, \boldsymbol{\gamma}^{(prop)})}, 1\right\},$$

where

$$q\left(\delta^{(prop)} \mid \delta, \boldsymbol{\gamma}^{(prop)}\right) = \frac{1}{\sqrt{2\pi}\tau}\left[\exp\left\{-\frac{1}{2\tau^2}(\delta^{(prop)} - \delta)^2\right\} + \exp\left\{-\frac{1}{2\tau^2}(2a_{\boldsymbol{\gamma}^{(prop)}} - \delta^{(prop)} - \delta)^2\right\}\right], \qquad \delta^{(prop)} \geq a_{\boldsymbol{\gamma}^{(prop)}}.$$

Note that, when move type 2 is selected, since $p_{\boldsymbol{\gamma}}$ remains unchanged, $q(\delta \mid \delta^{(prop)}, \boldsymbol{\gamma}) = q(\delta^{(prop)} \mid \delta, \boldsymbol{\gamma}^{(prop)})$ and the acceptance probability simplifies further. Similar simplification is observed under hyper-g/n prior since proposal reflection boundary, $a_{\boldsymbol{\gamma}} = 0$.

Once the model $(\boldsymbol{\gamma}, \delta)$ has been sampled, the regression coefficients can be updated using the fact that $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta, \boldsymbol{y}^*, \boldsymbol{\omega}, \boldsymbol{y} \sim \mathsf{N}(\boldsymbol{m}_{\boldsymbol{\gamma},\boldsymbol{\omega}}, \boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\omega}})$, where

$$\boldsymbol{m}_{\boldsymbol{\gamma},\boldsymbol{\omega}} = \boldsymbol{V}_{\boldsymbol{\omega}} \left(\boldsymbol{X}_{\boldsymbol{\gamma}} \boldsymbol{\Omega} \boldsymbol{z} + \frac{1}{\delta} \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \hat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \right), \quad \boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\omega}} = \left(\boldsymbol{X}_{\boldsymbol{\gamma}}^T \boldsymbol{\Omega} \boldsymbol{X}_{\boldsymbol{\gamma}} + \frac{1}{\delta} \boldsymbol{H}_{\boldsymbol{\gamma}}(\boldsymbol{y}^*) \right)^{-1}.$$

2. A posteriori, the entries of $\boldsymbol{\omega}$ are conditionally independent from each other. Following Polson et al. (2013), it is straightforward to see that $\omega_i \mid \boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta, \boldsymbol{y}^* \sim PG(1, \boldsymbol{x}_{i,\boldsymbol{\gamma}}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}})$. Implementations of the samplers for the Pòlya-Gamma distribution are available, for example, in the `R` package `BayesLogit`.

3. The conditional distribution of $\boldsymbol{y}^*$ is proportional to

$$\pi(\boldsymbol{y}^* \mid \delta, \boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto \pi_{\boldsymbol{\gamma}}^{LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta, \boldsymbol{y}^*) m^*(\boldsymbol{y}^*)$$

While this distribution is supported over the finite set $\{0,1\}^n$, a direct sampler is difficult to construct in part because of its (typically) large size of the support. Hence, we rely again on Metropolis-Hastings steps.

In order to ensure adequate mixing of the algorithm, we consider both local and global proposals. At each iteration, the algorithm selects local moves with probability 0.7 and global moves with probability 0.3.

- For the local moves, we propose new $\boldsymbol{y}^{*(prop)}$ by randomly flipping $d \in \{1,2,3,4,5\}$ components of $\boldsymbol{y}^*$ with probability $(0.5, 0.2, 0.15, 0.10, 0.05)$. The components of $\boldsymbol{y}^*$ to be flipped are selected uniformly at random given d. Because this proposal is symmetric, the acceptance probability for this move is simply

$$\min \left\{ 1, \frac{\pi_{\boldsymbol{\gamma}}^{LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta, \boldsymbol{y}^{*(prop)}) m^*(\boldsymbol{y}^{*(prop)})}{\pi_{\boldsymbol{\gamma}}^{LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta, \boldsymbol{y}^*) m^*(\boldsymbol{y}^*)} \right\}$$

- For the global moves, we use an independent proposal similar to that used by Fouskakis et al. (2018), $q(\boldsymbol{y}^*) = \prod_{i=1}^{n} \text{Berl}(\pi_i{}^*)$ where

$$\pi_i^* = \frac{\pi_0^{1/n} \pi_{i,\boldsymbol{\gamma}_{-1}}^{1/\delta}}{\pi_0^{1/n} \pi_{i,\boldsymbol{\gamma}_{-1}}^{1/\delta} + (1-\pi_0)^{1/n} (1-\pi_{i,\boldsymbol{\gamma}_{-1}})^{1/\delta}},$$

$\pi_0 = \frac{1}{1+\exp\{(-\beta_0)\}}$, $\pi_{i,\boldsymbol{\gamma}_{-1}}^* = \frac{1}{1+\exp\left\{(-\boldsymbol{x}_{i,\boldsymbol{\gamma}_{-1}}^T \boldsymbol{\beta}_{\boldsymbol{\gamma}_{-1}}\right\})}$, and $\boldsymbol{\beta}_{\boldsymbol{\gamma}_{-1}}$ and $\boldsymbol{\gamma}_{-1}$ represent the coefficient vector and indicator variable excluding the intercept term. The associated acceptance probability is then

$$\min\left\{1, \frac{\pi_{\boldsymbol{\gamma}}^{PEP}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta, \boldsymbol{y}^{*(prop)}\right) m^*\left(\boldsymbol{y}^{*(prop)}\right) q(\boldsymbol{y}^*)}{\pi_{\boldsymbol{\gamma}}^{PEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta, \boldsymbol{y}^*) m^*(\boldsymbol{y}^*) q\left(\boldsymbol{y}^{*(prop)}\right)}\right\}$$

4. The fact that δ is jointly sampled with $\boldsymbol{\gamma}$ in step 1 above means that the algorithm might be slow to mix. In order to address this issue, we incorporate an additional sampler for δ alone. The target full conditional distribution is given by:

$$\pi(\delta \mid \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{y}^*, \boldsymbol{\gamma}) \propto \pi_{\boldsymbol{\gamma}}^{PEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta, \boldsymbol{y}^*) f(\delta \mid \boldsymbol{\gamma})$$

For the prior distributions we discuss in this paper, this full posterior conditional distribution does not belong to a known family. Hence, we again use a Metropolis-Hastings algorithm to sample δ that mimics what we did in step 1. In particular, we propose new values for δ from a reflective Gaussian distribution centered around the current value of δ and with scale $\tau = n/2$ and a left reflective boundary $a_{\boldsymbol{\gamma}} = 0$ fr the hyper-g/n prior and $a_{\boldsymbol{\gamma}} = \frac{n-p_{\boldsymbol{\gamma}}}{p_{\boldsymbol{\gamma}}+1}$ for the robust prior. The proposed values are then accepted with probability:

$$\min\left\{1, \frac{\pi_{\boldsymbol{\gamma}}^{LPEP}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta^{(prop)}, \boldsymbol{y}^*\right) p\left(\delta^{(prop)}\right)}{\pi_{\boldsymbol{\gamma}}^{LPEP}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \delta, \boldsymbol{y}^*) p(\delta)} \frac{q(\delta \mid \delta^{(prop)}, \boldsymbol{\gamma})}{q(\delta^{(prop)} \mid \delta, \boldsymbol{\gamma})}\right\}.$$

C.6 Details of the model search algorithm based on a prior-based Bayesian Information Criteria

Using the Laplace approximation of GLM likelihood (4.3), the posterior distribution under hierarchical distribution of LPEP prior (4.7) can be (approximately) written as

$$\pi(\boldsymbol{\gamma}, \boldsymbol{y}^*, \delta \mid \boldsymbol{y}, \boldsymbol{X}) \propto f_{\boldsymbol{\gamma}}^L(\boldsymbol{y} \mid \delta, \boldsymbol{y}^*) f(\delta \mid \boldsymbol{\gamma}) f(\boldsymbol{\gamma}) m^*(\boldsymbol{y}^*).$$

Using the above, we can draw from full conditional distribution of parameters as follows:

1. As before, since the prior support of δ may be dependent on model $\boldsymbol{\gamma}$, we use a random walk Metropolis-Hasting step to jointly sample $(\boldsymbol{\gamma}, \delta)$ from

$$f(\boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{y}) \propto f(\boldsymbol{y} \mid \delta, \boldsymbol{\gamma}, \boldsymbol{y}^*) f(\delta \mid \boldsymbol{\gamma}) f(\boldsymbol{\gamma}),$$

   where $f(\boldsymbol{y} \mid \delta, \boldsymbol{\gamma}, \boldsymbol{y}^*)$ is given in (4.10). To generate proposed values $(\boldsymbol{\gamma}^{(prop)}, \delta^{(prop)})$ for the Metropolis-Hastings algorithm, we use the same procedure as described in step 1 of Appendix C.5. These proposed values are then accepted with probability

$$\min\left\{\frac{f\left(\boldsymbol{\gamma}^{(prop)}, \delta^{(prop)} \mid \boldsymbol{y}^*, \boldsymbol{y}\right)}{f\left(\boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{y}\right)} \frac{q\left(\delta \mid \delta^{(prop)}, \boldsymbol{\gamma}\right)}{q\left(\delta^{(prop)} \mid \delta, \boldsymbol{\gamma}^{(prop)}\right)}, 1\right\},$$

2. Posterior samples from $\boldsymbol{y}^*$ and δ are drawn the same way as described in steps 3 and 4 of Appendix C.5.

3. If (approximate) samples from the posterior distribution $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are desired, these can be obtained by sampling from $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \delta \mid \boldsymbol{y}^*, \boldsymbol{y} \sim \mathsf{N}\left(\tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}}, \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}}^{-1}\right)$, where $(\tilde{\boldsymbol{m}}_{\boldsymbol{\gamma}}, \tilde{\boldsymbol{V}}_{\boldsymbol{\gamma}})$ are given in (4.11). However, note that, unlike in Appendix C.5, this step is not required if the interest is exclusively on exploring the set of possible models.

C.7 Computational efficiency and accuracy - simulation study

To better understand the computational complexity associated with methods based on training sample and the trade-offs between computation and accuracy, we look in detail at two

additional simulation scenarios where $p_{\boldsymbol{\gamma}_T} = 10$, $r = 0.75$. The first one relies on the hyper-g/n prior on the scaling parameter, and the second one uses the unit information prior with $\delta = n$. For the purpose of these additional simulations, all approaches were implemented by us in R so the timings and computational efficiency are comparable.

As before, the results are based on a total of 100 datasets. Our primary metrics for comparing computational efficiencies are the effective sample size (ESS) for the vector $\boldsymbol{\gamma}$ and the corresponding effective sample rate (ESR), defined as ESS per hour of runtime. These metrics are meaningful for all four approaches, as even those that use a Laplace approximation for computing the marginal likelihoods are embedded into an MCMC scheme in order to explore the space of models. To compute the ESS we use the R package `coda` (Plummer et al., 2006). In all four cases, we calculate ESS and ESR from MCMC on the basis of 131,000 samples obtained after a burn-in of 10,000 samples. We report median ESS and ESR across all 100 simulated datasets. Additionally, we also report median F1-score and median MSE to compare the model selection and parameter estimation performance of different methods. A higher score is better for ESS, ESR and F1 score while a lower score is better for MSE.

	LPEPE	LPEPL	LCE	LCL
ESS	**297.200**	276.007	14.609	7.749
ESR	38.931	**57.231**	3.120	2.815
MSE	**13.124**	15.092	19.409	29.245
F1	**0.750**	**0.750**	0.598	0.250

Table C.1: Median summary of metrics evaluated across 100 simulated datasets ($r = 0.75, p_{\boldsymbol{\gamma}_T} = 10$) for different methods under the Hyper-g/n prior for the scaling parameter.

We consider first the results under the Hyper-g/n prior for the scaling parameter. Table

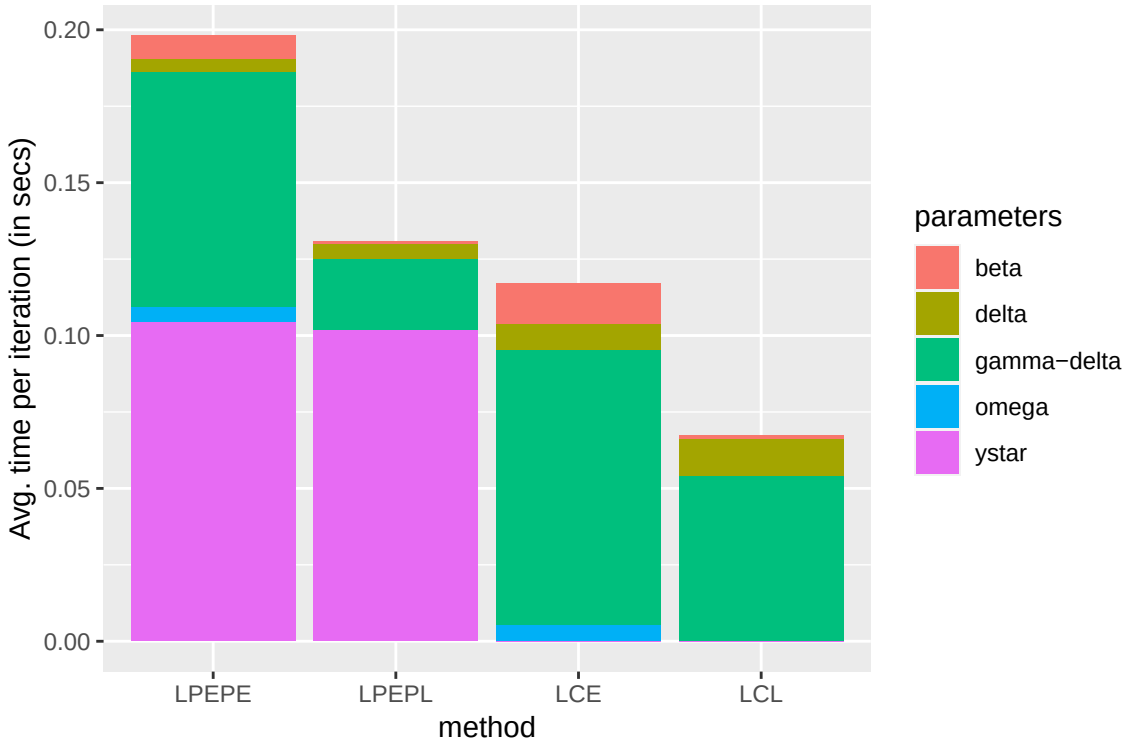


Figure C.1: Stacked bar plot showing average time per iteration taken by MCMC steps of different variables for each method.

	LPEPE	LPEPL	LCE	LCL
ESS	1085.585	1208.781	1102.643	**1252.395**
ESR	136.149	242.156	287.201	**1538.281**
MSE	16.861	**16.368**	17.613	17.122
F1	**0.750**	**0.750**	**0.750**	**0.750**

Table C.2: Median summary of metrics evaluated across 100 simulated datasets ($r = 0.75, p_{\boldsymbol{\gamma}_T} = 10$) for different methods under the UIP prior ($\delta = n$) prior for scaling parameter.

C.1 shows that the "exact" schemes that rely on the Pólya-Gamma auxiliary variables are more accurate than those based on Laplace approximations, as they yield substantially lower MSE values and higher F1 scores. "Exact" methods also tend to yield higher values for the ESS, although the picture around ESR is somewhat mixed: LPEPL has a higher median ESR than LPEPE, but LCE yields a higher median ESR than LCL. Overall, methods based on the LPEP prior led to more accurate estimates and to better values of the ESR than those based on the Li and Clyde (2018) prior. These two results would seem to be connected. The big difference in ESR between LPEP-based methods and those based on the prior by Li and Clyde (2018) is likely driven by the fact that both set of approaches exploring fundamentally different regions of the model space, with LCE and LCL focused on exploring the "wrong" region. These results suggest that the additional computational cost per iteration of the latent variable approach can be compensated by better mixing of the algorithm, leading to higher ESRs overall. They also further confirm that the method of Li and Clyde (2018) can have a very poor performance when a hyperprior is assigned to δ.

We consider now the same methods under the UIP prior for the scaling parameter. Table C.2 indicates that, in this case, methods based on LPEP priors yield only marginally better MSE values that those based on the prior of Li and Clyde (2018), and that all methods have

essentially the same variable selection performance. This suggests that, unlike the case of the Hyper-g/n prior, all four approaches are exploring the same regions of the model space. In this case, the values of the ESS are also all quite close. However, when looking at the ESR, the computational advantages provided by using a Laplace approximation are clear. In particular, the ESR under LCL is about one order of magnitude higher than the ESR under LCE or LPEPE. These results suggest that it is the use Laplace approximations for the marginal likelihood that has the biggest impact on computational efficiency. Indeed, LCE's ESR is only about twice the ESR for LPEPE.

Finally, to complement the previous discussions, we present a comparison of the distribution of average raw computing time per iteration for the various components of the MCMC samplers associated with each of the four methods under the hyper-g/n prior for the scaling parameter (please see Figure C.1). Sampling the imaginary samples required to train the prior (which is only required for the two methods based on the LPEP) is the most time consuming step. However, the computational burden seems manageable: for LPEPE, it represents about half of all the computational effort. On the other hand, the graph suggests that the burden associated with sampling the Pólya-Gamma latent variables (which appear only for the "Exact" methods) is relatively modest.

C.8 Additional Simulation study

We conduct a similar simulation study to section 4.6 where the total number of possible covariates is $p = 20$. All the other scenarios and coefficients follow the same structure to the simulation study discussed in.

p		20							
$p(\boldsymbol{\gamma})$		Beta-Binomial(1,1)							
$p_{\boldsymbol{\gamma}_T}$		0		5		10		20	
r		0	0.75	0	0.75	0	0.75	0	0.75
$\delta = n$	LPEPE	99	100	65	17	43	1	91	10
	LPEPL	99	100	68	18	42	1	92	10
	LCE	99	100	64	14	44	1	34	3
	LCL	99	100	65	15	42	1	40	2
$\delta \sim$ robust	LPEPE	99	100	62	21	34	4	100	100
	LPEPL	99	100	64	23	33	2	100	100
	LCE	99	100	61	23	7	3	100	100
	LCL	99	100	63	22	2	2	100	100
$\delta \sim$ hyper g/n	LPEPE	99	100	65	20	38	8	100	100
	LPEPL	98	100	64	20	34	5	100	100
	LCE	99	100	57	22	0	1	100	100
	LCL	99	100	57	20	0	0	100	100
	LASSO	66	64	0	0	0	0	75	42
	SCAD	74	67	19	2	4	1	64	34
	MCP	75	66	32	6	13	3	61	36
	AIC	75	80	7	5	0	0	100	100
	BIC	99	100	65	13	41	1	16	0

Table C.3: Number of times (**over 100 replications**) that the **MAP** model coincides with the true model in the logistic regression.

p		20							
$p(\boldsymbol{\gamma})$		Beta-Binomial(1,1)							
$p_{\boldsymbol{\gamma}_T}$		0		5		10		20	
r		0	0.75	0	0.75	0	0.75	0	0.75
$\delta = n$	LPEPE	0.53	**0.46***	10.65	29.57	28.11	59.16	54.58	92.20
	LPEPL	**0.52***	**0.46***	**9.76**	**29.13**	**22.26**	**57.53**	**35.55**	**87.54**
	LCE	0.53	**0.46***	11.28	30.27	32.18	62.77	72.08	100.45
	LCL	**0.52***	**0.46***	10.31	29.71	25.22	61.38	45.88	97.20
$\delta \sim$ robust	LPEPE	0.56	**0.46***	9.59	25.29	**21.48***	44.15	**29.95***	55.92
	LPEPL	0.55	0.47	**9.26***	**24.64***	22.39	**42.67***	44.56	**54.71**
	LCE	0.57	0.47	12.50	26.26	39.84	52.58	61.04	69.08
	LCL	**0.52***	**0.46***	10.78	25.79	27.99	47.38	37.15	58.29
$\delta \sim$ hyper g/n	LPEPE	0.68	0.69	9.77	25.25	22.09	45.43	30.91	57.60
	LPEPL	0.68	0.64	**9.37**	**25.07**	**21.04**	**44.58**	37.73	58.40
	LCE	0.67	0.61	11.40	25.19	26.64	46.26	32.64	54.10
	LCL	**0.60**	**0.49**	10.73	25.10	24.00	44.65	**30.16**	**53.33***
	LASSO	1.03	**1.15**	16.52	30.55	28.15	**49.72**	**30.81**	**58.63**
	SCAD	0.89	2.27	11.21	**30.44**	26.13	59.73	45.37	75.99
	MCP	**0.87**	2.65	**10.57**	32.27	**25.69**	59.97	45.05	75.43
	AIC	1.59	2.43	15.69	**29.68**	35.15	**55.28**	**44.65**	**62.84**
	BIC	**0.52**	**0.46**	**10.36**	29.78	**25.41**	61.52	46.25	97.43

Table C.4: 1000 times the **AMSE** for estimated coefficients over 100 replications; **BOLD** represent group minimum; * represent overall minimum.

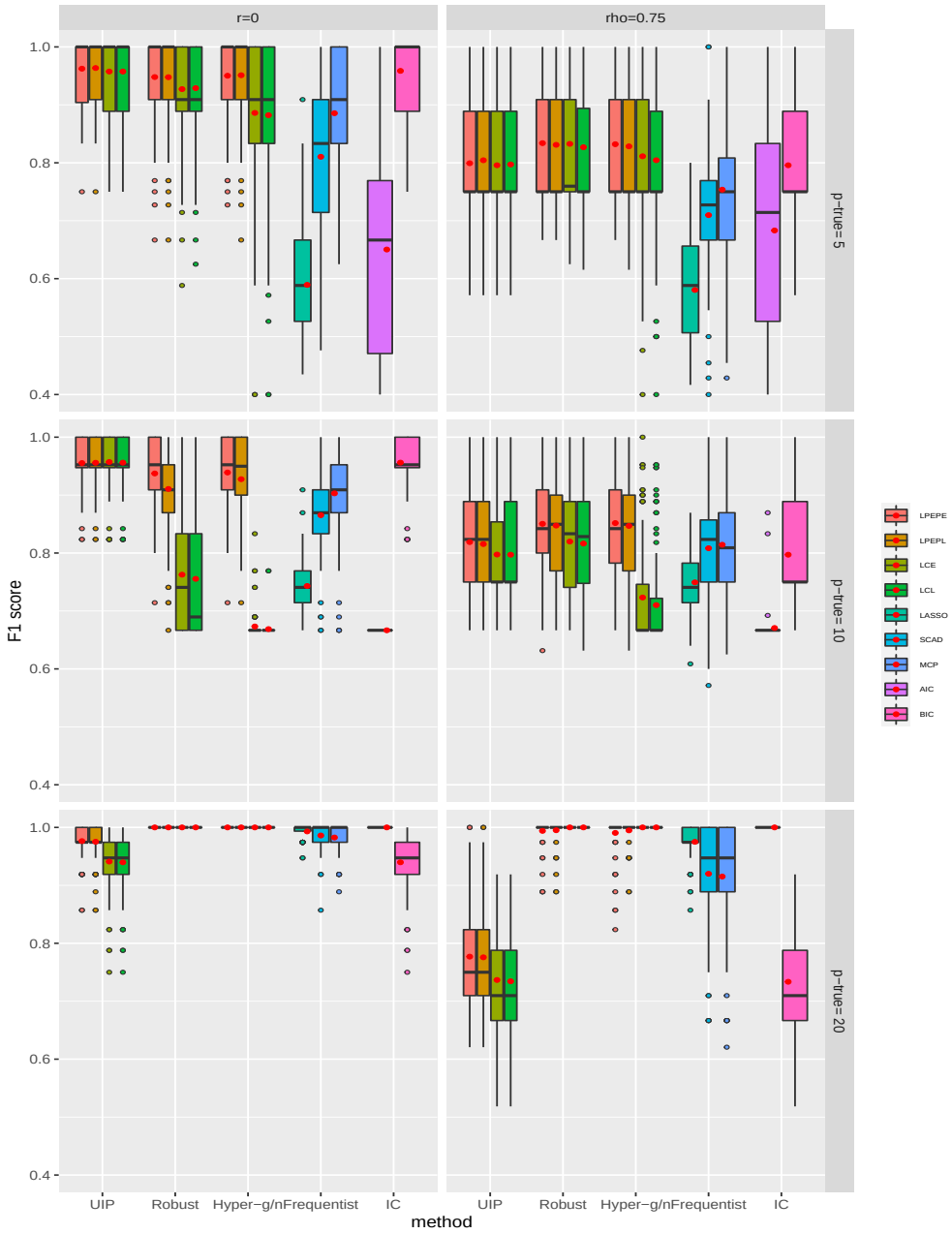


Figure C.2: F1 score for the MAP model estimated by various methods and prior combinations for 100 simulated datasets ($n = 500, p = 20$) under different scenarios of correlation ($r = 0$: left; $r = 0.75$: right) and true number of non-zero coefficients specified in rows ($p - true = p_{\mathcal{M}_T}$); Red dots represent the average F1 score across 100 simulated datasets.

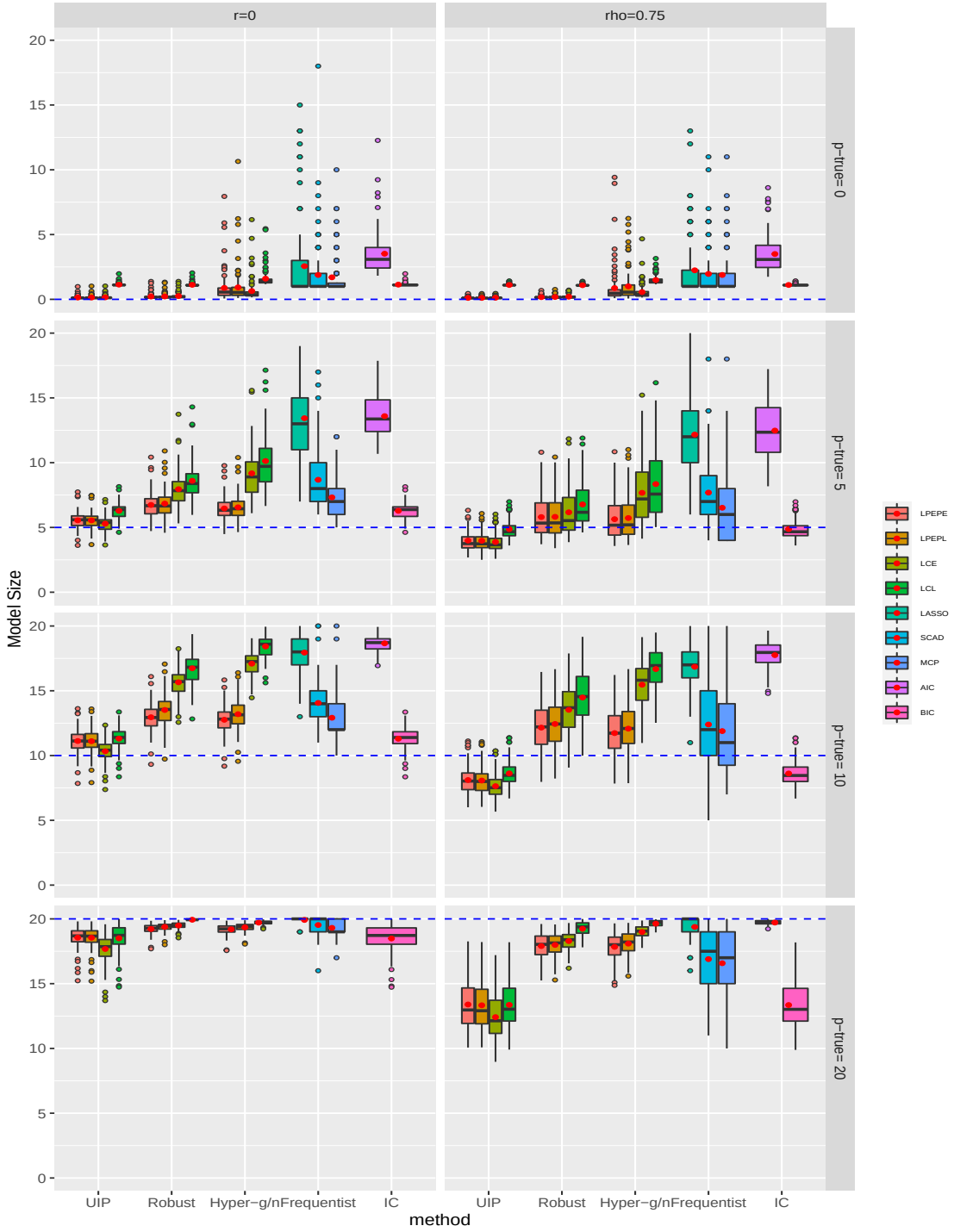


Figure C.3: Average size of models selected by various methods and prior combinations for 100 simulated datasets ($n = 500, p = 20$) under different scenarios of correlation ($r = 0$: left; $r = 0.75$: right) and true number of non-zero coefficients specified in rows ($p - true = p_{\mathcal{M}_T}$); Dotted blue line indicates the true model size $p_{\mathcal{M}_T}$ and red dots represent the average model size over 100 simulated datasets.

C.9 `Endometrial` *data set*

The second data set with full separation that we consider is `endometrial`, which contains data on histology grade and risk factors for 79 cases of endometrial cancer. This data set was first published in Heinze and Schemper (2002), and is also discussed in Agresti (2015). In this case, the first variable fully separates the three classes, and the maximum likelihood estimator for the corresponding coefficient is infinite as shown in table C.5.

Similar to the `urinary` dataset, the regression coefficients for various Bayesian and penalized likelihood methods are shown below along with 95% credible intervals for Bayesian techniques. For the first coefficient, we again observe that LCL and CR/DRPEP methods produces large estimates with very wide credible intervals under all hyper priors. The can be attributed to the ill-definition of the two priors when the MLE are not finite and data is separable.

We also show the posterior probabilities of all eight models in the model space in Table C.6. The models that are favored by all methods are the full model and model without the second variable. The former is the preferred model by LCL, except when $\delta = n$ while the latter is consistently the preferred model by CRPEP, DRPEP and LPEPE.

		β_0	β_1	β_2	β_3
	MLE	4.30	18.19	-0.04	-2.90
		(1.43 , 7.95)	(-69.23 , 506.19)	(-0.14 , 0.04)	(-4.79 , -1.44)
$\delta = n$	LPEPE	3.86	5.85	-0.02	-2.87
		(1.27 , 6.98)	(0.00 , 14.23)	(-0.11 , 0.02)	(-4.62 , -1.42)
	LCL	3.83	16.59	-0.01	-2.85
		(0.81 , 6.77)	(-3317.71 , 3429.10)	(-0.11, 0.02)	(-4.54 , -1.12)
	CRPEP	4.14	29.72	-0.02	-3.04
		(1.53 , 7.40)	(1.71 , 25.05)	(-0.11 , 0.02)	(-4.81 , -1.53)
	DRPEP	3.80	15.49	-0.02	-2.86
		(1.73 , 7.08)	(2.47 , 25.96)	(-0.13 , 0.03)	(-4.51 , -1.62)
$\delta \sim$ robust	LPEPE	3.84	8.25	-0.02	-2.88
		(1.31 , 7.08)	(0.00 , 32.42)	(-0.11 , 0.02)	(-4.75 , -1.41)
	LCL	3.92	16.85	-0.02	-2.80
		(0.83 , 6.86)	(-3303.78 , 3333.49)	(-0.12 , 0.03)	(-4.50 , -1.15)
$\delta \sim$ hyper g/n	LPEPE	3.86	6.03	-0.02	-2.89
		(1.25 , 7.13)	(0.00 , 19.80)	(-0.11 , 0.02)	(-4.67 , -1.40)
	LCL	3.94	15.53	-0.02	-2.54
		(0.84 , 7.02)	(-3158.79 , 3255.70)	(-0.11 , 0.03)	(-4.06 , -0.87)
	CRPEP	3.83	14.36	-0.02	-2.90
		(1.09 , 6.77)	(1.92 , 34.42)	(-0.11 , 0.02)	(-4.71 , -1.35)
	DRPEP	4.06	9.97	-0.02	-3.00
		(1.47 , 7.33)	(0.00 , 20.56)	(-0.12 , 0.01)	(-4.82 , -1.45)
	LASSO	2.39	2.16	0.00	-2.04
	SCAD	3.02	1.84	0.00	-2.44
	MCP	2.70	1.71	0.00	-2.23

Table C.5: Estimated BMA coefficients and 95% credible intervals for different Bayesian techniques for `endometrial` dataset; For frequentist techniques, estimated coefficient is displayed.

		$(\gamma_1, \gamma_2, \gamma_3)$							
		$(0,0,0)$	$(1,0,0)$	$(0,1,0)$	$(1,1,0)$	$(0,0,1)$	$(1,0,1)$	$(0,1,1)$	$(1,1,1)$
$\delta = n$	LPEP	0.00	0.00	0.00	0.00	0.02	**0.55**	0.00	0.42
	LCL	0.00	0.00	0.00	0.00	0.06	**0.60**	0.01	0.33
	CRPEP	0.00	0.00	0.00	0.00	0.01	**0.52**	0.00	0.47
	DRPEP	0.00	0.00	0.00	0.00	0.00	**0.57**	0.00	0.43
$\delta \sim$ robust	LPEPE	0.00	0.00	0.00	0.00	0.03	**0.59**	0.00	0.37
	LCL	0.00	0.00	0.00	0.00	0.04	0.47	0.01	**0.49**
$\delta \sim$ hyper g/n	LPEPE	0.00	0.00	0.00	0.00	0.03	**0.54**	0.01	0.42
	LCL	0.00	0.00	0.00	0.00	0.02	0.43	0.00	**0.54**
	CRPEP	0.00	0.00	0.00	0.00	0.01	**0.62**	0.00	0.38
	DRPEP	0.00	0.00	0.00	0.00	0.06	**0.59**	0.01	0.34
	LASSO	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
	SCAD	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00
	MCP	0.00	0.00	0.00	0.00	0.00	**1.00**	0.00	0.00

Table C.6: Posterior model probabilities for all the models in the model space for `endometrial` dataset.

C.10 `Pima` *data set*

We consider publicly available `Pima` Indians diabetes dataset (Smith et al., 1988), containing $n = 532$ women with complete records on diabetes presence, to illustrate model inference using LPEP prior. This data is publicly available in `MASS` package. It consists of seven potential covariates to model binary response variable `type` indicating diabetes presence; namely, number of pregnancies (`npreg`), plasma glucose concentration (`glu`), diastolic blood pressure (`bp`), triceps skin fold thickness (`skin`), body mass index (`bmi`), diabetes pedigree function (`ped`) and age (`age`). For Bayesian procedures LPEP, CRPEP and DRPEP, we run MCMC chain for 40000 iterations after a burn-in of 10000 iterations while for LCL, we use full enumeration of models for model selection inference.

The marginal posterior inclusion probabilities (PIPs) for each variable are presented in Figure C.4. Similar to simulation studies, we observe that frequentist techniques select denser model in comparison to Bayesian techniques. Among Bayesian techniques, methods with UIP prior selects more parsimonious model than hyper-g/n and robust counterparts. We also observe that PIP for `age` variable is not in agreement under the two CRPEP versions.

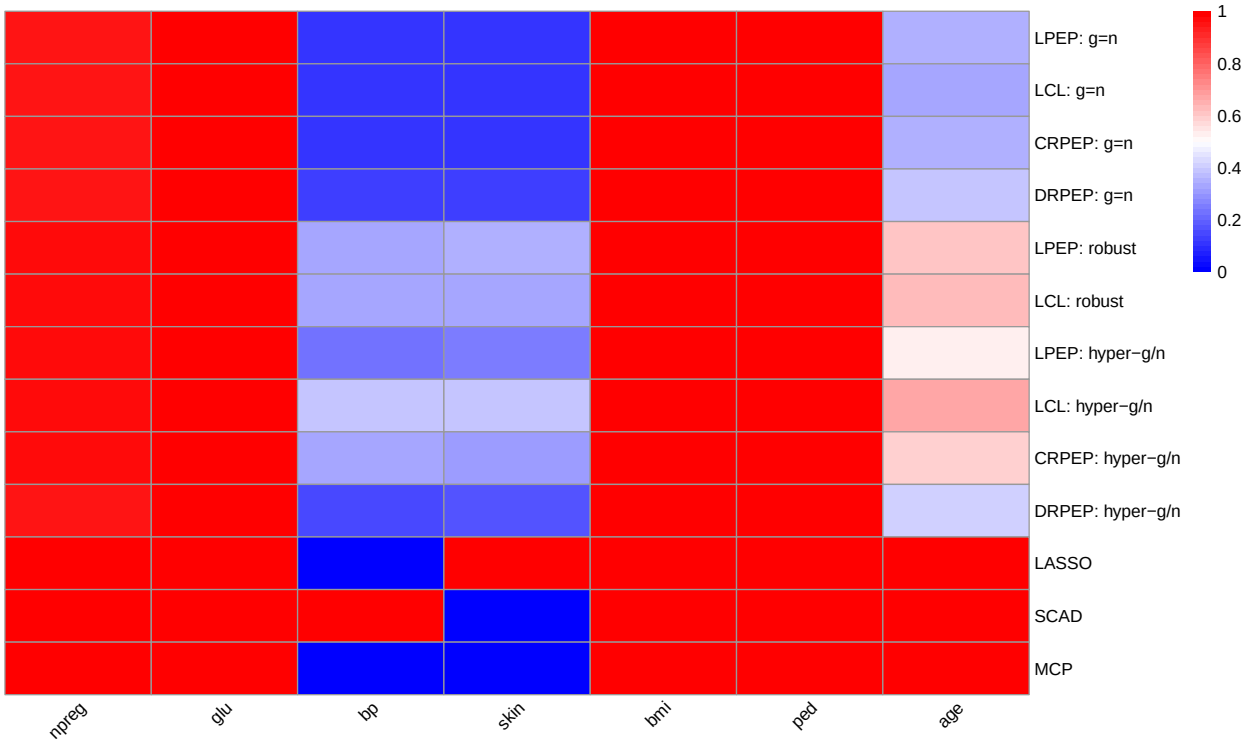


Figure C.4: Marginal posterior inclusion probabilities (PIPs) for `Pima` dataset (Bayesian procedures) and variables included in the model (penalized likelihood methods).

C.11 HOUSE107*: Determinant of legislator behavior in the 107th U.S. House of Representatives*

Most analyses of congressional voting treat all roll-call votes in the same way, independently of the type of vote. This might mask considerable variation in voting behavior across different types of votes. For example, Jessee and Theriault (2012) provide empirical evidence that the forces affecting legislators' voting on procedural and final passage matters have exhibited important changes over time, with differences between these two vote types becoming larger, particularly in recent congresses.

One shortcoming of the methodology presented in Jessee and Theriault (2012) is that it provides legislature-wide measures of agreement among vote types, but cannot ascribe observed differences to individual legislators. Recently, Lofland et al. (2017) and Moser et al. (2021) developed methodology that enables the identification of differences in voting behavior across votes types for individual legislators. In this section we analyze a dataset where the response variable corresponds to estimates of whether each legislator in the 107$^{\text{th}}$ U.S. House of Representatives share the same voting behavior across final passage, amendment and procedural votes. These estimates are obtained using the model introduced in Moser et al. (2021). Hence, in this case, $y_i = 1$ if the i-th legislator voting preferences remain unchanged across all three vote types, and $y_i = 0$ otherwise. The goal of this analysis is to understand which, among a group of 26 characteristics of the legislator or its constituency, affect the likelihood of such changes. Linking voting behavior with these underlying characteristics can provide important insights into the workings of a political system (e.g., see Facchini and Steinhardt, 2011, O'Roark and Wood, 2011 and Cragg et al., 2013).

Detailed description of the variables is available in Table C.7. All the results for the Bayesian procedures in this section are based on the same settings and number of iterations as those used in the previous Section for the GUSTO-I dataset. Figure C.5 shows the marginal PIPs for all 26 variables, along with a list of variables selected by the penalized likelihood methods. Note that, as in previous illustrations, the penalized likelihood methods tend to

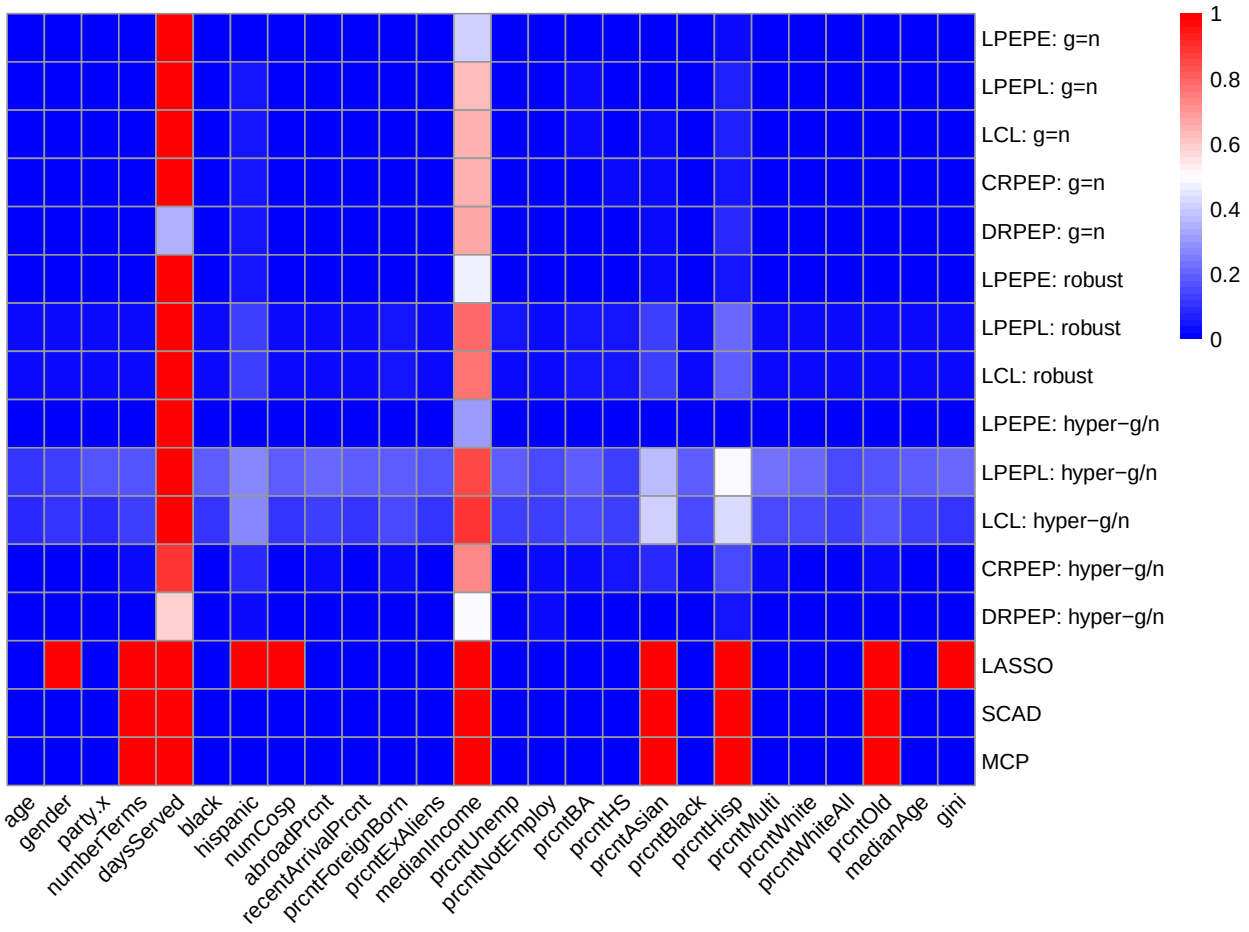


Figure C.5: Marginal posterior inclusion probabilities (PIPs) for `HOUSE107` dataset (Bayesian procedures) and variables included in the model (penalized likelihood methods).

select a superset of the variables selected by the Bayesian approaches, with LASSO selecting the largest superset. Furthermore, there is broad agreement among all Bayesian procedures. For example, all techniques, except both versions of DRPEP, assign high PIP to `daysServed`. Based on our experience with this application, including `daysServed` in the model is sensible. For example, length of tenure has been previously identified in the literature as an important predictor of legislator's effectiveness (e.g., see Miquel and Snyder Jr, 2006). One place where the different procedures do seem to disagree is whether `medianIncome` (the median income

in the district represented by the legislator) explains voting behavior. Procedures based on theLPEPE agree in providing weak to moderate evidence *against* the inclusion of this variable, while most other procedures provide weak to moderate evidence *in favor of it*. Interestingly, the two versions of the DRPEP seem to disagree, with the unit information version providing weak evidence in favor of its inclusion and the hyper-g/n providing weak evidence against it.

Variable	Description
y	Legislator voting preference remain unchanged across three vote types (Yes = 1, No = 0)
age	Age (in years)
gender	Gender (Female=0; Male =1)
party.x	Party of the legislator (0=Democrat; 1= Republican)
numberTerms	Number of terms served in the House
daysServed	Number of days served
black	Legislator is African American (Yes=1; No=0)
hispanic	Legislator is Hispanic (Yes=1; No=0)
numCosp	Number of bills an MC cosponsored in a term
abroadPrcnt	Percent of the district that lived abroad in the census used for this term
recentArrivalPrcnt	Percent of the district that recently moved into the district from another county or state
prcntForeignBorn	Percentage of the population in the district that are foreign born
prcntExAliens	Percentage of the foreign born population in the district that became a citizen
medianIncome	Median income of district represented by the legislator
prcntUnemp	Unemployment rate in the district represented by the legislator
prcntNotEmploy	Percentage of adults not employed in the district represented by the legislator
prcntBA	Percentage of the adult population in district with bachelor´s degree
prcntHS	Percentage of the adult population in district with a high school degree
prcntAsian	Percentage of Asian population in district represented by legislator
prcntBlack	Percentage of African American population in district represented by legislator
prcntHisp	Percentage of Hispanic population in district represented by legislator
prcntMulti	Percentage of population with multiple ethnicity in district represented by legislator
prcntWhite	Percentage of White population in district represented by legislator, excluding hispanics
prcntWhiteAll	Percentage of White population in district represented by legislator, including hispanics
prcntOld	Percentage of individual over 65 years of age in district represented by legislator
medianAge	Median age of district represented by legislator
gini	Gini Index of the district represented by the legislator

Table C.7: Description of variables in HOUSE107 dataset.

C.12 Details of GUSTO-I *dataset*

Variable	Description
y	Death within 30 days after acute myocardial infarction (Yes = 1, No = 0)
SEX	Gender (Female=1; Male =0)
AGE	Age (in years)
KILLIP	KILLIP class (4 categories)
DIA	Diabetes (Yes =1; No=0)
HYP	Hypotension (Yes=1; No=0)
HRT	Tachycardia (Yes=1; No=0)
ANT	Anterior infarct location (Yes=1; No=0)
PMI	Previous myocardial infarction (Yes=1, No=0)
HEI	Height (in cm)
WEI	Weight (in kg)
HTN	Hypertension history (Yes=1; No=0)
SMK	Smoking status (Never/ ex/ current)
LIP	Hypercholestrolaemia (Yes=1, No=0)
PAN	Previous angina pectoris (Yes=1, No=0)
FAM	Family history of myocardial infarctions (Yes=1, No=0)
STE	ST elevation on ECG: number of leads (0-11)
TTR	Time to relief of chest pain more than one hour (Yes=1, No=0)

Table C.8: Description of variables in GUSTO-I dataset.

Appendix D

APPENDIX D

D.1 Empirical Illustration of GL-g priors avoiding Conditional Lindley paradox

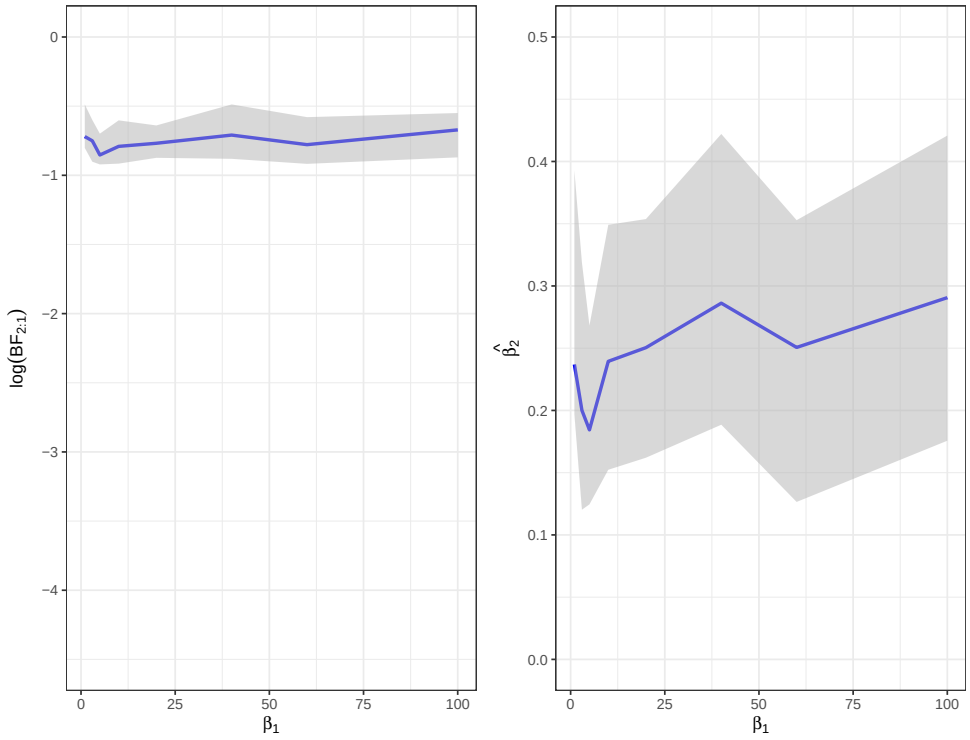


Figure D.1: Empirical illustration of GL-g prior with hyper-g prior on each g avoiding CLP; Left displays $log(BF_{2:1})$ and right figure displays $\hat{\beta}_2$ as β_1 increases under the asymptotic regime described in (5.3).

D.2 Details of MCMC sampler for DP block-g priors

Using the hierarchical representation of the DP block-g prior discussed at the start of Section 5.5.1, the posterior distribution for the augmented model can be written as

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \alpha, \boldsymbol{\xi}, \boldsymbol{\pi}, g_1^*, \ldots, g_k^* \mid \boldsymbol{Y}, \boldsymbol{X}) \propto f_{\boldsymbol{\gamma}}(\boldsymbol{Y} \mid \alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{X}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \sigma^2, \boldsymbol{G}^*, \boldsymbol{\xi}, \boldsymbol{\gamma}, \tau^2, \boldsymbol{X}) \\ f(\alpha, \sigma^2) f(\boldsymbol{\xi}|\pi) f(\boldsymbol{\pi}|a_0) \prod_{k=1}^{K} p(g_k^*) f(\tau) f(a_0).$$

From this, it is easy to devise samplers for the full conditional posterior distributions of various blocks of parameters. We focus below on the more general setting where $\boldsymbol{\gamma}, \boldsymbol{\xi}, \tau$ and a_0 has been assigned a hyperprior. The simplifications for the case where these parameters are fixed are straightforward and we do not discuss them explicitly. At each iteration, we sample from the following conditional posteriors as follows:

1. We sample model $\boldsymbol{\gamma}$ from the conditional posterior $f(\boldsymbol{\gamma}|\boldsymbol{G}^*, \boldsymbol{\xi}, \cdot)$ given by

$$f(\boldsymbol{\gamma}|\boldsymbol{G}^*, \boldsymbol{\xi}, \cdot) \propto f(\boldsymbol{\gamma}) \int \cdots \int f_{\boldsymbol{\gamma}}(\boldsymbol{Y}|\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{X}) f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \boldsymbol{G}^*, \boldsymbol{\xi}, \boldsymbol{\gamma}, \tau^2, \boldsymbol{X}) f(\sigma^2, \alpha) d\boldsymbol{\beta}_{\boldsymbol{\gamma}} d\alpha d\sigma^2 \\ \propto f(\boldsymbol{\gamma}) f(\boldsymbol{Y}|\boldsymbol{G}^*_{\boldsymbol{\xi},\boldsymbol{\gamma}}, \boldsymbol{X}),$$

   where $f(\boldsymbol{Y}|\boldsymbol{G}^*_{\boldsymbol{\xi},\boldsymbol{\gamma}}, \boldsymbol{X})$ is given by (5.5). We generate samples from the above distribution using metropolis Hastings algorithm for $\boldsymbol{\gamma}$ using a symmetric random walk proposal similar to equation (46) of George and McCulloch (1997) as follows:

   - We define two probability vectors $p_1 = (0.7, 0.3)$ and $p_2 = (0.4, 0.3, 0.2, 0.1)$.
   - Each time, we decide on one of two types of moves according to the probability vector p_1.
     - If a move type 1 is selected, then the proposed new model $\boldsymbol{\gamma}^{(prop)}$ is generated by randomly flipping $d \in \{1, 2, 3, 4\}$ components of $\boldsymbol{\gamma}$ with probability $p_{2,d}$. The components of $\boldsymbol{\gamma}$ to be flipped are selected uniformly at random given d.

– If a move type 2 is selected, then the proposed model $\boldsymbol{\gamma}^{(prop)}$ is generated by removing one variable currently included in the model and replacing it with a variable that is currently excluded, leaving the dimensionality of the model unchanged. The variables to add and remove are chosen uniformly at random within each set.

The proposed model is then accepted with probability

$$\min\left\{\frac{f(\boldsymbol{\gamma}^{(prop)}|\boldsymbol{G}^*,\boldsymbol{\xi})}{f(\boldsymbol{\gamma}|\boldsymbol{G}^*,\boldsymbol{\xi})},1\right\}.$$

2. Once the model is sampled, we can update α and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ using normal-normal conjugacy as follows:

$$\alpha|\cdot \sim \mathcal{N}(\bar{\boldsymbol{Y}},\frac{\sigma^2}{n}),$$
$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\cdot \sim \mathcal{N}(\boldsymbol{m}_{\boldsymbol{\gamma},\boldsymbol{\xi}},\boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}})$$

where

$$\boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}}=\sigma^2\left\{\frac{{\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{-1/2}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}{\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{-1/2}}{\tau^2}+\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{X}_{\boldsymbol{\gamma}}\right\}^{-1},\qquad \boldsymbol{m}_{\boldsymbol{\gamma},\boldsymbol{\xi}}=\frac{\boldsymbol{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}}\boldsymbol{X}_{\boldsymbol{\gamma}}^T\boldsymbol{Y}}{\sigma^2}.$$

3. We can sample sample variance as

$$\sigma^2|\cdot \sim \text{Inverse-Gamma}\left(\frac{n-1}{2},\frac{\boldsymbol{Y}^T(\boldsymbol{I}+\tau^2\boldsymbol{X}_{\boldsymbol{\gamma}}{\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}{\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{1/2}\boldsymbol{X}_{\boldsymbol{\gamma}}^T)^{-1}\boldsymbol{Y}-n\bar{\boldsymbol{Y}}^2}{2}\right)$$

4. Sample sequentially ξ_i for $i=1,\ldots,p$ according to the following

- If $\gamma_i=1$, sample according to

$$p(\xi_i=k|\cdot)\propto\pi_k\mathcal{N}(\boldsymbol{0},\sigma^2\tau^2{\tilde{\boldsymbol{G}}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{1/2}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}{\tilde{\boldsymbol{G}}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{1/2}),$$

where $\tilde{\boldsymbol{G}}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ is a diagonal matrix same as $\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}$ with $g^*_{\xi_i}$ replaced by g^*_k.

- If $\gamma_i=0$, sample according to

$$p(\xi_i=k|\cdot)\propto\pi_k$$

Then, let $S_{\boldsymbol{\gamma},k} = \{i : \gamma_i = 1 \text{ and } \xi_i = k\}$ for $k = 1, \ldots, K$ denote the set of indices that belong to cluster k and are included in the model and compute $c_{\boldsymbol{\gamma},k} = |S_{\boldsymbol{\gamma},k}|$.

5. Cluster probabilities $\boldsymbol{\pi}$ can then be updated as

$$\boldsymbol{\pi}|\cdot \sim \text{Dir}\left(\frac{a_0}{K} + c_{\boldsymbol{\gamma},1}, \ldots, \frac{a_0}{K} + c_{\boldsymbol{\gamma},K}\right).$$

6. Since clustering is performed only over variables included in the model, let $p_{\boldsymbol{\gamma}} = |\boldsymbol{\gamma}|$ denote the number of variables being clustered. Assuming $a_0 \sim \text{Gamma}(1,1)$ apriori, we can sample a_0 as follows: a) if $p_{\boldsymbol{\gamma}} = 0$, sample a_0 from the prior, i.e. $a_0 \sim \text{Gamma}(1,1)$; and b) if $p_{\boldsymbol{\gamma}} \neq 0$, use augmentation strategy by Escobar and West (1995) as follows:

$$a_0|\eta, K_0 \sim \pi_\eta \text{Gamma}(1 + K_0, 1 - \log(\eta)) + (1 - \pi_\eta)\text{Gamma}(1 + K_0 - 1, 1 - \log(\eta))$$
$$\eta|a_0 \sim \text{Beta}(a_0 + 1, p_{\boldsymbol{\gamma}})$$

with the weights π_η defined by $\frac{\pi_\eta}{1-\pi_\eta} = \frac{K_0}{p_{\boldsymbol{\gamma}}(1-\log(\eta))}$ where K_0 denote the number of non-empty clusters.

7. The conditional posterior distribution of g_k^* for $k = 1, \ldots, K$ is given by

$$\begin{aligned} f(g_k^*|\cdot) &\propto f(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\sigma^2, \boldsymbol{G}^*, \boldsymbol{\xi}, \boldsymbol{\gamma}, \boldsymbol{X}, \tau^2)p(g_k^*) \\ &\propto \mathcal{N}(\mathbf{0}, \sigma^2\tau^2 {\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{1/2} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}} {\boldsymbol{G}^*_{\boldsymbol{\gamma},\boldsymbol{\xi}}}^{1/2}) BP(a,b,1) \end{aligned}$$

Note that if $|S_{\boldsymbol{\gamma},k}| = 0$, then we sample corresponding $g_k^* \sim BP(a,b,1)$ since there is no likelihood contribution. If $|S_{\boldsymbol{\gamma},k}| \neq 0$, then we can simplify the above conditional posterior as

$$f(g_k^*|.) \propto (g_k^*)^{b - \frac{c_{\boldsymbol{\gamma},k}}{2}} (1 + g_k^*)^{-a-b-2} \exp\left(-\frac{v_k}{g_k^*} - \frac{w_k}{\sqrt{g_k^*}}\right),$$

where

$$v_k = \frac{1}{2\sigma^2\tau^2} \sum_{\substack{j \in S_{\boldsymbol{\gamma},k} \\ i \in S_{\boldsymbol{\gamma},k}}} \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\gamma},jj} \beta_{\boldsymbol{\gamma},j} \beta_{\boldsymbol{\gamma},i} \qquad w_k = \frac{1}{\sigma^2\tau^2} \sum_{\substack{j \in S_{\boldsymbol{\gamma},k} \\ i \notin S_{\boldsymbol{\gamma},k}}} \frac{\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\gamma},ji} \beta_{\boldsymbol{\gamma},j} \beta_{\boldsymbol{\gamma},i}}{\sqrt{g_{\xi_i}}}.$$

Using the transformation $t_k = \frac{v_k}{g_k^*}$, we can re-parametrize this density as

$$f(t_k|.) \sim t_k^{a+\frac{c_{\boldsymbol{\gamma},k}}{2}} \left(1 + \frac{t_k}{v_k}\right)^{-a-b-2} \exp\left(-t_k - \frac{w_k}{\sqrt{v_k}}\sqrt{t_k}\right).$$

Introduce the auxiliary variable u_k. Then, we can use slice sampling in conjunction with a modification of rejection sampler developed by Liu et al. (2012) to sample t_k from a truncated extended gamma distribution to sample t_k as follows

$$u_k|t_k \sim \mathcal{U}\left(0, \left(\frac{v_k}{v_k + t_k}\right)^{a+b+2}\right),$$
$$t_k|u_k, . \sim \text{Truncated-Extended-Gamma}\left(a + \frac{c_{\boldsymbol{\gamma},k}}{2} + 1, \frac{w_k}{2\sqrt{v_k}}, v_k(u_k^{\frac{-1}{a+b+2}} - 1)\right),$$

where Truncated-Extended-Gamma distribution is given by

$$f(t|a, b, c) \propto t^{a-1} \exp\Big(-t - 2\sqrt{t}b\Big) \mathbb{1}_{\{0<t<c\}}, \quad t > 0,$$

where $a > 0$ and $b \in \mathbb{R}$. Note that rejection sampler for untruncated extended gamma distributions developed by Liu et al. (2012) can be modified in a straightforward manner by using truncated proposals at the truncation level c. This can then be used to efficiently draw from truncated versions of extended Gamma distribution.

8. To sample τ, we use the parameter augmentation trick by Makalic and Schmidt (2015) as follows

$$\tau^2|. \sim \text{Inverse-Gamma}\left(\frac{p_{\boldsymbol{\gamma}} + 1}{2}, \frac{1}{\nu} + \frac{\boldsymbol{\beta}_{\boldsymbol{\gamma}}^T {\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\xi}}^*}^{-1/2} \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} {\boldsymbol{G}_{\boldsymbol{\gamma},\boldsymbol{\xi}}^*}^{-1/2} \boldsymbol{\beta}_{\boldsymbol{\gamma}}}{2\sigma^2}\right),$$
$$\nu|\tau^2 \sim \text{Inverse-Gamma}\left(1, 1 + \frac{1}{\tau^2}\right).$$