

UNIVERSITY OF CALGARY

Using prior-data conflict to tune Bayesian regularized regression models

by

Timofei Biziaev

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN MATHEMATICS AND STATISTICS

CALGARY, ALBERTA

SEPTEMBER, 2023

© Timofei Biziaev 2023

Abstract

In high-dimensional regression models, variable selection becomes challenging from a computational and theoretical perspective. Bayesian regularized regression via shrinkage priors like the Laplace or spike-and-slab prior are effective methods for variable selection in $p > n$ scenarios provided the shrinkage priors are configured adequately. We propose configuring shrinkage priors using checks for prior-data conflict: tests that assess whether there is disagreement in parameter information provided by the prior and data. We apply our proposed method to the Bayesian LASSO and spike-and-slab shrinkage priors and assess variable selection performance of our prior configurations against competing models through a linear and logistic high-dimensional simulation study. Additionally, we apply our method to proteomic data collected from patients admitted to the Albany Medical Center in Albany NY in April of 2020 with COVID-like respiratory issues. Simulation results suggest our proposed configurations may outperform competing models when the true regression effects are small.

Preface

This thesis is an original work by the author. No part of this thesis has been previously published.

Acknowledgements

I would like to thank my supervisors Dr. Karen Kopciuk and Dr. Thierry Chekouo for all their support, advice, and guidance throughout this degree.

Table of Contents

Abstract	ii
Preface	iii
Acknowledgements	iv
Table of Contents	vi
List of Figures	vii
List of Tables	viii
List of Symbols, Abbreviations, and Nomenclature	ix
1 Introduction	1
1.1 Variable selection methods	2
1.1.1 Classical variable selection and penalized regression	3
1.1.2 Bayesian regularized regression	5
1.2 Scope of the thesis	8
2 Methods	9
2.1 Existing regularized regression models	9
2.1.1 LASSO (Least absolute shrinkage and selection operator)	10
2.1.2 Elastic-Net	11
2.1.3 The Spike-and-Slab LASSO	11
2.1.4 The Bayesian LASSO	12
2.1.5 The Spike-and-Slab Prior	14
2.2 Prior-data conflict	16
2.3 Minimum-divergence prior expectation	18
2.4 Summary	21
3 Simulation	23
3.1 Design of the simulation	23
3.1.1 Competing Models	23
3.1.2 Configurations of the data	24
3.1.3 Performance metrics	26
3.1.4 Implementation of methods	26
3.2 Results	28
3.2.1 Linear regression simulation results	28
3.2.2 Logistic regression simulation results	34
3.3 Discussion	38
3.3.1 Linear regression	38
3.3.2 Logistic regression	40
3.3.3 Comparison between minimum-divergence models	41

4	Application to a proteomics COVID-19 dataset	44
4.1	Overview of data	44
4.2	Methods	45
4.3	Results and discussion	45
4.3.1	Analysis of COVID severity	45
4.3.2	Analysis of COVID presence	47
4.3.3	Conclusions	48
5	Conclusions and future work	50
5.1	Summary and discussion	50
5.2	Future work	53
	Bibliography	55

List of Figures

2.1	The point-mass spike-and-slab prior (left) with $\theta = 0.25$, $\tau^2 = 10$, the Laplace prior (used the Bayesian LASSO model, center) with $\lambda = 1$, and the Laplace spike-and-slab prior (used in the Spike-and-Slab LASSO model, right) with $\theta = 0.25$, $\lambda_0 = 3$, and $\lambda_1 = 1$	14
3.1	Correlation plots of the proteomic data. Left plot is of the original unperturbed data, and the right of the perturbed block-correlated data	25
3.2	Average simulated prior-to-posterior KL Divergences of τ^2 in the linear spike-and-slab model based on 90-100 replicates. Vertical dashed lines denote the first local minimum. For uncorrelated $p = 200$ scenario with $var(\tau^2) = 1$ (red) and $var(\tau^2) = 10$ (black)	28
3.3	Distribution of minimum-divergence prior means of τ^2 in the linear spike-and-slab model based on 90-100 replicates. For uncorrelated $p = 200$ scenario with $var(\tau^2) = 10$	29
3.4	Average simulated inverse prior-to-posterior KL Divergences of λ^2 in the linear Bayesian LASSO model based on 90-100 replicates. Vertical dashed lines denote the first local maximum. For uncorrelated $p = 200$ scenario with $var(\lambda^2) = 1$ (red) and $var(\lambda^2) = 10$ (black)	31
3.5	Distribution of minimum-divergence prior means of λ^2 in the linear Bayesian LASSO model based on 90-100 replicates. For uncorrelated $p = 200$ scenario with $var(\lambda^2) = 10$	31
3.6	Average simulated prior-to-posterior KL Divergences of τ^2 in the logistic spike-and-slab model based on 90-100 replicates. Vertical dashed lines denote the first local minimum. For uncorrelated $p = 200$ scenario with $var(\tau^2) = 1$ (red) and $var(\tau^2) = 10$ (black)	34
3.7	Distribution of minimum-divergence prior means of τ^2 in the logistic spike-and-slab model based on 90-100 replicates. For uncorrelated $p = 200$ scenario with $var(\tau^2) = 10$	34
3.8	Posterior distributions of τ^2 , λ^2 in the linear spike-and-slab and Bayesian LASSO model where $\pi(\tau)$, $\pi(\lambda) \sim \text{Uniform}(0, 100)$, for the 7 th replicate of uncorrelated simulation scenario with $p = 200$, $c = 4$, effects $+/-0.5$. Linear $\pi(\tau^2 \mathbf{y})$ is scaled down by a factor of 10 for comparability	42
3.9	Posterior distributions of τ^2 in the linear and logistic spike-and-slab model where $\pi(\tau) \sim \text{Uniform}(0, 100)$, for the 2 nd replicate of uncorrelated simulation scenario with $p = 200$, $c = 4$, effects $+/-0.5$. Linear $\pi(\tau^2 \mathbf{y})$ is scaled down by a factor of 10 for comparability	43
4.1	Standardized inverse prior-to-posterior KL Divergences of τ^2 in the linear (left) and logistic (right) spike-and-slab model. Vertical dashed lines denote the first local maximum	46
4.2	Posterior model inclusion probabilities of proteins for the MD-LinSS model (left) and MD-LogSS model (right). Red probabilities denote selected proteins	46

List of Tables

3.1	Factors varied and values considered in simulation study	26
3.2	Uncorrelated linear simulation study results using 90 to 100 replicates. Each entry refers to the average value of the metric and simulation standard error (in parentheses)	30
3.3	Uncorrelated linear simulation study results using 90 to 100 replicates, continued. Each entry refers to the average value of the metric and simulation standard error (in parentheses) . . .	32
3.4	Correlated linear simulation study results using 90 to 100 replicates. Each entry refers to the average value of the metric and simulation standard error (in parentheses)	33
3.5	Uncorrelated logistic simulation study results using 90 to 100 replicates for $p = 200$, and preliminary uncorrelated logistic simulation study results using 15 to 20 replicates for $p = 500$. Each entry refers to the average value of the metric and simulation standard error (in parentheses)	35
3.6	Uncorrelated logistic simulation study results using 90 to 100 replicates for $p = 200$, and preliminary uncorrelated logistic simulation study results using 15 to 20 replicates for $p = 500$, continued. Each entry refers to the average value of the metric and simulation standard error (in parentheses)	36
3.7	Preliminary correlated logistic simulation study results using 15 to 20 replicates. Each entry refers to the average value of the metric and simulation standard error (in parentheses) . . .	37

List of Symbols, Abbreviations, and Nomenclature

Symbol	Definition
p	Number of predictors
n	Number of observations
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
β	Vector of regression coefficients $(\beta_1, \beta_2, \dots, \beta_p)^T$
\mathbf{X}	$n \times p$ matrix of covariates (X_1, X_2, \dots, X_p)
\mathbf{y}	Vector of responses $(y_1, y_2, \dots, y_n)^T$
$\ \cdot\ _j$	L_j norm, where $j = 0, 1$, or 2
θ	Mixing weight for spike-and-slab priors
γ	Vector of p indicator variables for spike-and-slab priors $(\gamma_1, \gamma_2, \dots, \gamma_p)^T$
$\bar{\gamma}$	Posterior sample mean of γ
λ	Rate parameter for Laplace distribution
τ^2	Slab variance for the point-mass spike-and-slab prior
λ_0, λ_1	Rate parameters for the spike and slab, respectively, of the Spike-and-Slab LASSO
$E(\cdot)$	Expectation
$Var(\cdot)$	Variance
MC-EM	Monte-Carlo Expectation-Maximization algorithm
MCMC	Markov-Chain Monte-Carlo algorithm
$\delta_0(\cdot)$	Dirac delta function at 0
$KL(\mathbf{y})$	The prior-to-posterior Kullback-Leibler divergence of a parameter
$m(\mathbf{y})$	The prior-predictive distribution

μ_{τ^2}	Prior mean of slab variance τ^2 in the point-mass spike-and-slab
$\sigma_{\tau^2}^2$	Prior variance of slab variance τ^2 in the point-mass spike-and-slab
μ_{λ^2}	Prior mean of the squared tuning parameter λ in the Bayesian LASSO
$\sigma_{\lambda^2}^2$	Prior variance of the squared tuning parameter λ in the Bayesian LASSO
$\mu_{\tau^2}^{MD}$	Minimum-divergence prior mean of τ^2
$\mu_{\lambda^2}^{MD}$	Minimum-divergence prior mean of λ^2
MD-LinBLASSO	Minimum-divergence linear Bayesian LASSO
MD-LogBLASSO	Minimum-divergence logistic Bayesian LASSO
MD-LinSS	Minimum-divergence linear spike-and-slab model
MD-LogSS	Minimum-divergence logistic spike-and-slab model
NI-LinBLASSO	Non-informative linear Bayesian LASSO
NI-LogBLASSO	Non-informative logistic Bayesian LASSO
NI-LinSS	Non-informative linear spike-and-slab model
NI-LogSS	Non-informative logistic spike-and-slab model
c	Number of nonzero effects
s	Sparsity level, c/p
TPR	True-positive rate
FDR	False-discovery rate
FNDR	False-nondiscovery rate
F_1	F_1 score, harmonic mean of TPR and $1 - \text{FDR}$
JAGS	Just Another Gibbs Sampler
HFD	Hospital-free days
NET	Neutrophil extracellular trap
BFDR	Bayesian false-discovery rate

Chapter 1

Introduction

Advancements in information technology have made collection and storage of highly detailed data relatively inexpensive. As opposed to carefully selecting a small number of variables to measure, it is now easy in many scenarios to create a “snapshot” of each observation by collecting measurements of a large number of variables, leading to data with more predictors, p , than observations, n . Such data are referred to as high-dimensional data and examples include biological data like genomic, proteomic and metabolomic data, high-resolution imaging like magnetic resonance imaging (MRI), hyperspectral satellite imaging, high-frequency financial data, text data, and others (Donoho et al., 2000; Fan & Li, 2006). Whether the purpose of modelling is to make accurate predictions or to understand the underlying data-generating process, it is often important to identify which predictors are actually associated with the outcome. This is referred to as variable selection and is particularly important, and challenging, in high-dimensions. It is important because high-dimensional data necessarily contains a lot of irrelevant information (noise) that, when modelled, can decrease performance of the model (Fan & Fan, 2008; Fan & Li, 2006). Additionally, high-dimensional data can be collected and analyzed precisely for the purpose of identifying predictors associated with an outcome as in genome-wide association studies or quantitative trait loci mapping, where genes are identified to provide a genetic basis for disease or some measurable trait like body mass index (Fan & Lv, 2010; Myles & Wayne, 2008). In the analysis of health or biological data, regression models are often used for their interpretability and, when $p > n$, variable selection for such models becomes challenging from a theoretical, as well as practical, perspective. Many regression variable selection methods use p -values to test hypotheses concerning inclusion of certain variables. Traditional statistical techniques, however, assume $p < n$, and for asymptotic considerations assume $n \rightarrow \infty$, whereas in high-dimensions it may be more reasonable to consider $p \rightarrow \infty$ with $p > n$, or even n fixed (Donoho et al., 2000). As such, no asymptotically valid p -

values are available (Meinshausen et al., 2009). In high-dimensions severe correlation among predictors is expected which increases the chances that irrelevant variables are selected in addition to, or worse, instead of, the truly significant variable they are correlated with. Even if predictors are truly uncorrelated with one another, Fan and Lv, 2010 show in the case that $p = 10^3$ or $p = 10^4$ and $n = 50$, relevant variables can easily be approximated by irrelevant ones due to chance correlation. Some variable selection methods (Akaike, 1998; Foster & George, 1994; George & Foster, 2000; Mallows, 2000; Schwarz, 1978) identify the best subset of predictors by evaluation of a criterion such as the Akaike or Bayesian information criterion (AIC, BIC, respectively) for all possible submodels. This becomes infeasible for large p as it involves computation of the AIC or BIC for 2^p models, falling into a class of computational problems referred to as having NP-complexity (Fan & Li, 2006). Regularized regression, however, provides a computationally and theoretically viable alternative to p -value or criterion-based variable selection that is applicable in $p > n$ situations. In regularized regression, it is assumed that most variables are just noise and this assumption leveraged for improved prediction accuracy and variable selection performance. Novel regularization methods and theory are consistently being developed to deal with the challenges posed by high dimensionality, with some researchers posing high-dimensional data analysis as the most important statistical topic of this century (Donoho et al., 2000; Fan & Li, 2006).

In this chapter we outline methods for variable selection in regression models. We begin by outlining some benefits of variable selection and in Section 1.1.1 review classical variable selection methods as well as penalized regression methods and their applicability to high-dimensional data. In Section 1.1.2 we outline Bayesian regularized regression models and in Section 1.2 provide a scope of the thesis.

1.1 Variable selection methods

The benefits of variable selection depend on the purpose of the model. Statistical models can be used to test causal theories (explanatory modelling), to provide accurate predictions (prediction or prognostic modelling), or to investigate associations between the outcome and predictors (descriptive modelling) (Shmueli, 2010). In analysis of health and biological data, modelling is often prognostic or descriptive. If the purpose of a model is prognostic, i.e. to provide accurate predictions of a patient outcome like disease or survival (Altman, 2000), reducing the number of necessary predictors through variable selection reduces the number of patient measurements needed to produce a prediction. Additionally, overly complex models are more likely to become over-fit to development data (data used for model training) and thus perform poorly on new patients (Steyerberg, 2008). If the model is descriptive, for example modelling is done to identify risk factors associated with a disease, then variable selection is useful precisely to identify such risk factors. In

general, variable selection leads to simpler, more parsimonious models that are easier to interpret and easier to validate than larger ones (Chowdhury & Turin, 2020).

1.1.1 Classical variable selection and penalized regression

Regression models are widely used in health research due to the interpretability of regression coefficients (Heinze et al., 2018), with some commonly used models being the linear, logistic, probit, and proportional odds regression models. A variety of different variable selection methods for regression models exist, each with their own benefits and drawbacks. Selection methods can be based on significance of regression coefficients, model information criteria, regularization, and background information (Heinze et al., 2018). For selection using significance levels, tests like the Wald or likelihood-ratio test can be used to test for inclusion of a given variable into the model, with the variable being included if its p -value is less than a predefined type-I error rate α , such as $\alpha = 0.05$. Testing for inclusion of multiple covariates successively gives either forward selection or backward elimination, depending on whether the intercept-only or full model are used initially, respectively. Stepwise forward or backward algorithms successively alternate between forward selection and backward elimination steps. Stepwise and backward or forward selection continue until no inclusion or exclusion of additional variables are significant, or some model fit criteria is met. Several issues arise when doing forward or backward selection including, but not limited to, falsely narrow confidence intervals of regression effects, p -values that are too small, and regression effect estimates that are too large in magnitude, all resulting from unaccounted-for multiple testing (Harrell, 2001). When the number of observations is too low for multivariable modelling, univariable screening can be used, where each variable is tested for association with the outcome in a univariable model and discarded if its association is non-significant. This screening, however, forgoes the selection of covariates which are significantly associated with the outcome only after controlling for other covariates (Harrell, 2001). Selection based on model information criteria, as opposed to significance-based testing, compares multiple prespecified models and chooses the model that optimizes an information criterion like the Akaike information criterion (AIC) (Akaike, 1998) or Bayesian information criterion (BIC) (Schwarz, 1978), extensions of the model likelihood that incur a penalty with increasing model complexity. For a candidate model with $k \leq p$ predictors, the AIC and BIC are given as

$$\text{AIC} = -2 \log(L(\hat{\beta}|\mathbf{y})) + 2k \tag{1.1}$$

$$\text{BIC} = \log(L(\hat{\beta}|\mathbf{y})) - \frac{1}{2}k \log(n) , \tag{1.2}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum-likelihood estimate of the k -dimensional regression vector $\boldsymbol{\beta}$, $L(\hat{\boldsymbol{\beta}}|\mathbf{y})$ the likelihood evaluated at $\hat{\boldsymbol{\beta}}$, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ the observed data, and n is the sample size. AIC is meant to be minimized while BIC is meant to be maximized. Since including more variables in a model will increase the likelihood (Schwarz, 1978), the model likelihood is not a suitable information criterion for parsimonious variable selection, unless computed for new observations. Comparing all possible subsets of variables according to AIC or BIC leads to “best subset” selection, with selection using BIC favouring a smaller model (Schwarz, 1978).

Variable selection through regularization refers to parameter estimation under a constraint placed on the size of the regression vector, often referred to as penalized regression, which shrinks the size of negligible effects to 0 while aiming to preserve the size of larger effects. Constraining the regression vector with respect to different norms yields different penalized regression models such as the LASSO (least absolute selection operator) (Tibshirani, 1996), Elastic Net (Zou & Hastie, 2005), and SCAD (smoothly clipped absolute deviation) (Fan & Li, 2001). In penalized regression the regression vector $\boldsymbol{\beta}$ is chosen that maximizes

$$\log(L(\boldsymbol{\beta}|\mathbf{y})) - \text{pen}_\lambda(\boldsymbol{\beta}) \text{ ,} \quad (1.3)$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ is the vector of regression coefficients, $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ the vector of n responses, $\log(L(\boldsymbol{\beta}|\mathbf{y}))$ the log-likelihood, and $\text{pen}_\lambda(\boldsymbol{\beta})$ a penalty function with tuning parameter λ . The LASSO corresponds to penalization with respect to the L_1 norm $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$. In this case $\text{pen}_\lambda(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_1$. The Elastic-Net penalizes with respect to a combination of the L_1 and L_2 norms, with penalty $\text{pen}_\lambda(\boldsymbol{\beta}) = \lambda(\alpha\|\boldsymbol{\beta}\|_1 + (1 - \alpha)\|\boldsymbol{\beta}\|_2)$, where $\|\boldsymbol{\beta}\|_2 = \sum_{j=1}^p \beta_j^2$ and $\alpha \in (0, 1)$. The SCAD penalty applies the LASSO penalty to small coefficients, and a less aggressive penalty to larger coefficients, with penalty given by

$$\text{pen}_\lambda(\beta_j) = \begin{cases} \lambda|\beta_j| & |\beta_j| \leq \lambda \\ \frac{-|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)} & |\beta_j| \in (\lambda, a\lambda] \\ \frac{(a+1)\lambda^2}{2} & |\beta_j| > a\lambda \end{cases}$$

Additional penalization methods include the adaptive LASSO (Zou, 2006), Spike-and-Slab LASSO (Ročková & George, 2018), and Ridge regression (Hoerl & Kennard, 1970), among others. The strength of regularized regression as a variable selection tool lies in its ability to encode sparsity into the model: the assumption that only a small fraction of all covariates are associated with the outcome, which is often the case in analyses where the number of predictors exceeds the number of observations. Lastly, background information or subject-specific knowledge is recommended by many authors (Chowdhury & Turin, 2020; Harrell, 2001;

Heinze et al., 2018) to inform variable selection. This refers to inclusion of predictors known by experts to be relevant to the outcome, or predictors that appear significant in multiple separate studies. Additionally, stability selection (Meinshausen & Bühlmann, 2010) can be used with any data-driven variable selection protocol to control for false discoveries. This involves doing variable selection on multiple subsamples of the data and selecting those variables that are identified in a large portion of the subsamples.

As mentioned earlier, variable selection becomes particularly challenging in the context of high-dimensional data, where the number of covariates p exceeds the number of observations n , whether moderately or by orders of magnitude. In high-dimensional regression however, inference and asymptotic guarantees can be made when an assumption of sparsity is made, i.e. it is assumed that very few parameters are significant (Fan & Lv, 2010). Penalized or regularized regression is precisely the variable selection and modelling tool that makes and leverages this assumption and for this reason, in addition to its connection to best subset selection and Bayesian regression, we focus mainly on regularized regression as a variable selection tool. Penalized regression can be viewed as a computationally feasible extension of best subset selection. In best subset selection, a criterion such as AIC or BIC is computed for all possible models and the model with smallest AIC or largest BIC chosen. Minimizing AIC (equation (1.1)) for a fixed k corresponds to penalized regression with $\text{pen}_\lambda(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_0 = \lambda \sum_{j=1}^k I\{\beta_j \neq 0\}$ and $\lambda = 1$, i.e. constraining $\boldsymbol{\beta}$ with respect to the L_0 norm. Maximizing BIC (equation (1.2)) corresponds to the same penalty but with $\lambda = \frac{1}{2} \log(n)$. Assessing these information criterion for all possible $k \in \{1, 2, \dots, p\}$ is not feasible for large p , but replacing the L_0 norm with the L_1 norm, as in the LASSO, turns an extensive model search problem into the simple concave optimization $\max_{\boldsymbol{\beta}} \{\log(L(\boldsymbol{\beta}|\mathbf{y})) - \lambda \|\boldsymbol{\beta}\|_1\}$ for some $\lambda > 0$. In addition to this connection with best subset selection, it has been noted (Tibshirani, 1996) that the LASSO estimate corresponds to the posterior mode of $\boldsymbol{\beta}$ where each $\beta_j \in \{\beta_1, \beta_2, \dots, \beta_p\}$ is assumed to have an independent Laplace distribution. This however is not unique to the LASSO but is a Bayesian interpretation common to all penalized likelihood methods. In general, the solution to the penalized likelihood given in (1.3) corresponds to maximizing the posterior distribution $\pi(\boldsymbol{\beta}|\mathbf{y})$ when $\pi(\boldsymbol{\beta}) = e^{-\text{pen}_\lambda(\boldsymbol{\beta})}$ (Ročková & George, 2018). Equivalently, this means the maximum a posteriori (MAP) estimate of $\boldsymbol{\beta}$ corresponds to a penalized likelihood estimate with $\text{pen}_\lambda(\boldsymbol{\beta}) = -\log(\pi(\boldsymbol{\beta}))$. Another example includes the Spike-and-Slab LASSO penalty, which is the penalty corresponding to the spike-and-slab LASSO prior (Ročková & George, 2018).

1.1.2 Bayesian regularized regression

In Bayesian modelling unknown parameters are given probability distributions referred to as priors. Inference regarding a parameter is done through analysis of the posterior distribution: the parameter distribution

conditional on observed data. For instance, in a Bayesian regression model the regression vector β is given a distribution $\pi(\beta)$, and inference regarding β is done through analysis of $\pi(\beta|\mathbf{y})$, the posterior distribution of β , where $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ is the vector of n responses. Depending on $\pi(\beta)$ and the type of regression model, $\pi(\beta|\mathbf{y})$ may or may not be known in closed-form. In the latter case, it can be estimated or summarized using a sample from $\pi(\beta|\mathbf{y})$ obtained through Markov Chain Monte Carlo (MCMC) sampling methods. An example of an MCMC sampling algorithm is the Gibbs sampler, which samples from the joint posterior by iteratively drawing from each parameter's full-conditional distribution: its distribution conditional on the data and all other parameters. If full-conditional distributions are not known then posterior sampling can be done with a Metropolis or Metropolis-Hastings algorithm, where a candidate posterior draw is proposed and accepted if its posterior probability is high relative to the previous sample. If only parameter point estimates are desired, then sampling can be avoided by finding β that maximizes the posterior distribution $\pi(\beta|\mathbf{y})$, referred to as a maximum a posteriori or MAP estimate of β .

Interestingly, all the data-driven variable selection procedures outlined earlier have similar Bayesian counterparts. When multiple nested models are compared, as in backwards elimination or forward selection, Bayes factors (Kass & Raftery, 1995) or pseudo Bayes factors like the intrinsic Bayes factor (Berger & Pericchi, 1996) can be used to select the model with a higher posterior probability. For an arbitrary prior on the space of possible models, the BIC is asymptotically equivalent to selecting the model with highest posterior probability (Schwarz, 1978), and so is the Bayesian alternative to AIC. Lastly, regularized regression can be obtained using Bayesian shrinkage priors: priors placed on β that shrink the posterior distribution of β towards 0 like the Laplace prior, which corresponds to the LASSO. In general such priors have substantial mass around 0, but enough mass elsewhere to allow non-negligible effects to escape shrinkage. Bayesian regression with shrinkage priors offer some advantages over penalized regression techniques. Firstly, a Bayesian analysis provides automatic uncertainty quantification of effect estimates through the posterior distribution. Standard errors for penalized regression models like the LASSO can be obtained by approximating the covariance matrix or bootstrapping but are considered unreliable (Cassella et al., 2010). Secondly, Bayesian models easily accommodate data with more predictors than observations since, in a Bayesian model, inference does not rely on asymptotic considerations that assume $n > p$, but instead relies on posterior distributions. Lastly, shrinkage is attained through adjustment of the prior distribution and so fits naturally in a Bayesian framework. Some disadvantages of Bayesian regression include increased computation time and the necessity of specifying prior distributions, though it should be noted that penalized regression requires specification of the penalty parameter λ meaning neither the Bayesian nor frequentist option is tuning-free.

A variety of shrinkage priors exist in the literature but come in two main flavours: discrete and continuous shrinkage priors, with each differing in the way mass is placed around 0. In the discrete case the prior is

a mixture of a distribution highly concentrated about 0, referred to as the ‘spike,’ and a more diffuse distribution centered about 0, referred to as the ‘slab.’ For this reason discrete shrinkage priors are often referred to as ‘spike-and-slab’ priors. Different specifications of the spike and slab components of the mixture prior have been proposed. The Kuo & Mallick Spike-and-Slab (Kuo & Mallick, 1998) uses a point mass at 0 spike and a centered normally distributed slab. Stochastic Search Variable Selection (George & McCulloch, 1993) adopts a normally distributed spike and slab, with a small variance for the spike and large variance for the slab. Priors may be placed on variance parameters leading to non-normal spikes or slabs (Castillo & Mismar, 2018; Ročková & George, 2018). Continuous shrinkage priors aim to emulate spike-and-slab priors with a single distribution that has a tall mode at 0 and relatively thick tails elsewhere. Continuous shrinkage priors include the Laplace prior, often referred to as the Bayesian LASSO (Park & Casella, 2008), the Horseshoe prior (C. M. Carvalho et al., 2010), Dirichlet-Laplace prior (Bhattacharya et al., 2015), and the Double-Pareto prior (Armagan et al., 2013). All these priors vary in thickness of their tails and whether shrinkage is induced locally (per covariate) and/or globally (simultaneously for all covariates). Though continuous shrinkage priors are generally more computationally efficient than spike-and-slab priors, spike-and-slab priors have distinct advantages. As will be seen in Section 2.1.5, spike-and-slab priors are expressed with latent indicator variables that specify whether the spike, or slab, is used in the model. This allows spike-and-slab priors to output posterior model inclusion probabilities for each covariate, and allows for group information to be easily encoded into the prior (Tadesse & Vannucci, 2021). Additionally, spike-and-slab priors are shown to have good theoretical properties in regression models (Castillo et al., 2015), while the same results have not been obtained for continuous shrinkage priors (Tadesse & Vannucci, 2021).

Sparsity of variable selection in Bayesian regression models with shrinkage priors is controlled by the variance of the shrinkage prior or, equivalently, how much probability mass the prior places away from 0. As such, adequate specification of the variance has been investigated by researchers for a variety of priors (C. M. Carvalho et al., 2010; Chipman et al., 2001; Fernandez et al., 2001; Narisetty & Hel, 2014; Piironen & Vehtari, 2017). One consideration to make when specifying any prior distribution is whether a given prior is reasonable in light of the data being analyzed. A prior may disagree or be in conflict with the observed data when, under that prior, the assumed model would hypothetically generate data very different from the data observed ((Evans & Moshonov, 2006)). If such a conflict is present, the validity of the assumed model and results of the analysis may be called into question. We propose specifying the variance of Bayesian shrinkage priors by minimizing conflict between a variance parameter and data. A variety of checks have been developed to identify prior-data conflict and, motivated by a distributional divergence-based check, we propose a method for configuring the variance in the Bayesian LASSO and the point mass spike-and-slab prior in linear and logistic regression models.

1.2 Scope of the thesis

This thesis is organized as follows: In the first section of Chapter 2, we review some existing penalized regression models and shrinkage priors and approaches to tuning them. In the second section of Chapter 2, we outline the notion of prior-data conflict and tests available to check for it. In the third section of Chapter 2, we propose a prior-data conflict inspired technique for treatment of the variance parameters in the Bayesian LASSO and spike-and-slab priors. In Chapter 3 we compare performance of our proposed prior elicitation against other regularized regression models through a simulation study. In Chapter 4 we apply our method to data from a multi-omics study (Overmyer et al., 2021) which contains proteomic data for individuals admitted to the Albany Medical Center with COVID-like respiratory issues. In Chapter 5 we conclude with a summary and related future work.

Chapter 2

Methods

2.1 Existing regularized regression models

We consider some existing high-dimensional regularized regression methods for linear and logistic regression models, focusing on these two for their common use in health science. For penalized methods we consider the LASSO, Elastic-Net, and the Spike-and-Slab LASSO. We consider the first two because they are likely the most well known penalized regression methods, and the last because it uses a penalty derived from a spike-and-slab prior, thus being a somewhat hybrid of Bayesian and frequentist regularization approaches. For Bayesian regularized regression we consider regression using both discrete and continuous shrinkage priors: the spike-and-slab prior and the Bayesian LASSO prior, both for their wide-spread use and the latter as it is, to our knowledge, the first continuous shrinkage prior proposed. For a sample of n independent response observations $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ and $n \times p$ matrix of predictors $\mathbf{X} = (X_1, \dots, X_p)^T$ where possibly $p > n$, the linear regression model has the form

$$\mathbf{Y} = \beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} , \tag{2.1}$$

with error variance $\epsilon_i \sim N(0, \sigma^2)$ independent for $i = 1, 2, \dots, n$ where $N(0, \sigma^2)$ refers to the normal distribution with mean 0 and variance σ^2 , and vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$. The logistic regression model has the form

$$\log \frac{P(\mathbf{Y} = \mathbf{1}_n)}{\mathbf{1}_n - P(\mathbf{Y} = \mathbf{1}_n)} = \beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} , \tag{2.2}$$

where $P(\mathbf{Y} = \mathbf{1}_n) = (P(Y_1 = 1), P(Y_2 = 1), \dots, P(Y_n = 1))^T$ refers to the vector of success probabilities of a vector of n Bernoulli distributed responses $\mathbf{Y} = (Y_1, \dots, Y_n)^T$.

2.1.1 LASSO (Least absolute shrinkage and selection operator)

The LASSO (Tibshirani, 1996) places a constraint on the L_1 norm of the regression vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$, $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$. The LASSO estimate $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \left\{ (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j| \right\}, \quad (2.3)$$

where $\tilde{\mathbf{y}}$ is the vector of centered continuous responses. The degree of regularization is controlled by λ with increasing values of λ inducing more shrinkage. Constraining with respect to the L_1 norm forces many estimated coefficients to be exactly 0 allowing for automatic variable selection by selecting only those covariates with nonzero effects. Though the LASSO produces a biased estimate of $\boldsymbol{\beta}$ (biased towards 0) (C.-H. Zhang, 2010), it does enjoy good asymptotic properties including consistency of parameter estimates and desirable prediction error. If there exists a nonsingular matrix C such that $\lim_{n \rightarrow \infty} \sum_{i=1}^n X_i X_i^T = C$, then if $\lim_{n \rightarrow \infty} \frac{\lambda_n}{n} = 0$, the LASSO estimator $\hat{\boldsymbol{\beta}}$ is consistent, meaning that it converges in probability to the true parameter vector $\boldsymbol{\beta}$ as the sample size $n \rightarrow \infty$, for a fixed p (Fu & Knight, 2000). Additionally, when λ is of order $\sqrt{\log p/n}$, the LASSO prediction error is similar to the error associated with an estimate under knowledge of the true active covariates (Bühlmann & Van De Geer, 2011). Sparsity of the estimator $\hat{\boldsymbol{\beta}}$ is ensured with *oracle inequalities*: inequalities taking the form $\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_2/n \leq \text{constant} \times \lambda^2 s_0$ with high probability, where s_0 refers to the true sparsity level, the number of active covariates (van de Geer, 2010). Consistency of variable selection as opposed to estimation relates to the model's ability to recover the true active covariates. For the LASSO, this can be viewed as sign consistency, where $\lim_{n \rightarrow \infty} P(\text{sign}(\hat{\boldsymbol{\beta}}(\lambda_n)) = \text{sign}(\boldsymbol{\beta})) = 1$, for $\lambda_n = f(n)$, where the $\text{sign}(\cdot)$ function denotes the signs (± 1 or 0) of the vector. Consistency in this regard is only guaranteed for the LASSO with a strong condition, referred to as the irrepresentable condition, on the design matrix \mathbf{X} (Zhao & Yu, 2006). In the logistic case the squared error in (2.3) is replaced with the negative log-likelihood and similar consistency and oracle properties apply (Bühlmann & Van De Geer, 2011). The tuning parameter λ is generally chosen through cross validation minimizing mean-squared error or deviance for the linear and logistic regression, respectively. This value of λ , while optimal for prediction, often results in many false-positives (F. Wang et al., 2020).

2.1.2 Elastic-Net

The Elastic-Net builds on the LASSO by penalizing with respect to both the L_1 and L_2 norms encouraging grouped variable selection, with estimates given by

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta}} \left\{ (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \lambda(\alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2) \right\}, \quad (2.4)$$

where $\alpha \in (0, 1)$ denotes the elastic net mixing parameter. In high-dimensional settings high correlation among predictors is expected, and distinguishing between significant and non-significant correlated predictors is a challenge. For two highly correlated predictors of whom one is significant, the LASSO will typically select one but not both of the predictors. Inclusion of the L_2 penalty allows the Elastic-Net to safeguard against possible selection of the incorrect covariate by allowing for selection of both (Zou & Hastie, 2005). Selection is done as in the LASSO by selecting those covariates with nonzero effect estimates. Similar to the LASSO, strong restrictions on the design matrix \mathbf{X} ensure consistency of variable selection (Jia & Yu, 2010). In a variety of scenarios the Elastic-Net outperforms the LASSO with respect to prediction error (Zou & Hastie, 2005) though it has been noted that when the LASSO fails to identify the true model it is likely the Elastic-Net fails to as well (Jia & Yu, 2010). Application to logistic regression is done by replacing the the squared error in (2.4) with the negative log-likelihood.

2.1.3 The Spike-and-Slab LASSO

Maximizing a penalized likelihood corresponds to posterior mode estimation of $\boldsymbol{\beta}$ assuming a certain prior $\pi(\boldsymbol{\beta})$. The Spike-and-Slab LASSO is the penalized regression model corresponding to the Spike-and-Slab Laplace prior on $\boldsymbol{\beta}$ given by

$$\pi(\boldsymbol{\beta}|\theta) = \prod_{j=1}^p \theta f_1(\beta_j) + (1 - \theta) f_0(\beta_j), \quad (2.5)$$

where $f_1(\beta_j) = \frac{\lambda_1}{2} e^{-\lambda_1 |\beta_j|}$, $f_0(\beta_j) = \frac{\lambda_0}{2} e^{-\lambda_0 |\beta_j|}$ are Laplace densities with mean 0 and small and large variances, respectively, and $\theta \in (0, 1)$ (Ročková & George, 2018). The Spike-and-Slab LASSO estimate is given by

$$\hat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}} \left\{ -\frac{1}{2} (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^T (\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) + \text{pen}(\boldsymbol{\beta}|\theta) \right\}, \quad \text{where} \quad (2.6)$$

$$\begin{aligned} \text{pen}(\boldsymbol{\beta}|\theta) &= -\lambda_1 |\boldsymbol{\beta}| + \sum_{j=1}^p \log \frac{p_{\theta}^*(0)}{p_{\theta}^*(\beta_j)} \\ p_{\theta}^*(\beta_j) &= \frac{\theta f_1(\beta_j)}{\theta f_1(\beta_j) + (1 - \theta) f_0(\beta_j)} \end{aligned} \quad (2.7)$$

As the sparsity level θ usually is unknown, a prior $\pi(\theta)$ may be specified, leading to the non-separable Spike-and-Slab LASSO penalty

$$\text{pen}_{NS}(\boldsymbol{\beta}|\theta) = -\lambda_1|\boldsymbol{\beta}| + \log \left(\int \frac{\theta^p}{\prod_{j=1}^p p_{\theta}^*(\beta_j)} d\pi(\theta) \Big/ \int \frac{\theta^p}{p_{\theta}^*(0)^p} d\pi(\theta) \right) \quad (2.8)$$

Model selection is carried out by fixing $\lambda_1 = 1$ (corresponding to a slab variance of 2), and successively estimating $\boldsymbol{\beta}$ for an increasingly aggressive spike. The increasingly aggressive spike serves to filter out negligible effects, forcing them to 0, and preserving the effects of relevant predictors. The final model is obtained by selecting those covariates with nonzero effects after successive estimates of $\boldsymbol{\beta}$ no longer change with the change in spike, indicating stabilization of the solution-path (Ročková & George, 2018). The Spike-and-Slab LASSO is shown to have variable selection performance that mimics performance of a penalty that knows the true number of active covariates through similar oracle inequalities as in the LASSO. Additionally, the Spike-and-Slab LASSO prior has near optimal posterior concentration, which refers to an asymptotically vanishing expected posterior probability of the discrepancy between $\boldsymbol{\beta}$ and the true regression vector (Ročková & George, 2018).

2.1.4 The Bayesian LASSO

The linear LASSO estimate given in (2.3) corresponds to the posterior mode of $\boldsymbol{\beta}$ when β_1, \dots, β_p are assumed to have independent Laplace distributions with location 0 and scale λ . The Bayesian LASSO proposed by Park and Casella, 2008 is the corresponding Bayesian linear regression model with a Laplace prior on $\boldsymbol{\beta}|\sigma^2$. Conditioning on the error variance σ^2 ensures that the posterior distribution of $\boldsymbol{\beta}$ is unimodal. The Laplace density is a scale-mixture of normal distributions with exponential mixing density, meaning it may be expressed as

$$\begin{aligned} \beta_j | \tau_j^2, \sigma^2 &\sim N(0, \tau_j^2 \sigma^2) \\ \tau_j^2 &\sim \text{Exponential}(\lambda^2/2) \end{aligned} \quad (2.9)$$

independent for each $j = 1, 2, \dots, p$. Equivalently, after integrating out τ_j^2 from (2.9), $\beta_j|\sigma^2$ is Laplace distributed with variance $\frac{2\sigma^2}{\lambda^2}$. Larger values of λ correspond to a smaller prior variance on $\beta_j|\sigma^2$ and hence induce more shrinkage of regression effects towards 0. Variable selection is done through analysis of the posterior distribution $\pi(\boldsymbol{\beta}|\mathbf{y})$. This posterior is not known in closed form but can be sampled from using MCMC methods. In the linear case, due to the hierarchical form of the Laplace prior given in (2.9), the full-conditional distributions $\pi(\boldsymbol{\beta}|\boldsymbol{\tau}^2, \sigma^2, \lambda, \mathbf{y})$, $\pi(\boldsymbol{\tau}^2|\boldsymbol{\beta}, \sigma^2, \lambda, \mathbf{y})$ and $\pi(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\tau}^2, \lambda, \mathbf{y})$ are all known in closed-form, so a Gibbs sampler can be used, where a sample from $\pi(\boldsymbol{\beta}, \boldsymbol{\tau}^2, \sigma^2|\lambda, \mathbf{y})$ is drawn by iteratively sampling from

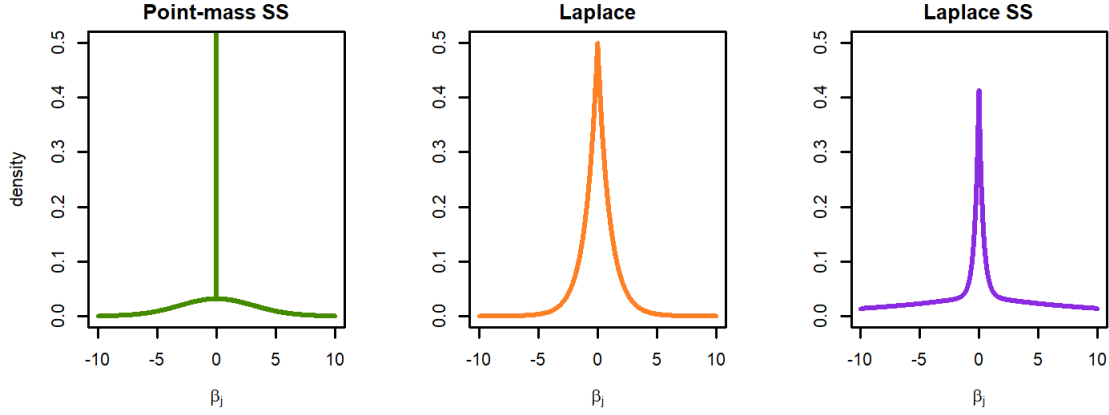
the full-conditional distributions of $\boldsymbol{\beta}$, $\boldsymbol{\tau}^2$, and σ^2 . The Bayesian LASSO for logistic regression is the same as in the linear case except that the Laplace distribution on each $\beta_1, \beta_2, \dots, \beta_p$ is not conditioned on any σ^2 . In the logistic case the full-conditional distributions are not known in closed-form so a Metropolis-Hastings algorithm may be used. The posterior sample of $\boldsymbol{\beta}$ can be used to construct a credible interval for each β_j , $j = 1, 2, \dots, p$, or to compute a posterior mode, median, or mean for each β_j . Slopes whose posterior credible interval exclude 0 or whose absolute mode, median, or mean exceed some user-defined threshold are selected for inclusion in the final model.

For the frequentist LASSO, λ is generally specified using cross validation and similar methods for the Bayesian LASSO are applicable if λ is assumed fixed. For a candidate λ_0 , the model may be fit for one or more data splits and an average prediction error computed, with the λ_0 that gives the smallest prediction error chosen. A λ value can also be chosen that maximizes a cross-validated log-likelihood (Genking et al., 2007). In either case, posterior means or modes of regression coefficients are used to generate predictions or compute log-likelihoods for the test data. Another approach is to compute the marginal maximum likelihood estimate of λ using a Monte-Carlo Expectation-Maximization (MC-EM) algorithm as proposed in Casella, 2001; Park and Casella, 2008. The MC-EM algorithm for maximizing the marginal likelihood $L(\lambda|\tilde{\mathbf{y}})$ iteratively updates the value of λ using $\lambda^{(k+1)} = \sqrt{2p/\sum_{j=1}^p \widehat{E}_{\lambda^{(k)}}(\tau_j^2|\tilde{\mathbf{y}})}$ until some convergence criterion is met, where $\widehat{E}_{\lambda^{(k)}}(\tau_j^2|\tilde{\mathbf{y}})$ is the posterior sample mean of τ_j^2 obtained from a run of the Gibbs sampler. The MC-EM algorithm requires repeatedly fitting the data, possibly hundreds of times, until convergence, making it computationally expensive. Atchadé, 2011 proposes a stochastic approximation method for estimating the MLE of λ with only a single run of the Gibbs sampler, but the algorithm requires defining a sequence of step sizes which is itself an important problem of stochastic approximation (Leng et al., 2014). Methodology based on efficient importance sampling schemes can also be used to estimate the MLE of λ (Vivekananda & Chakraborty, 2017). Genking et al., 2007 use a norm-based value of λ for the logistic Bayesian LASSO inspired by the regularization parameter in a Support Vector Machine model. Lastly, λ may be informed by previous analyses and fixed at a value known to be suitable for the data at hand.

If λ is not fixed, a Gamma(a, b) prior may be placed on λ^2 for which a variety of methods exist for specifying a, b . The gamma prior is placed on λ^2 as opposed to λ to obtain conjugacy of its full-conditional distribution (Park & Casella, 2008). A semi-Bayesian approach includes choosing hyperparameters that place sufficient mass around the MLE of λ^2 and have little mass as λ^2 becomes very large (Park & Casella, 2008). Alternatively, the shape parameter a may be fixed and rate b estimated using the MC-EM algorithm (Leng et al., 2014). Camli et al., 2022 proposes giving the parameters a, b flat priors, computing their posterior modes \hat{a}, \hat{b} , and using the prior $\lambda^2 \sim \text{Gamma}(\hat{a}, \hat{b})$ for analysis. The hyperparameters can also be chosen with cross-validation as described in the simulation results of Huang et al., 2013. If a fully Bayesian approach is

taken, where no part of the prior on λ^2 is estimated from the data, a relatively flat prior may be placed on λ^2 , or a, b can be chosen so that the regression coefficients are within a range known to be suitable to the data (Biswas & Lin, 2012).

Figure 2.1: The point-mass spike-and-slab prior (left) with $\theta = 0.25$, $\tau^2 = 10$, the Laplace prior (used the Bayesian LASSO model, center) with $\lambda = 1$, and the Laplace spike-and-slab prior (used in the Spike-and-Slab LASSO model, right) with $\theta = 0.25$, $\lambda_0 = 3$, and $\lambda_1 = 1$



2.1.5 The Spike-and-Slab Prior

The spike-and-slab prior is a mixture between a distribution highly concentrated at 0, and a more diffuse distribution about 0. For the regression models (2.1) or (2.2), the point-mass spike-and-slab prior (Kuo & Mallick, 1998) uses a point-mass at 0 spike and a centered normally distributed slab:

$$\begin{aligned} \beta_j | \gamma_j &\sim (1 - \gamma_j)\delta_0(\beta_j) + \gamma_j N(0, \tau^2) \\ \gamma_j &\sim \text{Bernoulli}(\theta) \end{aligned} \tag{2.10}$$

independent for each $j = 1, 2, \dots, p$, where δ_0 denotes the Dirac function at 0. The point-mass spike-and-slab allows selection of covariates based on their how well they can be distinguished from 0. Since some covariates may have negligible nonzero effects, selection based on a variable's meaningful effect size can be facilitated by replacing the point-mass at 0 with a narrow, normally distributed spike at 0. Selection using this prior gives Stochastic Search Variable Selection (SSVS) (George & McCulloch, 1993),

$$\begin{aligned} \beta_j | \gamma_j &\sim (1 - \gamma_j)N(0, \tau_0^2) + \gamma_j N(0, \tau_1^2) \\ \gamma_j &\sim \text{Bernoulli}(\theta) \end{aligned} \tag{2.11}$$

independent for each $j = 1, 2, \dots, p$, where $\tau_0^2 < \tau_1^2$. While the prior used in the Bayesian LASSO allows for individual shrinkage of regression effects through local variance parameters, the spike-and-slab priors considered here perform simultaneous shrinkage of regression effects through the global variance parameter τ^2 in the point-mass spike-and-slab, and through τ_0^2, τ_1^2 in SSVS. Increasing values of τ^2 or τ_1^2 correspond to increasing shrinkage of regression effects (Chipman et al., 2001). Sparsity is additionally controlled by θ , which can be viewed as the prior proportion of active covariates. In the absence of prior information regarding the level of sparsity, θ may be given a Beta(1, 1) prior which is a uniform distribution over (0, 1). Variable selection is done through analysis of the posterior distribution $\pi(\boldsymbol{\gamma}|\mathbf{y})$. For some $j \in \{1, 2, \dots, p\}$ and sample $\{\gamma_j^{(1)}, \gamma_j^{(2)}, \dots, \gamma_j^{(S)}\}$ of size S drawn from the posterior distribution $\pi(\gamma_j|\mathbf{y})$, the estimate of the posterior mean $\widehat{E}(\gamma_j|\mathbf{y}) = \bar{\gamma}_j = \frac{1}{S} \sum_{k=1}^S \gamma_j^{(k)}$ gives the proportion of times the j^{th} variable was included in the model throughout the sampling process. $\bar{\gamma}_j$ is referred to as the j^{th} variable's posterior inclusion probability. Median probability model selection (Maddalena & Berger, 2004) refers to selecting covariates whose posterior inclusion probability exceeds 0.5. A subset of predictors can also be selected based on which subset appears most often in the sampling process, referred to as the highest posterior probability model. A minimum of 2^p posterior samples would have to be drawn for each submodel to be visited once, but it is presumed that those models not visited often or at all in the sampling process have a low posterior probability and are not of interest (George & McCulloch, 1993). Aside from being easier to identify in high-dimensions, Barbieri and Berger, 2004 show that the median probability model is often optimal for prediction.

Different recommendations exist for treatment of the variance parameters in spike-and-slab models. In SSVS, individual variance parameters must be specified for both the spike and slab components and can be specified based on thresholds of practical significance (George & McCulloch, 1997). Correlation among regression effects may also be accounted for by using a $N_p(\mathbf{0}, c\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ distributed slab, referred to as Zellner's g-prior (Zellner, 1986), with $c > 0$. Chipman et al., 2001 recommend choosing c large enough that the slab is relatively flat over plausible values of $\boldsymbol{\beta}$, with Fernandez et al., 2001 recommending $c = \max\{p^2, n\}$. In simulations O'hara and Sillanpää, 2009 considered slab variances as small as 1 and as large as 10^9 , though note instability of parameters estimates for large slab variances when using the point-mass spike-and-slab prior. For the linear spike-and-slab model with $N_p(\mathbf{0}, c\sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$ distributed slab, it is possible to obtain marginal maximum-likelihood estimates of c and the mixing weight θ provided the number of covariates is not large, as it involves computation of the maximum-likelihood estimate of $\boldsymbol{\beta}$ for all 2^p possible submodels (Chipman et al., 2001). However, if \mathbf{X} is orthogonal, the marginal likelihood of (c, θ) simplifies and can be maximized with numerical methods even when p is moderately large (George & Foster, 2000).

The slab variance may not be fixed and given a prior distribution. In the case of SSVS, an exponentially distributed spike variance and slab variance yield the Spike-and-Slab LASSO (Ročková & George, 2018)

model. An Inverse-Gamma(a, b) prior may also be placed on the slab variance giving a scaled- t distributed slab (Ročková et al., 2012), with small values like $a = b = 0.1$ supporting a wide range of prior slab variances. For a Bayesian sparse group selection model using spike-and-slab priors, a Gamma($\frac{m_g+1}{2}, \frac{\eta^2}{2}$) prior is placed on the group-specific slab variance and η estimated with an MC-EM algorithm, where m_g denotes the size of group $g = 1, 2, \dots, G$ (Xu & Ghosh, 2015). As well, Uniform(a, b) distributed slab variances may be considered, as recommended in Gelman, 2006, with O’hara and Sillanpää, 2009 noting improved mixing when compared to a fixed slab variance.

2.2 Prior-data conflict

Bayesian models have two main components: the sampling model, also referred to as the likelihood, and the prior. Let $f(\mathbf{y}|\theta)$ denote the sampling model with parameter vector θ , and let $\pi(\theta)$ be its prior distribution. The sampling model assumes how the data might be generated, while the prior specifies distributions for the unknown parameters in the data-generating model. The choice of prior may be influenced by computational convenience as well as whether or not the analyst has prior knowledge they wish to include in the model. Regardless of what drives the choice of a prior, if the sampling model under a given prior gives rise to data that are very different than the observed data, the validity of the inferences made using that prior may be called into question (Evans & Moshonov, 2006). When this occurs there is said to be prior-data conflict.

Many methods have been developed to check for prior-data conflict. The basis for a number of these checks involves assessing the discrepancy between observed data and data generated by the sampling model $f(\mathbf{y}|\theta)$ assuming $\theta \sim \pi(\theta)$ (Nott et al., 2020). Data generated by the sampling model under the prior $\pi(\theta)$ are data drawn from the prior predictive distribution $m(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta$. To check for prior-data conflict, Evans and Moshonov, 2006 suggest comparing an observed minimal sufficient statistic for θ , $T_0 = T(\mathbf{y}_0)$, to its prior-predictive density, $m(T)$, to assess if T_0 is a surprising value from this distribution. They suggest doing so by computing the p -value $P(m(T(\mathbf{y}_0)) > m(T))$. If the p -value is small this indicates that T_0 is in the tails of $m(T)$ and hence the data observed is surprising assuming our prior, $\pi(\theta)$. A similar check, proposed primarily for use in regression models, builds on this previous check by comparing a sufficient statistic for a given regression coefficient to a pseudo-prior predictive distribution of that estimator, allowing for component-wise assessment of prior-data conflict (Leonardo Egidi & Torelli, 2022). A score-type statistic has also been proposed to check for prior-data conflict (Nott et al., 2021), as well as a prior-to-posterior divergence statistic (Nott et al., 2020). In both tests, the observed value of the statistic is compared to its prior-predictive distribution to assess whether it is a surprising value from this distribution. Another divergence-based prior-data conflict check has been proposed, but requires specification of a non-informative

prior (Bousquet, 2008).

The divergence-based check proposed in Nott et al., 2020 uses the prior-to-posterior Rényi divergence as a statistic, of which the Kullback-Leibler (KL) divergence is a special case. The prior-to-posterior Kullback-Leibler divergence for parameter θ is defined as

$$\text{KL}(\mathbf{y}) = \text{KL}(\pi(\theta|\mathbf{y}) || \pi(\theta)) = \int \log \frac{\pi(\theta|\mathbf{y})}{\pi(\theta)} \pi(\theta|\mathbf{y}) d\theta \quad (2.12)$$

For observed data \mathbf{y}_0 , the check computes the p -value

$$P(\text{KL}(\mathbf{Y}) > \text{KL}(\mathbf{y}_0)) , \quad (2.13)$$

where $\mathbf{Y} \sim m(\mathbf{y}) = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta$, the prior-predictive distribution. If the p -value is small this indicates that $\text{KL}(\mathbf{y}_0)$ is in the tails of the distribution of $\text{KL}(\mathbf{Y})$. This check is attractive in that it is intuitive; if a prior were to hardly change from prior to posterior, this would indicate that whatever information the data has regarding that parameter is already assumed in the prior. If the change from prior to posterior is very large, the prior is in some sense specified contrary to the data indicating a possible prior-data conflict (Nott et al., 2020).

Example 2.1: We illustrate the divergence-based check given by (2.13) with a simple example taken from Evans and Moshonov, 2006. Suppose \mathbf{y}_0 is a sample of size n drawn from $N(\mu, 1)$ and assume the prior $\mu \sim N(\mu_0, \sigma_0^2)$. We have that the posterior distribution of μ is $N(\alpha^2\rho, \alpha^2)$ where $\alpha^2 = (\frac{1}{\sigma_0^2} + n)^{-1}$ and $\rho = (\frac{\mu_0}{\sigma_0^2} + n\bar{y}_0)$ where \bar{y}_0 is the sample mean. Using the closed-form expression of the Kullback-Leibler divergence between two normal distributions (Gil et al., 2013) we have that

$$\text{KL}(\mathbf{y}_0) = \text{KL}(\pi(\mu|\mathbf{y}_0) || \pi(\mu)) = \frac{1}{2\sigma_0^2} \left[(\alpha^2\rho - \mu_0)^2 + \alpha^2 - 1 \right] + \log\left(\frac{1}{\alpha}\right) ,$$

where π is the prior distribution of μ . Only ρ depends on \mathbf{y}_0 and since

$$(\alpha^2\rho - \mu_0)^2 = (\alpha^2(\rho - \mu_0/\alpha^2))^2 = (\alpha^2n(\bar{y}_0 - \mu_0))^2$$

the p -value $P(\text{KL}(\mathbf{Y}) > \text{KL}(\mathbf{y}_0))$ simplifies to $P((\bar{Y} - \mu_0)^2 > (\bar{y}_0 - \mu_0)^2)$. Note that $\bar{Y} = \mu + Z$ where $Z \sim N(0, 1/n)$ is independent of μ (Evans & Moshonov, 2006) and so $\bar{Y} \sim N(\mu_0, \sigma_0^2 + 1/n)$. We then have that

$$P(\text{KL}(\mathbf{Y}) > \text{KL}(\mathbf{y}_0)) = P((\bar{Y} - \mu_0)^2 > (\bar{y}_0 - \mu_0)^2) = P\left(\chi^2 > \frac{(\bar{y}_0 - \mu_0)^2}{\sigma_0^2 + 1/n}\right)$$

where χ^2 is chi-squared distributed with 1 degree of freedom. Suppose we observe $\bar{y}_0 = 2.83$ from a sample

of size $n = 10$ and want to use the prior $\mu \sim N(5, 1)$. The prior-data conflict p -value is then $P(\chi^2 > \frac{(\bar{y}_0 - \mu_0)^2}{\sigma_0^2 + 1/n}) = P(\chi^2 > 4.30) = 0.04$, meaning that if a significance level of 0.05 is used, there is evidence of prior-data conflict. However, if we increase the spread of $\pi(\mu)$ by increasing its variance to 4, we get a p -value of 0.28, meaning that such a prior on μ would not conflict with the data.

As is often the case in more complex models, the posterior density $\pi(\theta|\mathbf{y})$ and/or the prior-predictive distribution $m(\mathbf{y})$ are not known in closed-form, making computation of the p -value (2.13) challenging. When $m(\mathbf{y})$ is not known, the distribution of $\text{KL}(\mathbf{Y})$ must be simulated by computing values of $\text{KL}(\mathbf{y}^{(i)})$ for a series of draws $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(B)}$ from $m(\mathbf{y})$. If $\pi(\theta|\mathbf{y})$ is known then computation of $\text{KL}(\mathbf{y}^{(i)})$ is straightforward. If $\pi(\theta|\mathbf{y})$ is not known, then $\text{KL}(\mathbf{y}^{(i)})$ must be estimated with $\tilde{\text{KL}}(\mathbf{y}^{(i)}) = \text{KL}(\tilde{\pi}(\theta|\mathbf{y}^{(i)}) || \pi(\theta))$, where $\tilde{\pi}(\theta|\mathbf{y}^{(i)})$ is a density that estimates $\pi(\theta|\mathbf{y}^{(i)})$ obtained through a sample from $\pi(\theta|\mathbf{y}^{(i)})$.

Regardless of the check being used, if there is evidence of prior-data conflict for some prior $\theta \sim \pi(\theta)$, the question of what to do about it arises. Though not entirely Bayesian, new hyperparameters for $\pi(\theta)$ based on the data could be used which would likely resolve the issue. However, if the prior $\pi(\theta)$ contains information that the analyst would like to include in the model, like a prior mean for θ , then discarding this information in order to avoid conflict is not appealing. Instead, as in Example 2.1, a more conservative prior, referred to as being weakly informative relative to π , may be used. Such a prior, say π_0 , avoids conflict not by abandoning available prior knowledge but by inputting a reduced amount of it into the model. Evans and Jang, 2011 outline a definition of weak informativity and provide methods for identifying a weakly informative prior when the original prior conflicts with the data. Lastly, it is also possible to proceed with a conflicting prior if it is not desirable to allow the data to have any bearing on the choice of prior. Though an opinion held by some pure Bayesians, Evans and Moshonov, 2006 argue that is it at least informative to check for prior-data conflict, particularly if results of the analysis are to be used by others.

2.3 Minimum-divergence prior expectation

The motivation behind prior-data conflict checks is to ensure that an elicited prior is sound in light of observed data. Our use of prior-data conflict is instead for the purpose of finding a prior that might conflict least with the data, in hopes that such a prior improves variable selection performance of the model, i.e. the models ability to identify relevant predictors. This is in line with empirical Bayes methodology, where the data are used to estimate certain hyperparameters in the model and so is not strictly Bayesian. For hierarchical models like the Bayesian LASSO and regression with spike-and-slab priors, methods exist for finding marginal maximum-likelihood estimates of the tuning parameters but these methods often involve MC-EM estimation which requires many repeated fits of the data. Depending on the initial values proposed,

hundreds or possibly thousands of repeated fits of the data may be required until convergence of the algorithm. Our thinking is that a good estimate of the shrinkage parameter in these priors may improve the models ability to distinguish between relevant and irrelevant predictors when compared to a non-informative prior placed on the shrinkage parameter. This is because, depending on how strongly identified the shrinkage parameter is in the data, a non-informative prior may result in no clearly determined level of shrinkage. Indeed, Cui and George, 2008 show that an empirical Bayes treatment of the the point-mass spike-and-slab with Zellner’s g-prior distributed slab outperforms the fully Bayes treatment, with Nott et al., 2007 showing promising results for an empirical Bayes variable selection method in micorarray experiments.

Suppose for the likelihood $f(\mathbf{y}|\theta)$ and prior $\pi(\theta|\alpha)$ with hyperparameter α , we wish to specify a value $\alpha = \alpha_0$ that conflicts least with the data using the divergence-based check outlined in the previous subsection. Asymptotically, the p -value (2.13) is a measure of how far out in the tails of $\pi(\theta)$ the true value of θ lies (Nott et al., 2020), so a point of least conflict could be seen as the α that maximizes this p -value. If $m(\mathbf{y})$, $\pi(\theta|\mathbf{y}, \alpha)$, $\text{KL}(\pi(\theta|\mathbf{y}, \alpha) || \pi(\theta|\alpha))$ are all tractable, then the function $g(\alpha) = P(\text{KL}(\mathbf{Y}|\alpha) > \text{KL}(\mathbf{y}_0|\alpha))$ could be maximized with respect to α directly. If some or none of those quantities are tractable then $g(\alpha)$ could be estimated by evaluating the p -value (2.13) at possible values of α .

In the Bayesian LASSO the only tuning parameter is λ so we aim to find a value or prior on λ that conflicts least with the data. In the point-mass spike-and-slab prior however, regularization is controlled by the slab variance τ^2 in addition to the prior inclusion probability θ . We focus only on configuring τ^2 for consistency with the Bayesian LASSO as well as because an empirical Bayes estimate of θ can be obtained by numerically optimizing the marginal likelihood of θ (Castillo & Mismar, 2018). Suppose we have either the linear or logistic regression model (2.1) or (2.2). Using the check given in (2.13), if $\boldsymbol{\beta}$ follows the Laplace prior defined in (2.9), finding a minimally conflicting fixed value of λ involves assessing conflict between either $\boldsymbol{\tau}^2 = (\tau_1^2, \tau_2^2, \dots, \tau_p^2)^T$ and the data, or between each τ_j^2 , $j = 1, 2, \dots, p$ and the data. In the former case a p -dimensional posterior density for $\boldsymbol{\tau}^2$ would have to be estimated which for large p is not feasible, or in the latter case conflict would be assessed per-covariate in which case a method for determining the overall conflict associated with a given value of λ would have to be proposed. For the point-mass spike-and-slab prior the same issue would arise except at the level of regression coefficients $\boldsymbol{\beta}$.

To avoid these issues we propose placing a prior on the parameter λ^2 in the Bayesian LASSO, and a prior on the parameter τ^2 in the point-mass spike-and-slab, fixing each of their prior variances to some small value, and identifying prior means that may conflict least with the data. We assume

$$\lambda^2 \sim \text{Gamma}(\mu_{\lambda^2}, \sigma_{\lambda^2}^2) \tag{2.14}$$

$$\tau^2 \sim \text{Inverse-Gamma}(\mu_{\tau^2}, \sigma_{\tau^2}^2) , \quad (2.15)$$

where $\mu_{\lambda^2}, \sigma_{\lambda^2}^2$ and $\mu_{\tau^2}, \sigma_{\tau^2}^2$ are the prior mean and variance of λ^2 and τ^2 , respectively. Ideally, the p -value $P(\text{KL}(\mathbf{Y}|\mu) > \text{KL}(\mathbf{y}_0|\mu))$ could be maximized with respect to μ , the prior mean of either $\pi(\lambda^2)$ or $\pi(\tau^2)$. As the quantities $\text{KL}(\mathbf{y}_0|\mu)$ and $\text{KL}(\mathbf{Y}|\mu)$ are not known in closed form, maximizing with respect to μ directly is not feasible for either model. This means in order to identify a μ that conflicts the least with the data, a sequence of prior means $\{\mu_1, \mu_2, \dots, \mu_l\}$ must be specified and a measure of conflict computed for each μ_i , $i = 1, 2, \dots, l$, where l is the length of the sequence. Computing the p -value (2.13) for just one μ_i requires simulation of the distribution of $\text{KL}(\mathbf{Y}|\mu_i)$, which requires fitting many draws from the prior predictive distribution $m(\mathbf{y})$. We decide instead to measure the conflict of each μ_i with only the test statistic of the p -value (2.13), the prior to posterior divergence. So for prior means $\mu_{\lambda^2}, \mu_{\tau^2}$ we compute

$$\text{KL}(\pi(\lambda^2|\mathbf{y}, \mu_{\lambda^2}) || \pi(\lambda^2|\mu_{\lambda^2})) \quad (2.16)$$

$$\text{KL}(\pi(\tau^2|\mathbf{y}, \mu_{\tau^2}) || \pi(\tau^2|\mu_{\tau^2})) \quad (2.17)$$

Our justification for computing only the test statistic for the divergence-based check as opposed to the p -value is similar to what motivates the check itself: given two candidate prior means, the prior mean that results in a larger prior-to-posterior divergence may be more likely to conflict with the data than the other.

Since the posterior distributions $\pi(\lambda^2|\mathbf{y})$, $\pi(\tau^2|\mathbf{y})$, and the prior-predictive distribution $m(\mathbf{y})$ are not known in closed form, the divergences (2.16), (2.17) must be estimated. We first estimate the posterior distributions $\pi(\lambda^2|\mathbf{y}, \mu_{\lambda^2})$, $\pi(\tau^2|\mathbf{y}, \mu_{\tau^2})$ with standard normal kernel density estimates $\tilde{\pi}(\lambda^2|\mathbf{y}, \mu_{\lambda^2})$, $\tilde{\pi}(\tau^2|\mathbf{y}, \mu_{\tau^2})$ computed with posterior samples using the function `density()` in R version 4.2.0, and then estimate (2.16), (2.17) with the estimated divergence between $\pi(\lambda^2|\mu_{\lambda^2})$ and $\tilde{\pi}(\lambda^2|\mathbf{y}, \mu_{\lambda^2})$, and between $\pi(\tau^2|\mu_{\tau^2})$ and $\tilde{\pi}(\tau^2|\mathbf{y}, \mu_{\tau^2})$. For a sample $\{x_1, x_2, \dots, x_m\}$ of size m from some distribution $f(x)$, the standard normal kernel density estimate of f is:

$$\tilde{f}(x) = \frac{1}{m} \sum_{i=1}^m \phi\left(\frac{x - x_i}{h}\right)$$

where $\phi(x)$ is the standard normal density and h a tuning parameter referred to as the bandwidth (Chen, 2017). The bandwidth can be specified using Silverman's rule of thumb, $h = 0.9 * \min\{\hat{\sigma}, \frac{\text{IQR}}{1.34}\}n^{-1/5}$, where $\hat{\sigma}$ is the sample standard deviation and IQR the interquartile range (Sheather, 2004). The divergence between prior and kernel-estimated posterior is done analytically through estimation of the integral (2.12). Other methods for estimating divergences include direct divergence estimation (Sugiyama et al., 2013), and estimation using minimal spanning trees (Hero & Michel, 1999). Direct divergence estimation is developed for

estimating the divergence between two samples from unknown distributions as an alternative to computing the divergence of estimates of their densities. Since, in our case, one of the densities is known (the prior), direct divergence estimation would not make use of this information. Estimation using minimal spanning trees estimates the divergence between a known reference density and a sample from an unknown distribution but are more computationally costly than our divergence estimation and useful particularly for evaluation of multidimensional divergences (Hero & Michel, 1999).

An issue arises when we observe that the divergences (2.16), (2.17) tend to sharply decrease and stabilize once the prior means μ_{λ^2} , μ_{τ^2} become large enough, meaning that for certain sets of candidate means it is likely that the largest mean will correspond to the smallest divergence. We note however that before the divergence sharply decreases a local minimum can usually be observed. As such for some sequence of prior means $\{\mu_1, \mu_2, \dots, \mu_l\}$ for either λ^2 or τ^2 , we denote the minimum-divergence prior means $\mu_{\lambda^2}^{MD}$, $\mu_{\tau^2}^{MD}$ to be the prior means corresponding to the first local minimum of the prior-to-posterior divergence (2.16) or (2.17). Let the minimum-divergence linear Bayesian LASSO (MD-LinBLASSO) and minimum-divergence logistic Bayesian LASSO (MD-LogBLASSO) denote the linear and logistic Bayesian LASSO using the Gamma prior in (2.14) on λ^2 with minimum-divergence prior mean $\mu_{\lambda^2}^{MD}$ and $\sigma_{\lambda^2}^2$ fixed at some small value like 1 or 10. Let the minimum-divergence linear spike-and-slab model (MD-LinSS) and minimum-divergence logistic spike-and-slab model (MD-LogSS) denote the linear and logistic point-mass spike-and-slab model using the Inverse-Gamma prior in (2.15) on τ^2 with minimum-divergence prior mean $\mu_{\tau^2}^{MD}$, where $\sigma_{\tau^2}^2$ is fixed at some reasonably small value like 1 or 10.

2.4 Summary

In this chapter we review regularized regression models like the LASSO, Elastic-Net, Spike-and-Slab LASSO, the Bayesian LASSO, and regression with spike-and-slab priors. Adequate performance of these models rests on proper configuration of the tuning parameters that control the degree of regularization. For penalized methods, tuning parameters can be specified that minimize cross-validated mean-squared error or, in the logistic case, deviance. For the Bayesian LASSO a variety of treatments of the tuning parameter λ are available, including estimation from data with an MC-EM or stochastic approximation algorithm, cross-validation, use of a non-informative prior, or a value of λ known to suit the data. Depending on the type of spike-and-slab prior used, methods including thresholds of practical significance, cross-validation, direct estimation from data (when p is small), priors on the spike or slab variance leading to non-normal spikes and slabs, estimation via MC-EM algorithm, and a variety of general recommendations can be applied to configure a spike-and-slab prior. To assess the suitability of a given prior in light of the data, checks referred

to as prior-data conflict checks have been proposed. Prior-data conflict occurs when prior information regarding a parameter is in disagreement with the data-information regarding a parameter. We propose a novel method for tuning the Bayesian LASSO and point-mass spike-and-slab prior that aims to minimize prior-data conflict associated with their squared tuning parameters λ^2 and τ^2 , respectively. To do so we place Gamma and Inverse-Gamma priors on λ^2 and τ^2 respectively, and, motivated by a prior-to-posterior divergence-based check for prior-data conflict (Nott et al., 2020), propose identifying prior means for those parameters that result in minimal prior-to-posterior divergence, for some small fixed prior variance.

In the next chapter we conduct a simulation study across multiple high-dimensional scenarios to assess variable selection performance of our proposed minimum-divergence configurations of the Bayesian LASSO and spike-and-slab priors.

Chapter 3

Simulation

In this chapter we conduct simulation studies to assess variable selection performance of the MD-LinBLASSO, MD-LinSS, and MD-LogSS against variable selection methods in both linear and logistic regression models. We do not consider the MD-LogBLASSO model as preliminary results show the joint Laplace prior on regression coefficients β induces too much shrinkage in the logistic regression model, resulting in no selection of covariates. We assess performance of the different variable selection methods for multiple sparsity levels, number of predictors, and effect sizes. We consider an uncorrelated as well as correlated predictor design, using real proteomic data from Overmyer et al., 2021 to simulate responses for the correlated design. The performance metrics include the true positive rate (TPR), false discovery rate (FDR), false nondiscovery rate (FNDR), and the F_1 score. Section 3.2.1 outlines the competing models considered, Section 3.1.2 outlines the different simulations scenarios, Section 3.1.3 provides definitions of the performance metrics, Section 3.1.4 describes how all methods are implemented, and Sections 3.2 and 3.3 present and discuss the simulation results.

3.1 Design of the simulation

3.1.1 Competing Models

As the motivation of the minimum-divergence priors is to specify least-conflicting prior distributions of the variance parameters λ^2 , τ^2 in the Bayesian LASSO and spike-and-slab prior, respectively, we compare variable selection performance of the minimum-divergence models against the Bayesian LASSO and spike-and-slab prior when λ^2 and τ^2 are given non-informative priors. We view a somewhat non-informative prior on λ^2 to be a Gamma distribution with a large variance like 10^3 , with a prior mean of 10 or 100, as

smaller prior means tended to lead to mixing issues. Variable selection results differed minimally between the prior means 10 and 100 so we set $\lambda^2 \sim \text{Gamma}(\alpha = 10, \beta = 0.1)$ as a non-informative prior for the Bayesian LASSO, corresponding to $E(\lambda^2) = 100$ and $\text{Var}(\lambda^2) = 10^3$. For the non-informative spike-and-slab prior we assume $\tau \sim \text{Uniform}(0, 100)$, as recommended in Gelman, 2006. Let NI-LinBLASSO denote the linear Bayesian LASSO with non-informative prior $\lambda^2 \sim \text{Gamma}(\alpha = 10, \beta = 0.1)$, and let NI-LinSS and NI-LogSS denote the linear and logistic spike-and-slab models, respectively, with non-informative prior $\tau \sim \text{Uniform}(0, 100)$. In all spike-and-slab models we assume a prior mixing probability of $\theta \sim \text{Beta}(1, 1)$, which corresponds to a uniform distribution over $(0, 1)$, reflecting an absence of prior knowledge regarding sparsity. No intercept parameter β_0 is assumed for linear models as continuous responses are centered and predictors centered. An intercept parameter is assumed for logistic models and given the prior $\beta_0 \sim \text{N}(0, 10^2)$. For linear models, the error variance is given the non-informative prior $\sigma^2 \sim \text{Inverse-Gamma}(0.1, 0.1)$. For models using spike-and-slab priors, covariates are selected if their posterior inclusion probability exceeds 0.5, and for Bayesian LASSO models a covariate is selected if its corresponding 95% posterior credible interval excludes 0. In addition to the non-informative versions of the Bayesian LASSO and spike-and-slab prior, we also compare performance of our models against the linear and logistic versions of the LASSO (LinLASSO, LogLASSO, respectively), Elastic-Net (LinEN, LogEN, respectively), and Spike-and-Slab LASSO (LinSSLASSO, LogSSLASSO, respectively). For these models a covariate is selected if it has a nonzero effect estimate.

3.1.2 Configurations of the data

We simulate continuous responses from the linear regression model (2.1) with $\sigma^2 = 1$ and binary responses from the logistic regression model (2.2), with $\beta_0 = -1$. We assign $c = s \times p$ nonzero effects to the p -dimensional regression vector β , where s denotes the sparsity level $s \in \{0.01, 0.02, 0.04, 0.08\}$, with effects of either ± 1 , or ± 0.5 . We consider uncorrelated and correlated designs, using synthetic data (both covariates and responses are simulated) for the uncorrelated design and semi-synthetic data (real-world covariates are used to simulate responses) for the correlated design. Covariates used for the semi-synthetic data are proteomic data from the multi-omics study Overmyer et al., 2021 which analyzes the relationship between biomolecules and the presence or severity of COVID-19. The data contained measurements of 517 proteins, 13,263 RNA transcripts, 646 lipids, and 110 metabolites of 128 individuals admitted to ICU with COVID-like symptoms (Overmyer et al., 2021). Based on the number of predictors and degree of correlation among them, we use only the proteomic data to simulate responses for the correlated design.

- *Uncorrelated design*: All predictor variables are drawn from independent and identically distributed

standard normal distributions. The c effects are assigned to the first c covariates X_1, X_2, \dots, X_c .

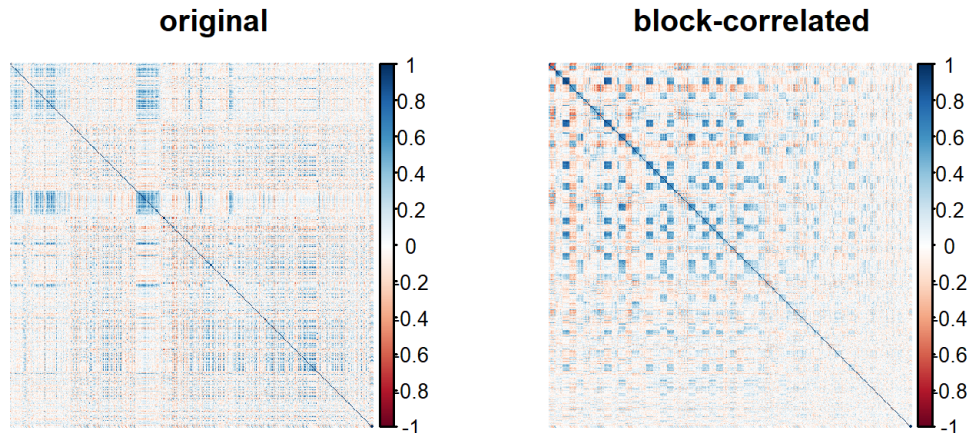
- *Correlated design:* The proteomic data is perturbed to form a block-correlated design, and the c effects are assigned to the initial covariate of the first c blocks, resulting in no more than one significant covariate per block. The block size is fixed at $B = 10$, and the blocking process follows from F. Wang et al., 2020:

(i) The two most highly-correlated predictors are selected to form the first two entries of the first block.

(ii) The remaining $B - 2$ covariates most highly correlated with the first form the remainder of the B -sized block.

(iii) Repeat steps (i)-(ii) with the remaining covariates for the desired number of blocks.

Figure 3.1: Correlation plots of the proteomic data. Left plot is of the original unperturbed data, and the right of the perturbed block-correlated data



We fix the number of responses across all uncorrelated scenarios to be $n = 100$ and consider combinations of $p \in \{200, 500\}$ with $s \in \{0.01, 0.02, 0.04, 0.08\}$, though no combination of p and s producing $s \times p = c < 4$ or $s \times p = c > 20$ active covariates is used. For the correlated data, we have $n = 128$ observations and $p = 517$ measured proteins and consider the number of active covariates to be $c = 5$ and $c = 10$. All configurations of n , p , and c are considered with ± 0.5 effects and with ± 1 effects. For linear regression, we replicate each scenario 100 times. For logistic regression, we do 100 replicates when $p = 200$, and obtain preliminary results from 20 replicates when $p \geq 500$ due to heavy computational demands.

Table 3.1: Factors varied and values considered in simulation study

Correlation	n	p	Number of significant covariates c	Effect sizes
Uncorrelated	100	200	2, 4, 8	± 0.5 and ± 1
	100	500	5, 10, 20	± 0.5 and ± 1
Correlated	128	517	5, 10	± 0.5 and ± 1

3.1.3 Performance metrics

To assess variable selection performance, we report the average true positive rate (TPR), false discovery rate (FDR), false non-discovery rate (FNDR), and F_1 score across all replicates of a given scenario:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}}; \text{FNDR} = \frac{\text{FN}}{\text{TN} + \text{FN}}; F_1 = \frac{2\text{TPR}(1 - \text{FDR})}{\text{TPR} + 1 - \text{FDR}} \quad (3.1)$$

The TPR captures the proportion of significant variables selected, the FDR gives the proportion of false selections out of the total number of selections and the FNDR gives the proportion of false non-selections out of the total number of non-selections, where non-selections are predictors not selected by the model. The F_1 score is the harmonic mean of TPR and precision, where precision is $1 - \text{FDR}$, or the positive predictive value (PPV).

3.1.4 Implementation of methods

JAGS (Just another Gibbs sampler), an open-source sampling program for Bayesian hierarchical models, is used to obtain posterior samples of parameters in all Bayesian regression models considered. All that needs to be supplied by the user to JAGS is the sampling model, prior distributions for parameters, and initial values. JAGS then compiles the model as a directed acyclic graph (DAG) and, using a combination of different sampling algorithms like Metropolis and Gibbs sampling, produces posterior samples of all parameters (Depaoli et al., 2016). JAGS is interfaced through the R package `runjags` in R version 4.2.0. As the minimum-divergence models require fitting the data multiple times, these computations are done in parallel in the Advanced Research Computing (ARC) cluster at the University of Calgary, resulting in no increase in computation time for repeated fits. Kernel density estimation of the posterior densities of λ^2 and τ^2 is done using the `density()` function in R with standard normal kernel and bandwidth $h = 0.9 * \min\{\hat{\sigma}, \frac{\text{IQR}}{1.34}\}n^{-1/5}$ (Silverman’s rule of thumb (Sheather, 2004)). For the proposed minimum divergence models we assess divergence at 12 prior means $\mu \in \{0.1, 0.5, 1, 2.5, 5, 10, 50, 100, 250, 500, 750, 1000\}$. This sequence was chosen based on preliminary simulations. For the Bayesian LASSO, 30,000 samples of the joint posterior are drawn with the first 10,000 discarded as burn-in for $p = 200$ scenarios. For $p \geq 500$

scenarios, 45,000 posterior samples are drawn with the first 15,000 discarded as burn-in. For the linear spike-and-slab models, 37,500 samples of the joint posterior are drawn with the first 12,500 discarded as burn-in for scenarios where $p = 200$. For $p \geq 500$ scenarios, 75,000 posterior samples are drawn with the first 25,000 discarded as burn-in. For logistic spike-and-slab models, 125,000 and 150,000 posterior samples are drawn with the first 25,000 discarded as burn-in for $p = 200$ and $p \geq 500$ scenarios respectively. All samples refer to samples drawn from each of two parallel chains. Convergence of continuous parameters is monitored using the potential scale reduction factor, referred to as R-hat (Gelman & Rubin, 1992), which is an estimate of how much the variance of the posterior distribution may decrease if infinitely many samples were drawn. As recommended in Gelman et al., 1995 convergence is indicated with an R-hat of < 1.1 . For the spike-and-slab prior, we assess agreement of $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_p)^T$ between the two chains by computing the correlation of the p posterior inclusion probabilities $\bar{\gamma}_1, \bar{\gamma}_2, \dots, \bar{\gamma}_p$ from each chain, with a correlation of > 0.98 indicating good agreement. For the Bayesian LASSO, initial values of $\beta_1, \beta_2, \dots, \beta_p$ are randomly drawn from independent standard normal distributions, initial values of $\tau_1^2, \tau_2^2, \dots, \tau_p^2$ randomly drawn from Exponential(rate = $\frac{1}{2}$), and initial λ^2 randomly drawn from its prior distribution. For the spike-and-slab models, initial values of $\beta_1, \beta_2, \dots, \beta_p$ are randomly drawn from independent standard normal distributions, except in the logistic case with $p \geq 500$ where they are drawn from independent $N(0, 0.2^2)$ distributions, due to too large of initial regression coefficients resulting in observation node errors in the DAG compiled by JAGS. The initial value of slab-variance τ^2 is drawn from the prior, except in the logistic case with $p \geq 500$ where τ^2 is initialized at 1 for each chain due to similar sampling issues that occur when initializing $\boldsymbol{\beta}$. Each indicator $\gamma_1, \gamma_2, \dots, \gamma_p$ is initialized from a Bernoulli(0.5) distribution and the mixing weight θ initialized at 0.25 and 0.75 for the first and second chain, respectively.

For the linear and logistic LASSO and EN models, the regularization parameter is computed via 10-fold cross validation, minimizing MSE and deviance for linear and logistic models, respectively. For the EN, a mixing weight of $\alpha = 0.6$ is used, encouraging a light grouping effect. Cross-validation and estimation is done in the R package `glmnet`. For the linear Spike-and-Slab LASSO (LinSSLASSO) a sequence of spike penalty parameters $\lambda_0 \in \{5, 6, 7, \dots, 100\}$ is used, which is similar to the sequence used in the simulation of the original paper (Ročková & George, 2018), and a slab penalty parameter fixed at the default $\lambda_1 = 1$. Estimation is done with the R package `SSLASSO`, and the selected model corresponds to the last model after stabilization of the solution-path. For the logistic Spike-and-Slab LASSO (LogSSLASSO) no stabilized solution-path is computed due to differences in available estimation methods in R, so the model corresponding to a 10-fold cross-validated λ_0 minimizing deviance is used. Deviance is computed for λ_0 values ranging from 1 to 100 with fixed slab penalty parameter $\lambda_1 = 1$. Estimation is done with the R package `BhGLM`.

3.2 Results

3.2.1 Linear regression simulation results

Results of the uncorrelated simulation scenarios for linear regression based on 90 to 100 replicates are shown in Tables 3.2 and 3.3, and results of the correlated linear scenarios in Table 3.4. Figures 3.2 and 3.4 denote mean prior-to-posterior KL-Divergences across all realizations of a scenario, as a function of the log-prior mean of τ^2 and λ^2 .

Figure 3.2: Average simulated prior-to-posterior KL Divergences of τ^2 in the linear spike-and-slab model based on 90-100 replicates. Vertical dashed lines denote the first local minimum. For uncorrelated $p = 200$ scenario with $\text{var}(\tau^2) = 1$ (red) and $\text{var}(\tau^2) = 10$ (black)

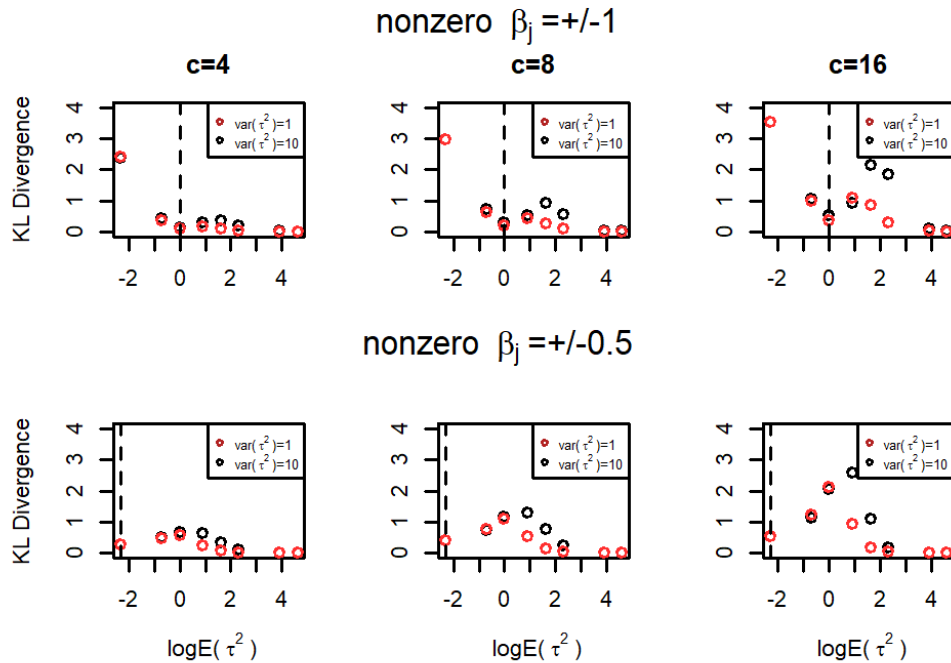


Figure 3.3: Distribution of minimum-divergence prior means of τ^2 in the linear spike-and-slab model based on 90-100 replicates. For uncorrelated $p = 200$ scenario with $var(\tau^2) = 10$

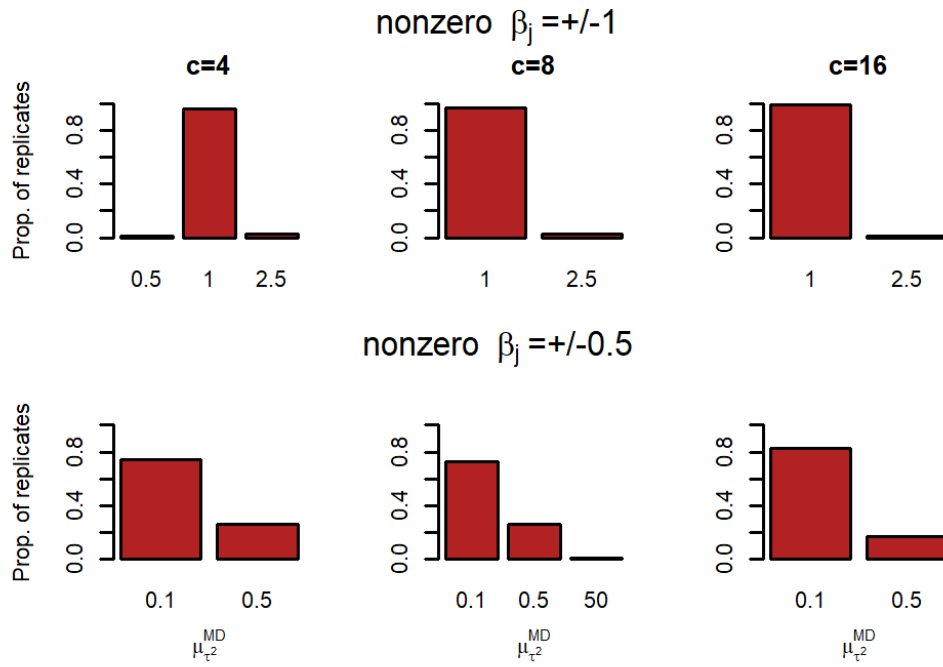


Table 3.2: Uncorrelated linear simulation study results using 90 to 100 replicates. Each entry refers to the average value of the metric and simulation standard error (in parentheses)

		$P = 200$, Uncorrelated Design											
		$C = 4, \pm 1$ EFFECT				$C = 8, \pm 1$ EFFECT				$C = 10, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNR	F1	TPR	FDR	FNR	F1	TPR	FDR	FNR	F1
MD-LinSS	1	1.00 (0.000)	0.05 (0.010)	0.00 (0.000)	0.97 (0.006)	1.00 (0.000)	0.04 (0.007)	0.00 (0.000)	0.98 (0.004)	1.00 (0.000)	0.04 (0.007)	0.00 (0.000)	0.98 (0.004)
	10	1.00 (0.000)	0.05 (0.010)	0.00 (0.000)	0.97 (0.006)	1.00 (0.000)	0.04 (0.007)	0.00 (0.000)	0.98 (0.004)	1.00 (0.000)	0.04 (0.007)	0.00 (0.000)	0.98 (0.004)
NI-LinSS		1.00 (0.000)	0.05 (0.009)	0.00 (0.000)	0.97 (0.005)	1.00 (0.000)	0.04 (0.007)	0.00 (0.000)	0.98 (0.004)	1.00 (0.000)	0.04 (0.007)	0.00 (0.000)	0.98 (0.004)
MD-LinBLASSO	1	0.99 (0.006)	0.06 (0.010)	0.00 (0.000)	0.96 (0.007)	0.90 (0.028)	0.09 (0.022)	0.00 (0.001)	0.89 (0.026)	0.90 (0.028)	0.09 (0.022)	0.00 (0.001)	0.89 (0.026)
	10	0.97 (0.013)	0.05 (0.010)	0.00 (0.000)	0.95 (0.010)	0.79 (0.038)	0.13 (0.028)	0.01 (0.002)	0.79 (0.035)	0.79 (0.038)	0.13 (0.028)	0.01 (0.002)	0.79 (0.035)
NI-LinBLASSO		1.00 (0.000)	0.02 (0.007)	0.00 (0.000)	0.99 (0.004)	0.98 (0.005)	0.00 (0.002)	0.00 (0.000)	0.99 (0.003)	0.98 (0.005)	0.00 (0.002)	0.00 (0.000)	0.99 (0.003)
LinSSLASSO		1.00 (0.000)	0.00 (0.003)	0.00 (0.000)	1.00 (0.002)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	1.00 (0.001)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	1.00 (0.001)
LinLASSO		1.00 (0.000)	0.78 (0.012)	0.00 (0.000)	0.35 (0.015)	1.00 (0.000)	0.76 (0.008)	0.00 (0.000)	0.38 (0.011)	1.00 (0.000)	0.76 (0.008)	0.00 (0.000)	0.38 (0.011)
LinEN		1.00 (0.000)	0.83 (0.008)	0.00 (0.000)	0.28 (0.001)	1.00 (0.000)	0.80 (0.006)	0.00 (0.000)	0.34 (0.008)	1.00 (0.000)	0.80 (0.006)	0.00 (0.000)	0.34 (0.008)
		$C = 4, \pm 0.5$ EFFECT				$C = 8, \pm 0.5$ EFFECT				$C = 10, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNR	F1	TPR	FDR	FNR	F1	TPR	FDR	FNR	F1
MD-LinSS	1	0.96 (0.013)	0.22 (0.024)	0.00 (0.000)	0.84 (0.019)	0.95 (0.009)	0.22 (0.018)	0.00 (0.000)	0.84 (0.013)	0.95 (0.009)	0.22 (0.018)	0.00 (0.000)	0.84 (0.013)
	10	0.96 (0.010)	0.22 (0.023)	0.00 (0.000)	0.84 (0.017)	0.94 (0.013)	0.23 (0.020)	0.00 (0.000)	0.83 (0.016)	0.94 (0.013)	0.23 (0.020)	0.00 (0.000)	0.83 (0.016)
NI-LinSS		0.90 (0.024)	0.21 (0.029)	0.00 (0.000)	0.81 (0.025)	0.88 (0.025)	0.24 (0.026)	0.00 (0.001)	0.79 (0.024)	0.88 (0.025)	0.24 (0.026)	0.00 (0.001)	0.79 (0.024)
MD-LinBLASSO	1	0.54 (0.034)	0.17 (0.035)	0.01 (0.001)	0.62 (0.032)	0.45 (0.026)	0.13 (0.029)	0.02 (0.001)	0.56 (0.027)	0.45 (0.026)	0.13 (0.029)	0.02 (0.001)	0.56 (0.027)
	10	0.53 (0.034)	0.20 (0.037)	0.01 (0.001)	0.60 (0.033)	0.41 (0.028)	0.19 (0.036)	0.02 (0.001)	0.52 (0.031)	0.41 (0.028)	0.19 (0.036)	0.02 (0.001)	0.52 (0.031)
NI-LinBLASSO		0.77 (0.022)	0.07 (0.012)	0.00 (0.000)	0.82 (0.016)	0.55 (0.017)	0.03 (0.007)	0.02 (0.001)	0.69 (0.015)	0.55 (0.017)	0.03 (0.007)	0.02 (0.001)	0.69 (0.015)
LinSSLASSO		0.65 (0.036)	0.12 (0.031)	0.01 (0.001)	0.72 (0.033)	0.54 (0.037)	0.16 (0.036)	0.02 (0.001)	0.61 (0.036)	0.54 (0.037)	0.16 (0.036)	0.02 (0.001)	0.61 (0.036)
LinLASSO		0.99 (0.004)	0.77 (0.014)	0.00 (0.000)	0.36 (0.016)	0.99 (0.004)	0.75 (0.009)	0.00 (0.000)	0.39 (0.011)	0.99 (0.004)	0.75 (0.009)	0.00 (0.000)	0.39 (0.011)
LinEN		1.00 (0.004)	0.80 (0.011)	0.00 (0.000)	0.31 (0.014)	0.99 (0.004)	0.77 (0.008)	0.00 (0.000)	0.37 (0.010)	0.99 (0.004)	0.77 (0.008)	0.00 (0.000)	0.37 (0.010)
		$C = 5, \pm 1$ EFFECT				$C = 10, \pm 1$ EFFECT							
Model	$Var(\tau^2)$	TPR	FDR	FNR	F1	TPR	FDR	FNR	F1				
MD-LinSS	1	1.00 (0.000)	0.04 (0.008)	0.00 (0.000)	0.98 (0.004)	1.00 (0.000)	0.04 (0.006)	0.00 (0.000)	0.98 (0.003)				
	10	1.00 (0.000)	0.05 (0.008)	0.00 (0.000)	0.97 (0.005)	1.00 (0.000)	0.04 (0.006)	0.00 (0.000)	0.98 (0.003)				
NI-LinSS		1.00 (0.000)	0.03 (0.008)	0.00 (0.000)	0.98 (0.004)	1.00 (0.001)	0.04 (0.007)	0.00 (0.000)	0.98 (0.004)				
MD-LinBLASSO	1	0.73 (0.028)	0.03 (0.017)	0.00 (0.000)	0.80 (0.024)	0.14 (0.012)	0.28 (0.045)	0.02 (0.000)	0.22 (0.018)				
	10	0.52 (0.035)	0.16 (0.037)	0.00 (0.000)	0.61 (0.035)	0.06 (0.010)	0.62 (0.049)	0.02 (0.000)	0.10 (0.015)				
NI-LinBLASSO		0.80 (0.020)	0.00 (0.000)	0.00 (0.000)	0.87 (0.015)	0.16 (0.012)	0.20 (0.040)	0.02 (0.000)	0.15 (0.017)				
LinSSLASSO		1.00 (0.000)	0.01 (0.004)	0.00 (0.000)	1.00 (0.002)	1.00 (0.000)	0.00 (0.000)	0.00 (0.000)	1.00 (0.000)				
LinLASSO		1.00 (0.000)	0.82 (0.008)	0.00 (0.000)	0.30 (0.011)	1.00 (0.000)	0.80 (0.006)	0.00 (0.000)	0.33 (0.008)				
LinEN		1.00 (0.000)	0.86 (0.005)	0.00 (0.000)	0.24 (0.007)	1.00 (0.000)	0.83 (0.005)	0.00 (0.000)	0.28 (0.007)				
		$C = 5, \pm 0.5$ EFFECT				$C = 10, \pm 0.5$ EFFECT							
Model	$Var(\tau^2)$	TPR	FDR	FNR	F1	TPR	FDR	FNR	F1				
MD-LinSS	1	0.91 (0.017)	0.22 (0.019)	0.00 (0.000)	0.82 (0.16)	0.77 (0.023)	0.28 (0.022)	0.00 (0.000)	0.73 (0.020)				
	10	0.91 (0.017)	0.22 (0.019)	0.00 (0.000)	0.82 (0.17)	0.76 (0.021)	0.27 (0.019)	0.00 (0.000)	0.73 (0.018)				
NI-LinSS		0.88 (0.025)	0.20 (0.025)	0.00 (0.000)	0.82 (0.023)	0.69 (0.036)	0.30 (0.035)	0.01 (0.001)	0.66 (0.033)				
MD-LinBLASSO	1	0.21 (0.018)	0.33 (0.047)	0.01 (0.000)	0.31 (0.024)	0.06 (0.008)	0.57 (0.050)	0.02 (0.000)	0.10 (0.013)				
	10	0.18 (0.019)	0.41 (0.049)	0.01 (0.000)	0.27 (0.025)	0.05 (0.008)	0.63 (0.049)	0.02 (0.000)	0.09 (0.014)				
NI-LinBLASSO		0.23 (0.018)	0.28 (0.045)	0.01 (0.000)	0.33 (0.024)	0.06 (0.008)	0.55 (0.050)	0.02 (0.000)	0.10 (0.013)				
LinSSLASSO		0.59 (0.037)	0.21 (0.038)	0.00 (0.000)	0.63 (0.034)	0.46 (0.040)	0.40 (0.044)	0.01 (0.001)	0.47 (0.038)				
LinLASSO		0.98 (0.006)	0.81 (0.009)	0.00 (0.000)	0.31 (0.012)	0.90 (0.017)	0.75 (0.012)	0.00 (0.000)	0.36 (0.012)				
LinEN		0.98 (0.006)	0.84 (0.007)	0.00 (0.000)	0.27 (0.010)	0.92 (0.014)	0.79 (0.009)	0.00 (0.000)	0.33 (0.010)				

Figure 3.4: Average simulated inverse prior-to-posterior KL Divergences of λ^2 in the linear Bayesian LASSO model based on 90-100 replicates. Vertical dashed lines denote the first local maximum. For uncorrelated $p = 200$ scenario with $\text{var}(\lambda^2) = 1$ (red) and $\text{var}(\lambda^2) = 10$ (black)

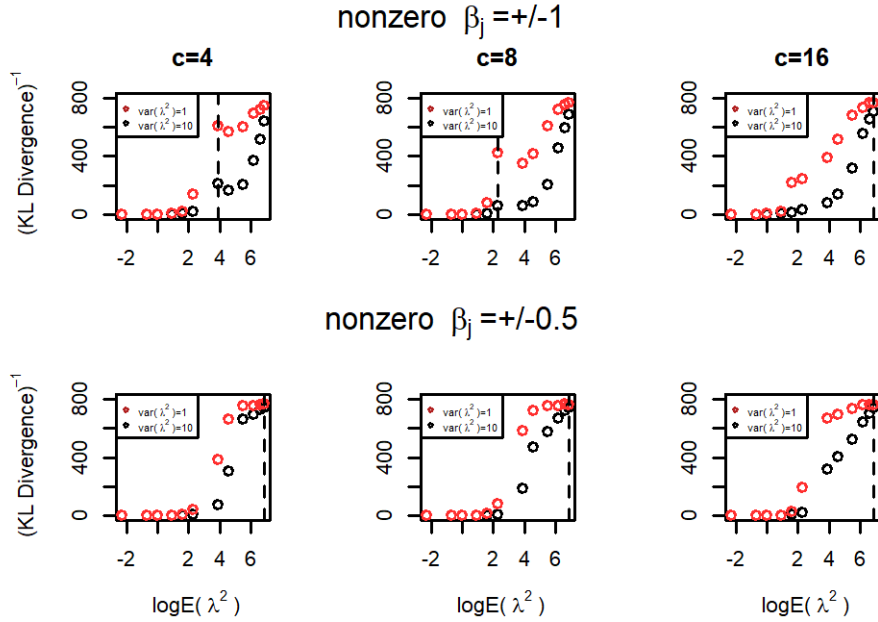


Figure 3.5: Distribution of minimum-divergence prior means of λ^2 in the linear Bayesian LASSO model based on 90-100 replicates. For uncorrelated $p = 200$ scenario with $\text{var}(\lambda^2) = 10$

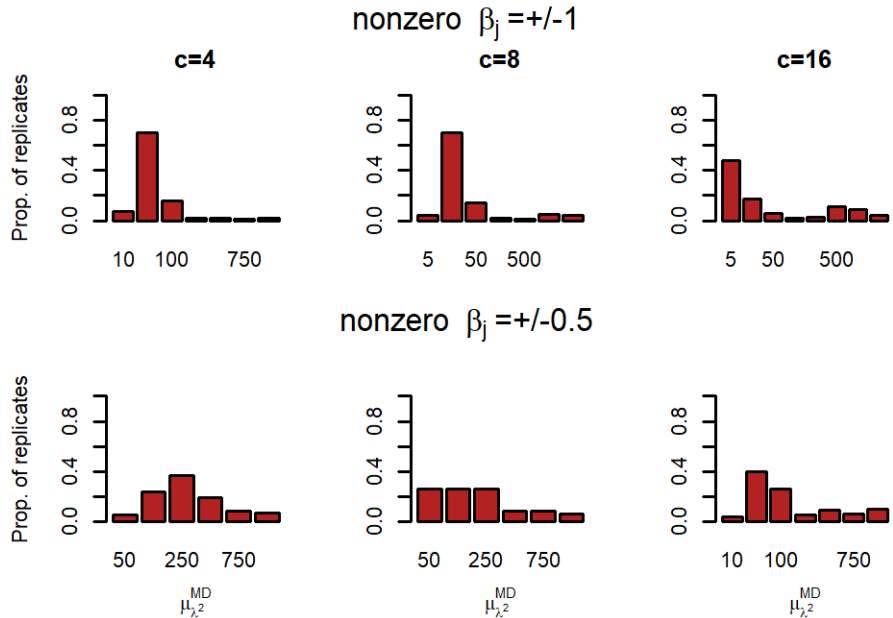


Table 3.3: Uncorrelated linear simulation study results using 90 to 100 replicates, continued. Each entry refers to the average value of the metric and simulation standard error (in parentheses)

		$P = 200$, Uncorrelated Design			
		$C = 16, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LinSS	1	1.00 (0.000)	0.06 (0.006)	0.00 (0.000)	0.97 (0.003)
	10	1.00 (0.000)	0.06 (0.006)	0.00 (0.000)	0.97 (0.003)
NI-LinSS	1	1.00 (0.000)	0.05 (0.006)	0.00 (0.000)	0.97 (0.003)
	10	1.00 (0.000)	0.05 (0.006)	0.00 (0.000)	0.97 (0.003)
MD-LinBLASSO	1	0.59 (0.036)	0.14 (0.033)	0.03 (0.003)	0.65 (0.037)
	10	0.27 (0.037)	0.61 (0.049)	0.06 (0.003)	0.30 (0.041)
NI-LinBLASSO	1	0.54 (0.013)	0.00 (0.002)	0.04 (0.001)	0.69 (0.011)
	10	0.54 (0.013)	0.00 (0.002)	0.04 (0.001)	0.69 (0.011)
LinSSLASSO		1.00 (0.000)	0.01 (0.003)	0.00 (0.000)	0.99 (0.002)
LinLASSO		1.00 (0.000)	0.73 (0.006)	0.00 (0.000)	0.42 (0.008)
LinEN		1.00 (0.000)	0.75 (0.005)	0.00 (0.000)	0.39 (0.006)
		$C = 16, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LinSS	1	0.90 (0.11)	0.38 (0.023)	0.04 (0.017)	0.71 (0.019)
	10	0.90 (0.010)	0.38 (0.022)	0.02 (0.010)	0.70 (0.019)
NI-LinSS	1	0.80 (0.026)	0.32 (0.025)	0.02 (0.002)	0.70 (0.024)
	10	0.80 (0.026)	0.32 (0.025)	0.02 (0.002)	0.70 (0.024)
MD-LinBLASSO	1	0.32 (0.018)	0.13 (0.029)	0.06 (0.001)	0.45 (0.022)
	10	0.24 (0.018)	0.29 (0.043)	0.06 (0.001)	0.35 (0.025)
NI-LinBLASSO	1	0.29 (0.011)	0.04 (0.013)	0.06 (0.001)	0.43 (0.014)
	10	0.29 (0.011)	0.04 (0.013)	0.06 (0.001)	0.43 (0.014)
LinSSLASSO		0.53 (0.040)	0.29 (0.042)	0.04 (0.003)	0.57 (0.041)
LinLASSO		0.97 (0.007)	0.71 (0.007)	0.00 (0.001)	0.44 (0.008)
LinEN		0.97 (0.008)	0.72 (0.007)	0.00 (0.001)	0.42 (0.008)
		$P = 500$, Uncorrelated Design			
		$C = 20, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LinSS	1	0.94 (0.017)	0.17 (0.028)	0.03 (0.017)	0.85 (0.027)
	10	0.96 (0.013)	0.17 (0.028)	0.05 (0.022)	0.86 (0.026)
NI-LinSS	1	0.95 (0.020)	0.08 (0.015)	0.00 (0.001)	0.93 (0.019)
	10	0.95 (0.020)	0.08 (0.015)	0.00 (0.001)	0.93 (0.019)
MD-LinBLASSO	1	0.02 (0.003)	0.69 (0.046)	0.04 (0.000)	0.04 (0.006)
	10	0.01 (0.003)	0.83 (0.038)	0.04 (0.000)	0.02 (0.005)
NI-LinBLASSO	1	0.02 (0.003)	0.60 (0.049)	0.04 (0.000)	0.04 (0.006)
	10	0.02 (0.003)	0.60 (0.049)	0.04 (0.000)	0.04 (0.006)
LinSSLASSO		0.92 (0.026)	0.06 (0.023)	0.00 (0.001)	0.92 (0.026)
LinLASSO		0.92 (0.014)	0.74 (0.006)	0.00 (0.001)	0.40 (0.001)
LinEN		0.90 (0.019)	0.77 (0.007)	0.00 (0.001)	0.36 (0.009)
		$C = 20, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LinSS	1	0.45 (0.015)	0.43 (0.019)	0.02 (0.001)	0.49 (0.014)
	10	0.46 (0.013)	0.39 (0.018)	0.02 (0.001)	0.51 (0.013)
NI-LinSS	1	0.39 (0.028)	0.53 (0.034)	0.05 (0.014)	0.39 (0.026)
	10	0.39 (0.028)	0.53 (0.034)	0.05 (0.014)	0.39 (0.026)
MD-LinBLASSO	1	0.01 (0.003)	0.73 (0.046)	0.04 (0.000)	0.03 (0.005)
	10	0.01 (0.003)	0.77 (0.065)	0.04 (0.000)	0.02 (0.006)
NI-LinBLASSO	1	0.01 (0.002)	0.74 (0.044)	0.04 (0.000)	0.03 (0.005)
	10	0.01 (0.002)	0.74 (0.044)	0.04 (0.000)	0.03 (0.005)
LinSSLASSO		0.11 (0.019)	0.80 (0.033)	0.04 (0.001)	0.12 (0.020)
LinLASSO		0.67 (0.020)	0.70 (0.013)	0.01 (0.001)	0.39 (0.010)
LinEN		0.70 (0.022)	0.71 (0.015)	0.01 (0.001)	0.36 (0.010)

Table 3.4: Correlated linear simulation study results using 90 to 100 replicates. Each entry refers to the average value of the metric and simulation standard error (in parentheses)

		$P = 517$, Correlated Semi-Synthetic Design			
		$C = 5, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LinSS	1	0.93 (0.012)	0.08 (0.013)	0.00 (0.000)	0.92 (0.012)
	10	0.92 (0.012)	0.08 (0.012)	0.00 (0.000)	0.92 (0.012)
NI-LinSS	1	0.91 (0.013)	0.09 (0.014)	0.00 (0.000)	0.91 (0.013)
	10	0.91 (0.013)	0.09 (0.014)	0.00 (0.000)	0.91 (0.013)
MD-LinBLASSO	1	0.12 (0.012)	0.49 (0.049)	0.01 (0.000)	0.19 (0.018)
	10	0.07 (0.010)	0.70 (0.046)	0.01 (0.000)	0.11 (0.016)
NI-LinBLASSO	1	0.13 (0.012)	0.42 (0.049)	0.01 (0.000)	0.21 (0.018)
	10	0.13 (0.012)	0.42 (0.049)	0.01 (0.000)	0.21 (0.018)
LinSSLASSO		0.87 (0.015)	0.14 (0.017)	0.00 (0.000)	0.86 (0.016)
LinLASSO		0.96 (0.009)	0.85 (0.006)	0.00 (0.000)	0.26 (0.009)
LinEN		0.99 (0.005)	0.88 (0.003)	0.00 (0.000)	0.21 (0.005)
		$C = 5, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LinSS	1	0.31 (0.017)	0.42 (0.033)	0.01 (0.000)	0.42 (0.019)
	10	0.31 (0.019)	0.41 (0.034)	0.01 (0.000)	0.43 (0.020)
NI-Lin SS	1	0.41 (0.024)	0.46 (0.034)	0.01 (0.000)	0.40 (0.022)
	10	0.41 (0.024)	0.46 (0.034)	0.01 (0.000)	0.40 (0.022)
MD-LinBLASSO	1	0.00 (0.000)	1.00 (0.000)	0.01 (0.000)	0.00 (0.000)
	10	0.00 (0.000)	1.00 (0.000)	0.01 (0.000)	0.00 (0.000)
NI-LinBLASSO	1	0.01 (0.003)	0.97 (0.017)	0.01 (0.000)	0.01 (0.001)
	10	0.01 (0.003)	0.97 (0.017)	0.01 (0.000)	0.01 (0.001)
LinSSLASSO		0.29 (0.020)	0.61 (0.03)	0.01 (0.000)	0.30 (0.02)
LinLASSO		0.66 (0.020)	0.87 (0.007)	0.00 (0.000)	0.21 (0.01)
LinEN		0.76 (0.018)	0.89 (0.004)	0.00 (0.000)	0.17 (0.01)
		$C = 10, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LinSS	1	0.90 (0.010)	0.09 (0.011)	0.00 (0.000)	0.90 (0.010)
	10	0.90 (0.011)	0.10 (0.010)	0.00 (0.000)	0.90 (0.010)
NI-LinSS	1	0.90 (0.011)	0.09 (0.010)	0.00 (0.000)	0.90 (0.010)
	10	0.90 (0.011)	0.09 (0.010)	0.00 (0.000)	0.90 (0.010)
MD-LinBLASSO	1	0.09 (0.008)	0.33 (0.047)	0.02 (0.000)	0.16 (0.013)
	10	0.09 (0.008)	0.33 (0.047)	0.02 (0.000)	0.16 (0.013)
NI-LinBLASSO	1	0.10 (0.008)	0.29 (0.048)	0.02 (0.000)	0.17 (0.013)
	10	0.10 (0.008)	0.29 (0.048)	0.02 (0.000)	0.17 (0.013)
LinSSLASSO		0.87 (0.011)	0.11 (0.011)	0.00 (0.000)	0.88 (0.010)
LinLASSO		0.88 (0.010)	0.85 (0.004)	0.00 (0.000)	0.26 (0.01)
LinEN		0.90 (0.009)	0.86 (0.004)	0.00 (0.000)	0.20 (0.003)
		$C = 10, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LinSS	1	0.36 (0.015)	0.47 (0.025)	0.01 (0.000)	0.42 (0.015)
	10	0.36 (0.015)	0.47 (0.026)	0.01 (0.000)	0.42 (0.015)
NI-LinSS	1	0.44 (0.018)	0.50 (0.027)	0.02 (0.011)	0.41 (0.017)
	10	0.44 (0.018)	0.50 (0.027)	0.02 (0.011)	0.41 (0.017)
MD-LinBLASSO	1	0.01 (0.002)	0.95 (0.022)	0.02 (0.000)	0.01 (0.004)
	10	0.00 (0.002)	0.97 (0.017)	0.02 (0.000)	0.01 (0.003)
NI-LinBLASSO	1	0.01 (0.003)	0.89 (0.032)	0.02 (0.000)	0.02 (0.006)
	10	0.01 (0.003)	0.89 (0.032)	0.02 (0.000)	0.02 (0.006)
LinSSLASSO		0.32 (0.016)	0.48 (0.027)	0.01 (0.000)	0.37 (0.017)
LinLASSO		0.59 (0.013)	0.85 (0.005)	0.01 (0.000)	0.23 (0.007)
LinEN		0.64 (0.013)	0.86 (0.004)	0.01 (0.000)	0.22 (0.007)

3.2.2 Logistic regression simulation results

Figure 3.6: Average simulated prior-to-posterior KL Divergences of τ^2 in the logistic spike-and-slab model based on 90-100 replicates. Vertical dashed lines denote the first local minimum. For uncorrelated $p = 200$ scenario with $\text{var}(\tau^2) = 1$ (red) and $\text{var}(\tau^2) = 10$ (black)

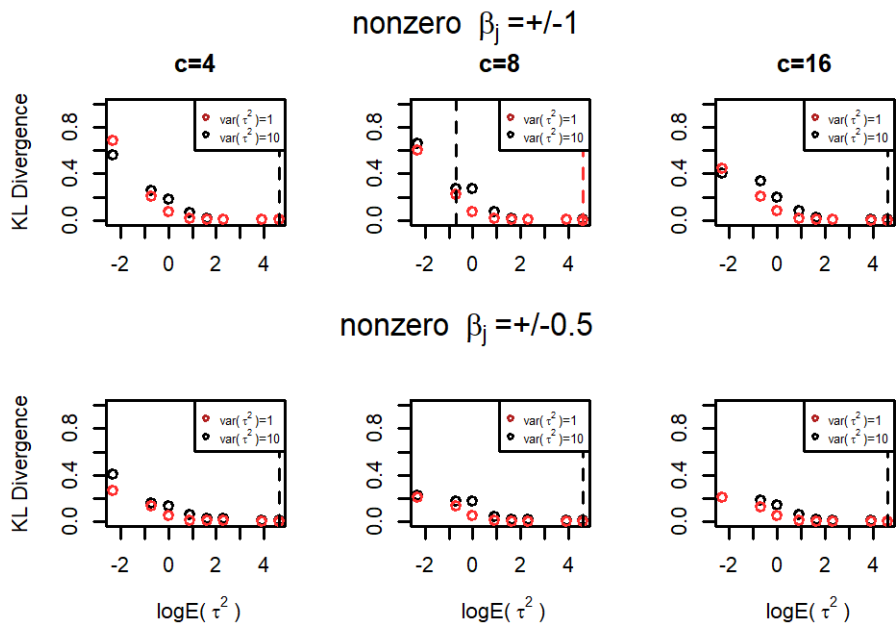


Figure 3.7: Distribution of minimum-divergence prior means of τ^2 in the logistic spike-and-slab model based on 90-100 replicates. For uncorrelated $p = 200$ scenario with $\text{var}(\tau^2) = 10$

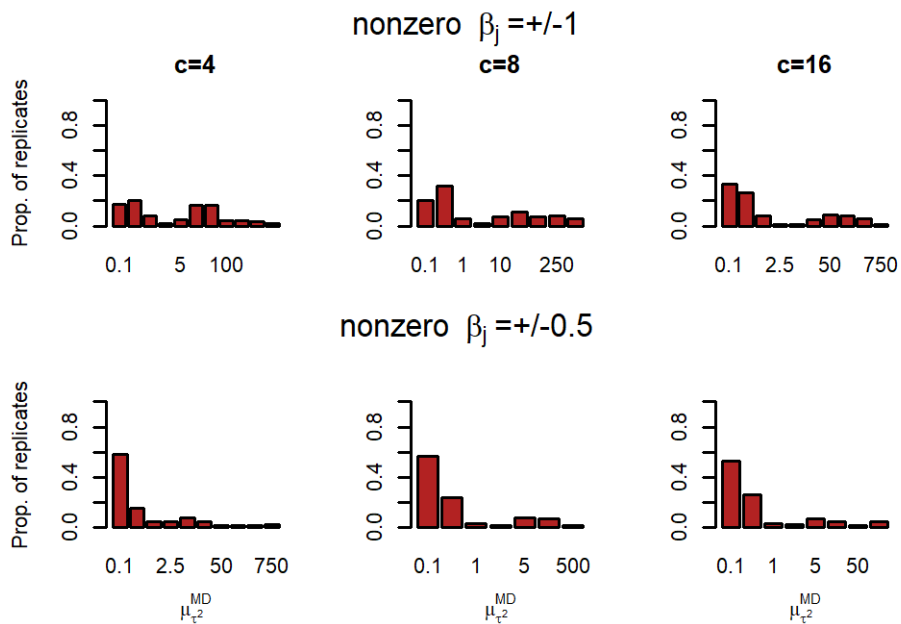


Table 3.5: Uncorrelated logistic simulation study results using 90 to 100 replicates for $p = 200$, and preliminary uncorrelated logistic simulation study results using 15 to 20 replicates for $p = 500$. Each entry refers to the average value of the metric and simulation standard error (in parentheses)

		$P = 200$, Uncorrelated Design											
		$C = 4, \pm 1$ EFFECT					$C = 8, \pm 1$ EFFECT						
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1	TPR	FDR	FNDR	F1	TPR	FDR	FNDR	F1
MD-LogSS	1	0.90 (0.019)	0.63 (0.027)	0.09 (0.030)	0.45 (0.026)	0.80 (0.019)	0.68 (0.025)	0.17 (0.036)	0.38 (0.024)				
	10	0.88 (0.021)	0.57 (0.029)	0.07 (0.026)	0.51 (0.027)	0.79 (0.020)	0.65 (0.026)	0.20 (0.040)	0.40 (0.024)				
NI-LogSS		0.90 (0.020)	0.65 (0.023)	0.04 (0.020)	0.45 (0.025)	0.78 (0.018)	0.63 (0.024)	0.11 (0.030)	0.44 (0.022)				
LogSSLASSO		0.97 (0.008)	0.89 (0.003)	0.00 (0.000)	0.20 (0.004)	0.87 (0.011)	0.80 (0.005)	0.01 (0.001)	0.32 (0.007)				
LogLASSO		0.90 (0.020)	0.72 (0.019)	0.00 (0.000)	0.38 (0.015)	0.78 (0.020)	0.68 (0.015)	0.01 (0.001)	0.42 (0.012)				
LogEN		0.94 (0.017)	0.81 (0.014)	0.00 (0.000)	0.30 (0.015)	0.83 (0.018)	0.76 (0.013)	0.01 (0.001)	0.35 (0.011)				
		$C = 4, \pm 0.5$ EFFECT					$C = 8, \pm 0.5$ EFFECT						
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1	TPR	FDR	FNDR	F1	TPR	FDR	FNDR	F1
MD-LogSS	1	0.68 (0.036)	0.85 (0.026)	0.26 (0.045)	0.16 (0.018)	0.65 (0.033)	0.80 (0.024)	0.31 (0.045)	0.20 (0.015)				
	10	0.64 (0.038)	0.82 (0.028)	0.29 (0.047)	0.19 (0.019)	0.61 (0.034)	0.79 (0.027)	0.28 (0.045)	0.21 (0.015)				
NI-LogSS		0.61 (0.040)	0.90 (0.014)	0.21 (0.040)	0.13 (0.014)	0.62 (0.033)	0.83 (0.021)	0.17 (0.037)	0.20 (0.014)				
LogSSLASSO		0.68 (0.023)	0.86 (0.011)	0.01 (0.001)	0.22 (0.01)	0.65 (0.015)	0.87 (0.004)	0.02 (0.001)	0.22 (0.005)				
LogLASSO		0.40 (0.033)	0.76 (0.027)	0.01 (0.001)	0.23 (0.019)	0.35 (0.024)	0.69 (0.027)	0.03 (0.001)	0.26 (0.015)				
LogEN		0.44 (0.035)	0.78 (0.027)	0.01 (0.001)	0.23 (0.019)	0.41 (0.027)	0.74 (0.024)	0.03 (0.001)	0.24 (0.014)				
		$C = 4, \pm 1$ EFFECT					$C = 10, \pm 1$ EFFECT						
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1	TPR	FDR	FNDR	F1	TPR	FDR	FNDR	F1
MD-LogSS	1	0.62 (0.066)	0.58 (0.088)	0.17 (0.090)	0.34 (0.054)	0.60 (0.067)	0.70 (0.061)	0.11 (0.068)	0.30 (0.044)				
	10	0.73 (0.061)	0.67 (0.080)	0.14 (0.091)	0.33 (0.057)	0.67 (0.076)	0.68 (0.072)	0.23 (0.100)	0.29 (0.056)				
NI-LogSS		0.59 (0.057)	0.53 (0.067)	0.06 (0.055)	0.46 (0.060)	0.54 (0.059)	0.64 (0.058)	0.08 (0.066)	0.36 (0.038)				
LogSSLASSO		0.85 (0.029)	0.90 (0.007)	0.00 (0.000)	0.17 (0.011)	0.71 (0.034)	0.84 (0.006)	0.01 (0.001)	0.26 (0.010)				
LogLASSO		0.68 (0.062)	0.63 (0.070)	0.00 (0.001)	0.38 (0.040)	0.47 (0.068)	0.69 (0.058)	0.01 (0.001)	0.31 (0.042)				
LogEN		0.73 (0.049)	0.73 (0.056)	0.00 (0.000)	0.31 (0.032)	0.55 (0.064)	0.79 (0.028)	0.01 (0.001)	0.28 (0.031)				
		$C = 5, \pm 0.5$ EFFECT					$C = 10, \pm 0.5$ EFFECT						
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1	TPR	FDR	FNDR	F1	TPR	FDR	FNDR	F1
MD-LogSS	1	0.76 (0.064)	0.90 (0.050)	0.35 (0.109)	0.09 (0.021)	0.81 (0.068)	0.89 (0.053)	0.55 (0.113)	0.08 (0.017)				
	10	0.80 (0.062)	0.90 (0.051)	0.50 (0.114)	0.09 (0.026)	0.87 (0.048)	0.95 (0.015)	0.55 (0.113)	0.08 (0.020)				
NI-LogSS		0.56 (0.055)	0.89 (0.021)	0.05 (0.050)	0.16 (0.024)	0.60 (0.060)	0.85 (0.042)	0.16 (0.081)	0.16 (0.025)				
LogSSLASSO		0.53 (0.044)	0.90 (0.011)	0.00 (0.000)	0.17 (0.017)	0.44 (0.032)	0.91 (0.007)	0.01 (0.001)	0.15 (0.010)				
LogLASSO		0.27 (0.062)	0.73 (0.080)	0.01 (0.001)	0.18 (0.034)	0.13 (0.032)	0.76 (0.067)	0.02 (0.001)	0.14 (0.031)				
LogEN		0.32 (0.061)	0.79 (0.068)	0.01 (0.001)	0.16 (0.026)	0.22 (0.044)	0.80 (0.050)	0.02 (0.001)	0.17 (0.029)				

Table 3.6: Uncorrelated logistic simulation study results using 90 to 100 replicates for $p = 200$, and preliminary uncorrelated logistic simulation study results using 15 to 20 replicates for $p = 500$, continued. Each entry refers to the average value of the metric and simulation standard error (in parentheses)

		$P = 200$, Uncorrelated Design			
		$C = 16, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LogSS	1	0.80 (0.024)	0.74 (0.021)	0.41 (0.048)	0.31 (0.017)
	10	0.82 (0.023)	0.77 (0.021)	0.49 (0.049)	0.28 (0.016)
NI-LogSS		0.75 (0.022)	0.71 (0.020)	0.29 (0.043)	0.35 (0.017)
LogSSLASSO		0.71 (0.011)	0.70 (0.006)	0.03 (0.001)	0.42 (0.007)
LogLASSO		0.53 (0.024)	0.61 (0.017)	0.04 (0.002)	0.40 (0.013)
LogEN		0.62 (0.023)	0.64 (0.018)	0.04 (0.002)	0.40 (0.010)
		$C = 16, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LogSS	1	0.68 (0.035)	0.78 (0.022)	0.37 (0.047)	0.22 (0.010)
	10	0.63 (0.036)	0.76 (0.025)	0.36 (0.046)	0.23 (0.011)
NI-LogSS		0.64 (0.035)	0.80 (0.019)	0.34 (0.046)	0.22 (0.011)
LogSSLASSO		0.54 (0.013)	0.79 (0.005)	0.05 (0.001)	0.30 (0.007)
LogLASSO		0.22 (0.018)	0.66 (0.028)	0.07 (0.001)	0.22 (0.015)
LogEN		0.28 (0.021)	0.70 (0.026)	0.06 (0.001)	0.24 (0.016)
		$P = 500$, Uncorrelated Design			
		$C = 20, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LogSS	1	0.62 (0.084)	0.78 (0.054)	0.42 (0.109)	0.18 (0.029)
	10	0.81 (0.069)	0.88 (0.041)	0.56 (0.118)	0.14 (0.026)
NI-LogSS		0.61 (0.070)	0.77 (0.049)	0.23 (0.094)	0.23 (0.026)
LogSSLASSO		0.47 (0.020)	0.80 (0.011)	0.02 (0.001)	0.28 (0.013)
LogLASSO		0.23 (0.042)	0.69 (0.060)	0.03 (0.002)	0.22 (0.034)
LogEN		0.31 (0.043)	0.69 (0.059)	0.03 (0.002)	0.24 (0.027)
		$C = 20, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LogSS	1	0.74 (0.074)	0.91 (0.019)	0.56 (0.111)	0.12 (0.016)
	10	0.73 (0.083)	0.89 (0.030)	0.54 (0.115)	0.12 (0.018)
NI-LogSS		0.40 (0.076)	0.80 (0.034)	0.19 (0.088)	0.18 (0.022)
LogSSLASSO		0.34 (0.022)	0.84 (0.011)	0.03 (0.001)	0.21 (0.014)
LogLASSO		0.15 (0.032)	0.82 (0.035)	0.03 (0.001)	0.15 (0.030)
LogEN		0.19 (0.034)	0.77 (0.050)	0.03 (0.001)	0.17 (0.025)

Table 3.7: Preliminary correlated logistic simulation study results using 15 to 20 replicates. Each entry refers to the average value of the metric and simulation standard error (in parentheses)

		$P = 517$, Correlated Semi-Synthetic Design			
		$C = 5, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LogSS	1	0.78 (0.068)	0.92 (0.023)	0.30 (0.114)	0.11 (0.026)
	10	0.84 (0.078)	0.95 (0.016)	0.31 (0.119)	0.07 (0.017)
NI-LogSS		0.73 (0.080)	0.91 (0.021)	0.15 (0.097)	0.13 (0.025)
LogSSLASSO		0.50 (0.059)	0.95 (0.008)	0.01 (0.001)	0.10 (0.015)
LogLASSO		0.43 (0.060)	0.90 (0.014)	0.01 (0.001)	0.16 (0.021)
LogEN		0.63 (0.062)	0.91 (0.011)	0.00 (0.001)	0.16 (0.018)
		$C = 5, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LogSS	1	0.36 (0.108)	0.99 (0.005)	0.17 (0.090)	0.04 (0.016)
	10	0.30 (0.086)	0.90 (0.053)	0.21 (0.091)	0.13 (0.037)
NI-LogSS		0.26 (0.080)	0.96 (0.015)	0.06 (0.050)	0.05 (0.018)
LogSSLASSO		0.24 (0.034)	0.98 (0.003)	0.01 (0.000)	0.04 (0.006)
LogLASSO		0.17 (0.030)	0.90 (0.029)	0.01 (0.000)	0.10 (0.021)
LogEN		0.26 (0.046)	0.92 (0.025)	0.01 (0.000)	0.09 (0.017)
		$C = 10, \pm 1$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LogSS	1	0.46 (0.085)	0.84 (0.046)	0.17 (0.085)	0.18 (0.032)
	10	0.55 (0.084)	0.87 (0.037)	0.26 (0.098)	0.15 (0.029)
NI-LogSS		0.35 (0.078)	0.80 (0.044)	0.07 (0.058)	0.18 (0.032)
LogSSLASSO		0.38 (0.034)	0.92 (0.007)	0.01 (0.001)	0.13 (0.011)
LogLASSO		0.33 (0.035)	0.86 (0.019)	0.01 (0.001)	0.19 (0.023)
LogEN		0.47 (0.031)	0.88 (0.009)	0.01 (0.001)	0.18 (0.013)
		$C = 10, \pm 0.5$ EFFECT			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-LogSS	1	0.67 (0.098)	0.91 (0.050)	0.46 (0.113)	0.07 (0.011)
	10	0.70 (0.094)	0.91 (0.049)	0.61 (0.111)	0.07 (0.010)
NI-LogSS		0.57 (0.087)	0.93 (0.026)	0.18 (0.089)	0.07 (0.009)
LogSSLASSO		0.25 (0.020)	0.96 (0.004)	0.02 (0.000)	0.07 (0.006)
LogLASSO		0.20 (0.017)	0.91 (0.010)	0.02 (0.000)	0.12 (0.012)
LogEN		0.26 (0.020)	0.92 (0.009)	0.02 (0.000)	0.13 (0.012)

3.3 Discussion

3.3.1 Linear regression

Minimum-divergence spike-and-slab: Performance of the minimum-divergence spike-and-slab model (MD-LinSS) is generally better or as good as the noninformative spike-and-slab model (NI-LinSS) in most uncorrelated scenarios (Tables 3.2,3.3). When the effect is ± 1 , there is little difference in variable selection performance between the two models however when the effect decreases to ± 0.5 , the minimum-divergence model outperforms the noninformative model in almost all uncorrelated scenarios. The MD-LinSS model in this case has a higher TPR and lower FDR resulting in an F_1 score nearer to 1 except in the case where $p = 200$, $c = 4$ and $p = 500$, $c = 20$, where the MD-LinSS has a higher FDR resulting in nearly the same F_1 score as the noninformative model. Both models have similar small FNDR values of ≤ 0.04 across all uncorrelated scenarios. In the correlated scenarios performance between the minimum-divergence and noninformative spike-and-slab models are comparable (Table 3.4). When the effect size is 1 both models have metrics that differ by at most 0.01 and when the effect size is 0.5 the minimum-divergence model has a smaller TPR and FDR but F_1 scores slightly nearer to 1. Both models have similarly small FNDR values of ≤ 0.02 . Overall, the MD-LinSS model performs as well as the NI-LinSS model but shows improved performance against the noninformative model when the active covariates' effects are ± 0.5 and the data uncorrelated. The MD-LinSS model outperforms the Spike-and-Slab LASSO (SSLASSO) in all scenarios except in uncorrelated scenarios with ± 1 effects, in which case it performs similarly or slightly worse, with the minimum-divergence model tending to select more irrelevant variables than the SSLASSO. Compared to the LASSO and Elastic-Net (EN), the MD-LinSS model performs much better based on smaller FDR values. In the absence of strong correlation between predictors and with small effect sizes of 0.5, the MD-LinSS outperforms all competing models by obtaining a better trade-off between true-positives and false-discoveries as reflected in larger F_1 scores. Additionally, the MD-LinSS model was not sensitive to the choice of slab hypervariance $var(\tau^2)$, though we note that very small hypervariances of 0.05, 0.1 resulted in poorer performance (results not shown).

Figure 3.2 shows the average prior-to-posterior divergence of τ^2 as a function of log-prior means $\mu_{\tau^2} = 0.1, 0.5, 1, 2.5, 5, 10, 50, 100$ for uncorrelated cases at $p = 200$. A local minimum is observed at $\mu_{\tau^2} = 1$ for all c values when effects are ± 1 (top of Figure 3.2), regardless of whether $var(\tau^2) = 1$ or 10, indicating that a prior for τ^2 centered about 1 may conflict least with data under those data configurations. When the effect is lowered to ± 0.5 , the plots (bottom of Figure 3.2) indicate that a prior for τ^2 centered about 0.1 may conflict least with data. When p is increased to 500 the plots in Figure 3.2 exhibit a similar relationship between μ_{τ^2} and divergence of τ^2 , as well as when predictors are correlated (plots not shown). Based on the plots in

Figure 3.2, the prior mean μ_{τ^2} at which there is locally minimal divergence seems to be determined largely by the effect size of active covariates and not by the average effect size of all covariates. This is reasonable since the slab is used to model nonzero effects and τ^2 reflects our belief about the size of those effects. The histograms in Figure 3.3 illustrate that, regardless of which replication of the dataset is being analyzed, the prior mean $\mu_{\tau^2}^{MD}$ used in the MD-LinSS model is largely either 0.1 or 1, depending on whether effects are of size 0.5 or 1. This is consistent with the average prior-to-posterior divergence plots in Figure 3.2.

Minimum-divergence Bayesian LASSO: Performance of the minimum-divergence Bayesian LASSO (MD-LinBLASSO) is generally slightly worse than its noninformative counterpart (NI-LinBLASSO) in almost all scenarios (Tables 3.2,3.3,3.4). The MD-LinBLASSO on average detects fewer significant covariates than the NI-LinBLASSO as seen in a smaller TPR, except in the cases where $p = 200$, $c = 16$ and $p = 500$, $c \geq 10$, where it is slightly larger when $var(\lambda^2) = 1$. Note however that when $p = 500$ and $c \geq 10$ the Bayesian LASSO fails to make any selection most of the time. The MD-LinBLASSO in all cases but one has a higher FDR than the NI-LinBLASSO and a smaller F_1 score. The FNDR is equally low for both Bayesian LASSO models with the highest being 0.06. Compared to the frequentist models, the MD-LinBLASSO tends to perform worse than the SSLASSO, though better than the LASSO and EN. The LASSO and EN tend to select $> c$ covariates, resulting in high FDR's. Additionally, performance of the MD-LinBLASSO was sensitive to the hypervariance $var(\lambda^2)$. For instance, when $p = 200$, $c = 16$, and effects ± 1 , using the hypervariance $var(\lambda^2) = 1$ gave an F_1 score of 0.65 and using $var(\lambda^2) = 10$ gave a score of 0.30. Overall, the MD-LinBLASSO tended to select less covariates than other models and often times failed to make any selection at all, resulting in low TPR's and relatively high FDR's. This is explained when we note that the prior mean of λ^2 that gave minimal prior-to-posterior divergence was often larger than the prior mean that gave the best selection, resulting in too much shrinkage of the regression vector β towards 0.

Figure 3.4 shows the average inverse prior-to-posterior divergence of λ^2 as a function of log-prior means $\mu_{\lambda^2} = 0.1, 0.5, 1, 2.5, 5, 10, 50, 100, 250, 500, 1000$ for uncorrelated cases at $p = 200$. The inverse divergence is reported as opposed to the divergence since, for this model, a local maximum of the inverse divergence is easier to identify visually than the local minimum of the divergence. A locally minimal average divergence is identified only for $c = 4, 8$ with effects ± 1 , at prior means 50 and 10, respectively. This is consistent with histograms in figure 3.4 (top left-most 2 plots), where $\mu_{\lambda^2}^{MD} = 50$ and 10 for close to 80% and 60% of simulated datasets, respectively, for those scenarios at $var(\lambda^2) = 10$. The remaining $p = 200$ scenarios show greater variability in values of $\mu_{\lambda^2}^{MD}$ across different replicates, resulting in no clear region of μ_{λ^2} where the average divergence across replicates is locally small (Figure 3.4). When $p \geq 500$ no clear region of μ_{λ^2} where the average divergence across replicates is locally small is identified (plots not shown). As effect sizes drop from 1 to 0.5, the values of $\mu_{\lambda^2}^{MD}$ across replicates are generally larger, corresponding to a smaller variance

of the joint Laplace prior on β . Additionally, as c increases, the values of $\mu_{\lambda^2}^{MD}$ across replicates generally decrease. This is reasonable since, unlike the spike-and-slab prior, the variance of the Laplace prior on β reflects our belief about the strength of any effect, not just an active effect. The result is that as c increases, the average effect size $\sum_{j=1}^p |\beta_j|/p$ increases, necessitating less shrinkage i.e. a smaller value of $\mu_{\lambda^2}^{MD}$. As p increases, the variability of $\mu_{\lambda^2}^{MD}$ across replicates increases (plots not shown). This difference in $\mu_{\lambda^2}^{MD}$ among different generated datasets from a given scenario may be due to lack of data-information regarding the parameter λ^2 at the sample size $n = 100$.

3.3.2 Logistic regression

Minimum-divergence spike-and-slab: In uncorrelated scenarios (Tables 3.5,3.6), performance of the logistic minimum-divergence spike-and-slab model (MD-LogSS) relative to other models varies. Compared to the NI-LogSS model, when $p = 200$ with ± 0.5 effects, the MD-LogSS model compares favourably, with a lower FDR and similar TPR, FNDR, and F_1 scores. The comparison is similar when $p = 200$, $c = 4$ and effects ± 1 . In all other uncorrelated scenarios the MD-LogSS tends to have a higher TPR, but also a higher FDR, FNDR, and lower F_1 score than the NI-LogSS model. When predictors are correlated the MD-LogSS obtains a higher average TPR and lower or equal FDR than the NI-LogSS when effects are ± 0.5 , though with a larger FNDR. The generally higher TPR's, FDR's, and FNDR's of the MD-LogSS model in most uncorrelated scenarios can be attributed to the model's tendency to select all covariates for inclusion into the model which results in $TPR = 1$, $FDR = 1 - \frac{c}{p}$, and $FNDR = 1$. Compared to the LogLASSO and LogEN, the MD-LogSS model has a higher TPR in most uncorrelated scenarios. When effects are ± 1 and $c = 4, 5, 8$, or $c = 10$, i.e. the sparsity, s , is not at it's highest, the MD-LogSS has a lower FDR and higher F_1 score than the LogEN. Again, when sparsity s is not at it's highest and effects ± 1 , the MD-LogSS has a better trade-off between false discovery and true positives (higher F_1 score) than the LogSSLASSO. When effects are ± 0.5 , the minimum-divergence model is generally able to recover more of the active covariates than the LogSSLASSO, particularly when $c = 10, 16$, though with a lower average F_1 score. When predictors are correlated and effects ± 0.5 , the MD-LogSS model boasts a higher average TPR and lower FDR than frequentist models. Compared to the Bayesian models, frequentist models tend to have a much lower FNDR, though this is not due to a tendency of the Bayesian models to falsely categorize important variables as unimportant, but due to a tendency of the Bayesian models to categorize all variables as important, resulting in no false-nondiscoveries ($FNDR = 1$). Overall, performance of the MD-LogSS is comparable or better than its non-informative counterpart when $p = 200$ or $p = 517$ with effects ± 0.5 , and has a better trade-off between TPR and FDR than the LogLASSO and LogEN when $c = 4, 5, 8, 10$ with

effects ± 1 , but in many scenarios selects too many covariates. When the number of covariates is $p = 200$ the MD-LogSS model is not very sensitive to choice of hypervariance $var(\tau^2)$, though becomes more sensitive as the number of covariates increases to $p = 500$ or $p = 517$.

The average prior-to-posterior divergence of τ^2 across all replicates of uncorrelated $p = 200$ scenarios is decreasing in μ_{τ^2} , except when $c = 16$ for ± 0.5 effects, where the average divergence is locally minimal at $\mu_{\tau^2} = 5$ (Figure 3.6). This relationship between μ_{τ^2} and divergence is consistent with the histograms in Figure 3.7, where $\mu_{\tau^2}^{MD}$ varies across replicates. As effect sizes drop from 1 to 0.5, the values of $\mu_{\tau^2}^{MD}$ across replicates are generally smaller, corresponding to smaller slab variances of the MD-LogSS model, on average. When $p \geq 500$ the average prior-to-posterior divergence of τ^2 across all replicates of a given scenario is strictly decreasing in μ_{τ^2} , and $\mu_{\tau^2}^{MD}$ varies across replicates (results not shown). Like the MD-LinBLASSO, the variability in $\mu_{\tau^2}^{MD}$ across replicates of a given scenario may be due to a lack of data-information concerning τ^2 at $n = 100$.

3.3.3 Comparison between minimum-divergence models

In the linear case, the minimum-divergence configuration of τ^2 results in better variable selection at a weak effect size (0.5) than a non-informative prior on τ^2 , but the same is not true for the Bayesian LASSO, where the NI-LinBLASSO outperforms the MD-LinBLASSO in almost all scenarios, with the MD-LinBLASSO often inducing too much shrinkage on β . A noticeable difference in behaviour between the MD-LinSS and MD-LinBLASSO models is in the distribution of their prior means $\mu_{\lambda^2}^{MD}$ and $\mu_{\tau^2}^{MD}$ across realizations of a given scenario. The prior mean that locally minimizes divergence of τ^2 for the spike-and-slab model is generally the same across datasets for a single scenario, where as the prior that locally minimizes divergence of λ^2 may differ between simulated datasets for a single scenario. This difference in behaviour may be a result of difference in data-information regarding the parameters λ^2 , τ^2 since, if a parameter is not identified strongly in the data, it may be difficult to determine a prior for it that conflicts least with the data. Figure 3.3.3 compares the posterior distributions of λ^2 and τ^2 when λ , τ are given the non-informative prior distribution $\pi(\lambda)$, $\pi(\tau) \sim \text{Uniform}(0, 100)$, for the 7th replicate of the the uncorrelated simulation scenario with $p = 200$, $c = 4$ and ± 0.5 effects. Since this prior distribution is highly non-informative, these posterior distributions largely reflect what information the data has regarding λ^2 , τ^2 . We see in Figure 3.3.3 that the posterior distribution of λ^2 is multimodal and quite diffuse relative to posterior of τ^2 . To illustrate why this is problematic for the MD-LinBLASSO model we note firstly that, for this simulated dataset, the MD-LinBLASSO model uses the prior mean $\mu_{\lambda^2}^{MD} = 750$ on λ^2 . The divergence at this mean is, however, only very slightly smaller than at neighbouring means indicating that this is likely not an actual point of minimal

prior-data conflict but is just a point where the prior-to-posterior divergence of λ^2 is underestimated, or its neighbours' overestimated. Due to the wide range of λ^2 values supported by the data we are unable to identify a mean of minimal prior-data conflict and, as a result, choose a prior mean that results in a divergence that is only spuriously small relative to its neighbours. Here $\mu_{\lambda^2}^{MD} = 750$ is unsurprisingly large since a spuriously minimal divergence is likely to occur at large μ_{λ^2} due to stabilization of the divergence. The main issue with the MD-LinBLASSO model is its use of too large a prior mean on λ^2 so instances like this may be responsible for this model's poor performance relative to the NI-LinBLASSO model. The reason why the data seems to support a wide range of λ values is possibly due to the assumption that each $\beta_1, \beta_2, \dots, \beta_p$ in the MD-LinBLASSO is Laplace distributed only when conditioned on the error variance σ^2 , i.e. $\beta_j | \lambda, \sigma \sim \text{Laplace}(\frac{\sigma}{\lambda})$ resulting in $\text{var}(\beta_j | \lambda, \sigma) = 2\sigma^2 / \lambda^2$ for all $j = 1, 2, \dots, p$. Additional preliminary simulations show that when we assume $\beta_j | \lambda \sim \text{Laplace}(\frac{1}{\lambda})$ for all $j = 1, 2, \dots, p$, the MD-LinBLASSO performance improves (results not shown).

Similar to the MD-LinSS model, performance of the MD-LogSS model is comparable or better than its non-informative counterpart in uncorrelated cases when effects are ± 0.5 , though in the logistic case this is restricted only to $p = 200$ scenarios. In contrast to the MD-LinSS model, the value of $\mu_{\tau^2}^{MD}$ in the logistic case varies with different simulated datasets of a single scenario (Figure 3.7), more so than even the value of $\mu_{\lambda^2}^{MD}$ varies in the MD-LinBLASSO. The reason for this may also be a lack of data-information regarding the parameter τ^2 , as seen in a diffuse posterior distribution of τ^2 in the logistic spike-and-slab model when $\pi(\tau) \sim \text{Uniform}(0, 100)$, for the 2nd replicate of the uncorrelated scenario with $p = 200$, $c = 4$, and ± 0.5 effects (Figure 3.3.3). The reason for this is likely due to increased model complexity of the logistic regression model relative to the linear model.

Figure 3.8: Posterior distributions of τ^2 , λ^2 in the linear spike-and-slab and Bayesian LASSO model where $\pi(\tau), \pi(\lambda) \sim \text{Uniform}(0, 100)$, for the 7th replicate of uncorrelated simulation scenario with $p = 200$, $c = 4$, effects $+/- 0.5$. Linear $\pi(\tau^2 | \mathbf{y})$ is scaled down by a factor of 10 for comparability

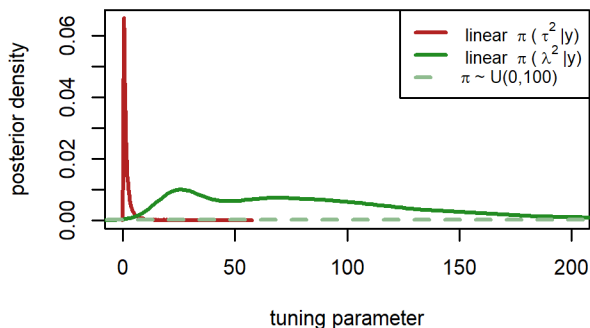
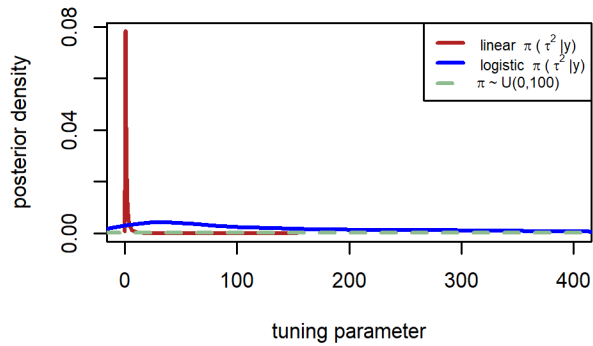


Figure 3.9: Posterior distributions of τ^2 in the linear and logistic spike-and-slab model where $\pi(\tau) \sim \text{Uniform}(0, 100)$, for the 2nd replicate of uncorrelated simulation scenario with $p = 200$, $c = 4$, effects $+/- 0.5$. Linear $\pi(\tau^2|\mathbf{y})$ is scaled down by a factor of 10 for comparability



Chapter 4

Application to a proteomics

COVID-19 dataset

In this chapter we apply the minimum-divergence models outlined in Chapter 2 to proteomic data predicting presence or severity of COVID-19. The data are the same as those used to generate responses for the correlated simulation scenarios in Chapter 3. We predict severity of COVID-19 with linear regression and presence or absence of COVID-19 with logistic regression.

4.1 Overview of data

The data comes from the multi-omics study by Overmyer et al., 2021, where RNA-sequencing and mass spectrometry was performed on blood samples from 128 individuals admitted to the Albany Medical Center in Albany NY with COVID-like respiratory issues between April 6 and May 1 of 2020. This study began less than 3 months after recording of the first COVID-19 case in America and months before development of any vaccines (T. Carvalho et al., 2021). At the time of enrollment patients were tested for the SARS-CoV-2 infection and designated to positive or negative COVID-status groups, with 102 testing positive and 26 testing negative. At this time the alpha variant had already been detected in all 50 states (Webster, 2021), so it is possible that individuals in the study were infected with a variant of the original virus. Various clinical data like age, sex, Charleston comorbidity index, whether the patient was admitted to the intensive care unit, and the number of days spent on a ventilator, were collected. Severity of illness was measured by recording the number of days a patient spent out of hospital in a 45 day period following admittance. If a patient remained in the hospital after the 45th day a score of 0 was given, indicating very

severe illness. Blood samples were drawn at the time of enrollment and mass spectrometry of blood plasma used to obtain abundance measurements of 517 proteins, 646 lipids, and 110 metabolites. RNA-sequencing yielded measurements of 13,263 leukocyte mRNA transcripts (Overmyer et al., 2021). The measurements of all molecules were normalized and transformed with a base 2 logarithm (Overmyer et al., 2021). The proteomic data has dimension most comparable to our simulation so we restrict our analysis to this data set. Due to some missing data we exclude 2 of the COVID patients, leaving 100 patients diagnosed with COVID and 26 without. To improve mixing, we remove all proteins with sample variances of < 0.25 , as in Lipman et al., 2022, leaving 441 proteins for analysis of disease severity and 443 for analysis of disease presence.

4.2 Methods

We use linear and logistic regression to model the relationship between protein measurements and hospital-free days (HFD) and presence of COVID-19, respectively. We apply the MD-LinBLASSO and MD-LinSS models to the linear case, and the MD-LogSS model to the logistic case. We center and scale all proteins to have mean 0 and variance 1, and center HFD's. For the linear model, we assign the prior $\sigma^2 \sim \text{Inverse-Gamma}(0.1, 0.1)$ to the error variance. For the MD-LinSS and MD-LogSS models we assume a prior mixing probability of $\theta \sim \text{Beta}(1, 1)$. Since we center HFD and \mathbf{X} we assume no intercept in the linear model. For the logistic model, we assume the prior $\beta_0 \sim N(0, 10^2)$ on the intercept. Preliminary analysis identified a minimally conflicting region for the prior mean of τ^2 in both the linear and logistic case to be ≤ 10 , so we define the sequence of prior means for the spike-and-slab models to be $\{0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 7.5, 10\}$. For the MD-LinBLASSO, we use the same sequence of prior means as used in the simulation, namely $\{0.1, 0.5, 1, 2.5, 5, 10, 50, 100, 250, 500, 750, 1000\}$. We assume hypervariances for λ^2 , τ^2 of 10, based on their performance in the simulation. We run 2 parallel MCMC chains for each model. For the MD-LinSS and MD-LogSS models, we draw 375,000 and 475,000 posterior samples with the first 25,000 discarded as burn-in for each chain, respectively, to obtain convergence of the sampling algorithm. For the MD-LinBLASSO, we draw 75,000 posterior samples and discard the first 15,000 as burn-in for each chain to obtain convergence.

4.3 Results and discussion

4.3.1 Analysis of COVID severity

We apply both MD-LinSS and MD-LinBLASSO models to the COVID severity data but present and discuss results only for the MD-LinSS model as the Bayesian LASSO fails to make any selection at any of the prior

Figure 4.1: Standardized inverse prior-to-posterior KL Divergences of τ^2 in the linear (left) and logistic (right) spike-and-slab model. Vertical dashed lines denote the first local maximum

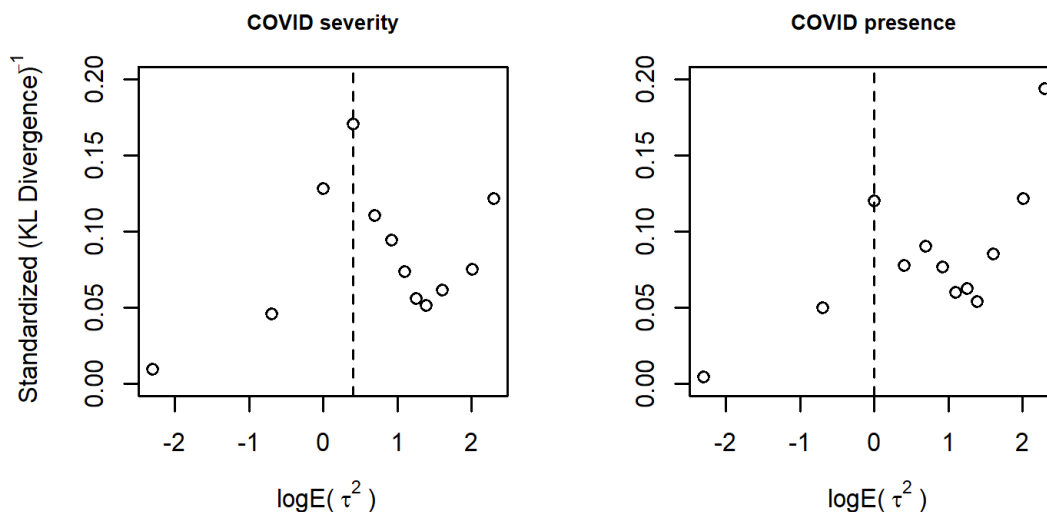
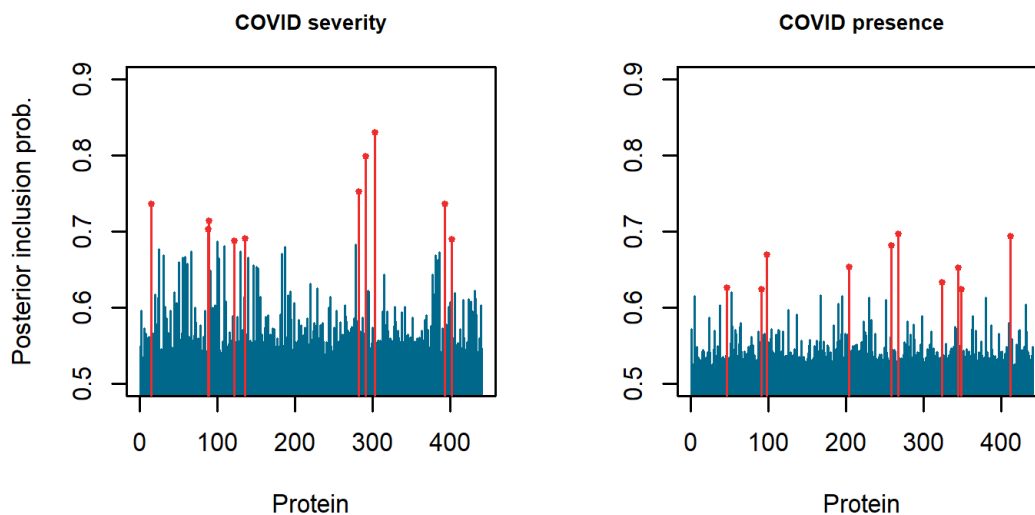


Figure 4.2: Posterior model inclusion probabilities of proteins for the MD-LinSS model (left) and MD-LogSS model (right). Red probabilities denote selected proteins



means considered. Figure 4.1 (left) shows the inverse prior-to-posterior divergences of the slab variance τ^2 for each prior mean $\mu_{\tau^2} \in \{0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 7.5, 10\}$, identifying $\mu_{\tau^2}^{MD} = 1.5$ as the prior mean corresponding to a locally minimal divergence. We plot the inverse divergence as opposed to the divergence as it more clearly identifies this prior mean. The correlation of posterior inclusion probabilities between the two MCMC chains was > 0.98 , indicating good agreement of inclusion probabilities. Figure 4.2 (left) denotes the posterior inclusion probabilities of each protein using the MD-LinSS model. All proteins have inclusion probabilities of > 0.5 meaning that if a cut-off of 0.5 were used, all proteins would be selected

as important. As opposed to this, we select the 10 proteins with highest inclusion probabilities which corresponds to selection with a cut-off of 0.6738. We select proteins P24821, P20742, P15814, Q15166, A0A075B7D0, C9JF17, Q14766, O95445, Q5JP53, and K7ER74, which correspond to genes TNC, PZP, IGLL1, PON3, IGHV1OR15-1, APOD, LTBP1, APOM, TUBB, and APOC4-APOC2. Of these 10 proteins, all but IGLL1 and LTBP1 were also identified as being related to COVID severity in a separate analysis of the same data that used univariable filtering followed by elastic-net stability selection (Lipman et al., 2022). In that analysis, the genes TNC, PZP, APOD, and TUBB were found to be functionally associated with neurological disease, supporting findings that patients infected with the SARS-CoV-2 virus are more likely to develop certain neurological or psychiatric illnesses like ischaemic stroke and anxiety than those infected with influenza viruses, particularly in those who require hospitalization (Lipman et al., 2022; Taquet et al., 2021). Identification of IGLL1, a gene coding for a protein critical in B-cell development (a type of white blood cell) (Gemayel et al., 2016), is consistent with research showing a significant decrease in pre-B cells among patients with severe COVID-19 compared to healthy patients, where IGLL1 is used to characterize clusters of pre-B cells (X. Wang et al., 2021). Selection of the gene LTBP1, involved in the TGF- β pathway, supports other findings that TGF- β is potentially important in COVID symptom control (Wu et al., 2022). The remaining genes were associated with inflammation response, infectious or metabolic disease, or cell function and maintenance, based on the analysis of Lipman et al., 2022.

4.3.2 Analysis of COVID presence

We use the MD-LogSS model to identify proteins associated with COVID-19 status. Figure 4.1 (right) shows the inverse prior-to-posterior divergences of the slab variance τ^2 for each prior mean μ_{τ^2} considered, identifying $\mu_{\tau^2}^{MD} = 1$ as the prior mean corresponding to a locally minimal divergence. Posterior inclusion probabilities between MCMC chains have a correlation of > 0.98 indicating good agreement of inclusion probabilities. Figure 4.2 (right) shows the posterior inclusion probabilities of each protein. As with the MD-LinSS model, all proteins have inclusion probabilities of > 0.5 so we select the proteins with 10 highest inclusion probabilities, corresponding to selection with a cut-off of 0.625. We select proteins corresponding to genes HSPD1, H3C13, IGHV1-8, CSF1R, HRG, LUM, HLA-B, CRTAC1, PLTP, and SFTPB. All selected genes but H3C13 and IGHV1-8 were also found to be associated with COVID status in a separate analysis of the same data (Lipman et al., 2022). As with the COVID severity analysis, some of the selected proteins here (HSPD1, CSF1R, HRG, LUM, HLA-B) are abundant in pathways associated with neurological conditions (Lipman et al., 2022), supporting the growing body of evidence concerning neurological manifestations of SARS-CoV-2 infection. Conditions including encephalitis, encephalopathy, psychosis, stroke, among others,

have been documented in a large number of patients (Krasemann et al., 2022; Mao et al., 2020), in addition to cognitive impairment following infection (Woo et al., 2020). The potential for central nervous system infection has been documented for most coronaviruses like COVID-19, SARS, and MERS (Ellul et al., 2020), but the mechanism and consequences of infection are not entirely understood. Some research shows support for the blood brain barrier as an entry route for infection (Krasemann et al., 2022), and that potential for central nervous system infection can even play a role in respiratory failure (Y.-C. Li et al., 2020). It has been observed that the cognitive impairment experienced by some patients following COVID infection resembles that of those undergoing chemotherapy, and show that even mild infection can result in ongoing neural inflammation, dysregulation of neural cell types, and reduction in myelin (Fernández-Castañeda et al., 2022). The gene IGHV1-8 codes for a protein that allows binding of antibodies to antigens and so is essential in immune response (Gaudet et al., 2011). H3C13 refers to a particular protein of the third histone cluster H3. This cluster is used as a marker to identify remnants of neutrophil extracellular traps (NETs): webs expelled by neutrophils (a type of white blood cell) to contain infections (Zuo et al., 2020). NETs, however, contain many toxic molecules and, when dysregulated, can heighten inflammation and compromise the operation of blood vessels in the lungs of COVID patients, making NETs possible novel targets for COVID treatment (Zuo et al., 2020). The remaining selected proteins are associated with organismal injury, metabolic disease, or cell death and survival (Lipman et al., 2022).

4.3.3 Conclusions

In our analysis we apply the MD-LinSS and MD-LogSS models to the proteomic data and identify the 10 proteins most associated with COVID severity, and the 10 proteins most associated with COVID presence, respectively. The choice to select only the top 10 proteins in either case is a result of posterior inclusion probabilities all exceeding the threshold of 0.5, meaning all proteins are selected if a cut-off of 0.5 is used. Choosing a cut-off > 0.5 can be done visually or using a prespecified Bayesian false-discovery rate (BFDR), where $\text{BFDR}_k = \frac{1}{n_k} \sum_{\{j \mid \bar{\gamma}_j > k\}} (1 - \bar{\gamma}_j)$, with n_k referring to the number of variables selected using the cut-off k . Since the BFDR can be viewed as a kind of average type I error rate, a prespecified BFDR might be ≤ 0.1 . Figure 4.2 shows no obvious cut-off for inclusion and the BFDR is > 0.1 for all cut-offs $0 < k < 1$, so we instead choose the 10 proteins that have highest probability of being associated with COVID presence or severity, depending on the model. The selected proteins are largely consistent with a previous analysis of the same data that used univariable filtering coupled with elastic-net stability selection (Lipman et al., 2022), and the proteins not discovered in the previous analysis (IGLL1 and LTBP1 for linear, IGHV1-8 and H3C13 for logistic) have been shown to be associated with COVID in the literature. For instance, our

analysis is consistent with Wu et al., 2022, where LTBP1 is found to be increased in asymptomatic patients, implying that LTBP1 or the pathway it is associated with, TGF- β , may be a novel target for treatment. The protein coded by gene H3C13 was also a novel protein identified in this data set and is used as a marker to identify the remnants of NETs whose pathogenic role in conditions like sepsis, blood clots, and lung failure have been revealed in the past decade (Zuo et al., 2020). The ratio between neutrophil and lymphocyte has also been identified as a promising tool for distinguishing between severe and non-severe COVID cases and for predicting COVID patient outcomes (B. Zhang et al., 2020). Interestingly, though it was not detected in Overmyer et al., 2021 or Lipman et al., 2022, the gene H3C13 had the 2nd highest model inclusion probability in our model. In both linear and logistic analysis, proteins associated with neurological conditions (TNC, PZP, APOD, and TUBB for linear, HSPD1, CSF1R, HRG, LUM, and HLA-B for logistic) had high probabilities of inclusion, contributing to the growing body of evidence showing that COVID-19 infection is not strictly a respiratory disease but has complex negative neurological and psychiatric effects. We note also that, of the 10 proteins selected by each model, only 2 (APOD and APOM for linear, CSF1R and CRTAC1 for logistic) in each case were present in the 52 relevant proteins identified in the original analysis by Overmyer et al., 2021. These 52 proteins however were selected based on their relevance to COVID severity *and* presence, so it is not surprising that only a few of the proteins we selected were selected in the original analysis.

A limitation of our analysis is that we restrict our attention to the proteomic dataset as opposed to analysis of all omic datasets from this study. Like in previous analyses of this data, significant lipids, metabolomes, and RNA transcripts may be identified and cross-ome correlation analysis done to assess whether selected biomolecules from different datasets are associated with one another. Additionally, we do not assess the relationship between selected proteins and clinical covariates like age, comorbidity, white blood cell count, and others. A strength of our analysis however is that we are able to probabilistically rank the importance of proteins in order to select the most relevant ones.

Chapter 5

Conclusions and future work

In this chapter we summarize the thesis, discuss results obtained in the simulation study, and outline any future work.

5.1 Summary and discussion

Classical statistical methods of variable selection based on information criteria and significance tests are not applicable to high-dimensional regression models for theoretical as well as computational reasons. Bayesian regularized regression, however, provides a viable alternative to classical variable selection that is readily applicable in high-dimensional scenarios. The degree of regularization is controlled by configuration of the shrinkage priors placed on regression coefficients. For the Laplace prior (Bayesian LASSO) and the point-mass spike-and-slab prior, two of the most common shrinkage priors, shrinkage is controlled, in part, by the parameters λ^2 and τ^2 , respectively. When prior knowledge regarding the level of shrinkage is not known, as is often the case, these parameters may be given uninformative priors or estimated from the data. In the Bayesian LASSO, estimating λ^2 , to our knowledge, requires either a costly MC-EM algorithm (Casella, 2001; Park & Casella, 2008) or a stochastic approximation method that is sensitive to a choice of step size parameter (Atchadé, 2011). Estimating τ^2 in the spike-and-slab prior, to our knowledge, requires either computation of $\hat{\beta}_{MLE}$ for all 2^p possible submodels, or using a computationally expensive MC-EM algorithm. In this thesis we propose placing priors on λ^2 and τ^2 and eliciting prior means that aim, for some fixed prior variance, to minimize prior-data-conflict: disagreement between the prior and likelihood's information regarding the unknown parameter. Motivated by a divergence-based check for prior-data conflict, we aim to minimize conflict between the prior of λ^2 or τ^2 and data by finding a prior mean that results in minimal prior-to-posterior Kullback-Leibler divergence. For an increasing sequence of candidate priors means for λ^2 or τ^2 ,

we select the prior mean that corresponds to the first locally minimal divergence, denoted $\mu_{\lambda^2}^{MD}$ or $\mu_{\tau^2}^{MD}$, depending on the model. We don't select the prior mean corresponding to the global minimum because it seems that the prior-to-posterior divergence of λ^2 or τ^2 generally decreases and stabilizes as the prior means μ_{λ^2} or μ_{τ^2} increase.

We assess performance of our proposed minimum-divergence configurations of the Laplace and spike-and-slab priors in linear and logistic regression models (denoted MD-LinBLASSO, MD-LogBLASSO, MD-LinSS, and MD-LogSS) in a simulation study covering multiple high-dimensional scenarios, and then apply the models to real proteomic data predicting severity and presence of COVID-19. In the linear simulation study, the MD-LinSS model is generally comparable to its non-informative counterpart NI-LinSS, but outperforms the other when effect sizes are 0.5 as opposed to 1 and predictors uncorrelated, with the MD-LinSS often obtaining larger TPR's, smaller FDR's, and larger F_1 scores. When the data are correlated, the MD-LinSS still performs slightly better in terms of F_1 score. The MD-LinSS model also outperforms both the LinLASSO and LinEN models in all scenarios in terms of trade-off between true positive and false discovery as seen in larger F_1 scores. This occurs due to the LASSO and Elastic-Net tending to select too many covariates, resulting in a good true positive rate but many false discoveries. The MD-LinSS however only outperforms the linear Spike-and-Slab LASSO when effects are 0.5 regardless of correlation structure of \mathbf{X} , as seen in F_1 scores nearer to 1, and larger TPR's. The MD-LinBLASSO generally does not perform favourably relative to the NI-LinBLASSO due to often times failing to select any covariates, with both models generally failing to make much selection when $p \geq 500$. When $p = 200$ however, the MD-LinBLASSO still provides a better trade off on average of TPR and FDR as seen in a larger mean F_1 score.

Based on our simulations, recommendations regarding which regularized linear regression model to use depend on the data and objectives of the analysis. Assuming $n \approx 100$, if there is relatively little correlation among the predictors X_1, X_2, \dots, X_p , $p \leq 500$, and the size of effects are presumed to be relatively large (≈ 1), then the Spike-and-Slab LASSO provides a very fast way of identifying relevant predictors with a low false discovery and nondiscovery rate. When effects are smaller and predictors relatively uncorrelated, then the choice between Bayesian and non-Bayesian methods rests on the objective of the analysis. If the goal of the analysis is simply to "cast a wide net" and select a group of covariates for further analysis, then penalized regression methods like the LASSO and Elastic-Net are computationally efficient methods for doing so. If, for instance, it is necessary to identify targets for a costly experiment, then greater certainty regarding the selected variables' association with the outcome is needed, meaning a good trade-off between TPR and FDR. In this small-effect situation the MD-LinSS model would be recommended as it obtains the highest average F_1 scores. In addition to this, since the spike-and-slab prior produces inclusion probabilities for predictors, there is a probabilistic method for obtaining, say, the top 5 or 10 *most* relevant predictors, a feature the

Laplace prior or frequentist models do not have. In high-dimensional data like genomic or proteomic data however, there is likely correlation among predictors, with the proteomic data analyzed in Chapter 4 having correlations between predictors as high as 0.98. Depending on the objective of the analysis frequentist or Bayesian methods may be best. If a few variables are to be selected with certainty, then the MD-LinSS, NI-LinSS, or LinSSLASSO models all perform similarly, with the MD-LinSS performing best in terms of F_1 , particularly when effects are smaller. If computation is not an obstacle, our simulations imply that the MD-LinSS model with a relatively fine sequence of prior means $0.1 \leq \mu_{\tau^2} \leq 10$ would likely be best. As before, if it is more important to ensure selection of all relevant predictors than to minimize false discoveries, then the LASSO and Elastic-Net are attractive for their high TPR and computational efficiency.

In the logistic simulation study, the MD-LogSS compares favourably to the NI-LogSS model when $p = 200$ or $p = 517$ and effects ± 0.5 , obtaining a similar or larger F_1 and TPR score, and smaller FDR. Compared to the LogLASSO and LogEN models, the MD-LogSS obtains a larger F_1 score when effects are ± 1 , predictors uncorrelated, and the number of significant covariates < 16 , though this does not hold for comparison with the LogLASSO when $p \geq 500$. When the number of significant covariates is ≥ 16 , the MD-LogSS model is generally outperformed by other models as it tends to select all or almost all covariates often times. In the presence of correlation, the MD-LogSS model outperforms all models when effects are ± 0.5 in terms of average TPR and FDR. In terms of F_1 score, the MD-LogSS model seems to improve with increased sparsity when effects are $+/- 0.5$, but worsen with increased sparsity when effects are $+/- 1$.

Recommendations regarding which regularized logistic regression model to use based on our simulation results are more difficult to provide than in the linear case. Based on the performance metrics, the Bayesian models have a slightly worse trade-off between true positive and false discovery than the frequentist models in most uncorrelated scenarios, a better TPR and FDR in correlated scenarios, and across all scenarios a seemingly high probability of “missing” relevant predictors as seen in a high FNDR. This relatively high FNDR is mainly a result of the Bayesian model’s tendency, particularly the MD-LogSS, to select all covariates for model inclusion (resulting in FNDR= 1), meaning that posterior inclusion probabilities of covariates are often all > 0.5 . However, if all covariates have inclusion probabilities of > 0.5 , this does not mean the model is performing poorly but only that an inclusion cut-off of 0.5 is too low and a larger cut-off should be used. Indeed, even if throughout the sampling process a given X_j , $1 \leq j \leq p$, is selected for inclusion into the model more than half the time, it’s effect in the model at those draws may still be extremely close to 0, since the slab has prior mass around 0. Recommendations regarding use of the MD-LogSS or NI-LogSS models made based on performance of those models at a cut-off of 0.5 then are not meaningful to make, since in any real scenario where all posterior inclusion probabilities exceed 0.5, a larger cut-off would be used, perhaps based on a predefined Bayesian false-discovery rate. Using a larger cut-off than 0.5 may likely make the MD-LogSS

model more competitive against the penalized regression models since, in some simulation scenarios, they already perform equally well. In terms of penalized regression models, if selecting all relevant covariates is more important than mitigating false-discoveries, the Logistic Spike-and-Slab LASSO seems to be preferable to the LASSO or Elastic-Net as it boasts a larger TPR. Note however that all results concerning simulations with $p \geq 500$ are only preliminary as they summarize performance over only 20 replicates as opposed to 100.

It should be noted that though the minimum-divergence models require repeated fits of the data at varying prior means μ_{λ^2} or μ_{τ^2} , these multiple fits may all be done in parallel if there are sufficient computational resources, like a cluster. For instance, in the uncorrelated simulation scenario where $p = 500$, $c = 5$, and effects ± 0.5 , fitting of the MD-LinSS model took on average 3.3 hours using 24 CPU cores (2 chains for each of 12 values of μ_{τ^2}), and fitting of the NI-LinSS model took the same amount of time though it used 2 CPU cores. If, as opposed to using a minimum-divergence or non-informative configuration of τ^2 one wanted to obtain the marginal maximum-likelihood estimate of τ^2 , an MC-EM algorithm would need to be used. In this case the sampler would need to be run possibly hundreds of times *successively*, meaning that these computations could not be parallelized, resulting in possibly hundreds of hours of computation time. Additionally, for the logistic version of the uncorrelated scenario where $p = 500$, $c = 5$, and effects ± 0.5 , fitting the MD-LogSS model took on average 11.1 hours with all computations parallelized, but fitting of the NI-LinSS model took an average of 14.8 hours with computations parallelized meaning that, unless there are insufficient resources to parallelize computations, the MD-LogSS model seems to take less time than the NI-LogSS model to fit.

5.2 Future work

One immediate area of future work includes extending the present simulation study. It would be of interest to assess performance of the minimum-divergence models at cut-offs greater than 0.5, for scenarios with larger p like $p = 1000$, using a finer sequence of prior means for λ^2 or τ^2 , and using a different prior on the mixing weight θ , or even estimating it from the data. Additionally it would be of interest to assess more thoroughly the performance of the MD-LinBLASSO if each β_j , $j = 1, 2, \dots, p$ is assumed Laplace unconditional on the error variance σ^2 . As well, due to the computational cost associated with obtaining marginal maximum-likelihood estimates of λ^2 and τ^2 , we were unable to compare our minimum-divergence versions of the Bayesian LASSO and spike-and-slab prior with the versions using those estimates. Future work could involve seeing, in general, how variable selection is influenced by the degree to which the tuning parameters in shrinkage priors are fixed. To our knowledge, a thorough study of this in a high-dimensional context has not been undertaken. Future work could also involve minimizing conflict between λ^2 or τ^2 and data by

other means, such as the score-based check proposed in Nott et al., 2021, or a modified Kullback-Leibler divergence statistic that does not tend to decrease with an increasing prior mean μ_{λ^2} or μ_{τ^2} . Additionally, computing the actual divergence-based p -value given in (2.13) as opposed to only the statistic could be made computationally feasible if the posterior distribution of λ^2 or τ^2 was estimated using Laplace approximation or Variational Bayes methods: computationally efficient alternatives to MCMC sampling. The Bayesian LASSO and point-mass spike-and-slab prior are only 2 of many shrinkage priors that our minimum-divergence method could be applied to. Other shrinkage priors include the Horseshoe prior (C. M. Carvalho et al., 2010), the Dirichlet-Laplace prior (Bhattacharya et al., 2015), the Double Pareto prior (Armagan et al., 2013), and spike-and-slab priors like Stochastic Search Variable Selection (SSVS) (George & McCulloch, 1993), the Spike-and-Slab LASSO (Ročková & George, 2018), or spike-and-slab priors with nonlocal slabs (Johnson & Rossell, 2012; W. Li & Chekouo, 2022) or g-prior distributed slabs (Chipman et al., 2001; Zellner, 1986). As noted in Section 2.3, our proposed method uses the data to configure the prior and so is not fully Bayesian. It would be meaningful to extend it to the fully Bayesian context, for which the notion of weak informativity of one prior relative to another would be useful (Evans & Jang, 2011). Lastly, the prior-to-posterior divergence of a parameter is closely related to a measure of statistical evidence referred to as the relative belief ratio (Baskurt & Evans, 2013; Evans, 2016). Future work could involve using the relative belief ratio in high-dimensional variable selection problems, as investigated in Evans and Tomal, 2016.

Bibliography

- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer.
- Altman, D. G. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, *19*(4), 453–473.
- Armagan, A., Dunson, D. B., & Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*, *23*(1), 119.
- Atchadé, Y. F. (2011). A computational framework for empirical Bayes inference. *Statistics and Computing*, *21*(4), 463–473.
- Barbieri, M. M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *2*(3), 870–897.
- Baskurt, Z., & Evans, M. (2013). Hypothesis assessment and inequalities for bayes factors and relative belief ratios.
- Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, *91*(433), 109–122.
- Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, *110*(512), 1479–1490.
- Biswas, S., & Lin, S. (2012). Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics*, *68*(2), 587–597.
- Bousquet, N. (2008). Diagnostics of prior-data agreement in applied Bayesian analysis. *Journal of Applied Statistics*, *35*(9), 1011–1029.
- Bühlmann, P., & Van De Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer Science & Business Media.
- Camli, O., Kalaylioglu, Z., & SenGupta, A. (2022). Variable selection in linear-circular regression models. *Journal of Applied Statistics*, 1–22.

- Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe prior for sparse signals. *Biometrika*, *97*(2), 465–480.
- Carvalho, T., Krammer, F., & Iwasaki, A. (2021). The first 12 months of COVID-19: A timeline of immunological insights. *Nature Reviews Immunology*, *21*(4), 245–256.
- Casella, G. (2001). Empirical Bayes Gibbs sampling. *Biostatistics*, *2*(4), 485–500.
- Cassella, G., Gosh, M., Gill, J., & Kyung, M. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411.
- Castillo, I., & Misner, R. (2018). Empirical Bayes analysis of spike and slab posterior distributions. *Electronic Journal of Statistics*, *12*(2), 3953–4001.
- Castillo, I., Schmidt-Hieber, J., & Van der Vaart, A. (2015). Bayesian linear regression with sparse priors.
- Chen, Y.-C. (2017). A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, *1*(1), 161–187.
- Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., & Stine, R. A. (2001). The practical implementation of Bayesian model selection. *Lecture Notes-Monograph Series*, 65–134.
- Chowdhury, M. Z. I., & Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1).
- Cui, W., & George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, *138*(4), 888–900.
- Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just another Gibbs sampler (JAGS) flexible software for MCMC implementation. *Journal of Educational and Behavioral Statistics*, *41*(6), 628–649.
- Donoho, D. L., et al. (2000). High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture*, *1*(2000), 32.
- Ellul, M. A., Benjamin, L., Singh, B., Lant, S., Michael, B. D., Easton, A., Kneen, R., Defres, S., Sejvar, J., & Solomon, T. (2020). Neurological associations of COVID-19. *The Lancet Neurology*, *19*(9), 767–783.
- Evans, M. (2016). Measuring statistical evidence using relative belief. *Computational and structural biotechnology journal*, *14*, 91–96.
- Evans, M., & Jang, G. H. (2011). Weak informativity and the information in one prior relative to another.
- Evans, M., & Moshonov, H. (2006). Checking for prior-data conflict. *Bayesian Analysis*, *1*(4), 893–914.
- Evans, M., & Tomal, J. (2016). Multiple testing via relative belief ratios. *arXiv preprint arXiv:1609.06418*.
- Fan, J., & Fan, Y. (2008). High dimensional classification using features annealed independence rules. *The Annals of Statistics*, *36*(6), 2605.
- Fan, J., & Li, R. (2006). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *arXiv preprint math/0602133*.

- Fan, J., & 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., & Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, *20*(1), 101.
- Fernandez, C., Ley, E., & Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, *100*(2), 381–427.
- Fernández-Castañeda, A., Lu, P., Geraghty, A. C., Song, E., Lee, M.-H., Wood, J., O’Dea, M. R., Dutton, S., Shamardani, K., Nwangwu, K., et al. (2022). Mild respiratory COVID can cause multi-lineage neural cell and myelin dysregulation. *Cell*, *185*(14), 2452–2468.
- Foster, D. P., & George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, *22*(4), 1947–1975.
- Fu, W., & Knight, K. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, *28*(5), 1356–1378.
- Gaudet, P., Livstone, M. S., Lewis, S. E., & Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the Gene Ontology consortium. *Briefings in Bioinformatics*, *12*(5), 449–462.
- Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–534.
- Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman; Hall/CRC.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.
- Gemayel, K. T., Litman, G. W., & Sriaroon, P. (2016). Autosomal recessive agammaglobulinemia associated with an IGLL1 gene missense mutation. *Annals of Allergy, Asthma & Immunology*, *117*(4), 439–441.
- Genking, A., Lewis, D. D., & Madigan, D. (2007). Large-scale Bayesian logistic regression for text categorization. *Technometrics*, *49*(3), 291–304.
- George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 339–373.
- George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.
- George, E., & Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, *87*(4), 731–747.
- Gil, M., Alajaji, F., & Linder, T. (2013). Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, *249*, 124–131.

- Harrell, F. (2001). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression and survival analysis*. (2nd ed.). Springer Series in Statistics.
- Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection—a review and recommendations for the practicing statistician. *Biometrical Journal*, *60*(3), 431–449.
- Hero, A., & Michel, O. J. (1999). Estimation of Rényi information divergence via pruned minimal spanning trees. *Proceedings of the IEEE Signal Processing Workshop on Higher-Order Statistics. SPW-HOS'99*, 264–268.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.
- Huang, A., Xu, S., & Cai, X. (2013). Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. *BMC Genetics*, *14*(1), 1–14.
- Jia, J., & Yu, B. (2010). On model selection consistency of the elastic net when $p > n$. *Statistica Sinica*, *20*, 595–611.
- Johnson, V. E., & Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association*, *107*(498), 649–660.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.
- Krasemann, S., Haferkamp, U., Pfefferle, S., Woo, M. S., Heinrich, F., Schweizer, M., Appelt-Menzel, A., Cubukova, A., Barenberg, J., Leu, J., et al. (2022). The blood-brain barrier is dysregulated in COVID-19 and serves as a CNS entry route for SARS-CoV-2. *Stem Cell Reports*, *17*(2), 307–320.
- Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, 65–81.
- Leng, C., Tran, M.-N., & Nott, D. (2014). Bayesian adaptive Lasso. *Annals of the Institute of Statistical Mathematics*, *66*(2), 221–244.
- Leonardo Egidi, F. P., & Torelli, N. (2022). Avoiding prior–data conflict in regression models via mixture priors. *Canadian Journal of Statistics*, *50*(2), 491–510.
- Li, W., & Chekouo, T. (2022). Bayesian group selection with non-local priors. *Computational Statistics*, *37*(1), 287–302.
- Li, Y.-C., Bai, W.-Z., & Hashikawa, T. (2020). The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. *Journal of Medical Virology*, *92*(6), 552–555.
- Lipman, D., Safo, S. E., & Chekouo, T. (2022). Multi-omic analysis reveals enriched pathways associated with COVID-19 and COVID-19 severity. *PLOS One*, *17*(4), e0267047.

- Maddalena, M., & Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, *32*, 870–897.
- Mallows, C. L. (2000). Some comments on Cp. *Technometrics*, *42*(1), 87–94.
- Mao, L., Jin, H., Wang, M., Hu, Y., Chen, S., He, Q., Chang, J., Hong, C., Zhou, Y., Wang, D., et al. (2020). Neurologic manifestations of hospitalized patients with coronavirus disease 2019 in Wuhan, China. *JAMA Neurology*, *77*(6), 683–690.
- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *72*(4), 417–473.
- Meinshausen, N., Meier, L., & Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, *104*(488), 1671–1681.
- Myles, C., & Wayne, M. (2008). Quantitative trait locus (QTL) analysis. *Nature Education* *1* (1), 208.
- Narisetty, N. N., & Hel, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*, *42*(2), 789–817.
- Nott, D. J., Seah, M., Al-Labadi, L., Evans, M., Ng, H. K., & Englert, B.-G. (2021). Using Prior Expansions for Prior-Data Conflict Checking. *Bayesian Analysis*, *16*(1), 203–231.
- Nott, D. J., Wang, X., Evans, M., & Englert, B.-G. (2020). Checking for Prior-Data Conflict using prior-to-posterior divergences. *Statistical Science*, *35*(2), 243–253.
- Nott, D. J., Yu, Z., Chan, E., Cotsapas, C., Cowley, M. J., Pulvers, J., Williams, R., & Little, P. (2007). Hierarchical Bayes variable selection and microarray experiments. *Journal of Multivariate Analysis*, *98*(4), 852–872.
- O’hara, R. B., & Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, *4*(1), 85–117.
- Overmyer, K. A., Shishkova, E., Miller, I. J., Balnis, J., Bernstein, M. N., Peters-Clarke, T. M., Meyer, J. G., Quan, Q., Muehlbauer, L. K., Trujillo, E. A., et al. (2021). Large-scale multi-omic analysis of COVID-19 severity. *Cell Systems*, *12*(1), 23–40.
- Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.
- Piironen, J., & Vehtari, A. (2017). On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. *Artificial Intelligence and Statistics, PMLR*.
- Ročková, V., & George, E. I. (2018). The spike-and-slab lasso. *Journal of the American Statistical Association*, *113*(521), 431–444.
- Ročková, V., Lesaffre, E., Luime, J., & Löwenberg, B. (2012). Hierarchical Bayesian formulations for selecting variables in regression models. *Statistics in Medicine*, *31*(11-12), 1221–1237.

- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.
- Sheather, S. J. (2004). Density estimation. *Statistical Science*, 588–597.
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310.
- Steyerberg, E. W. (2008). *Clinical prediction models: A practical approach to development, validation, and updating*. Springer Science; Business Media.
- Sugiyama, M., Liu, S., du Plessis, M. C., Yamanaka, M., Yamada, M., Suzuki, T., & Kanamori, T. (2013). Direct divergence approximation between probability distributions and its applications in machine learning. *Journal of Computing Science and Engineering*, 7(2), 99–111.
- Tadesse, M. G., & Vannucci, M. (2021). Handbook of bayesian variable selection.
- Taquet, M., Geddes, J. R., Husain, M., Luciano, S., & Harrison, P. J. (2021). 6-month neurological and psychiatric outcomes in 236 379 survivors of COVID-19: A retrospective cohort study using electronic health records. *The Lancet Psychiatry*, 8(5), 416–427.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1), 267–288.
- van de Geer, S. (2010). L1-regularization in high-dimensional statistical models. *Proceedings of the International Congress of Mathematicians 2010 (ICM 2010) (In 4 Volumes) Vol. I: Plenary Lectures and Ceremonies Vols. II–IV: Invited Lectures*, 2351–2369.
- Vivekananda, R., & Chakraborty, S. (2017). Selection of tuning parameters, solution paths and standard errors for Bayesian lassos. *Bayesian Analysis*, 12(3), 753–778.
- Wang, F., Mukherjee, S., Richardson, S., & Hill, S. M. (2020). High-dimensional regression in practice: An empirical study of finite-sample prediction, variable selection and ranking. *Statistics and Computing*, 30, 697–719.
- Wang, X., Wen, Y., Xie, X., Liu, Y., Tan, X., Cai, Q., Zhang, Y., Cheng, L., Xu, G., Zhang, S., et al. (2021). Dysregulated hematopoiesis in bone marrow marks severe COVID-19. *Cell Discovery*, 7(1), 60.
- Webster, P. (2021). COVID-19 timeline of events. *Nature Medicine*, 27(12), 2054–2055.
- Woo, M. S., Malsy, J., Pöttgen, J., Seddiq Zai, S., Ufer, F., Hadjilaou, A., Schmiedel, S., Addo, M. M., Gerloff, C., Heesen, C., et al. (2020). Frequent neurocognitive deficits after recovery from mild COVID-19. *Brain Communications*, 2(2), fcaa205.
- Wu, S., Xu, Y., Zhang, J., Ran, X., Jia, X., Wang, J., Sun, L., Yang, H., Li, Y., Fu, B., et al. (2022). Longitudinal serum proteome characterization of COVID-19 patients with different severities revealed potential therapeutic strategies. *Frontiers in Immunology*, 13, 893943.
- Xu, X., & Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 4, 909–936.

- Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*.
- Zhang, B., Zhou, X., Zhu, C., Song, Y., Feng, F., Qiu, Y., Feng, J., Jia, Q., Song, Q., Zhu, B., et al. (2020). Immune phenotyping based on the neutrophil-to-lymphocyte ratio and IgG level predicts disease severity and outcome for patients with COVID-19. *Frontiers in Molecular Biosciences*, 7, 157.
- Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty.
- Zhao, P., & Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7, 2541–2563.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.
- Zou, H., & 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.
- Zuo, Y., Yalavarthi, S., Shi, H., Gockman, K., Zuo, M., Madison, J. A., Blair, C., Weber, A., Barnes, B. J., Egeblad, M., et al. (2020). Neutrophil extracellular traps in COVID-19. *JCI Insight*, 5(11).