Statistics and Computing (2025) 35:53
https://doi.org/10.1007/s11222-025-10582-1

ORIGINAL PAPER

Using prior-data conflict to tune Bayesian regularized regression models

Timofei Biziaev[1] · Karen Kopciuk[1,3,4] · Thierry Chekouo[1,2]

Received: 25 December 2023 / Accepted: 3 February 2025
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2025

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 an empirical Bayes configuration 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 in the linear regression model and assess the variable selection performance of our prior configurations through a 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.

Keywords Kullback–Leibler divergence · MCMC · Bayesian variable selection · COVID-19 · High-dimensional data

1 Introduction

Advancements in information technology have made the collection and storage of highly detailed data relatively inexpensive, leading to data with more predictors, p, than observations, n. Such data are referred to as being high-dimensional and examples include biological data like genomic, proteomic, and metabolomic data, high-resolution imaging like magnetic resonance imaging (MRI), hyperspectral satellite imaging, and others (Fan and Li 2006; Donoho et al. 2000).

Whether the modelling purpose is to make accurate predictions or to understand the underlying data-generating process, it is often necessary to identify which predictors are 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 model performance (Fan and Li 2006; Fan and Fan 2008). Additionally, high-dimensional data can be collected and analyzed precisely to identify predictors associated with an outcome as in genome-wide association studies or quantitative trait loci mapping (Myles and Wayne 2008; Fan and Lv 2010). When p exceeds n, variable selection becomes challenging due to severe collinearity of predictors and inapplicability of classical variable selection techniques, such as those based on successive significance tests and information criteria.

Regularized regression is a popular variable selection method applicable in $p > n$ scenarios. In the frequentist framework, regularization is obtained by maximizing a penalized likelihood. In the Bayesian framework, it is

✉ Thierry Chekouo
tchekouo@umn.edu

Timofei Biziaev
timofei.biziaev@ucalgary.ca

Karen Kopciuk
kakopciu@ucalgary.ca

[1] Department of Mathematics and Statistics, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada

[2] Division of Biostatistics and Health Data Science, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA

[3] Cancer Epidemiology and Prevention Research, Cancer Care Alberta, Alberta Health Services, 3395 Hospital Drive N.W., Calgary, AB T2N 5G2, Canada

[4] Department of Oncology, Community Health Sciences, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada

Published online: 20 February 2025

obtained through shrinkage priors: prior distributions that shrink the posterior distribution towards 0. The benefits of Bayesian regularization include automatic uncertainty quantification via the posterior distribution and ranking of importance of variables. Additionally, shrinkage is attained through adjustment of the prior distribution and so fits naturally in a Bayesian framework. Limitations of the Bayesian framework include increased computation time when compared to frequentist methods and the necessity for prior configuration controlling the degree of shrinkage. Indeed, the issue of adequate configuration of shrinkage priors has been investigated by researchers for a variety of priors (Carvalho et al. 2010; Piironen and Vehtari 2017; Fernandez et al. 2001; Chipman et al. 2001; Narisetty and Hel 2014). In this manuscript, we propose a novel empirical Bayes methodology for configuring the degree of shrinkage applied to the Laplace and point-mass spike-and-slab priors.

1.1 Variable selection methods

A variety of frequentist and Bayesian methods for performing data-driven variable selection have been developed. In the frequentist case variable selection can be done by evaluating an information criterion like the Akaike information criterion (AIC) (Akaike 1998) for all possible subsets of predictors, but becomes infeasible for large p as it involves computation of the AIC for 2^p models. Variable selection can also be done using successive significance tests like backward and forward selection (Harrell 2001) but are are inapplicable in $p > n$ scenarios as there are no asymptotically valid p-values when $p > n$ (Meinshausen et al. 2009). Applicable in high-dimensional scenarios however are regularized regression methods like the LASSO (Tibshirani 1996), Elastic-Net (Zou and Hastie 2005), and SCAD (smoothly clipped absolute deviation) (Fan and Li 2001), also referred to as penalized regression methods.

Similar Bayesian methods have been developed. For an arbitrary prior on the space of possible models, the Bayesian information criterion (BIC) is asymptotically equivalent to selecting the model with highest posterior probability (Schwarz 1978). When multiple nested models are compared, as in backwards elimination or forward selection, Bayes factors (Kass and Raftery 1995) or pseudo Bayes factors like the intrinsic Bayes factor (Berger and Pericchi 1996) can be used. Lastly, regularized regression can be obtained using Bayesian shrinkage priors: priors placed on the regression effects $\beta_1, \beta_2, ..., \beta_p$ that shrink the posterior distribution of $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ towards 0 like the Laplace prior, which corresponds to the LASSO. A second Bayesian methodology related to penalized regression is Bayesian variable selection from a decision theoretical perspective (Hahn and Carvalho 2015; Kowal 2022). In this context, variables are selected by minimizing a posterior loss function subject to a penalty for including irrelevant variables and can be used with a variety of priors (Hahn and Carvalho 2015).

In this paper we focus on shrinkage priors and propose a methodology for determining the level of shrinkage based on the data being analyzed, contributing to the literature concerning hyperparameter tuning of shrinkage priors (Carvalho et al. 2010; Piironen and Vehtari 2017; Fernandez et al. 2001; Chipman et al. 2001; Narisetty and Hel 2014; Vivekananda and Chakraborty 2017). We do so by aiming to configure a shrinkage parameter prior that is minimally affected by the data, as measured by the Kullback–Leibler divergence, motivated by checks for prior-data conflict (Evans and Moshonov 2006; Nott et al. 2020). This idea is related, though opposite, to the notion of an objective or reference prior (Berger et al. 2009): a prior which aims to be maximally affected by the data, as measured by a related quantity referred to as the Shannon expected information (Lindley 1956).

1.2 Bayesian regularized regression

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 and Mallick (1998) Spike-and-Slab uses a point mass at 0 spike and a centered normally distributed slab. Stochastic Search Variable Selection (George and 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 (Ročková and George 2018; Castillo and Mismer 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 used in the Bayesian LASSO (Park and Casella 2008), the Horseshoe prior (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 Sect. 2, 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 and Vannucci 2021; Chekouo et al. 2015; Chekouo and Safo 2022; Chekouo et al. 2016). 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 and 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. One of two approaches may be adopted when specifying parameters of a prior: either the data is used to estimate the parameter, or the parameter is elicited using prior knowledge or absence thereof. The former refers to empirical Bayes methods, while the latter to fully Bayes methods. In this paper we adopt the empirical Bayes approach. 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 observed data (Evans and 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, yielding a prior heavily informed by the 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 point mass spike-and-slab prior in the linear regression model.

2 Methods

For a sample of n independent response observations $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)^T$ and $n \times p$ matrix of predictors $\boldsymbol{X} = (X_1, ..., X_p)^T$ where possibly $p > n$, the linear regression model has the form

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

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

2.1 The Bayesian LASSO

The Bayesian LASSO proposed by Park and Casella (2008) is the 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

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

independent for each $j = 1, 2, ..., p$. Equivalently, after integrating out τ_j^2 from (2), $\beta_j|\sigma^2$ 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. Figure 1 (center) shows a centered Laplace density with $\lambda = 1$. It should be noted that unlike the frequentist LASSO, variable selection in the Bayesian LASSO is not automatic, but is done through analysis of the posterior distribution $\pi(\boldsymbol{\beta}|\boldsymbol{y})$. This posterior is not known in closed form but samples can be drawn using a Gibbs sampler as all the full-conditional distributions $\pi(\boldsymbol{\beta}|\boldsymbol{\tau}^2, \sigma^2, \lambda, \boldsymbol{y})$, $\pi(\boldsymbol{\tau}^2|\boldsymbol{\beta}, \sigma^2, \lambda, \boldsymbol{y})$ and $\pi(\sigma^2|\boldsymbol{\beta}, \boldsymbol{\tau}^2, \lambda, \boldsymbol{y})$ are known in closed-form. The posterior sample of $\boldsymbol{\beta}$ can be used to construct a credible interval for each β_j, $j = 1, 2, ..., p$, 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. In general, variable selection performance based on either credible intervals or posterior summary statistics are influenced by the size of the true underlying effects, degree of shrinkage, and threshold selection (size and construction of interval when using credible intervals, threshold for significant effect when using a summary statistic) (Tadesse and Vannucci 2021; van der Pas et al. 2017; Castillo et al. 2015; Lykou and Ntzoufras 2013). Indeed, when independent identical Laplace priors are placed on $\beta_1, \beta_2, ..., \beta_p$ unconditional on error variance σ^2 (corresponding to the frequentist LASSO), posterior credible intervals, while honest (contain the true parameter value), are too large to be used for effective uncertainty quantification (Castillo et al. 2015).

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 Park and Casella (2008);

Casella (2001). The MC-EM algorithm for maximizing the marginal likelihood $L(\lambda|\tilde{\mathbf{y}})$, where $\tilde{\mathbf{y}}$ is the centered response vector, 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 maximum-likelihood estimate (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 and 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 in the Bayesian LASSO, 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 and 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 and 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 $\lambda^2 \sim$ 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 and Lin 2012).

2.2 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 linear regression model (1), we may impose spike-and-slab priors on the regression effects $\beta_1, \beta_2, ..., \beta_p$ to perform variable selection (George and McCulloch 1993). For instance, the point-mass spike-and-slab prior uses a point-mass at 0 spike and a centered normally distributed slab:

$$\begin{aligned} &\beta_j|\gamma_j \sim (1-\gamma_j)\delta_0(\beta_j) + \gamma_j \mathrm{N}(0, \tau^2)\,, \\ &\gamma_j \sim \mathrm{Bernoulli}(\theta_0) \end{aligned} \tag{3}$$

independent for each $j = 1, 2, ..., p$, where δ_0 denotes the Dirac function at 0, γ_j the variable selection indicator for covariate j, where $\gamma_j = 1$ indicates covariate j is selected and 0 otherwise, and θ_0 is the prior probability of feature inclusion. Figure 1 (left) shows a point-mass spike-and-slab density with $\theta_0 = 0.25$ and $\tau^2 = 10$. The point-mass spike-and-slab allows the selection of covariates based on how well their regression effects 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. Hence, the prior of each β_j can be defined as

$$\begin{aligned} &\beta_j|\gamma_j \sim (1-\gamma_j)\mathrm{N}(0, \tau_0^2) + \gamma_j \mathrm{N}(0, \tau_1^2) \\ &\gamma_j \sim \mathrm{Bernoulli}(\theta_0) \end{aligned} \tag{4}$$

independent for each $j = 1, 2, ..., p$, where $\tau_0^2 < \tau_1^2$. Selection using spike and slab priors defined in (3) or (4) gives Stochastic Search Variable Selection (SSVS) where latent variables are used to identify subsets of important features (George and McCulloch 1993). 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_0, which can be viewed as the prior proportion of active covariates. In the absence of prior information regarding the level of sparsity, θ_0 may be given a Beta(1, 1) prior. Moreover, unlike the Bayesian LASSO, spike-and-slab priors (3) or (4) do not undershrink negligible coefficients and overshrink large coefficients as pointed out in Bai et al. (2021) and Ghosh et al. (2016). As with the Bayesian LASSO, variable selection is done through analysis of the marginal posterior distribution $\pi(\boldsymbol{\gamma}|\mathbf{y})$. For some $j \in \{1, 2, ..., p\}$, let $\{\gamma_j^{(1)}, \gamma_j^{(2)}, ..., \gamma_j^{(S)}\}$ be a sample of size S drawn from the marginal 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 and is referred to as the j^{th} variable's posterior inclusion probability. Median probability model selection (Maddalena and Berger 2004) refers to selecting covariates whose posterior inclusion probability exceeds 0.5. A subset of predictors can also be selected based on which

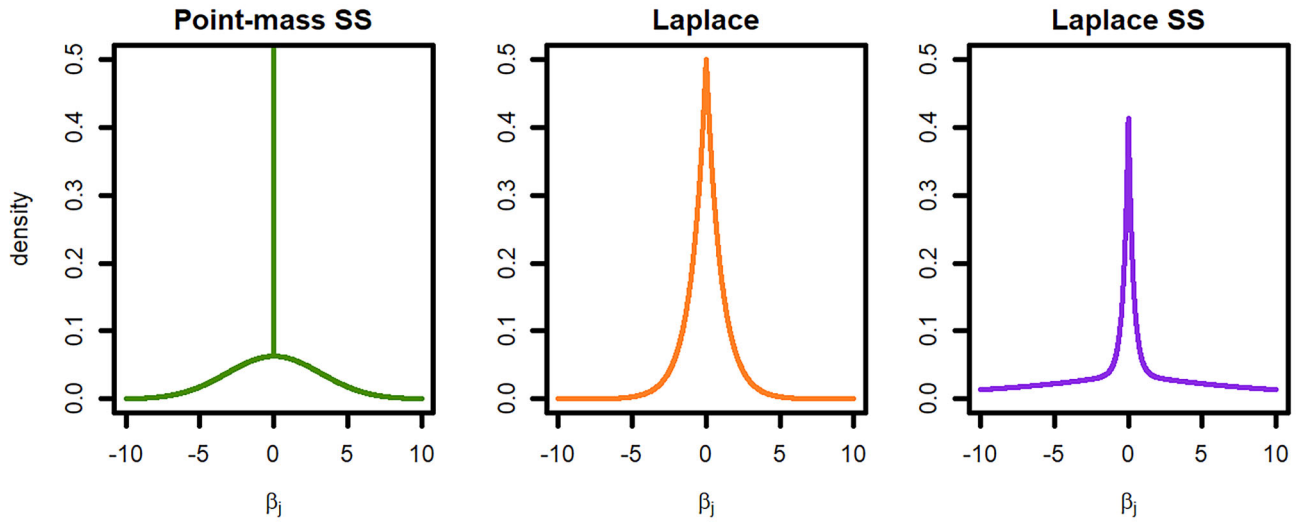

Fig. 1 The point-mass spike-and-slab prior (left) with $\theta_0 = 0.25,\ \tau^2 = 10$, the Laplace prior (used the Bayesian LASSO model, center) with $\lambda = 1$, and the Laplace spike-and-slab prior (used in the Spike-and-Slab LASSO model, right) with $\theta_0 = 0.25,\ \lambda_0 = 3$, and $\lambda_1 = 1$

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 and 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 the treatment of the variance parameters in spike-and-slab models. In SSVS, individual variance parameters must be specified for both the spike (if normal) and slab components and can be specified based on thresholds of practical significance (George and McCulloch 1997). Correlation among regression effects may also be accounted for by using a $\mathrm{N}_p(\mathbf{0}, c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ 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 $\mathrm{N}_p(\mathbf{0}, c\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$ distributed slab, it is possible to obtain marginal maximum-likelihood estimates of c and the mixing weight θ_0 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 $\boldsymbol{X}$ is orthogonal, the marginal likelihood of (c, θ_0) simplifies and can be maximized with numerical methods even when p is moderately large (George and Foster 2000).

The slab variance may not be fixed and given a prior distribution. In the case of SSVS with a normal spike, an exponentially distributed spike variance and slab variance yield the Spike-and-Slab LASSO (Ročková and 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 $g = 1, 2, \ldots, G$ (Xu and 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.

2.3 Prior-data conflict

Let $f(\boldsymbol{y}|\theta)$ 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 and 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 $f(\boldsymbol{y}|\theta)$ assuming $\theta \sim \pi(\theta)$ (Nott et al. 2020) where $\pi(\theta)$ is the prior distribution. Data generated by the sampling model under the prior $\pi(\theta)$ are data drawn from the prior predictive distribution $m(\boldsymbol{y}) = \int f(\boldsymbol{y}|\theta)\pi(\theta)d\theta$. To check for prior-data conflict, Evans and Moshonov (2006) suggest comparing an observed minimal sufficient statistic for θ, $T_0 = T(\boldsymbol{y}_0)$ for observed data $\boldsymbol{y}_0$, 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 $P(m(T(\boldsymbol{y}_0)) > m(T))$. 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 (Egidi et al. 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

$$\begin{aligned} \mathrm{KL}(\boldsymbol{y}) &= \mathrm{KL}(\pi(\theta|\boldsymbol{y}) \,||\, \pi(\theta)) = \\ &= \int \log \frac{\pi(\theta|\boldsymbol{y})}{\pi(\theta)} \pi(\theta|\boldsymbol{y}) d\theta, \end{aligned} \tag{5}$$

where $\pi(\theta|\boldsymbol{y})$ is the posterior distribution of θ. For observed data $\boldsymbol{y}_0$, the check computes the p-value

$$P(\mathrm{KL}(\boldsymbol{Y}) > \mathrm{KL}(\boldsymbol{y}_0)), \tag{6}$$

where $\boldsymbol{Y} \sim m(\boldsymbol{y}) = \int f(\boldsymbol{y}|\theta)\pi(\theta)d\theta$, the prior-predictive distribution. If the p-value is small this indicates that $\mathrm{KL}(\boldsymbol{y}_o)$ is in the tails of the distribution of $\mathrm{KL}(\boldsymbol{Y})$. 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 (6) with a simple example taken from Evans and Moshonov (2006). Suppose $\boldsymbol{y}_0$ is a sample of size n drawn from $\mathrm{N}(\mu, 1)$ and assume the prior $\mu \sim \mathrm{N}(\mu_0, \sigma_0^2)$. The posterior distribution of μ is also normal and so the prior-to-posterior divergence of μ can be obtained in closed-form. Using the divergence between two normal distributions (Gil et al. 2013) and results from Example 3.1 of Nott et al. (2020), we get that

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

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 $\mu \sim \mathrm{N}(5, 1)$. 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 $\pi(\theta|\boldsymbol{y})$ and/or the prior-predictive distribution $m(\boldsymbol{y})$ are intractable, making computation of the p-value (6) challenging. When $m(\boldsymbol{y})$ is not known, the distribution of $\mathrm{KL}(\boldsymbol{Y})$ must be simulated by computing values of $\mathrm{KL}(\boldsymbol{y}^{(i)})$ for a series of draws $\boldsymbol{y}^{(1)}, \boldsymbol{y}^{(2)}, ..., \boldsymbol{y}^{(B)}$ from $m(\boldsymbol{y})$. However when $\pi(\theta|\boldsymbol{y})$ is not known, then $\mathrm{KL}(\boldsymbol{y}^{(i)})$ must be estimated with $\tilde{\mathrm{KL}}(\boldsymbol{y}^{(i)}) = \mathrm{KL}(\tilde{\pi}(\theta|\boldsymbol{y}^{(i)}) \,||\, \pi(\theta))$, where $\tilde{\pi}(\theta|\boldsymbol{y}^{(i)})$ is a density that estimates $\pi(\theta|\boldsymbol{y}^{(i)})$ obtained through a sample from $\pi(\theta|\boldsymbol{y}^{(i)})$.

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 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 it is at least informative to check for prior-data conflict, particularly if the results of the analysis are to be used by others.

2.4 Minimum-divergence prior expectation

The motivation behind prior-data conflict checks is to ensure that an elicited prior is sound in light of the 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. Since such a prior would be heavily dependent on data, what we propose is an empirical Bayes methodology. Assuming there is no reliable prior information regarding shrinkage level, our idea is that a good estimate of the shrinkage parameter may improve the model's 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 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 microarray experiments.

Suppose for the likelihood $f(\boldsymbol{y}|\theta)$ 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 (6) 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 $m(\boldsymbol{y}), \pi(\theta|\boldsymbol{y}, \alpha), \mathrm{KL}(\pi(\theta|\boldsymbol{y}, \alpha) \,||\, \pi(\theta|\alpha))$ are all tractable, then the function $g(\alpha) = P(\mathrm{KL}(\boldsymbol{Y}|\alpha) > \mathrm{KL}(\boldsymbol{y}_0|\alpha))$ 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 (6) 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_0. We focus only on configuring τ^2 for consistency with the Bayesian LASSO as well as because an empirical Bayes estimate of θ_0 can be obtained by numerically optimizing the marginal likelihood of θ_0 (Castillo and Mismer 2018). Assuming the linear regression model (1) and using the check given in (6), if $\boldsymbol{\beta}|\sigma^2$ follows the Laplace prior defined in (2), finding a minimally conflicting fixed value of λ involves assessing conflict between either $\boldsymbol{\tau}^2 = (\tau_1^2, \tau_2^2, ..., \tau_p^2)^T$ and the data or between each τ_j^2, $j = 1, 2, ..., p$ 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. The same issue would arise for the point-mass spike-and-slab prior 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

$$\lambda^2 \sim \mathrm{Gamma}(\mu_{\lambda^2}, \sigma^2_{\lambda^2}) \tag{7}$$

$$\tau^2 \sim \text{Inverse-Gamma}(\mu_{\tau^2}, \sigma^2_{\tau^2}) \ , \tag{8}$$

where $\mu_{\lambda^2}, \sigma^2_{\lambda^2}$ and $\mu_{\tau^2}, \sigma^2_{\tau^2}$ are the prior means and variances of λ^2 and τ^2, respectively. We use a Gamma distributed λ^2 in the Bayesian LASSO as it preserves Gibbs sampling, and an Inverse-Gamma distributed τ^2 in the spike-and-slab model as it is a conjugate prior and a popular configuration (Ročková et al. 2012; Ishwaran and Rao 2003; Malsiner-Walli and Wagner 2018). Ideally, the p-value $P(\mathrm{KL}(\boldsymbol{Y}|\mu) > \mathrm{KL}(\boldsymbol{y}_0|\mu))$ could be maximized with respect to μ, the prior mean of either $\pi(\lambda^2)$ or $\pi(\tau^2)$. As the quantities $\mathrm{KL}(\boldsymbol{y}_0|\mu)$ and $\mathrm{KL}(\boldsymbol{Y}|\mu)$ are intractable, 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 $\{\mu_1, \mu_2, ..., \mu_l\}$ must be specified and a measure of conflict computed for each $\mu_i, i = 1, 2, ..., l$, where l is the length of the sequence. Computing the p-value (6) for just one μ_i requires simulation of the distribution of $\mathrm{KL}(\boldsymbol{Y}|\mu_i)$, which requires fitting many draws from the prior predictive distribution $m(\boldsymbol{y})$. As a computationally feasible alternative, we decide instead to measure the conflict of each μ_i with only the test statistic of the p-value (6), the prior to posterior divergence. While computing a prior-to-posterior divergence for each $\mu_i, i = 1, 2, ..., l$ requires running an MCMC algorithm for each μ_i, it avoids simulation of the distribution $\mathrm{KL}(\boldsymbol{Y}|\mu_i)$ which would require running tens if not hundreds of MCMC algorithms for each μ_i under investigation. So for prior means μ_{λ^2}, μ_{τ^2} we compute

$$\mathrm{KL}(\pi(\lambda^2|\boldsymbol{y}, \mu_{\lambda^2}) \,||\, \pi(\lambda^2|\mu_{\lambda^2})) \tag{9}$$

$$\mathrm{KL}(\pi(\tau^2|\boldsymbol{y}, \mu_{\tau^2}) \,||\, \pi(\tau^2|\mu_{\tau^2})) \tag{10}$$

Our justification for computing only the divergence is that, since we are aiming to compare the level of conflict at different prior means as opposed to testing for significant prior data conflict at a given mean, a prior-to-posterior divergence is informative.

Since the posterior distributions $\pi(\lambda^2|\boldsymbol{y})$, $\pi(\tau^2|\boldsymbol{y})$, and the prior-predictive distribution $m(\boldsymbol{y})$ are intractable, the divergences (9), (10) must be estimated. We first estimate the posterior distributions $\pi(\lambda^2|\boldsymbol{y}, \mu_{\lambda^2})$, $\pi(\tau^2|\boldsymbol{y}, \mu_{\tau^2})$ with standard normal kernel density estimates $\tilde{\pi}(\lambda^2|\boldsymbol{y}, \mu_{\lambda^2})$, $\tilde{\pi}(\tau^2|\boldsymbol{y}, \mu_{\tau^2})$ computed with posterior samples using the function `density()` in R version 4.2.0, and then estimate (9), (10) with the estimated divergence between $\pi(\lambda^2|\mu_{\lambda^2})$ and $\tilde{\pi}(\lambda^2|\boldsymbol{y}, \mu_{\lambda^2})$, and between $\pi(\tau^2|\mu_{\tau^2})$ and $\tilde{\pi}(\tau^2|\boldsymbol{y}, \mu_{\tau^2})$. Kernel density estimation is used as it does not assume a known distribution for the density. For a sample $\{x_1, x_2, ..., x_m\}$ of size m from some distribution $f(x)$, the standard normal kernel density estimate of f is:

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

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 (5).

An issue arises when we observe that the divergences (9), (10) 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. The intuition behind this is that for large values of μ_{λ^2}, μ_{τ^2}, the level of shrinkage is high enough that new observations are interpreted by the model as having occurred due to chance or error, resulting in large posterior error variance σ^2 and a posterior of $\boldsymbol{\beta}$ similar to its prior (see Supplementary Material Section 1.4). We note however that before the divergence sharply decreases a local minimum can usually be observed. As such for some increasing sequence of prior means $\{\mu_1, \mu_2, ..., \mu_l\}$ for either λ^2 (when using the Bayesian LASSO) or τ^2 (when using spike-and-slab regression), we denote the minimum-divergence prior means $\mu_{\lambda^2}^{MD}$, $\mu_{\tau^2}^{MD}$ to be the smallest μ_k, $1 \leq k < l$, such that $\text{KL}_k \leq \text{KL}_{k+1}$, where KL_k refers to the prior-to-posterior KL-divergence of either λ^2 or τ^2 at prior mean μ_k. Equivalently, $\mu_{\lambda^2}^{MD}$, $\mu_{\tau^2}^{MD}$ refer to the prior means corresponding to the first local minimum of the prior-to-posterior divergence (9) or (10). If no such k exists then a finer sequence of prior means may be defined or k is just chosen based on the value that minimizes the KL in the specified sequence. Let the minimum-divergence Bayesian LASSO (MD-BLASSO) denote the Bayesian LASSO using the prior $\lambda^2 \sim \text{Gamma}(\mu_{\lambda^2}^{MD}, \sigma_{\lambda^2}^2)$ where $\sigma_{\lambda^2}^2$ fixed at some small value like 1 or 10. Let the minimum-divergence spike-and-slab (MD-SS) model denote the linear regression model with point-mass spike-and-slab prior where $\tau^2 \sim \text{Inverse-Gamma}(\mu_{\tau^2}^{MD}, \sigma_{\tau^2}^2)$ where $\sigma_{\tau^2}^2$ fixed at some small value like 1 or 10. The general algorithm for identifying the minimum-divergence prior mean for λ^2 in the Bayesian LASSO is given in Algorithm 1. The procedure for τ^2 in the spike-and-slab prior is identical, with λ^2 replaced with τ^2.

Some preliminary analyses may be required to arrive at a suitable sequence of prior means for λ^2 or τ^2. Such analyses can involve assessing prior-to-posterior divergence, convergence of the MCMC algorithm, and variable selection at an initially coarse sequence of means covering a wide range of values. This can identify prior means past which no variables are selected, as well as prior means that may result in convergence issues, like very small values of μ_{λ^2}. With this knowledge, a finer grid of values can be defined until an area of locally minimal divergence is identified.

For a sampling model $f(\boldsymbol{y}|\theta)$ and prior $\pi(\theta)$ on parameter vector θ, prior-data conflict checks are used to check if $\pi(\theta)$ lies in a region where the likelihood for θ is very low. As such, a prior on θ that conflicts minimally with data should place much of its mass in a region where the likelihood for θ is high, i.e. near the maximum-likelihood estimate of θ. An example of this using prior-to-posterior divergences in the Bayesian LASSO can be found in the Supplementary Material (in Section 1.2).

Algorithm 1 Identifying $\mu_{\lambda^2}^{MD}$

Require: Increasing sequence of prior means of λ^2, $\{\mu_1, \mu_2, ..., \mu_l\}$, and hypervariance $\sigma_{\lambda^2}^2$

Ensure: The minimum-divergence prior mean of λ^2, $\mu_{\lambda^2}^{MD}$

1: **for** i in 1 to l **do**

2: Run the MCMC algorithm to obtain posterior samples of λ^2

3: Compute an estimate of the prior-to-posterior KL-divergence of λ^2, denoted $\widehat{\text{KL}}(\mu_i)$

1. Obtain a kernel density estimate of the posterior distribution of λ^2, $\tilde{\pi}(\lambda^2|\mu_i, \boldsymbol{y}_0)$
2. For a sequence of equidistant values of λ^2, $\{\lambda_1^2, \lambda_2^2, ..., \lambda_d^2\}$, obtain the estimated posterior probabilities $\{\tilde{\pi}(\lambda_1^2|\mu_i, \boldsymbol{y}_0), \tilde{\pi}(\lambda_2^2|\mu_i, \boldsymbol{y}_0), ..., \tilde{\pi}(\lambda_d^2|\mu_i, \boldsymbol{y}_0)\}$
3. Compute

$$\widehat{\text{KL}}(\mu_i) = \Delta\lambda^2 \sum_{j=1}^{d} \tilde{\pi}(\lambda_j^2|\mu_i, \boldsymbol{y}_0)\left\{\log \frac{\tilde{\pi}(\lambda_j^2|\mu_i, \boldsymbol{y}_0)}{\pi(\lambda_j^2|\mu_i)}\right\} \quad (11)$$

where $\Delta\lambda^2$ denotes the distance between each λ^2 in $\{\lambda_1^2, \lambda_2^2, ..., \lambda_d^2\}$

4: **end for**

5: Set $\mu_{\lambda^2}^{MD}$ to be the smallest $\mu_k \in \{\mu_1, \mu_2, ..., \mu_{l-1}\}$ such that $\widehat{\text{KL}}(\mu_k) \leq \widehat{\text{KL}}(\mu_{k+1})$. If no such k exists either use k corresponding to the smallest divergence or define a larger/finer sequence of prior means for λ^2, repeating steps 1 to 4.

6: **return** The minimum-divergence prior mean $\mu_{\lambda^2}^{MD}$

3 Simulation

In this Section, we conduct simulation studies to assess the variable selection performance of the MD-BLASSO and MD-SS against their non-informative counterparts. We assess the performance of the different variable selection methods for multiple sparsity levels and number of predictors. 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, as defined in Sect. 3.2.

3.1 Design of the simulation

3.1.1 Competing models

We compare the 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. Let NI-BLASSO denote the Bayesian LASSO with non-informative prior $\lambda^2 \sim \text{Gamma}(\alpha = 10, \beta = 0.1)$ corresponding to $E(\lambda^2) = 100$ and $Var(\lambda^2) = 10^3$, and let NI-SS denote the spike-and-slab model with non-informative prior $\tau \sim \text{Uniform}(0, 100)$ as in Gelman (2006); O'hara and Sillanpää (2009). 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 as continuous responses are centered and predictors centered. The error variance is given the non-informative prior $\sigma^2 \sim$ 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 regression effect β_j has a 95% posterior credible interval that excludes 0. Comparisons with penalized models like the LASSO, Elastic-Net, and the Spike-and-Slab LASSO can be found in Supplemental Material (Tables S1-S5).

3.1.2 Configurations of the data

We simulate continuous responses from the linear regression model (1) with $\sigma^2 = 1$ and $\beta_0 = -1$. We assign $c = s \times p$ nonzero effects to the p-dimensional regression vector $\boldsymbol{\beta}$, where s denotes the sparsity level $s \in \{0.01, 0.02, 0.04, 0.08\}$, with effects of ± 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 from 128 individuals admitted to ICU with COVID-19 like symptoms (Overmyer et al. 2021). We use only the proteomic data to simulate responses for the correlated design with correlation assessed using the Pearson correlation coefficient.

- *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 $X_1, X_2, ..., X_c$ out of a total of p covariates.
- *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 as shown in Fig. 2. The block size is fixed at $B = 10$, and the blocking process follows from 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.

We fix the number of samples 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 (i.e., proteins) to be $c = 5$ and $c = 10$. All configurations of n, p, and c are considered with ± 0.5 small effects. Each scenario is replicated 100 times. Additional scenarios with ± 1 effects and sample sizes different from $n = 100$ can be found in Supplemental Material.

3.2 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

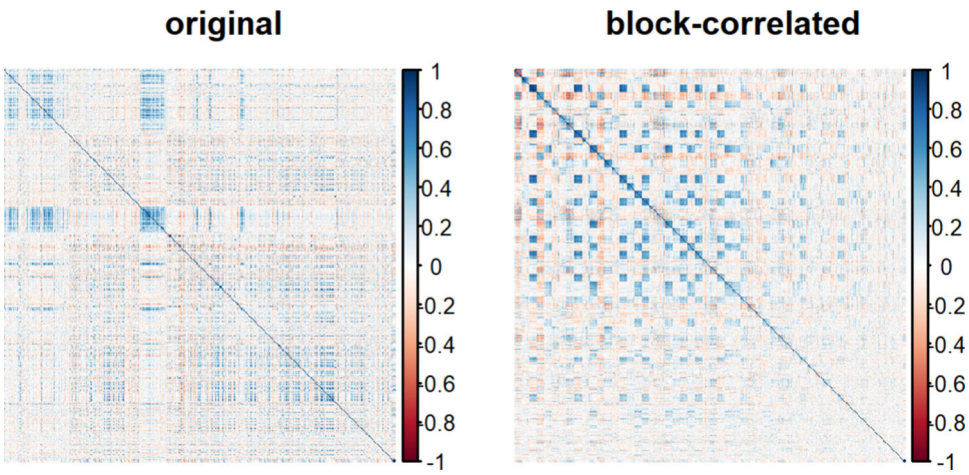

Fig. 2 Correlation plots of the proteomic data. The left plot is of the original unperturbed data, and the right of the perturbed block-correlated data

Table 1 Factors varied and values considered in simulation study

Correlation	n	p	Number of significant covariates c	Effect size
Uncorrelated	100	200	4, 8, 16	±0.5
	100	500	5, 10, 20	±0.5
Correlated	128	517	5, 10	±0.5

of a given scenario:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}; \quad \text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}};$$
$$\text{FNDR} = \frac{\text{FN}}{\text{TN} + \text{FN}}; \quad \text{F}_1 = \frac{2\text{TPR}(1 - \text{FDR})}{\text{TPR} + 1 - \text{FDR}}$$

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).

3.3 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, 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 a computer cluster, 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 which also evaluated the sensitivity of our method to other common choices of kernel density bandwidth (Sheather and Jones 1991; Sheather 2004) for the Bayesian LASSO at this sequence of prior means. The choice of bandwidth did not affect identifying the first local minimum. Results of the sensitivity to bandwidth are in the Supplementary Material (in Sect. 1.3). For the Bayesian LASSO, 30,000 samples of the joint posterior are drawn with the first 10,000 discarded as burn-in for scenarios with $p = 200$. For $p \geq 500$ scenarios, 45,000 posterior samples are drawn with the first 15,000 discarded as burn-in. For 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. All samples refer to samples drawn from each of the two parallel chains. Convergence of continuous parameters is monitored using the potential scale reduction factor, referred to as R-hat (Gelman and Rubin 1992). 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 $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..., \gamma_p)^T$ between the two chains by computing the correlation of the p posterior inclusion probabilities $\bar{\gamma}_1, \bar{\gamma}_2, ..., \bar{\gamma}_p$ from each chain, with a correlation of > 0.98 indicating good agreement. For the Bayesian LASSO, initial values of $\beta_1, \beta_2, ..., \beta_p$ are randomly drawn from independent standard normal distributions, initial values of $\tau_1^2, \tau_2^2, ..., \tau_p^2$ are randomly drawn

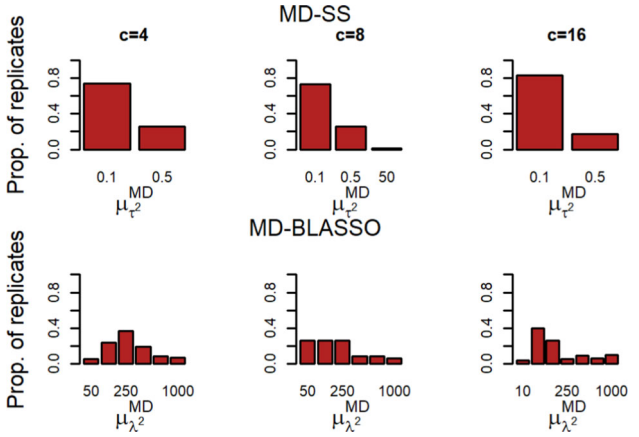

Fig. 3 Distribution of minimum-divergence prior means of τ^2 and λ^2 in the MD-SS and MD-BLASSO models based on 100 replicates. For uncorrelated $p = 200$ scenario with $var(\tau^2) = var(\lambda^2) = 10$

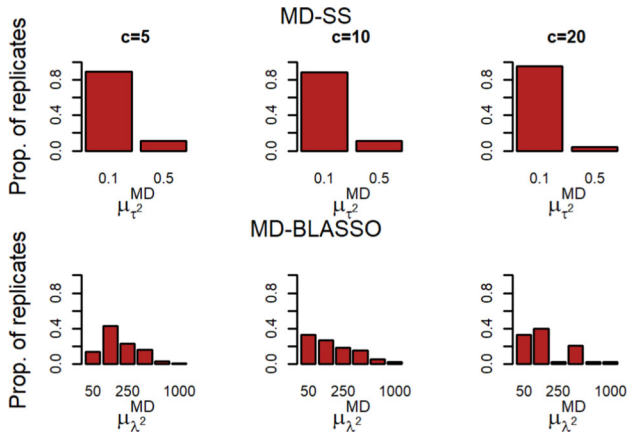

Fig. 4 Distribution of minimum-divergence prior means of τ^2 and λ^2 in the MD-SS and MD-BLASSO models based on 100 replicates. For uncorrelated $p = 500$ scenario with $var(\tau^2) = var(\lambda^2) = 10$

from Exponential(rate $= \frac{1}{2}$), and initial λ^2 are randomly drawn from its prior distribution. For the spike-and-slab models, initial values of $\beta_1, \beta_2, ..., \beta_p$ are randomly drawn from independent standard normal distributions. The initial value of slab-variance τ^2 is drawn from the prior and each indicator $\gamma_1, \gamma_2, ..., \gamma_p$ is initialized from a Bernoulli(0.5) distribution and the mixing weight θ_0 initialized at 0.25 and 0.75 for the first and second chain, respectively.

3.4 Simulation results

Results of the uncorrelated simulation scenarios based on 100 replicates are shown in Tables 2 and 3, and results of the correlated scenarios in Table 4.

Minimum-divergence spike-and-slab: Performance of the minimum-divergence spike-and-slab model (MD-SS) is generally better or as good as the noninformative spike-and-slab model (NI-SS) (Tables 2,3). In most uncorrelated scenarios, the minimum-divergence model outperforms the noninformative model by obtaining a higher TPR, lower FDR, and F_1 score nearer to 1. In the cases where $p = 200$, $c = 4,\ 16$ and $p = 500$, $c = 5$ however, the MD-SS has a slightly 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 4), with the minimum-divergence model obtaining 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-SS model performs as well as the NI-SS model but shows improved performance against the noninformative model when the data are uncorrelated by obtaining a better trade-off between true-positives and false-discoveries as reflected in larger F_1 scores. Additionally, the MD-SS 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). As seen in Figs. 3, 4, the minimally-divergent prior mean for τ^2 was identified as $\mu_{\tau^2}^{MD} = 0.1$ in the majority of replicates across all sparsity levels $s = p \times c$, i.e. $\mu_{\tau^2}^{MD}$ does not depend on s. As the slab variance τ^2 is used to model nonzero effects, and nonzero effects are fixed at ± 0.5 across all scenarios, it is reasonable that the minimum-divergence prior mean $\mu_{\tau^2}^{MD}$ does not vary greatly.

Minimum-divergence Bayesian LASSO: Performance of the minimum-divergence Bayesian LASSO (MD-BLASSO) is generally slightly worse than its noninformative counterpart (NI-BLASSO) in almost all scenarios (Tables 2, 3, 4). The MD-BLASSO on average detects fewer significant covariates than the NI-BLASSO as seen in a smaller TPR, except in the cases where $p = 200$ or $p = 500$ and $c \geq 10$, where it is slightly larger when $var(\lambda^2) = 1$. However when $p = 500$ and $c \geq 10$ the Bayesian LASSO fails to make any selection most of the time. The MD-BLASSO in all cases but one has a higher FDR than the NI-BLASSO and a smaller F_1 score. The FNDR is equally low for both Bayesian LASSO models with the highest being 0.06. Additionally, the performance of the MD-BLASSO was slightly sensitive to the hypervariance $var(\lambda^2)$. For instance, when $p = 200$, $c = 16$, using the hypervariance $var(\lambda^2) = 1$ gave an F_1 score of 0.45 and using $var(\lambda^2) = 10$ gave a score of 0.35. Overall, the MD-BLASSO tended to select fewer covariates than the NI-BLASSO and often failed to make any selection at all, resulting in low TPRs and relatively high FDRs. 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 $\boldsymbol{\beta}$ towards 0. As seen in Figs. 3 and 4, the minimally-divergent prior mean $\mu_{\lambda^2}^{MD}$ varied greatly across replicates

Table 2 Uncorrelated simulation study results for $p = 200$, $c = 4,\ 8$, and 16 using 100 replicates. Each entry refers to the average value of the performance metric and simulation standard error (in parentheses). MD-SS: Minimum-Divergence Spike-and-Slab, NI-SS: Noninformative Spike-and-Slab, MD-BLASSO: Minimum-Divergence Bayesian LASSO, NI-BLASSO: Bayesian LASSO with non-informative prior

		$C = 4$			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-SS	1	0.96 (0.013)	0.22 (0.024)	0.00 (0.000)	0.84 (0.019)
	10	0.96 (0.010)	0.22 (0.023)	0.00 (0.000)	0.84 (0.017)
NI-SS		0.90 (0.024)	0.21 (0.029)	0.00 (0.000)	0.81 (0.025)
MD-BLASSO	1	0.54 (0.034)	0.17 (0.035)	0.01 (0.001)	0.62 (0.032)
	10	0.53 (0.034)	0.20 (0.037)	0.01 (0.001)	0.60 (0.033)
NI-BLASSO		0.77 (0.022)	0.07 (0.012)	0.00 (0.000)	0.82 (0.016)
		$C = 8$			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-SS	1	0.95 (0.009)	0.22 (0.018)	0.00 (0.000)	0.84 (0.013)
	10	0.94 (0.013)	0.23 (0.020)	0.00 (0.000)	0.83 (0.016)
NI-SS		0.88 (0.025)	0.24 (0.026)	0.00 (0.001)	0.79 (0.024)
MD-BLASSO	1	0.45 (0.026)	0.13 (0.029)	0.02 (0.001)	0.56 (0.027)
	10	0.41 (0.028)	0.19 (0.036)	0.02 (0.001)	0.52 (0.031)
NI-BLASSO		0.55 (0.017)	0.03 (0.007)	0.02 (0.001)	0.69 (0.015)
		$C = 16$			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-SS	1	0.90 (0.110)	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-SS		0.80 (0.026)	0.32 (0.025)	0.02 (0.002)	0.70 (0.024)
MD-BLASSO	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-BLASSO		0.29 (0.011)	0.04 (0.013)	0.06 (0.001)	0.43 (0.014)

of any given scenario, which may be due to lack of data-information regarding the parameter λ^2 at the sample size $n = 100$.

3.5 Comparison between minimum-divergence models

The minimum-divergence configuration of τ^2 generally results in better variable selection than a non-informative prior on τ^2, but the same is not true for the Bayesian LASSO, where the NI-BLASSO outperforms the MD-BLASSO in almost all scenarios, with the MD-BLASSO often inducing too much shrinkage on $\boldsymbol{\beta}$. A noticeable difference in behaviour between the MD-SS and MD-BLASSO models is in the distribution of their prior means $\mu_{\lambda^2}^{MD}$ and $\mu_{\tau^2}^{MD}$ across realizations of a given scenario (Figs. 3, 4). 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, whereas the prior that locally minimizes divergence of λ^2 may differ between simulated datasets for a single scenario. This is further seen in Figures S1-S4 (Supplementary material), where the average prior-to-posterior divergence (averaged across replicates) of each prior mean for a given scenario achieves a local minimum for the spike-and-slab model, but generally achieves no local minimum for the Bayesian LASSO model, indicating that a locally minimal divergence is not consistent across replicates of a single scenario for the Bayesian LASSO. This difference in behaviour may be a result of the difference in data information regarding the parameters $\lambda^2,\ \tau^2$ 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 5 compares the posterior distributions of λ^2 and τ^2 when $\lambda,\ \tau$ are given the non-informative prior distribution $\pi(\lambda),\ \pi(\tau) \sim \text{Uniform}(0, 100)$, for the 7^{th} replicate of the uncorrelated simulation scenario with $p = 200,\ c = 4$. Since this prior distribution is highly non-informative, these posterior distributions largely reflect what information the data has regarding $\lambda^2,\ \tau^2$. We see in Fig. 5 that the posterior distribution of λ^2 is multimodal and quite diffuse relative to the posterior of τ^2. To illustrate why this is problematic for the MD-BLASSO model we note firstly that, for this simulated dataset, the MD-BLASSO 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

Table 3 Uncorrelated simulation study results for $p = 500$, $c = 5,\ 10$, and 20 using 100 replicates. Each entry refers to the average value of the performance metric and simulation standard error (in parentheses). MD-SS: Minimum-Divergence Spike-and-Slab, NI-SS: Noninformative Spike-and-Slab, MD-BLASSO: Minimum-Divergence Bayesian LASSO, NI-BLASSO: Bayesian LASSO with non-informative prior

		$C = 5$			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-SS	1	0.91 (0.017)	0.22 (0.019)	0.00 (0.000)	0.82 (0.016)
	10	0.91 (0.017)	0.22 (0.019)	0.00 (0.000)	0.82 (0.017)
NI-SS		0.88 (0.025)	0.20 (0.025)	0.00 (0.000)	0.82 (0.023)
MD-BLASSO	1	0.21 (0.018)	0.33 (0.047)	0.01 (0.000)	0.31 (0.024)
	10	0.18 (0.019)	0.41 (0.049)	0.01 (0.000)	0.27 (0.025)
NI-BLASSO		0.23 (0.018)	0.28 (0.045)	0.01 (0.000)	0.33 (0.024)
		$C = 10$			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-SS	1	0.77 (0.023)	0.28 (0.022)	0.00 (0.000)	0.73 (0.020)
	10	0.76 (0.021)	0.27 (0.019)	0.00 (0.000)	0.73 (0.018)
NI-SS		0.69 (0.036)	0.30 (0.035)	0.01 (0.001)	0.66 (0.033)
MD-BLASSO	1	0.06 (0.008)	0.57 (0.050)	0.02 (0.000)	0.10 (0.013)
	10	0.05 (0.008)	0.63 (0.049)	0.02 (0.000)	0.09 (0.014)
NI-BLASSO		0.06 (0.008)	0.55 (0.050)	0.02 (0.000)	0.10 (0.013)
		$C = 20$			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-SS	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-SS		0.39 (0.028)	0.53 (0.034)	0.05 (0.014)	0.39 (0.026)
MD-BLASSO	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-BLASSO		0.01 (0.002)	0.74 (0.044)	0.04 (0.000)	0.03 (0.005)

Table 4 Correlated simulation study results for $p = 517$, $c = 5$ and 10 using 100 replicates. Each entry refers to the average value of the performance metric and simulation standard error (in parentheses). MD-SS: Minimum-Divergence Spike-and-Slab, NI-SS: Noninformative Spike-and-Slab, MD-BLASSO: Minimum-Divergence Bayesian LASSO, NI-BLASSO: Bayesian LASSO with non-informative prior

		$C = 5$			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-SS	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-SS		0.41 (0.024)	0.46 (0.034)	0.01 (0.000)	0.40 (0.022)
MD-BLASSO	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-BLASSO		0.01 (0.003)	0.97 (0.017)	0.01 (0.000)	0.01 (0.001)
		$C = 10$			
Model	$Var(\tau^2)$	TPR	FDR	FNDR	F1
MD-SS	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-SS		0.44 (0.018)	0.50 (0.027)	0.02 (0.011)	0.41 (0.017)
MD-BLASSO	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-BLASSO		0.01 (0.003)	0.89 (0.032)	0.02 (0.000)	0.02 (0.006)

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 as λ^2 increases. The main issue with the MD-BLASSO 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-BLASSO model. The reason why the data seems to support a wide range of λ values is possibly due to the assumption that each $\beta_1, \beta_2, ..., \beta_p$ in the MD-BLASSO 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 $var(\beta_j|\lambda, \sigma) = 2\sigma^2/\lambda^2$ for all $j = 1, 2, ..., p$. Preliminary simulations show that when we assume $\beta_j|\lambda, \sim \text{Laplace}(\frac{1}{\lambda})$ for all $j = 1, 2, ..., p$, the MD-BLASSO performance improves (results not shown). Another possible reason for this may be that τ^2 in the spike-and-slab prior controls the effect size variability of active covariates, whereas λ^2 in the Laplace prior is used in modelling the effect size variability of all covariates. There may be less data information regarding the average effect size of all covariates than those of covariates that can be distinguished from 0. Lastly, selection using the Bayesian LASSO is generally worse than with a spike-and-slab prior, with the Bayesian LASSO models having lower TPR's across all scenarios and larger FDR's when $p \geq 500$, possibly relating to the coverage of posterior credible intervals arising from the Laplace prior (Bhadra et al. 2019), and the shrinking issue of Bayesian LASSO mentioned in Sect. 2.2.

Additional simulation scenarios with effects of ± 1 and $p = 200$, $c = 4$ active covariates were explored to determine whether the MD-BLASSO model could outperform the non-informative Bayesian LASSO. We find that if the prior variance on λ^2 is set to 0.005 as opposed to 1 or 10, the MD-BLASSO is better able to select relevant covariates and achieves a lower false-discovery rate than the NI-BLASSO when the number of observations is reduced to $n = 50$ (Table 5). This illustrates that a small hypervariance of the tuning parameter may allow the divergence to better detect the subtle differences in possible prior-data conflict that arise when the tuning parameter λ^2 is not strongly identified in the data. Similar performance is seen when n is set to 200 and c increased to 16 (see Supplementary Material).

4 Application to a proteomics COVID-19 dataset

In this Section, we apply the minimum-divergence spike-and-slab model to proteomic data predicting the severity of COVID-19. The data are the same as those used to generate responses for the correlated simulation scenarios in section 3.

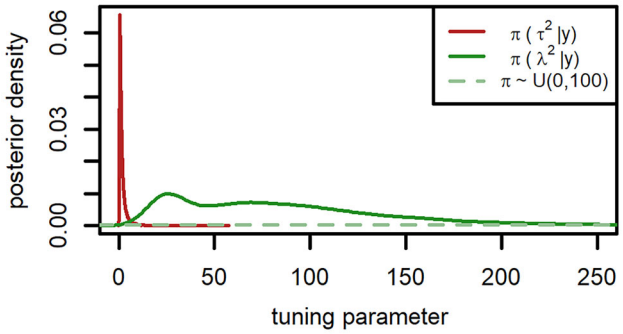

Fig. 5 Posterior distributions of τ^2, λ^2 in the spike-and-slab and Bayesian LASSO model where $\pi(\tau)$, $\pi(\lambda) \sim \text{Uniform}(0, 100)$, for the 7th replicate of uncorrelated simulation scenario with $p = 200$, $c = 4$. $\pi(\tau^2|\boldsymbol{y})$ is scaled down by a factor of 10 for comparability

4.1 Overview of proteomic data

The data comes from the multi-omics study by Overmyer et al. (2021), where RNA-sequencing and mass spectronomy were performed on blood samples from 128 individuals admitted to the Albany Medical Center in Albany NY with COVID-19-like respiratory issues between April 6 and May 1 of 2020. This study began less than 3 months after the recording of the first COVID-19 case in America and months before the development of any vaccines (Carvalho et al. 2021). At the time of enrollment, patients were tested for the SARS-CoV-2 infection, of whom 102 tested positive. 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. The severity of illness was measured by recording the number of days a patient spent out of the 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 spectronomy of blood plasma was 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 a 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-19 patients, leaving 100 patients diagnosed with COVID-19. To improve mixing, we remove all proteins with sample variances of

Table 5 Uncorrelated simulation study results for $n = 50$, $p = 200$, $c = 4$ significant effects of size ± 1 using 100 replicates. Each entry refers to the average value of the metric and simulation standard error (in parentheses). MD-BLASSO: Minimum-Divergence Bayesian LASSO, NI-BLASSO: Bayesian LASSO with non-informative prior

Model	$Var(\lambda^2)$	TPR	FDR	FNDR	F1
MD-BLASSO	0.005	0.39 (0.029)	0.22 (0.042)	0.01 (0.001)	0.49 (0.032)
NI-BLASSO	10^3	0.24 (0.023)	0.39 (0.049)	0.02 (0.000)	0.33 (0.029)

< 0.25, as in Lipman et al. (2022), leaving 441 proteins for analysis.

4.2 Methods

We use linear regression to model the relationship between protein measurements and hospital-free days (HFD). We apply only the MD-SS model as the Bayesian LASSO fails to make any selection regardless of prior configuration. We center and scale all proteins to have mean 0 and variance 1, and center HFDs. We assign the prior $\sigma^2 \sim$ Inverse-Gamma(0.1, 0.1) to the error variance and assume a prior mixing probability of $\theta_0 \sim$ Beta(1, 1) for the MD-SS model. Since we center HFDs and $\boldsymbol{X}$ we assume no intercept. Preliminary analysis identified a minimally conflicting region for the prior mean of τ^2 to be ≤ 10, so we define the sequence of prior means for the minimum-divergence spike-and-slab model to be $\{0.1, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 5, 7.5, 10\}$. We assume a hypervariance for τ^2 of 10, based on simulation performance. We run 2 MCMC chains for each prior mean, drawing 375, 000 posterior samples with the first 25, 000 discarded as burn-in for each chain to obtain convergence of the sampling algorithm. As in the simulation study, all computations are done in parallel in a computer cluster, resulting in no increase in computation time for repeated fits.

4.3 Results

Figure 6 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, with a total run-time of 13 hours and 3 minutes. Figure 7 denotes the posterior inclusion probabilities of each protein using the MD-SS 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. Figure 7 shows no obvious larger cut-off for inclusion and the Bayesian false-discovery rate

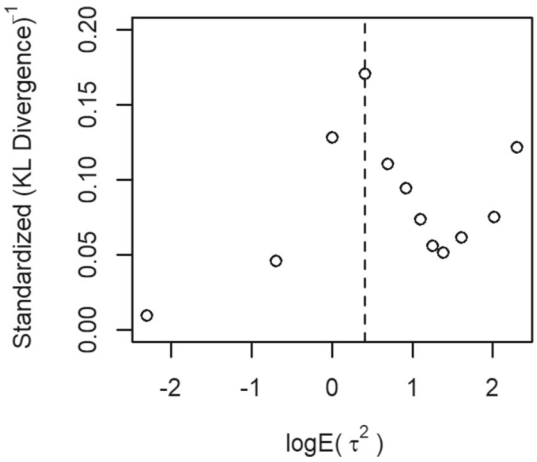

Fig. 6 Standardized inverse prior-to-posterior KL Divergences of τ^2. Vertical dashed lines denote the first local maximum

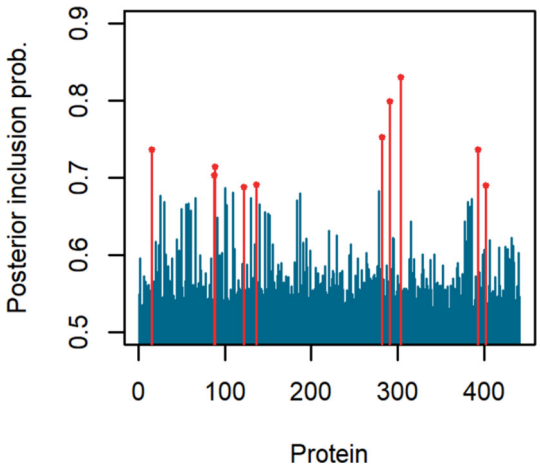

Fig. 7 Posterior model inclusion probabilities of proteins for the MD-SS model. Red probabilities denote selected proteins

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-19 severity, corresponding to 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-19 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 (Taquet et al. 2021; Lipman et al. 2022). 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 (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-19 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).

For comparison, we also apply the NI-SS model and find that, of the top 10 proteins it identifies as associated with severe COVID-19 infection, 2 differ from those identified by the MD-SS model. These correspond to genes HLA-C and APOM, both of which have been shown to be down-regulated in patients with severe COVID infection when compared to those with no or mild infection (Vigón et al. 2022; Shen et al. 2020; Overmyer et al. 2021).

A limitation of our analysis is that we restrict our attention to the proteomic dataset as opposed to the analysis of all omic datasets from this study. Like 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 can probabilistically rank the importance of proteins to select the most relevant ones.

5 Conclusions

Bayesian regularization through shrinkage priors requires specification of the level of shrinkage. One way to achieve this is through empirical Bayes methods, where parameters are estimated from data. Empirical Bayes estimates of the shrinkage parameters τ^2 and λ^2 in the spike-and-slab and Bayesian LASSO priors, respectively, require, to our knowledge, either a computationally expensive MC-EM algorithm or stochastic approximation with the latter option sensitive to a step-size parameter (Leng et al. 2014). As an alternative, we propose placing a narrow prior on τ^2 or λ^2 with a prior mean that results in minimal prior-to-posterior Kullback–Leibler divergence. Simulation results show that the minimum-divergence configuration of the spike-and-slab linear regression model outperforms a non-informative configuration for a weak effect size, providing evidence for the use of empirical Bayes methods in high-dimensional variable selection problems. For the Bayesian LASSO the minimum-divergence configuration largely results in over-shrinkage of regression effects, possibly due to λ^2 often being weakly identified in the data.

The performance of the MD-BLASSO model illustrates a limitation of the minimum-divergence methodology, namely, that the prior-to-posterior divergence may not be able to detect the subtle differences in possible prior-data conflict that arise when the parameter being tuned is not strongly identified in the data. While using a small hypervariance of $var(\lambda^2) = 0.005$ may remedy this issue, more investigation is needed to identify when the minimum-divergence configuration of the Bayesian LASSO is beneficial. Additionally, selection using the Bayesian LASSO is done here using 95% credible intervals, but credible intervals arising from shrinkage priors are not necessarily honest (contain the true parameter value) (Bhadra et al. 2019). Indeed, Li (1989) show there is fundamental disagreement between honesty and adaptation to sparsity, and Van Erp et al. (2019) illustrate in simulation that a credible interval of 95% is often too wide for variable selection when using the Bayesian LASSO. A drawback of our proposed methodology is that it requires running as many MCMC algorithms as there are prior means for λ^2 or τ^2 under consideration. Though this is less computationally intensive than computing the prior-data conflict p-value given in (6) for each prior mean, computation of (6) would make our proposed methodology more rigorous, in particular when the sample size is small.

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, which would be a meaningful area of future work. Another meaningful comparison would be with variable selection done from a Bayesian decision theoretical perspective, which does not necessitate the use of shrinkage priors (Hahn and Carvalho 2015). 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). Additionally, computing the actual divergence-based p-value given in (6) 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.

The Bayesian LASSO and point-mass spike-and-slab prior are only two of many shrinkage priors that our minimum-divergence method could be applied to. One extension would be to a grouped variable selection scenario using group-level indicator variables, as in Li and Chekouo (2022), with shrinkage priors placed on individual effects. For a spike-and-slab prior we may use either a common shrinkage parameter τ^2 or a group-specific shrinkage parameter τ_g^2, $g = 1, ..., G$, where g represents the group index, and G is the number of groups. In the first case our algorithm is directly applicable, while in the second case a grid of prior specifications of $\{\tau_g^2, g = 1, ..., G\}$ could be defined. As our methodology requires running an MCMC algorithm for each prior specification, this would be computationally expensive which, as mentioned earlier, is a drawback of our approach. Another application is to multiple-response data, where a different level of shrinkage would be defined for each response. The novel R2D2 prior (Zhang et al. 2022) induces shrinkage through a distribution on R^2, the coefficient of determination. This prior contains only one tuning parameter, b, and so could be configured with minimum-divergence by assessing the prior-to-posterior divergence of R^2 at different values of b. Our simulation tested performance of our methodology at a sample size of $n = 100$, but assessing performance at smaller and larger sample sizes would be of interest. A limitation of our work, as in empirical Bayes methods, is that it uses the same data to estimate hyperparameters as it does to derive a posterior distribution for inference; it would be meaningful to extend our method to a fully Bayesian context, for which the notion of weak informativity of one prior relative to another would be useful (Evans and 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 and 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).

Supplementary information

Additional plots and further simulation results can be found in the Supplementary Material. Tables S1 and S3 are expanded versions of Tables 2 and 3 that include the performance of penalized regression models like the LASSO (Tibshirani 1996), Elastic-Net (Zou and Hastie 2005), and the Spike-and-Slab LASSO (Ročková and George 2018). For the LASSO and Elastic-Net, tuning parameters are selected via 10-fold cross validation, and a mixing weight of $\alpha = 0.6$ is used in the Elastic-Net, encouraging a light grouping effect. Cross-validation and estimation are done in the R package `glmnet`. For the Spike-and-Slab LASSO a sequence of spike penalty parameters $\lambda_0 \in \{5, 6, 7, ..., 100\}$ is used, which is similar to the sequence used in the simulation of the original paper (Ročková and George 2018), and a slab penalty parameter fixed at the default $\lambda_1 = 1$. Estimation is done with the R package `SSLASSO`. Tables S2 and S4 are simulation study results for additional uncorrelated simulation scenarios where the active effect size is ± 1 as opposed to ± 0.5, including the performance of penalized regression models. Figures S1-S4 show the average prior-to-posterior divergences of λ^2 or τ^2 across replicates of a scenario, similar to Fig. 6. Simulation results for the Bayesian LASSO at samples sizes different from $n = 100$, sensitivity analysis of the prior-to-posterior KL-divergence to kernel density bandwidth, KL-divergence of λ^2 for large prior means μ_{λ^2}, and insight into the relationship between minimal divergence of λ^2 and the MMLE of λ^2 are also provided. Included also is R code for performing linear spike-and-slab regression with $\tau^2 \sim$ Inverse-Gamma$(\mu_{\tau^2}, \sigma_{\tau^2})$ and Beta distributed prior inclusion probability, as well as implementing the Bayesian LASSO with $\lambda^2 \sim$ Gamma$(\mu_{\lambda^2}, \sigma_{\lambda^2})$, both using JAGS interfaced with the R package `runjags`.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-025-10582-1.

Acknowledgements Thierry Chekouo was supported by an NSERC Discovery Grant (RGPIN-2019-04810), a start-up grant offered by the University of Minnesota, and a National Institutes of Health (NIH) grant: 1R35GM150537-01. Thierry Chekouo thanks Medtronic Inc. for their support in the form of a faculty fellowship. Karen Kopciuk was supported by the Canadian Institutes for Health Research (Primary Investigators Kopciuk, Robson, and Shack), grant number PJT-148774. The authors wish to thank Prof. Michael Evans from the University of Toronto for his thoughtful suggestions on the thesis work that this manuscript is based on.

References

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Selected papers of Hirotugu Akaike, pp. 199–213. Springer (1998)

Armagan, A., Dunson, D.B., Lee, J.: Generalized double Pareto shrinkage. Stat. Sin. **23**(1), 119 (2013)

Atchadé, Y.F.: A computational framework for empirical Bayes inference. Stat. Comput. **21**(4), 463–473 (2011)

Bai, R., Ročková, V., George, E.I.: Spike-and-slab meets lasso: a review of the spike-and-slab lasso. In: Tadesse, M.G., Vannucci, M. (eds.) Handbook of bayesian variable selection. CRC Press, Boca Raton, FL (2021)

Barbieri, M.M., Berger, J.O.: Optimal predictive model selection. Ann. Stat. **2**(3), 870–897 (2004)

Baskurt, Z., Evans, M.: Hypothesis assessment and inequalities for bayes factors and relative belief ratios (2013)

Berger, J.O., Bernardo, J.M., Sun, D.: The formal definition of reference priors. Ann. Stat. **37**(2), 905–938 (2009)

Berger, J.O., Pericchi, L.R.: The intrinsic Bayes factor for model selection and prediction. J. Am. Stat. Assoc. **91**(433), 109–122 (1996)

Bhadra, A., Datta, J., Polson, N.G., Willard, B.: Lasso meets horseshoe. Stat. Sci. **34**(3), 405–427 (2019)

Bhattacharya, A., Pati, D., Pillai, N.S., Dunson, D.B.: Dirichlet-Laplace priors for optimal shrinkage. J. Am. Stat. Assoc. **110**(512), 1479–1490 (2015)
Biswas, S., Lin, S.: Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age-related macular degeneration. Biometrics **68**(2), 587–597 (2012)
Bousquet, N.: Diagnostics of prior-data agreement in applied Bayesian analysis. J. Appl. Stat. **35**(9), 1011–1029 (2008)
Camli, O., Kalaylioglu, Z., SenGupta, A.: Variable selection in linear-circular regression models. J. Appl. Stat., 1–22 (2022)
Carvalho, C.M., Polson, N.G., Scott, J.G.: The horseshoe prior for sparse signals. Biometrika **97**(2), 465–480 (2010)
Carvalho, T., Krammer, F., Iwasaki, A.: The first 12 months of COVID-19: a timeline of immunological insights. Nat. Rev. Immunol. **21**(4), 245–256 (2021)
Casella, G.: Empirical Bayes Gibbs sampling. Biostatistics **2**(4), 485–500 (2001)
Castillo, I., Mismer, R.: Empirical Bayes analysis of spike and slab posterior distributions. Electron. J. Stat. **12**(2), 3953–4001 (2018)
Castillo, I., Schmidt-Hieber, J., van der Vaart, A.: Bayesian linear regression with sparse priors. Ann. Stat. **43**(5), 1986–2018 (2015)
Chekouo, T., Safo, S.E.: Bayesian integrative analysis and prediction with application to atherosclerosis cardiovascular disease. Biostatistics **24**(1), 124–139 (2022)
Chekouo, T., Stingo, F.C., Doecke, J.D., Do, K.-A.: miRNA-target gene regulatory networks: a Bayesian integrative approach to biomarker selection with application to kidney cancer. Biometrics **71**(2), 428–438 (2015)
Chekouo, T., Stingo, F.C., Guindani, M., Do, K.-A.: A bayesian predictive model for imaging genetics with application to schizophrenia. The Ann. Appl. Stat. **10**(3), 1547–1571 (2016)
Chen, Y.-C.: A tutorial on kernel density estimation and recent advances. Biostat. Epidemiol. **1**(1), 161–187 (2017)
Chipman, H., George, E. I., McCulloch, R. E., Clyde, M., Foster, D. P., Stine, R. A.: The practical implementation of Bayesian model selection. Lecture Notes-Monograph Series, 65–134 (2001)
Cui, W., George, E.I.: Empirical Bayes vs. fully Bayes variable selection. J. Stat. Plann. Inference **138**(4), 888–900 (2008)
Donoho, D.L., et al.: High-dimensional data analysis: The curses and blessings of dimensionality. AMS math challenges lecture **1**(2000), 32 (2000)
Egidi, L., Pauli, F., Torelli, N.: Avoiding prior-data conflict in regression models via mixture priors. Can. J. Stat. **50**(2), 491–510 (2022)
Evans, M.: Measuring statistical evidence using relative belief. Comput. Struct. Biotechnol. J. **14**, 91–96 (2016)
Evans, M., Jang, G.H.: Weak informativity and the information in one prior relative to another. Stat. Sci. **26**(3), 423–439 (2011)
Evans, M., Moshonov, H.: Checking for prior-data conflict. Bayesian Anal. **1**(4), 893–914 (2006)
Evans, M., Tomal, J.: Multiple testing via relative belief ratios (2016)
Fan, J., Fan, Y.: High dimensional classification using features annealed independence rules. Ann. Stat. **36**(6), 2605 (2008)
Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
Fan, J., Li, R.: Statistical challenges with high dimensionality: feature selection in knowledge discovery. arXiv preprint arXiv:math/0602133 (2006)
Fan, J., Lv, J.: A selective overview of variable selection in high dimensional feature space. Stat. Sin. **20**(1), 101 (2010)
Fernandez, C., Ley, E., Steel, M.F.: Benchmark priors for Bayesian model averaging. J. Econ. **100**(2), 381–427 (2001)
Gelman, A.: Prior distributions for variance parameters in hierarchical models. Bayesian Anal. **1**(3), 515–534 (2006)
Gelman, A., Carlin, J. B., Stern, H. S., Rubin, D. B.: Bayesian data analysis. Chapman and Hall/CRC (1995)
Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **7**(4), 457–472 (1992)
Gemayel, K.T., Litman, G.W., Sriaroon, P.: Autosomal recessive agammaglobulinemia associated with an IGLL1 gene missense mutation. Ann. Allergy Asthma Immunol. **117**(4), 439–441 (2016)
Genking, A., Lewis, D.D., Madigan, D.: Large-scale Bayesian logistic regression for text categorization. Technometrics **49**(3), 291–304 (2007)
George, E.I., Foster, D.P.: Calibration and empirical Bayes variable selection. Biometrika **87**(4), 731–747 (2000)
George, E.I., McCulloch, R.E.: Variable selection via Gibbs sampling. J. Am. Stat. Assoc. **88**(423), 881–889 (1993)
George, E.I., McCulloch, R.E.: Approaches for Bayesian variable selection. Stat. Sin. **7**(2), 339–373 (1997)
Ghosh, P., Tang, X., Ghosh, M., Chakrabarti, A.: Asymptotic properties of Bayes risk of a general class of shrinkage priors in multiple hypothesis testing under sparsity. Bayesian Anal. **11**(3), 753–796 (2016)
Gil, M., Alajaji, F., Linder, T.: Rényi divergence measures for commonly used univariate continuous distributions. Inf. Sci. **249**, 124–131 (2013)
Hahn, P.R., Carvalho, C.M.: Decoupling shrinkage and selection in bayesian linear models: a posterior summary perspective. J. Am. Stat. Assoc. **110**(509), 435–448 (2015)
Harrell, F.: Regression modeling strategies: with applications to linear models, logistic and ordinal regression and survival analysis. (2 ed.). Springer Series in Statistics (2001)
Huang, A., Xu, S., Cai, X.: Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. BMC Genet. **14**(1), 1–14 (2013)
Ishwaran, H., Rao, J.S.: Detecting differentially expressed genes in microarrays using bayesian model selection. J. Am. Stat. Assoc. **98**(462), 438–455 (2003)
Kass, R.E., Raftery, A.E.: Bayes factors. J. Am. Stat. Assoc. **90**(430), 773–795 (1995)
Kowal, D.R.: Fast, optimal, and targeted predictions using parameterized decision analysis. J. Am. Stat. Assoc. **117**(540), 1875–1886 (2022)
Kuo, L., Mallick, B.: Variable selection for regression models. Sankhya: The Indian J. Stat. Series B, 65–81 (1998)
Leng, C., Tran, M.-N., Nott, D.: Bayesian adaptive Lasso. Ann. Inst. Stat. Math. **66**(2), 221–244 (2014)
Li, K.-C.: Honest confidence regions for nonparametric regression. Ann. Stat. **17**(3), 1001–1008 (1989)
Li, W., Chekouo, T.: Bayesian group selection with non-local priors. Comput. Stat. **37**(1), 287–302 (2022)
Lindley, D.V.: On a measure of the information provided by an experiment. Ann. Math. Stat. **27**(4), 986–1005 (1956)
Lipman, D., Safo, S.E., Chekouo, T.: Multi-omic analysis reveals enriched pathways associated with COVID-19 and COVID-19 severity. PLoS ONE **17**(4), e0267047 (2022)
Lykou, A., Ntzoufras, I.: On bayesian lasso variable selection and the specification of the shrinkage parameter. Stat. Comput. **23**, 361–390 (2013)
Maddalena, M., Berger, J.O.: Optimal predictive model selection. Ann. Stat. **32**, 870–897 (2004)
Malsiner-Walli, G., Wagner, H.: Comparing spike and slab priors for bayesian variable selection. arXiv preprintarXiv:1812.07259 (2018)
Meinshausen, N., Meier, L., Bühlmann, P.: P-values for high-dimensional regression. J. Am. Stat. Assoc. **104**(488), 1671–1681 (2009)
Myles, C., Wayne, M.: Quantitative trait locus (QTL) analysis. Nat. Educ. **1**(1), 208 (2008)
Narisetty, N.N., Hel, X.: Bayesian variable selection with shrinking and diffusing priors. Ann. Stat. **42**(2), 789–817 (2014)

Nott, D.J., Seah, M., Al-Labadi, L., Evans, M., Ng, H.K., Englert, B.-G.: Using Prior Expansions for Prior-Data Conflict Checking. Bayesian Anal. **16**(1), 203–231 (2021)
Nott, D.J., Wang, X., Evans, M., Englert, B.-G.: Checking for Prior-Data Conflict using prior-to-posterior divergences. Stat. Sci. **35**(2), 243–253 (2020)
Nott, D.J., Yu, Z., Chan, E., Cotsapas, C., Cowley, M.J., Pulvers, J., Williams, R., Little, P.: Hierarchical Bayes variable selection and microarray experiments. J. Multivar. Anal. **98**(4), 852–872 (2007)
O'hara, R.B., Sillanpää, M.J.: A review of Bayesian variable selection methods: what, how and which. Bayesian Anal. **4**(1), 85–117 (2009)
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.: Large-scale multi-omic analysis of COVID-19 severity. Cell Syst. **12**(1), 23–40 (2021)
Park, T., Casella, G.: The bayesian lasso. J. Am. Stat. Assoc. **103**(482), 681–686 (2008)
Piironen, J., Vehtari, A.: On the hyperprior choice for the global shrinkage parameter in the horseshoe prior. Artif. Intell. Stat. *PMLR* (2017)
Ročková, V., George, E.I.: The spike-and-slab lasso. J. Am. Stat. Assoc. **113**(521), 431–444 (2018)
Ročková, V., Lesaffre, E., Luime, J., Löwenberg, B.: Hierarchical Bayesian formulations for selecting variables in regression models. Stat. Med. **31**(11–12), 1221–1237 (2012)
Schwarz G.: Estimating the dimension of a model. The Ann. Stat. 461–464 (1978)
Sheather, S.J.: Density Estimation. Stat. Sci. **19**(4), 588–597 (2004)
Sheather, S.J., Jones, M.C.: A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Stat. Soc.: Ser. B (Methodol.) **53**(3), 683–690 (1991)
Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., Quan, S., Zhang, F., Sun, R., Qian, L., et al.: Proteomic and metabolomic characterization of covid-19 patient sera. Cell **182**(1), 59–72 (2020)
Tadesse, M.G., Vannucci, M.: Handbook of bayesian variable selection. CRC Press (2021)
Taquet, M., Geddes, J.R., Husain, M., Luciano, S., Harrison, P.J.: 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 (2021)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat Methodol. **58**(1), 267–288 (1996)
van der Pas, S., Szabó, B., van der Vaart, A.: Uncertainty quantification for the horseshoe (with discussion). Bayesian Anal. **12**(4), 1221–1274 (2017)
Van Erp, S., Oberski, D.L., Mulder, J.: Shrinkage priors for bayesian penalized regression. J. Math. Psychol. **89**, 31–50 (2019)
Vigón, L., Galán, M., Torres, M., Martín-Galiano, A.J., Rodríguez-Mora, S., Mateos, E., Corona, M., Malo, R., Navarro, C., Murciano-Antón, M.A., et al.: Association between hla-c alleles and covid-19 severity in a pilot study with a spanish mediterranean caucasian cohort. PLoS ONE **17**(8), e0272867 (2022)
Vivekananda, R., Chakraborty, S.: Selection of tuning parameters, solution paths and standard errors for Bayesian lassos. Bayesian Anal. **12**(3), 753–778 (2017)
Wang, F., Mukherjee, S., Richardson, S., Hill, S.M.: High-dimensional regression in practice: an empirical study of finite-sample prediction, variable selection and ranking. Stat. Comput. **30**, 697–719 (2020)
Wang, X., Wen, Y., Xie, X., Liu, Y., Tan, X., Cai, Q., Zhang, Y., Cheng, L., Xu, G., Zhang, S., et al.: Dysregulated hematopoiesis in bone marrow marks severe COVID-19. Cell Discov. **7**(1), 60 (2021)
Webster, P.: COVID-19 timeline of events. Nat. Med. **27**(12), 2054–2055 (2021)
Wu, S., Xu, Y., Zhang, J., Ran, X., Jia, X., Wang, J., Sun, L., Yang, H., Li, Y., Fu, B., et al.: Longitudinal serum proteome characterization of COVID-19 patients with different severities revealed potential therapeutic strategies. Front. Immunol. **13**, 893943 (2022)
Xu, X., Ghosh, M.: Bayesian variable selection and estimation for group lasso. Bayesian Anal. **4**, 909–936 (2015)
Zellner, A.: On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian inference and decision techniques (1986)
Zhang, Y.D., Naughton, B.P., Bondell, H.D., Reich, B.J.: Bayesian regression using a prior on the model fit: The r2–d2 shrinkage prior. J. Am. Stat. Assoc. **117**(538), 862–874 (2022)
Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B Stat Methodol. **67**(2), 301–320 (2005)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.