Journal of Mathematical Psychology 89 (2019) 31–50

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Review

Shrinkage priors for Bayesian penalized regression

Sara van Erp [a,*,1], Daniel L. Oberski [b,1], Joris Mulder [a,1]

[a] Department of Methodology and Statistics, Tilburg University, Warandelaan 2, 5037 AB Tilburg, The Netherlands
[b] Department of Methodology and Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands

HIGHLIGHTS

- Various shrinkage priors have distinctive theoretical characteristics.
- Most priors have a similar prediction accuracy unless $p > n$.
- Different shrinkage priors vary in variable selection accuracy.

ARTICLE INFO

Article history:
Received 4 April 2018
Received in revised form 20 December 2018
Available online 28 January 2019

Keywords:
Bayesian
Shrinkage priors
Penalization
Empirical Bayes
Regression

ABSTRACT

In linear regression problems with many predictors, penalized regression techniques are often used to guard against overfitting and to select variables relevant for predicting an outcome variable. Recently, Bayesian penalization is becoming increasingly popular in which the prior distribution performs a function similar to that of the penalty term in classical penalization. Specifically, the so-called *shrinkage priors* in Bayesian penalization aim to shrink small effects to zero while maintaining true large effects. Compared to classical penalization techniques, Bayesian penalization techniques perform similarly or sometimes even better, and they offer additional advantages such as readily available uncertainty estimates, automatic estimation of the penalty parameter, and more flexibility in terms of penalties that can be considered. However, many different shrinkage priors exist and the available, often quite technical, literature primarily focuses on presenting one shrinkage prior and often provides comparisons with only one or two other shrinkage priors. This can make it difficult for researchers to navigate through the many prior options and choose a shrinkage prior for the problem at hand. Therefore, the aim of this paper is to provide a comprehensive overview of the literature on Bayesian penalization. We provide a theoretical and conceptual comparison of nine different shrinkage priors and parametrize the priors, if possible, in terms of scale mixture of normal distributions to facilitate comparisons. We illustrate different characteristics and behaviors of the shrinkage priors and compare their performance in terms of prediction and variable selection in a simulation study. Additionally, we provide two empirical examples to illustrate the application of Bayesian penalization. Finally, an R package `bayesreg` is available online (https://github.com/sara-vanerp/bayesreg) which allows researchers to perform Bayesian penalized regression with novel shrinkage priors in an easy manner.

Contents

* Corresponding author.
E-mail addresses: s.vanerp_1@tilburguniversity.edu (S. van Erp), d.l.oberski@uu.nl (D.L. Oberski), J.Mulder3@tilburguniversity.edu (J. Mulder).

[1] Declarations of interest: none.

https://doi.org/10.1016/j.jmp.2018.12.004
0022-2496/

1. Introduction

Regression analysis is one of the main statistical techniques often used in the field of psychology to determine the effect of a set of predictors on an outcome variable. The number of predictors is often large, especially in the current "Age of Big Data". For example, the Kavli HUMAN project (Azmak, Bayer, Caplin, Chun, Glimcher, Koonin, & Patrinos, 2015) aims to collect longitudinal data on all aspects of human life for 10,000 individuals. Measurements include psychological assessments (e.g., personality, IQ), health assessments (e.g., genome sequencing, brain activity scanning), social network assessment, and variables related to education, employment, and financial status, resulting in an extremely large set of variables. Furthermore, personal tracking devices allow the collection of large amounts of data on various topics, including for example mood, in a longitudinal manner (Fawcett, 2015). The problem with regular regression techniques such as ordinary least squares (OLS) is that they quickly lead to overfitting as the ratio of predictor variables to observations increases (see for example, McNeish, 2015, for an overview of the problems with OLS).

Penalized regression is a statistical technique widely used to guard against overfitting in the case of many predictors. Penalized regression techniques have the ability to select variables out of a large set of variables that are relevant for predicting some outcome. Therefore, a popular setting for penalized regression is in high-dimensional data, where the number of predictors p is larger than the sample size n. Furthermore, in settings where the number of predictors p is smaller than the sample size n (but still relatively large), penalized regression can offer advantages in terms of avoiding overfitting and achieving model parsimony compared to traditional variable selection methods such as null-hypothesis testing or stepwise selection methods (Derksen & Keselman, 1992; Tibshirani, 1996). The central idea of penalized regression approaches is to add a penalty term to the minimization of the sum of squared residuals, with the goal of shrinking small coefficients towards zero while leaving large coefficients large, i.e.,

$$\underset{\beta_0,\,\boldsymbol{\beta}}{\text{minimize}} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \beta_0 \mathbf{1} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_c \|\boldsymbol{\beta}\|_q \right\}, \tag{1}$$

$$\text{where } \|\boldsymbol{\beta}\|_q = \left(\sum_{j=1}^{p} |\beta_j|^q \right)^{\frac{1}{q}},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is an n-dimensional vector containing the observations on the outcome variable, β_0 reflects the intercept, $\mathbf{1}$ is an n-dimensional vector of ones, $\boldsymbol{X}$ is an $(n \times p)$ matrix of the observed scores on the p predictor variables, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional parameter vector of regression coefficients. λ_c reflects the penalty parameter, with large values resulting in more shrinkage towards zero while $\lambda_c = 0$ leads to the ordinary least squares solution. The choice of q determines the type of penalty induced, for example, $q = 1$ results in the well-known least absolute shrinkage and selection operator (lasso; Tibshirani, 1996) solution and $q = 2$ results in the ridge solution (Hoerl & Kennard, 1970). We refer to Hastie, Tibshirani, and Wainwright (2015) for a comprehensive introduction and overview of various penalized regression methods in a frequentist framework.

It is well known that many solutions to the penalized minimization problem in Eq. (1) can also be obtained in the Bayesian framework by using a specific prior combined with the posterior mode estimate, which has been shown to perform similar to or better than their classical counterparts (Hans, 2009; Kyung, Gill, Ghosh, & Casella, 2010; Li & Lin, 2010).[2] Adopting a Bayesian perspective on penalized regression offers several advantages. First, penalization fits naturally in a Bayesian framework since a prior distribution is needed anyway and shrinkage towards zero can be straightforwardly achieved by choosing a specific parametric form for the prior. Second, parameter uncertainty and standard errors follow naturally from the posterior standard deviations. As shown by Kyung et al. (2010) classical penalized regression procedures can result in estimated standard errors that suffer from multiple problems, such as variances estimated to be 0 (in the case of sandwich estimates), and unstable or poorly performing variance estimates (in the case of bootstrap estimates). Third, with Bayesian penalization it is possible to estimate the penalty parameter(s) λ simultaneously with the model parameters in a single step. This is especially advantageous when there are multiple penalty parameters (e.g., in the elastic net; Zou & Hastie, 2005), since sequential cross-validation procedures to determine multiple penalty parameters induce too much shrinkage (i.e., the double shrinkage problem; see e.g., Zou & Hastie, 2005). Fourth, Bayesian penalization relies on Markov Chain Monte Carlo (MCMC) sampling rather than optimization, which provides more flexibility in the sense that priors that would correspond to non-convex penalties (i.e., $q < 1$ in (1)) are easier to implement. Non-convex penalties would result in multiple modes, making them difficult to implement in an optimization framework. The cost of the flexibility of MCMC, however, is that it requires more computation time compared to standard optimization procedures. Finally, Bayesian estimates have an intuitive interpretation. For example, a 95%

[2] Note that from a Bayesian perspective, however, there is no theoretical justification for reporting the posterior mode estimate (Tibshirani, 2011).

Bayesian credibility interval can simply be interpreted as the interval in which the true value lies with 95% probability (e.g., Berger, 2006).

Due to these advantages, Bayesian penalization is becoming increasingly popular in the literature (see e.g., Alhamzawi, Yu, & Benoit, 2012; Andersen, Vehtari, Winther, & Hansen, 2017; Armagan, Dunson, & Lee, 2013; Bae & Mallick, 2004; Bhadra, Datta, Polson, & Willard, 2016; Bhattacharya, Pati, Pillai, & Dunson, 2012; Bornn, Gottardo, & Doucet, 2010; Caron & Doucet, 2008; Carvalho, Polson, & Scott, 2010; Feng, Wang, Lu, & Song, 2017; Griffin & Brown, 2017; Hans, 2009; Ishwaran & Rao, 2005; Lu, Chow, & Loken, 2016; Peltola, Havulinna, Salomaa, & Vehtari, 2014; Polson & Scott, 2011; Roy & Chakraborty, 2016; Zhao, Gao, Mukherjee, & Engelhardt, 2016). An active area of research investigates theoretical properties of priors for Bayesian penalization, such as the Bayesian lasso prior (for a recent overview, see Bhadra, Datta, Polson, & Willard, 2017). In addition to the Bayesian counterparts of classical penalized regression solutions, many other priors have been proposed that have desirable properties in terms of prediction and variable selection. However, the extensive (and often technical) literature and subtle differences between the priors can make it difficult for researchers to navigate the options and make sensible choices for the problem at hand. Therefore, the aim of this paper is to provide a comprehensive overview of the priors that have been proposed for penalization in (sparse) regression. We use the term *shrinkage priors* to emphasize that these priors aim to shrink small effects towards zero. We place the shrinkage priors in a general framework of scale mixtures of normal distributions to emphasize the similarities and differences between the priors. By providing insight in the characteristics and behaviors of the priors, we aid researchers in choosing a prior for their specific problem. Additionally, we present a straightforward method to obtain empirical Bayes (EB) priors for Bayesian penalization. We conduct a simulation study to compare the performance of the priors in terms of prediction and variable selection in a linear regression model, and provide two empirical examples to further illustrate the Bayesian penalization methods. Finally, the shrinkage priors have been implemented in the R package `bayesreg`, available from https://github.com/sara-vanerp/bayesreg, to allow general utilization.

The remainder of this paper is organized as follows: Section 2 introduces Bayesian penalized regression. A theoretical overview of the different shrinkage priors can be found in Section 3. Further insights into the priors are provided through illustrations in Section 4 and the priors are compared in a simulation study in Section 5. Section 6 presents the empirical applications, followed by a discussion in Section 7.

2. Bayesian penalized regression

The likelihood for the linear regression model is given by:

$$y_i|\beta_0, \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2 \sim \text{Normal}(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j, \sigma^2), \tag{2}$$

where β_0 represents the intercept, β_j the regression coefficient for predictor j, and σ^2 is the residual variance.

In a Bayesian analysis, a prior distribution is specified for each parameter in the model, e.g., $p(\beta_0, \boldsymbol{\beta}, \sigma^2, \lambda) = p(\beta_0)p(\boldsymbol{\beta}|\sigma^2, \lambda)p(\sigma^2)p(\lambda)$. Note that the prior for $\boldsymbol{\beta}$ is conditioned on the residual variance σ^2, as well as on λ. The conditioning on σ^2 is necessary in certain cases to obtain a unimodal posterior (Park & Casella, 2008). In Bayesian penalized regression, λ is a parameter in the prior (i.e., a hyperparameter) but has a similar role as the penalty parameter in classical penalized regression. Since this penalty parameter λ is used to penalize the regression coefficient, it only appears in the prior for $\boldsymbol{\beta}$. Throughout this paper we will focus on priors for the regression coefficients $\beta_1, \ldots, \beta_j$ and we will assume noninformative improper priors for the nuisance parameters, specifically, $p(\beta_0) = 1$ and a uniform prior on $\log(\sigma^2)$, i.e., $p(\sigma^2) = \sigma^{-2}$. Please note that these priors are chosen as noninformative choices for the linear regression model considered in this paper. However, other choices (including informative priors when prior information is available) are possible and might be preferred in other applications. We refer the reader to van Erp, Mulder, and Oberski (2018) for general recommendations on specifying prior distributions. We generally assume that the priors for the regression coefficients are independent, unless stated otherwise.

The prior distribution is then multiplied by the likelihood of the data to obtain the posterior distribution, i.e.,

$$p(\beta_0, \boldsymbol{\beta}, \sigma^2, \lambda|\boldsymbol{y}, X) \propto p(\boldsymbol{y}|X, \beta_0, \boldsymbol{\beta}, \sigma^2)p(\beta_0)p(\boldsymbol{\beta}|\sigma^2, \lambda)p(\sigma^2)p(\lambda). \tag{3}$$

Here, the normalizing constant is not included such that the right-hand side is proportional to the posterior. The only difference with the unpenalized problem (e.g., Bayesian linear regression) is the introduction of the penalty parameter λ. As a result of the shrinkage prior the posterior in (3) is generally more concentrated, or "shrunk towards", zero in comparison to the likelihood of the model.

An important choice is how to specify the penalty parameter λ. There are different possibilities for this.

1. *Full Bayes.* Treat λ as an unknown model parameter for which a prior needs to be specified. Typically, a vague prior $p(\lambda)$ is specified for λ. Due to its similarity with multilevel (or hierarchical) modeling, full Bayes (FB) is also known as "hierarchical Bayes" (see e.g., Wolpert & Strauss, 1996). This results in a fully Bayesian solution that incorporates the uncertainty about λ. The advantage of this approach is that the model can be estimated in one step. Throughout this paper, we will consider the half-Cauchy prior on λ, i.e., $\lambda \sim$ half-Cauchy(0, 1), which is a robust alternative and a popular prior distribution in the Bayesian literature (see e.g., Gelman, 2006; Mulder & Pericchi, 2018; Polson & Scott, 2012).
2. *Empirical Bayes.* Empirical Bayes (EB) methods, also known as the "evidence" procedure (see e.g., Wolpert & Strauss, 1996), first estimate the penalty parameter λ from the data and then plug in this EB estimate for λ in the model (see van de Wiel, Beest, & Münch, 2017, for an overview of EB methodology in high-dimensional data). The resulting prior is called an EB prior. Since an EB estimate is used for λ, the EB approach does not require the specification of a prior $p(\lambda)$ as in the FB approach. Since the exact choice of this prior can sometimes have a serious effect on the Bayesian estimates (Roy & Chakraborty, 2016), the EB approach would avoid sensitivity of the results to the exact choice of the prior $p(\lambda)$, while keeping the advantages of the Bayesian approach.
   Empirical Bayes is a two-step approach: first, the empirical Bayes choice for λ needs to be determined; second, the model is fitted using the EB prior. In order to obtain an EB estimate for λ, we need to find the solution that maximizes the marginal likelihood,[3] i.e.,

$$\lambda^{EB} = \arg\max p(\boldsymbol{y}|\lambda). \tag{4}$$

[3] The marginal likelihood quantifies the probability of observing the data given the model. Therefore, plugging in the EB estimate for λ will result in a prior that predicts the observed data best.

To obtain λ^{EB}, first note that the marginal likelihood is the product of the likelihood and prior integrated over the model parameters, i.e.,

$$p(\boldsymbol{y}|\lambda) = \iiint p(\boldsymbol{y}|X, \beta_0, \boldsymbol{\beta}, \sigma^2) \times p(\beta_0)p(\boldsymbol{\beta}|\sigma^2, \lambda)p(\sigma^2)\, d\beta_0\, d\boldsymbol{\beta}\, d\sigma^2. \quad (5)$$

Instead of directly optimizing, we achieve (4) by sampling from the posterior with a noninformative prior for λ.[4] The EB estimate λ^{EB} is the mode of the marginal posterior for λ, i.e., $p(\lambda|\boldsymbol{y})$. This corresponds to the maximum of the marginal likelihood $p(\boldsymbol{y}|\lambda)$ because of the noninformative prior for λ.

3. *Cross-validation.* For cross-validation (CV), the data is split into a training, validation, and test set. The goal is to find a value for λ which results in a model that is accurate in predicting new data, i.e., a generalizable model that captures the signal in the data, but does not overfit (Hastie et al., 2015). To find this value for λ, a range of values is considered, using the training data $\boldsymbol{y}^{train}$ to fit all models with the different λ values. Next, each resulting model is used to predict the responses in the validation set $\boldsymbol{y}^{val}$. The value for λ that minimizes some loss function is selected, i.e.,

$$\lambda^{CV} = \arg\min L(\boldsymbol{y}^{train}, \boldsymbol{y}^{val}). \quad (6)$$

Given that the loss function is the negative of the log likelihood, this is equivalent to:

$$\lambda^{CV} = \arg\max p(\boldsymbol{y}^{val}|\boldsymbol{y}^{train}, \lambda). \quad (7)$$

Finally, λ^{CV} is used to fit the model on the test set. Generally, the prediction mean squared error (PMSE) is used to determine λ^{CV}, which corresponds to a quadratic loss function.

In practice k-fold cross-validation is often used. k-fold cross-validation is a specific implementation of cross-validation in which the data is split in only a training and a test set. The training set is split in K parts (usually $K = 5$ or $K = 10$) and the range of λ values is applied K times on $K - 1$ parts of the training set, each time with a different part as validation set. The K estimates of the PMSE are then averaged and a standard error is computed.

Frequentist penalization approaches often rely on cross-validation. In the Bayesian literature, full and empirical Bayes are often employed, although cross-validation is also possible in a Bayesian approach (see for example the `loo` package in R; Vehtari, Gabry, Yao, & Gelman, 2018). The intuition behind empirical Bayes and cross-validation is similar: empirical Bayes aims to choose the value for λ that is best in predicting the full data set, while cross-validation aims to choose the value for λ that is best in predicting the validation set given a training set. A possible disadvantage of empirical Bayes and cross-validation is that the (marginal) likelihood can be flat or multimodal when there are multiple penalty parameters (van de Wiel et al., 2017).[5]

Throughout this paper, we will focus on the full and empirical Bayes approach to determine λ, and only consider cross-validation for the frequentist penalization methods we will compare the priors to.

[4] Specifically, we use $\lambda \sim$ half-Cauchy(0, 10000) to ensure a stable MCMC sampler.

[5] In the initial empirical Bayes approach we used a uniform prior for λ and this problem became evident through non-convergence of the sampler or extreme estimates for λ^{EB}. The problem was solved by using the half-Cauchy prior instead.

3. Overview shrinkage priors

In this section we will give a general overview of shrinkage priors that have been proposed in the literature. Given the extensive number of shrinkage priors that has been investigated, we will limit the overview to priors that are related to well-known classical penalization methods and shrinkage priors that are popular in the Bayesian literature. Given that most shrinkage priors fall into these categories, the resulting overview, while not exhaustive, is intended to be comprehensive and will help researchers to navigate through this literature. In total, we will discuss nine different shrinkage priors.

Many continuous, unimodal, and symmetric distributions can be parametrized as a scale mixture of normals meaning that the distribution is rewritten as a normal distribution (i.e., Normal(μ, σ^2)) where the scale parameter is given a mixing density $h(\sigma^2)$ (see e.g., West, 1987). Where possible, we will present the different priors in a common framework by providing the scale mixture of normals formulation for each prior. Using this formulation, the theoretical differences and similarities between the priors become more clear and, additionally, the scale mixture of normals formulation can be computationally more efficient.

We will now describe each prior in turn. The densities for several of the shrinkage priors are presented in Table 1 and plotted in Fig. 1. We will consider a full and empirical Bayes approach to obtain the penalty parameter λ (see Section 2) for all shrinkage priors, unless stated otherwise. For the full Bayesian approach, we will consider standard half-Cauchy priors for the penalty parameters as a robust default prior choice. We have included this choice for the prior on λ in the descriptions below, but note that other choices are possible as well.

3.1. Ridge

The ridge prior corresponds to normal priors centered around 0 on the regression coefficients, i.e., (see e.g., Hsiang, 1975)

$$\beta_j|\lambda, \sigma^2 \sim \text{Normal}(0, \frac{\sigma^2}{\lambda}), \text{ for } j = 1, \ldots, p. \quad (8)$$
$$\lambda \sim \text{half-Cauchy}(0, 1)$$

The posterior mean estimates under this prior will correspond to estimates obtained using the ridge penalty or l_2 norm, i.e., $q = 2$ in Eq. (1) (Hoerl & Kennard, 1970). The penalty parameter λ determines the amount of shrinkage, with larger values resulting in smaller prior variation and thus more shrinkage of the coefficients towards zero.

3.2. Local Student's t

We can extend the ridge prior in Eq. (8) by making the prior variances predictor-specific, thereby allowing for more variation, i.e.,

$$\beta_j|\tau_j^2 \sim \text{Normal}(0, \sigma^2\tau_j^2) \quad (9)$$
$$\tau_j^2|\nu, \lambda \sim \text{Inverse-Gamma}(\frac{\nu}{2}, \frac{\nu}{2\lambda}), \text{ for } j = 1, \ldots, p,$$
$$\lambda \sim \text{half-Cauchy}(0, 1)$$

When integrating τ_j^2 out, the following conditional prior distribution for the regression coefficients is obtained:

$$\beta_j|\nu, \lambda, \sigma^2 \sim \text{Student}(\nu, 0, \frac{\sigma^2}{\lambda}), \quad (10)$$

where Student(ν, 0, $\frac{\sigma^2}{\lambda}$) denotes a non-standardized Student's t distribution centered around 0 with ν degrees of freedom and scale parameter $\frac{\sigma^2}{\lambda}$. A smaller value for ν results in a distribution with

Table 1
Conditional prior densities for the regression coefficients $\boldsymbol{\beta}$ implied by the various shrinkage priors and references for each shrinkage prior.

Shrinkage prior	Conditional prior density $p(\beta_j|\lambda,\ldots)$	Reference
Ridge	$p(\beta_j\|\sigma^2,\lambda) = \sqrt{\frac{\lambda}{2\pi\sigma^2}}\exp\left\{-\frac{\lambda\beta_j^2}{2\sigma^2}\right\}$	Hsiang (1975)
Local Student's t	$p(\beta_j\|\sigma^2,\lambda) = \frac{\sigma^2}{\pi\lambda}\left(1+(\frac{\sigma^2}{\lambda\beta_j})^2\right)$	Griffin and Brown (2005) and Meuwissen, Hayes, and Goddard (2001)
Lasso	$p(\beta_j\|\sigma^2,\lambda) = \frac{\lambda}{2\sqrt{\sigma^2}}\exp\left\{\frac{-\lambda\|\beta_j\|}{\sqrt{\sigma^2}}\right\}$	Park and Casella (2008)
Elastic net	$p(\beta_j\|\sigma^2,\lambda_1,\lambda_2) = C\exp\left\{-\frac{1}{2\sigma^2}(\lambda_1\|\beta_j\|+\lambda_2\beta_j^2)\right\}$	Li and Lin (2010)
Group lasso	$p(\beta_j\|\sigma,\lambda) = C\exp\left\{-\frac{\lambda}{\sigma}\sum_{g=1}^{G}\|\|\boldsymbol{\beta}_g\|\|\right\}$	Kyung et al. (2010)
Hyperlasso	$p(\beta_j\|\lambda) = \lambda(2\pi)^{\frac{1}{2}}\left[1-\frac{(\lambda\|\beta_j\|)\{1-\Phi(\lambda\|\beta_j\|)\}}{\phi(\lambda\|\beta_j\|)}\right]$	Griffin and Brown (2011)
Horseshoe	Not analytically tractable	Carvalho et al. (2010)
Discrete normal mixture	$p(\beta_j\|\gamma_j,\phi_j^2) = (1-\gamma_j)\left(\frac{1}{\sqrt{2\pi\phi_j^2}}\exp\left\{-\frac{\beta_j^2}{2\phi_j^2}\right\}\right)+\gamma_j\left(\frac{1}{\pi(1+\beta_j^2)}\right)$	George and McCulloch (1993) and Mitchell and Beauchamp (1988)

Note. C denotes a normalization constant. $\Phi()$ and $\phi()$ in the hyperlasso are the cumulative density function and the probability density function of the standard normal distribution.

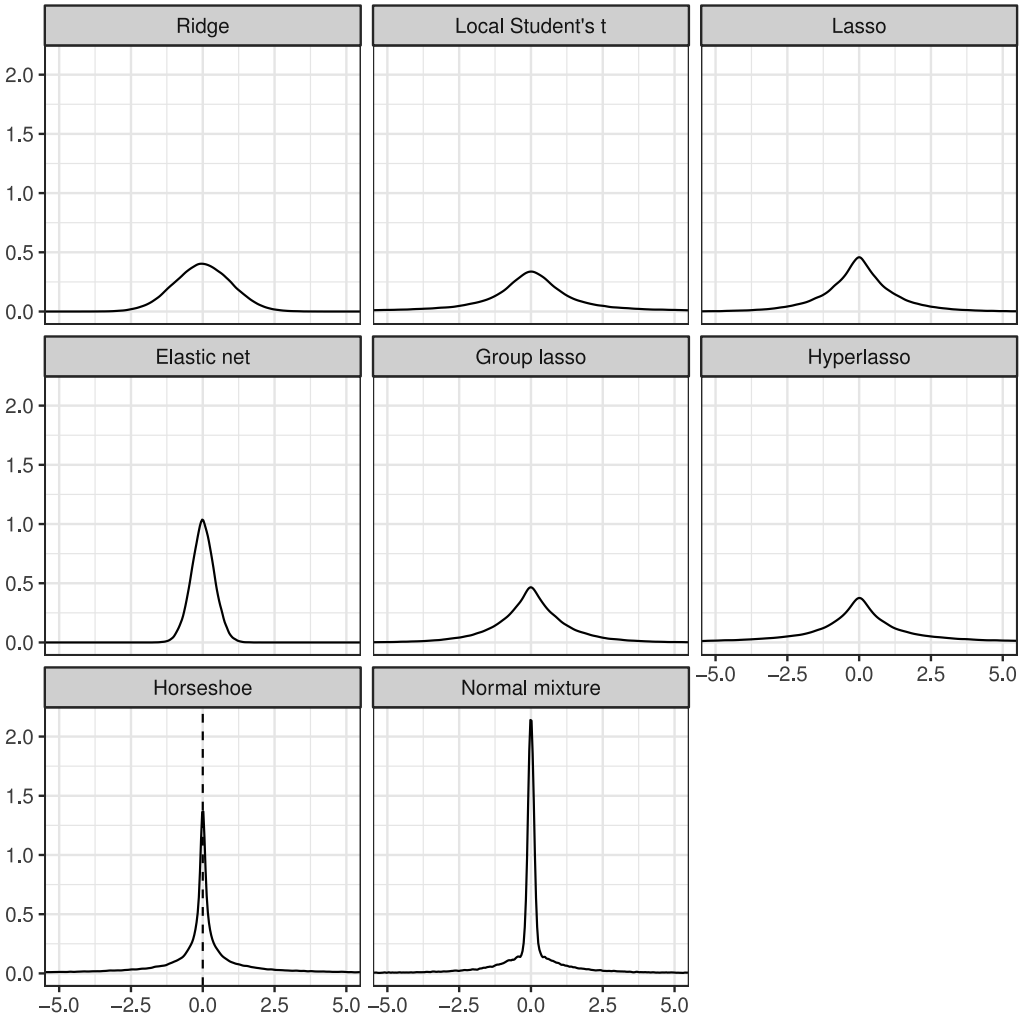

Fig. 1. Densities of the shrinkage priors.

heavier tails, with $\nu = 1$ implying a Cauchy prior for β_j. Larger (smaller) values for λ result in more (less) shrinkage towards m. This prior has been considered, among others, by Griffin and Brown (2005) and Meuwissen et al. (2001). Compared to the ridge prior in (8), the local Student's t prior has heavier tails. Throughout this paper, we will consider $\nu = 1$, such that the prior has Cauchy-like tails.

3.3. Lasso

The Bayesian counterpart of the lasso penalty was first proposed by Park and Casella (2008). The Bayesian lasso can be obtained as a scale mixture of normals with an exponential mixing density, i.e.,

$$\beta_j|\tau_j^2,\sigma^2 \sim \text{Normal}(0,\sigma^2\tau_j^2) \tag{11}$$

$$\tau_j^2|\lambda^2 \sim \text{Exponential}(\frac{\lambda^2}{2}),\ \text{for } j = 1, \ldots, p,$$
$$\lambda \sim \text{half-Cauchy}(0, 1)$$

Integrating τ_j^2 out results in double-exponential or Laplace priors on the regression coefficients, i.e.,

$$\beta_j|\lambda, \sigma \sim \text{Double-exponential}(0, \frac{\sigma}{\lambda}),\ \text{for } j = 1, \ldots, p. \tag{12}$$

With this prior, the posterior mode estimates are similar to estimates obtained under the lasso penalty or l_1 norm, i.e., $q = 1$ in Eq. (1) (Tibshirani, 1996). In addition to the overall shrinkage parameter λ, the lasso prior has an additional predictor-specific shrinkage parameter τ_j. Therefore, the lasso prior is more flexible than the ridge prior which only relies on the overall shrinkage parameter in (8). Fig. 1 clearly shows that the lasso prior has a sharper peak around zero compared to the ridge prior.

Disadvantages of the lasso

The popularity of the classical lasso lies in its ability to shrink coefficients to zero, thereby automatically performing variable selection. However, there are several disadvantages to the classical lasso. Specifically, (i) it cannot select more predictors than observations, which is problematic when $p > n$; (ii) when a group of predictors is correlated, the lasso generally selects only one predictor of that group; (iii) the prediction error is higher for the lasso compared to the ridge when $n > p$ and the predictors are highly correlated; (iv) it can lead to overshrinkage of large coefficients (see e.g., Polson & Scott, 2011); and (v) it does not always have the oracle property, which implies it does not always perform as well in terms of variable selection as if the true underlying model has been given (Fan & Li, 2001). The lasso only enjoys the oracle property under specific and stringent conditions (Fan & Li, 2001; Zou, 2006). These disadvantages have sparked the development of several generalizations of the lasso. We will now discuss the Bayesian counterparts of several of these generalizations, including the elastic net, group lasso, and hyperlasso. Note that for the Bayesian lasso, coefficients cannot become exactly zero and thus a criterion is needed to select the relevant variables. Depending on the criterion used, more predictors than observations could be selected. However, the Bayesian lasso does not allow a grouping structure to be included, it overshrinks large coefficients, and it does not have the oracle property since the tails for the prior on β_j are not heavier than exponential tails (Polson, Scott, & Windle, 2014).

3.4. Elastic net

The most popular generalization of the lasso is the elastic net (Zou & Hastie, 2005). The elastic net can be seen as a combination of the ridge and lasso. The elastic net resolves issues (i), (ii), and (iii) of the ordinary lasso. The elastic net prior can be obtained as the following scale mixture of normals (Li & Lin, 2010):

$$\beta_j|\lambda_2, \tau_j, \sigma^2 \sim \text{Normal}\left(0, (\frac{\lambda_2}{\sigma^2}\frac{\tau_j}{\tau_j - 1})^{-1}\right) \tag{13}$$
$$\tau_j|\lambda_2, \lambda_1, \sigma^2 \sim \text{Truncated-Gamma}\left(\frac{1}{2}, \frac{8\lambda_2\sigma^2}{\lambda_1^2}\right),$$
$$\text{for } j = 1, \ldots, p,$$
$$\lambda_1 \sim \text{half-Cauchy}(0, 1)$$
$$\lambda_2 \sim \text{half-Cauchy}(0, 1)$$

where the truncated Gamma density has support $(1, \infty)$. This implies the following conditional prior distributions for the regression coefficients:

$$p(\beta_j|\sigma^2, \lambda_1, \lambda_2) = C(\lambda_1, \lambda_2, \sigma^2)\exp\left\{-\frac{1}{2\sigma^2}(\lambda_1|\beta_j| + \lambda_2\beta_j^2)\right\}, \tag{14}$$
$$\text{for } j = 1, \ldots, p,$$

where $C(\lambda_1, \lambda_2, \sigma^2)$ denotes the normalizing constant. The corresponding posterior modes for β_j are equivalent to the estimates from the classical elastic net penalty. Expression (14) illustrates how the elastic net prior offers a combination of the double-exponential prior, i.e., the lasso penalty $\lambda|\beta_j|$, and the normal prior, i.e., the ridge penalty $\lambda\beta_j^2$. Specifically, the two penalty parameters λ_1 and λ_2 determine the relative influence of the lasso and ridge penalty, respectively. This can also be seen in Fig. 1: the elastic net is not as sharply peaked as the lasso prior, but it is sharper than the ridge prior. As mentioned in the Introduction, a disadvantage of the classical elastic net is that the sequential cross-validation procedure used to determine the penalty parameters results in overshrinkage of the coefficients. This problem is resolved in the Bayesian approach by estimating both penalty parameters simultaneously through a full or empirical Bayes approach.

3.5. Group lasso

The group lasso (Yuan & Lin, 2006) is a generalization of the lasso primarily aimed at improving performance when predictors are grouped in some way, for example when qualitative predictors are coded as dummy or one-hot variables (as is often implicitly done in ANOVA, for instance). Similarly to the elastic net, the penalty function induced by the group lasso lies between the l_1 penalty of the lasso in (12) and the l_2 penalty of the ridge in (8). To apply the group lasso, the vector of regression coefficients $\boldsymbol{\beta}$ is split in G vectors $\boldsymbol{\beta}_g$, where each vector represents the coefficients of predictors in that group. Denote by m_g the dimension of each vector $\boldsymbol{\beta}_g$. The group lasso corresponds to the following scale mixture of normals (Kyung et al., 2010):

$$\boldsymbol{\beta}_g|\tau_g^2, \sigma^2 \sim \text{MVN}(\mathbf{0}, \sigma^2\tau_g^2 I_{m_g}) \tag{15}$$
$$\tau_g^2|\lambda^2 \sim \text{Gamma}(\frac{m_g + 1}{2}, \frac{\lambda^2}{2}),\ \text{for } g = 1, \ldots, G,$$
$$\lambda \sim \text{half-Cauchy}(0, 1)$$

where MVN denotes the multivariate normal distribution with dimension m_g and I_{m_g} denotes an $(m_g \times m_g)$ identity matrix. Note that, contrary to the priors considered thus far, the group lasso prior does not consist of independent priors on the regression coefficients β_j, but rather independent priors on the groups of regression coefficients $\boldsymbol{\beta}_g$. If there is no grouping structure, $m_g = 1$ and the Bayesian group lasso in (15) reduces to the Bayesian lasso in (11). The scale mixture of normals in (15) leads to the following conditional prior for the regression coefficients (Kyung et al., 2010):

$$p(\beta_j|\sigma^2, \lambda) = C\exp\left\{-\frac{\lambda}{\sqrt{\sigma^2}}\sum_{g=1}^{G}\|\boldsymbol{\beta}_g\|\right\},$$
$$\text{for } g = 1, \ldots, G, \text{ and } j = 1, \ldots, p, \tag{16}$$

where $\|\boldsymbol{\beta}_g\| = (\boldsymbol{\beta}_g'\boldsymbol{\beta}_g)^{\frac{1}{2}}$ and C denotes the normalizing constant. Due to the simultaneous penalization of all coefficients in one group, all estimated regression coefficients in one group will be either zero or nonzero, depending on the value for λ.

3.6. Hyperlasso

Zou (2006) proposes the adaptive lasso as a generalization of the lasso that enjoys the oracle property (limitation (v) of the lasso),

i.e., it performs as well as if the true underlying model has been given. The central idea of the adaptive lasso is to separately weigh the penalty for each coefficient based on the observed data. A Bayesian adaptive lasso has been proposed, among others, by Alhamzawi et al. (2012) and Feng, Wu, and Song (2015). However, as noted by Griffin and Brown (2011), the weights included in the adaptive lasso place great demands on the data, which can lead to poor performance in terms of prediction and variable selection when the sample size is small. Therefore, Griffin and Brown (2011) propose the hyperlasso as a Bayesian alternative to the adaptive lasso, which is obtained through the following mixture of normals:

$$\begin{aligned} \beta_j|\phi_j^2 &\sim \text{Normal}(0, \phi_j^2) \\ \phi_j^2|\tau_j &\sim \text{Exponential}(\tau_j) \\ \tau_j|\nu, \lambda^2 &\sim \text{Gamma}(\nu, \frac{1}{\lambda^2}) \text{ for } j = 1, \ldots, p. \\ \lambda &\sim \text{half-Cauchy}(0, 1) \end{aligned} \tag{17}$$

This is equivalent to placing a Gamma mixing density on the hyperparameter of the double-exponential prior:

$$\begin{aligned} \beta_j|\tau_j &\sim \text{Double-Exponential}(0, (2\tau_j)^{1/2}) \\ \tau_j|\nu, \lambda^2 &\sim \text{Gamma}(\nu, \frac{1}{\lambda^2}), \text{ for } j = 1, \ldots, p. \end{aligned} \tag{18}$$

Note that the density of the hyperlasso prior strongly resembles the density of the lasso prior (Fig. 1), the main difference being that the hyperlasso has heavier tails than the lasso. Contrary to the priors considered thus far, this prior corresponds to a penalty that is non-convex implying that multiple posterior modes can exist. Therefore, care must be taken to ensure that the complete posterior distribution is explored. In addition, the hyperlasso prior for β is not conditioned on the error variance σ^2. Following Griffin and Brown (2011), we will consider the specific case of $\nu = 0.5$. However, whereas Griffin and Brown (2011) use cross-validation to choose λ, we will rely on a full and empirical Bayes approach.

3.7. Horseshoe

A popular shrinkage prior in the Bayesian literature is the horseshoe prior (Carvalho et al., 2010):

$$\begin{aligned} \beta_j|\tau_j^2 &\sim \text{Normal}(0, \tau_j^2) \\ \tau_j|\lambda &\sim \text{Half-Cauchy}(0, \lambda), \text{ for } j = 1, \ldots, p \\ \lambda|\sigma &\sim \text{Half-Cauchy}(0, \sigma). \end{aligned} \tag{19}$$

Note that Carvalho et al. (2010) explicitly include the half-Cauchy prior for λ in their specification, thereby implying a full Bayes approach. This formulation results in a horseshoe prior that is automatically scaled by the error standard deviation σ. The half-Cauchy prior can be written as a mixture of inverse Gamma and Gamma densities, so that the horseshoe prior in (19) can be equivalently specified as:

$$\begin{aligned} \beta_j|\tau_j^2 &\sim \text{Normal}(0, \tau_j^2) \\ \tau_j^2|\omega &\sim \text{inverse Gamma}(\frac{1}{2}, \omega) \\ \omega|\lambda^2 &\sim \text{Gamma}(\frac{1}{2}, \lambda^2) \\ \lambda^2|\gamma &\sim \text{inverse Gamma}(\frac{1}{2}, \gamma) \\ \gamma|\sigma^2 &\sim \text{Gamma}(\frac{1}{2}, \sigma^2) \end{aligned} \tag{20}$$

An expression for the marginal prior of the regression coefficients β_j is not analytically tractable. The name "horseshoe" prior arises from the fact that for fixed values $\lambda = \sigma = 1$, the implied prior for the shrinkage coefficient $\kappa_j = \frac{1}{1+\tau_j^2}$ is similar to a horseshoe shaped Beta(0.5, 0.5) prior. Large coefficients will lead to a shrinkage coefficient κ_j that is close to zero such that there is practically no shrinkage, whereas small coefficients will have a κ_j close to 1 and will be shrunken heavily. Note that the horseshoe prior is the only prior with an asymptote at zero (Fig. 1). Combined with the heavy tails, this ensures that small coefficients are heavily shrunken towards zero while large coefficients remain large. The horseshoe prior has also been termed a global-local shrinkage prior (e.g., Polson & Scott, 2011) because it has a predictor-specific local shrinkage component τ_j as well as a global shrinkage component λ. The basic intuition is that the global shrinkage parameter λ performs shrinkage on all coefficients and the local shrinkage parameters τ_j loosen the amount of shrinkage for truly large coefficients. Many global-local shrinkage priors (including the horseshoe and hyperlasso) are special cases of the general class of hypergeometric inverted-beta distributions (Polson & Scott, 2012). In addition to the full Bayes approach implied by the specification in (19), we will also consider an empirical Bayes approach to determine λ.

3.8. Regularized horseshoe

The horseshoe prior in Section 3.7 has the characteristic that large coefficients will not be shrunken towards zero too heavily. Indeed, this is one of the advertised qualities of the horseshoe prior (Carvalho et al., 2010). Although this property is desirable in theory, it can be problematic in practice, especially when parameters are weakly identified. In this situation, the posterior means of the regression coefficients might not exist and even if they do, the horseshoe prior can result in an unstable MCMC sampler (Ghosh, Li, & Mitra, 2017). To solve these problems (Piironen & Vehtari, 2017) propose the regularized horseshoe, which is defined as follows:

$$\begin{aligned} \beta_j|\tilde{\tau}_j^2, \lambda &\sim \text{Normal}(0, \tilde{\tau}_j^2\lambda), \text{ with } \tilde{\tau}_j^2 = \frac{c^2\tau_j^2}{c^2 + \lambda^2\tau_j^2} \\ \lambda|\lambda_0^2 &\sim \text{half-Cauchy}(0, \lambda_0^2), \text{ with } \lambda_0 = \frac{p_0}{p - p_0}\frac{\sigma}{\sqrt{n}} \\ \tau_j &\sim \text{half-Cauchy}(0, 1) \\ c^2|\nu, s^2 &\sim \text{inverse Gamma}(\nu/2, \nu s^2/2), \end{aligned} \tag{21}$$

where p_0 represents a prior guess of the number of relevant variables. The resulting prior will shrink small coefficients in the same way as the horseshoe,[6] but unlike the horseshoe, large coefficients will be shrunken towards zero by a Student's t distribution with ν degrees of freedom and scale s^2. Piironen and Vehtari (2017) use a Student's t distribution with $\nu = 4$ and $s^2 = 2$ and we use the same hyperparameters, although other choices are possible. As long as the degrees of freedom ν are small enough, the tails will be heavy enough to ensure a robust shrinkage pattern for large coefficients.

It is possible to specify a half-Cauchy prior for the global shrinkage parameter with a scale equal to 1 or the error standard deviation, i.e., $\lambda \sim \text{half-Cauchy}(0, 1)$ or $\lambda \sim \text{half-Cauchy}(0, \sigma)$, for example when no prior information is available regarding the number of relevant variables. However, as noted by Piironen and Vehtari (2017), the scale based on the a priori number of relevant variables will generally be much smaller than 1 or σ. In addition, even if the prior guess for the number of relevant parameters p_0 is incorrect, the results are robust to this choice as long as a half-Cauchy prior is used. Following the recommendations of Piironen

[6] Because of the similarity to the horseshoe, the density and contour plots for the regularized horseshoe are not substantially different from those of the horseshoe and are therefore not included in Figs. 1 and 3.

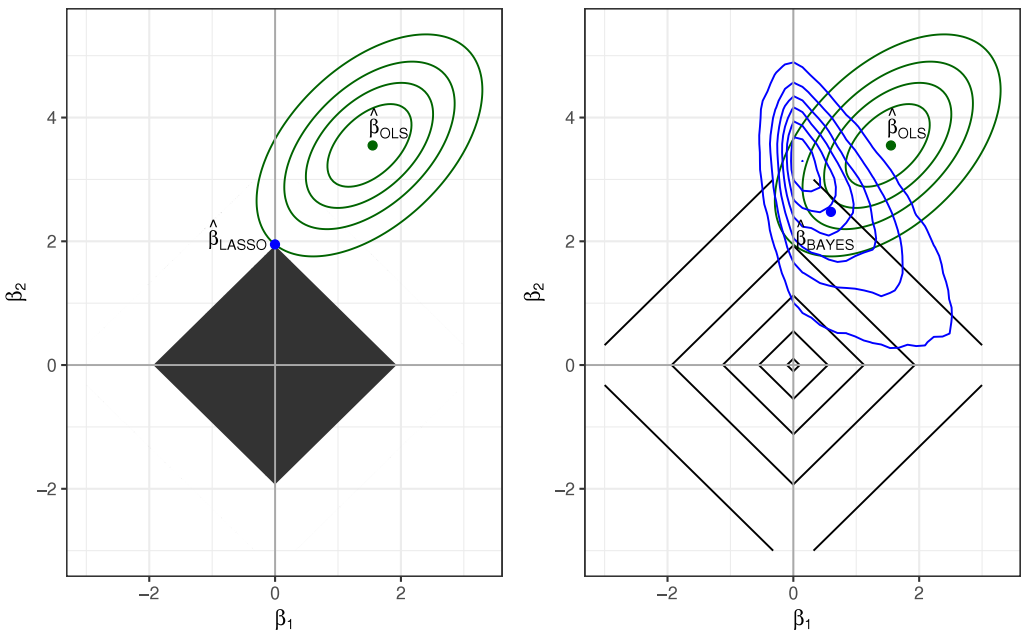

Fig. 2. Contour plot representing the sum of squared residuals, classical lasso constraint region (left), bivariate lasso prior and posterior distribution (right), and the classical and Bayesian penalized point estimates.

and Vehtari (2017), we will only consider a full Bayes approach to determine λ in the regularized horseshoe.

3.9. Discrete normal mixture

The normal mixture prior is a discrete mixture of a peaked prior around zero (the spike) and a vague proper prior (the slab); it is therefore also termed a spike-and-slab prior. It is substantially different from the priors considered thus far, which are all continuous mixtures of normal densities. Based on the data, regression coefficients close to zero will be assigned to the spike, resulting in shrinkage towards 0, while coefficients that deviate substantially from zero will be assigned to the slab, resulting in (almost) no shrinkage. Early proposals of mixture priors can be found in George and McCulloch (1993) and Mitchell and Beauchamp (1988), and a scale mixture of normals formulation can be found in Ishwaran and Rao (2005). We will consider the following specification of the mixture prior:

$$\begin{aligned}\beta_j|\gamma_j, \tau_j^2, \phi_j^2 &\sim (\gamma_j)\text{Normal}(0, \tau_j^2) + (1-\gamma_j)\text{Normal}(0, \phi_j^2) \qquad (22)\\ \tau_j^2 &\sim \text{inverse Gamma}(0.5, 0.5), \ \text{for } j = 1, \ldots, p,\end{aligned}$$

where τ_j is given a vague prior so that the variance of the slab is estimated based on the data and ϕ_j^2 is fixed to a small number, say $\phi_j^2 = 0.001$, to create the spike. By assigning an inverse Gamma(0.5, 0.5) prior on τ_j^2, the resulting marginal distribution of the slab component of the mixture is a Cauchy distribution.

There are several options for the prior on the mixing parameter γ_j. In this paper, we will consider the following two options: (1) γ_j as a Bernoulli distributed variable taking on the value 0 or 1 with probability 0.5, i.e., $\gamma_j \sim$ Bernoulli(0.5); and (2) γ_j uniformly distributed between 0 and 1, i.e., $\gamma_j \sim$ Uniform(0, 1). In the first option, which we label the Bernoulli mixture, each coefficient β_j is given either the slab or the spike as prior. The second option, labeled the uniform mixture, is more flexible in that each coefficient is given a prior consisting of a mixture of the spike and slab, with each component weighted by the uniform probabilities γ_j

The density of the normal mixture prior is presented in Fig. 1, which clearly shows the prior is a combination of two densities. The representation in Fig. 1 is based on a normal mixture with equal mixing probabilities, rather than a Bernoulli or uniform prior on the mixing probabilities. Note that the mixture prior is not conditioned on the error variance σ^2. We will only consider a full Bayesian approach for the mixture priors.

4. Illustrating the behavior of the shrinkage priors

4.1. Contour plots

Contour plots provide an insightful way to illustrate the behavior of classical penalties and Bayesian shrinkage priors. First, consider Fig. 2 which shows the frequentist and Bayesian contour plots for the lasso. In both plots, the green elliptical lines represent the contours of the sum of squared residuals, centered around the regular OLS estimate $\hat{\beta}_{OLS}$. The solid black diamond in the left plot represents the constraint region for the classical lasso penalty function for two predictors β_1 and β_2. The classical penalized regression solution $\hat{\beta}_{LASSO}$ is the point where the contour of the sum of squared residuals meets the constraint region. This point corresponds to the minimum of the penalized regression equation in (1). In the right plot, the diamond shaped contours reflect the shape of the lasso prior (Section 3.3). The contour of the Bayesian posterior distribution based on the lasso prior is shown in blue. As can be seen, the posterior distribution is located between the sum of squared residuals contour and the prior contour. The Bayesian posterior median estimate $\hat{\beta}_{BAYES}$ is added in blue and shrunken towards zero compared to the OLS estimate $\hat{\beta}_{OLS}$. Note that the posterior mode would correspond to the classical penalized regression solution, if the same value for the penalty parameter λ is used.

Fig. 3 shows the contour plots of the different shrinkage priors for two predictors β_1 and β_2, while Fig. 4 shows the contour plots for the lasso and group lasso for three predictors. From a classical penalization perspective, the lasso and elastic net penalties have sharp corners at $\beta_1 = \beta_2 = 0$. As a result, the contour of the sum of squared residuals will meet the contours of these penalties more easily at a point where one of the coefficients equals zero, which explains why these penalties can shrink coefficients to exactly zero. The ridge penalty, on the other hand, does not show these sharp corners and can therefore not shrink coefficients to exactly zero. From a Bayesian penalization perspective, the bivariate prior

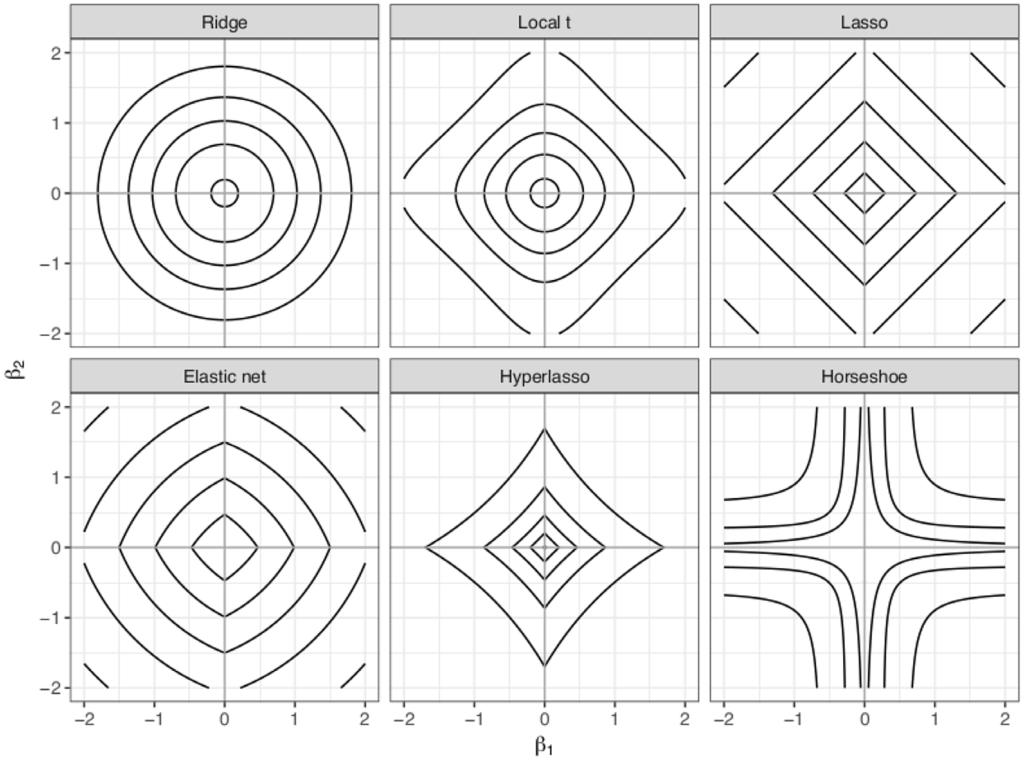

Fig. 3. Contour plots representing the bivariate prior distribution of the shrinkage priors.

contour plots illustrate the shrinkage behavior of the priors. For example, the hyperlasso and horseshoe have a lot of prior mass where at least one element is close to zero, while the ridge has most prior mass where both elements are close to zero. Fig. 3 also shows that the ridge, local Student's *t*, lasso, and elastic net are convex. This can be seen when drawing a straight line from one point to another point on a contour. For a convex distribution, the line lies completely within the contour. The hyperlasso and horseshoe prior are non-convex, which can be seen from the starlike shape of the contour. Frequentist penalization has generally focused on convex penalties, due to their computational convenience for optimization procedures. In the Bayesian framework, which relies on sampling (MCMC) techniques, the use of convex and non-convex priors is computationally similar. It is recommendable, however, to use multiple starting values in the case of non-convex priors due to possible multimodality of the posterior distribution (Griffin & Brown, 2011).

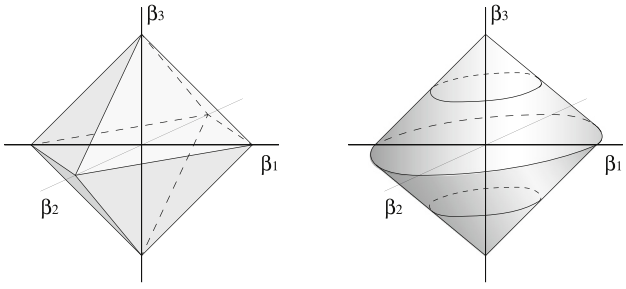

Fig. 4. Contour plots of the lasso (left) and group lasso (right) in $\mathbf{R}^3$, with β_1 and β_2 belonging to group 1 and β_3 belonging to group 2. For the group lasso, if we consider only β_1 and β_2, which belong to the same group, the contour resembles that of the ridge with most prior mass if both β_1 and β_2 are close to zero. On the other hand, if we consider β_1 and β_3, which belong to different groups, the contour is similar to that of the lasso, which has more prior mass where only one element is close to zero. This illustrates how the group lasso simultaneously shrinks elements belonging to the same group.

4.2. Shrinkage behavior

Prior shrinkage of small effects towards zero is important to obtain sparse solutions. Fig. 5 illustrates the shrinkage behavior of the priors in a simple normal model: $y \sim \text{Normal}(\beta, 1)$. We estimate β based on a single observation y, which is varied from 0 to 50. Using only a single observation is possible because the variance is known. The penalty parameter λ for each shrinkage prior is fixed to 1. The resulting difference between the posterior mean estimates and true means is shown in Fig. 5. The behavior of the priors varies greatly. Specifically, for the ridge and elastic net priors, the difference between the estimated and true effect increases as the true mean increases. For the lasso prior, the difference increases for small effects and then remains constant. Note how the difference for the elastic net lies between the difference obtained under the ridge and lasso priors, illustrating that the elastic net is a combination of the ridge and lasso priors. The other shrinkage priors all show some differences between estimated and true means for small effects, indicating shrinkage of these effects towards zero, but the difference is practically zero for large effects. The right column of Fig. 5 provides the same figure, but zoomed in on the small effects. Note how the regularized horseshoe shrinks large effects more than the horseshoe prior, but goes to zero eventually.

A similar illustration is presented in Fig. 6, but based on a full Bayes approach where λ is freely estimated. Thus, instead of fixing λ to a specific value, it is given a standard half-Cauchy prior distribution and estimated simultaneously with the other parameters in the model. Overall, all shrinkage priors show differences between true and estimated means for small effects, which decrease towards zero as the effect grows. Note that the difference is negative, indicating that the estimated mean is smaller than the true mean. Thus, all shrinkage priors heavily pull small effects towards zero, while asserting almost no influence on larger effects, although

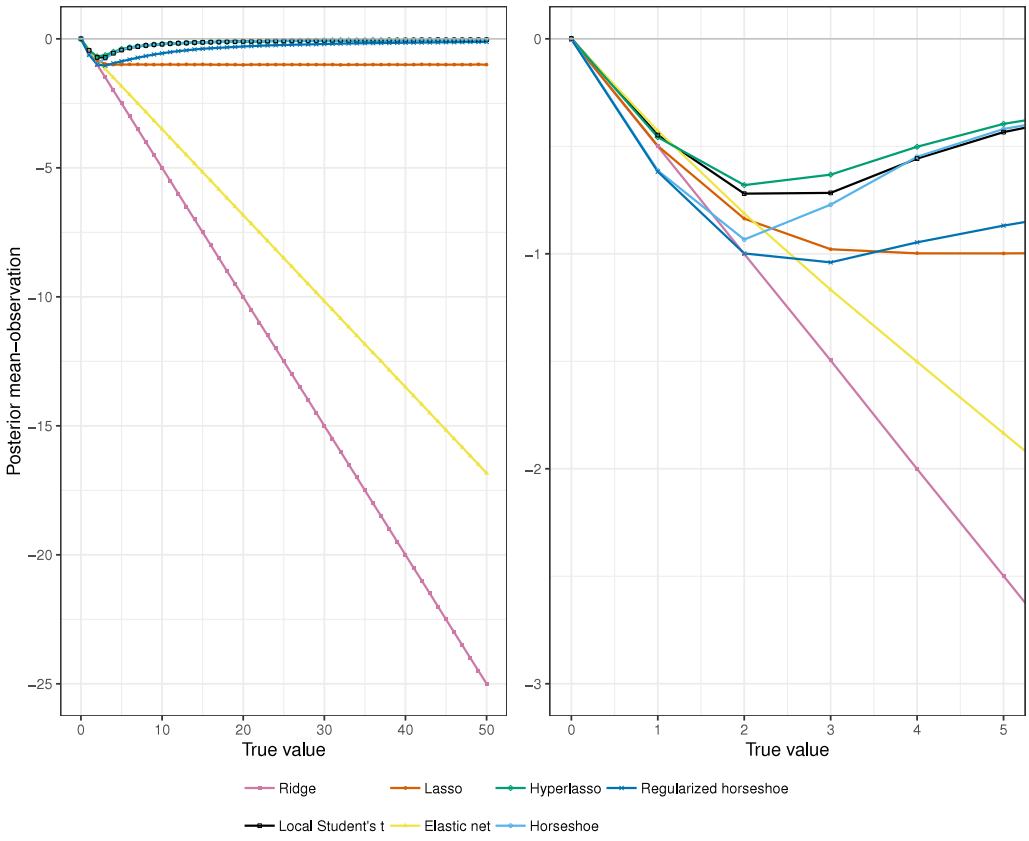

Fig. 5. Difference between the estimated and true effect for the shrinkage priors in a simple normal model with the penalty parameter λ fixed to 1.

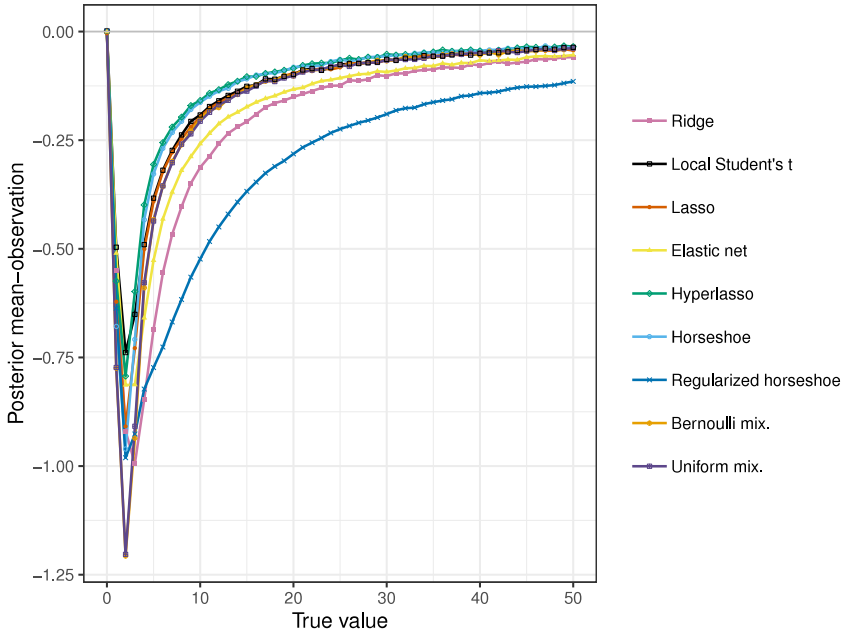

Fig. 6. Difference between the estimated and true effect for the shrinkage priors in a simple normal model with a half-Cauchy hyperprior specified for the penalty parameter λ.

some shrinkage still occurs even when the true mean equals 50. The mixture priors result in the largest differences between true and estimated small effects, indicating the most shrinkage, and the local Student's *t* prior shows the smallest difference for small effects. As the effect grows, the regularized horseshoe prior results in estimates farthest from the true effects, indicating the most shrinkage for large effects.

These illustrations indicate that when the penalty parameter is fixed, only the local Student's *t*, hyperlasso, and (regularized) horseshoe priors allow for shrinkage of small effects while

estimating large effects correctly. However, if a prior is specified for the penalty parameter, so that the uncertainty in this parameter is taken into account, all shrinkage priors show this desirable behavior.

5. Simulation study

5.1. Conditions

We conduct a Monte Carlo simulation study to compare the performance of the shrinkage priors and several frequentist penalization methods. We simulate data from the linear regression model, given by: $\boldsymbol{y} = \beta_0\mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\epsilon_i \sim \text{Normal}(0, \sigma^2)$. We consider six simulation conditions. Conditions are equal to the conditions considered in Li and Lin (2010). In addition, condition have also been considered in Kyung et al. (2010), Roy and Chakraborty (2016), Tibshirani (1996) and Zou and Hastie (2005). Condition has been included to investigate a setting in which $p > N$. The conditions are as follows[7]:

1. $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$; $\sigma^2 = 9$; $\boldsymbol{X}$ generated from a multivariate normal distribution with mean vector $\mathbf{0}$, variances equal to 1, and pairwise correlations between predictors equal to 0.5. The number of observations is $n = 240$, with 40 observations for training and 200 observations for testing the model.
2. $\boldsymbol{\beta} = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)'$; the other settings are equal to those in condition
3. $\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{15})$; $\sigma^2 = 225$; $\boldsymbol{x}_j = Z_1 + \omega_j$, for $j = 1, \ldots, 5$; $\boldsymbol{x}_j = Z_2 + \omega_j$, for $j = 6, \ldots, 10$; $\boldsymbol{x}_j = Z_3 + \omega_j$, for $j = 11, \ldots, 15$; and $\boldsymbol{x}_j \sim \text{Normal}(0, 1)$, for $j = 16, \ldots, 30$. Here, Z_1, Z_2, and Z_3 are independent standard normal variables and $\omega_j \sim \text{Normal}(0, 0.01)$. The number of observations is $n = 600$, with 200 observations for training and 400 observations for testing the model.
4. The number of observations is $n = 800$, with 400 observations for training and 400 observations for testing the model; the other settings are equal to those in condition
5. $\boldsymbol{\beta} = (\underbrace{3, \ldots, 3}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{3, \ldots, 3}_{10})$; the number of observations is $n = 440$, with 40 observations for training and 400 observations for testing the model; the other settings are equal to those in condition
6. The number of observations is $n = 55$, with 25 observations for training and 30 observations for testing the model; the other settings are equal to those in condition

We simulate 500 data sets per condition. All Bayesian methods have been implemented in the software package Stan (Stan development team, 2017c), which we call from R using `Rstan` (Stan development team, 2017a). We include the classical penalization methods available in the R-packages `glmnet` (Friedman, Hastie, & Tibshirani, 2010) and `grpreg` (Breheny & Huang, 2015), i.e., the ridge, lasso, elastic net, and group lasso, for comparison. For the classical penalization methods, the penalty parameter λ is selected based on cross-validation using 10 folds. We also include classical forward selection from the `leaps` (Lumley, 2017) package and we select the model based on three different criteria: the adjusted R^2, Mallows' C_p, and the BIC. For both the Bayesian and the classical group lasso, a grouping structure should be supplied for the analysis. We have used the grouping structure under which the data was simulated. Thus, for conditions 3 until 6, we have four groups with the following regression coefficient belonging to each group: $G_1 = \beta_1, \ldots, \beta_5$, $G_2 = \beta_6, \ldots, \beta_{10}$, $G_3 = \beta_{11}, \ldots, \beta_{15}$, and $G_4 = \beta_{16}, \ldots, \beta_{30}$. All code for the simulation study is available at https://osf.io/bf5up/.

5.2. Outcomes

The two main goals of regression analysis are: (1) to select variables that are relevant for predicting the outcome, and (2) to accurately predict the outcome. Therefore, we will focus on the performance of the shrinkage priors in terms of variable selection and prediction accuracy. Unlike frequentist penalization methods, Bayesian penalization methods do not automatically shrink regression coefficients to be exactly zero. A criterion is thus needed to select the relevant variables, for which we will use the credibility interval criterion.[8] Using the credibility interval criterion, a predictor is excluded when the credibility interval for β_j covers 0, and it is included when 0 is not contained in the credibility interval. This criterion thus depends on the percentage of posterior probability mass included in the credibility interval. We will investigate credibility intervals ranging from 0 to 100%, with steps of 10%. The optimal credibility interval is selected using the distance criterion (see e.g., Perkins & Schisterman, 2006), i.e.,

$$\text{distance} = \sqrt{(1 - \text{correct inclusion rate})^2 + (\text{false inclusion rate})^2}, \tag{23}$$

The credibility interval with the lowest distance is optimal in terms of the highest correct inclusion rate and lowest false inclusion rate. For the selected credibility interval, we will report Matthews' correlation coefficient (MCC; Matthews, 1975), which is a measure indicating the quality of the classification. MCC ranges between -1 and +1 with MCC $= -1$ indicating complete disagreement between the observed and predicted classifications and MCC $= +1$ indicating complete agreement.

To assess the prediction accuracy of the shrinkage priors, we will consider the prediction mean squared error (PMSE) for each replication. To compute the PMSE, we first estimate the regression coefficients $\hat{\boldsymbol{\beta}}$ on the training data only. These estimates are then used to predict the responses on the outcome variable of the test set, $\boldsymbol{y}^{gen}$, for which the actual responses, $\boldsymbol{y}$, are available. Prediction of $\boldsymbol{y}^{gen}$ occurs within the "generated quantities" block in Stan, meaning that for each MCMC draw, y_i^{gen} is generated such that we obtain the full posterior distribution for each y_i^{gen}. The mean of this posterior distribution is used as estimate for y_i^{gen}. The PMSE for each replication can then be computed as: $\frac{1}{N}\Sigma_{i=1}^{N}(y_i^{gen} - y_i)^2$. For each condition, this will result in 500 PMSEs, one for each replication, of which we will compute the median. Furthermore, to assess the uncertainty in the median PMSE estimate, we will bootstrap the standard error (SE) by resampling 500 PMSEs from the obtained PMSE values and computing the median. This process is repeated 500 times and the standard deviation of the 500 bootstrapped median PMSEs is used as SE of the median PMSE.

[7] We have also considered two additional conditions in which $p > n$ and the predictors are not highly correlated. Unfortunately, most shrinkage priors resulted in too much non-convergence to trust the results. A description of these additional conditions and the available results for the priors that did obtain enough convergence is available at https://osf.io/nveh3/. Additionally, we would like to refer to Kaseva (2018) where a more sparse, modified version of condition 1 is considered.

[8] We have also considered the scaled neighborhood criterion (Li & Lin, 2010) and a fixed cut-off value to select the predictors. The scaled neighborhood criterion excludes a predictor if the posterior probability contained in $[-\sqrt{\text{var}(\beta_p|\mathbf{y})}, \sqrt{\text{var}(\beta_p|\mathbf{y})}]$ exceeds a certain threshold. However, this criterion generally performed worse than the credibility interval criterion. For the fixed cut-off value we excluded predictors when the posterior estimate $|\hat{\beta}| \leq 0.1$ based on Feng et al. (2015). However, the choice of this threshold is rather arbitrary and resulted in very high false inclusion rates.

Table 2
Median prediction mean squared error (PMSE) with bootstrapped standard errors in brackets for the shrinkage priors.

Prior	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6
Full Bayes						
Ridge	10.95 (0.08)	10.49 (0.09)	243.09 (0.86)	236.07 (0.95)	319.84 (1.6)	371.61 (7.14)
Local Student's t	10.83 (0.09)	10.71 (0.08)	242.79 (0.8)	236.04 (0.89)	317.72 (2.31)	359.32 (6.67)
Lasso	10.78 (0.07)	10.53 (0.09)	238.89 (0.84)	234.29 (0.99)	316.43 (2.14)	360.37 (5.84)
Elastic net	10.93 (0.08)	10.62 (0.08)	243.23 (0.86)	236.1 (0.93)	323.24 (1.86)	387.47 (6.12)
Group lasso	NA[a]	NA[a]	241.06 (0.84)	235.35 (0.87)	316.23 (2.14)	358.17 (7.14)
Hyperlasso	10.77 (0.09)	10.52 (0.08)	238.69 (0.78)	234.24 (0.97)	316.29 (2.31)	356.45 (5.28)
Horseshoe	10.68 (0.07)	10.88 (0.09)	231.97 (0.87)	230.20 (0.82)	316.56 (2.43)	355.69 (4.18)
Regularized horseshoe true p_0[b]	10.69 (0.07)	10.58 (0.08)	233.42 (0.97)	230.43 (0.84)	316.51 (2.03)	356.93 (4.37)
Bernoulli mixture	***10.57*** (0.10)	11.26 (0.09)	***230.34*** (0.78)	***229.12*** (0.94)	322.35 (2.27)	357.62 (4.89)
Uniform mixture	***10.57*** (0.10)	11.24 (0.09)	230.42 (0.75)	229.36 (0.91)	322.40 (2.03)	359.25 (4.76)
Empirical Bayes						
Ridge	10.96 (0.08)	*10.29* (0.1)	242.71 (0.9)	235.91 (0.92)	317.85 (2.04)	421.86 (11.93)
Local Student's t	10.97 (0.08)	10.30 (0.08)	242.85 (0.89)	236.05 (0.88)	317.33 (2.12)	375.99 (6.85)
Lasso	10.78 (0.09)	10.49 (0.08)	238.86 (0.77)	234.10 (0.95)	317.38 (2.02)	424.82 (12.56)
Elastic net	10.95 (0.09)	10.31 (0.09)	242.67 (0.9)	235.92 (0.88)	316.64 (1.79)	365.52 (5.91)
Group lasso	NA[a]	NA[a]	240.89 (0.78)	235.29 (0.92)	315.05 (2.21)	369.00 (6)
Hyperlasso	10.79 (0.08)	10.43 (0.07)	238.62 (0.8)	234.02 (0.99)	***314.44*** (2.55)	436.24 (12.02)
Horseshoe	10.67 (0.08)	10.87 (0.08)	231.55 (0.89)	229.79 (0.82)	320.73 (3.08)	***354.69*** (4.13)
Classical penalization						
Ridge	10.96 (0.07)	**10.11** (0.06)	241.64 (1.13)	235.25 (1.04)	318.44 (2.48)	494.50 (7.29)
Lasso	10.70 (0.09)	11.06 (0.07)	235.76 (0.90)	231.23 (1.07)	339.09 (3.44)	410.84 (7.33)
Elastic net	10.72 (0.08)	10.89 (0.07)	235.96 (0.88)	231.54 (0.93)	335.50 (3.28)	394.76 (7.78)
Group lasso	NA[a]	NA[a]	233.11 (0.70)	229.76 (0.98)	343.06 (3.14)	407.14 (7.47)
Forward selection						
BIC	11.01 (0.07)	13.08 (0.11)	681.09 (3.95)	679.33 (2.85)	379.59 (4.48)	417.35 (8.34)
Mallows' C_p	11.06 (0.10)	12.45 (0.11)	464.94 (3.81)	466.25 (3.09)	478.00 (33.33)	687.46 (12.26)
Adjusted R^2	11.31 (0.12)	11.83 (0.10)	248.51 (2.35)	241.03 (1.98)	357.79 (3.20)	402.58 (8.73)

Note.
[a]No results are available for the group lasso in condition 1 and 2, since no grouping structure is present in these conditions.
[b]p_0 denotes the prior guess for the number of relevant variables, which was set to the true number of relevant variables, except in condition 2 where all eight variables are relevant, so we set $p_0 = 7$.
The smallest median PMSE per condition across methods is shown in bold and the smallest median PMSE per condition for the Bayesian methods is shown in italics.

5.3. Convergence

Convergence will be assessed using split $\hat{R}$, which is a version of the often used potential scale reduction factor (PSRF; Gelman & Rubin, 1992) that is implemented in Stan (Stan development team, 2017b, p. 370–373). Additionally, Stan reports the number of divergent transitions. A divergent transition indicates that the approximation error in the algorithm accumulates (Betancourt, 2017; Monnahan, Thorson, & Branch, 2016), which can be caused by a too large step size, or because of strong curvature in the posterior distribution. As a result, it can be necessary to adapt the settings of the algorithm or to reparametrize the model. For the simulation, we initially employed a very small step size (0.001) and high target acceptance rate (0.999), however, these settings result in much slower sampling. Therefore, in the later conditions we used the default step size and a lower target acceptance rate (0.85) and only reran the replications that did not converge with the stricter settings (i.e., smaller step size and higher target acceptance rate). Only if all parameters had a PSRF < 1.1 and there were no divergent transitions, did we consider a replication as converged.[9] We have only included those conditions in the results with at least 50% convergence (i.e., at least 250 converged replications). The convergence rates are available at https://osf.io/nveh3/.

5.4. Prediction accuracy

Table 2 shows the median PMSE per condition for the shrinkage priors and classical penalization methods. For the regularized horseshoe, the prior guess for the number of relevant variables p_0 was based on the data-generating model, however, the results were comparable when no prior guess or an incorrect prior guess was used. For all methods, the median PMSE increases as the condition becomes more complex. The smallest median PMSE per condition across methods is shown in bold and the smallest median PMSE per condition for the Bayesian methods is shown in italics. In condition 1, 3, and 4 the full Bayesian Bernoulli mixture prior performs best; in condition 2, the classical ridge performs best; in condition 5, the empirical Bayesian hyperlasso performs best; and in condition 6, the empirical Bayesian horseshoe performs best. However, the differences between the methods are relatively small. Only in condition 6, where the number of predictors is larger than the number of observations, the differences between the methods in terms of PMSE become more pronounced. As expected, forward selection performs the worst, especially when Mallows' C_p or the BIC is used to select the best model. This illustrates the advantage of using penalization, even when $p < n$. Overall, we can conclude that in terms of prediction accuracy the penalization methods perform quite similarly, except when $p > n$.[10]

[9] For the horseshoe prior, all replications in all conditions resulted in one or more divergent transitions, despite reparametrization of the model. The regularized horseshoe also resulted in divergent transitions for most replications, although the percentage of divergent transitions was on average much lower for the regularized horseshoe compared to the horseshoe. The percentages divergent transitions are available at https://osf.io/nveh3/. To be able to include these priors in the overview, we have only considered the PSRF to assess convergence and manually checked the traceplots. However, see Kaseva (2018) for a deeper investigation into the divergent transitions and alternative parametrizations of the horseshoe prior.

[10] We have also computed the PMSE for a large test set with 1,000,000 observations as an approximation to the theoretical prediction error. In general, the theoretical PMSEs did not differ substantially from the PMSE in Table 2, except in condition 6 where the theoretical PMSE was generally larger. The theoretical PMSEs are available online at https://osf.io/nveh3/.

5.5. Variable selection accuracy

Table 3 shows MCC and the correct and false inclusion rates for the optimal CIs for the shrinkage priors and MCC and the inclusion rates for the classical penalization methods, which automatically select predictors. The bold values indicate the best inclusion rates across all methods, whereas the italic values indicate the best inclusion rates across the Bayesian methods. Again, for the regularized horseshoe the results were comparable regardless of whether a correct, incorrect, or no prior guess was used. In the first condition, the classical penalization methods outperform the Bayesian methods in terms of correct inclusion rates, but at the cost of higher false inclusion rates. This is a well known problem of the lasso and elastic net when cross-validation is used to select the penalty parameter λ. A solution to this problem is to use stability selection to determine λ (Meinshausen & Bühlmann, 2010). The optimal Bayesian methods in the first condition based on the highest value for MCC are the mixture priors, both of which have reasonable correct and false inclusion rates. Note that, generally, the differences with the other Bayesian methods are relatively small in condition 1. In condition 3 and 4, the correct inclusion rates are generally high and the false inclusion rates are increased as well. As a result, the optimal Bayesian methods in condition 3 show a trade-off between correct and false inclusion rates, with the empirical Bayes group lasso having the highest value for MCC. However, the differences in MCC between most Bayesian methods are small and MCC is generally lower compared to condition 1 due to the increased false inclusion rates. In condition 4, multiple methods show a correct inclusion rate of 1, combined with a high false inclusion rate. In terms of MCC, the empirical Bayes ridge prior performs best. In condition 5, both rates and thus the MCC values are slightly lower across all methods, which is a result of the optimal CI being smaller. The full Bayes lasso and regularized horseshoe perform best in terms of MCC, although the other shrinkage priors show comparable MCC values. Condition 6 shows the most pronounced differences between the methods and the greatest trade-off between correct and false inclusion rates. None of the Bayesian methods attain a value for the MCC greater than 0.51, and some shrinkage priors (i.e., the empirical Bayes ridge and lasso) result in a MCC value of only 0.28. In conclusion, although there exist differences between the methods in terms of variable selection accuracy, there is not one method that performs substantially better than the other methods in terms of both correct and false inclusion rates.

6. Empirical applications

We will now illustrate the shrinkage priors on two empirical data sets. An R package `bayesreg` is available online (https://github.com/sara-vanerp/bayesreg) that can be used to apply the shrinkage priors. The first illustration (math performance) shows the benefits of using shrinkage priors in a situation where the number of predictors is smaller than the number of observations. In the second illustration (communities and crime), the number of predictors is larger than the number of observations, and it is necessary to use some form of regularization in order to fit the model.

6.1. Math performance

In this illustration, we aim to predict the final math grade of 395 Portugese students in secondary schools (Cortez & Silva, 2008), obtained from the UCL machine learning repository[11] (Lichman, 2013). The data set includes 30 predictors covering demographic, social and school related characteristics, such as parents' education and the time spent studying. The continuous predictors were standardized and dummy variables were used for the categorical predictors, resulting in a total of 39 predictors. We split the data into an approximately equal training ($n = 197$) and test ($n = 198$) set.

Table 4 presents the computation time in seconds for each method, the prediction mean squared error, and the number of included predictors. We have not included the results for the horseshoe prior because this prior resulted in divergent transitions, which in turn led to instable results (specifically, the PMSE varied greatly when rerunning the analysis).

It is clear that the Bayesian methods are computationally much more intensive than the classical penalization methods, especially the regularized horseshoe and mixture priors. The advantage brought by this increased computation time, however, is the more straightforward interpretation of results such as credibility intervals and the automatic computation of uncertainty estimates. This can be seen in Fig. 7 which shows the posterior density for one regression coefficient β_1 using the lasso prior and its 95% credibility interval (i.e., the shaded dark blue area). The bootstrapped 95% confidence interval obtained using the `HDCI` package (Liu, Xu, & Li, 2017) in R is shown by the dashed black lines and can be seen to underestimate the uncertainty. This problem is often observed using classical lasso estimation (Kyung et al., 2010). The PMSE clearly illustrates the advantage of penalization, even though the number of predictors is not greater than the sample size. Compared to regression using OLS, all penalization methods show lower PMSEs. Moreover, all penalization methods outperform forward selection in terms of PMSE. Between the different penalization methods, differences in PMSE are small.

The last column in Table 4 reports the number of included predictors for each method. Fig. 8 shows which predictors are included for each method. Each point indicates an included predictor, based on the optimal CI from condition 5 in the simulation study. OLS does not exclude any predictors, neither does the classical ridge generally although in this data set one coefficient was estimated to be zero. Most shrinkage priors included 22 predictors, with the hyperlasso and regularized horseshoe resulting in a slightly sparser solution. The mixture priors selected much less predictors (9) compared to the other methods. The number of included predictors for the forward selection method ranged from 4 to 25, depending on the criterion used to select the best model.

Based on the predicted errors and the number of included predictors, we conclude that essentially all Bayesian methods and the classical penalization methods performed best. The computation time for the Bayesian methods was considerably larger than for the classical methods. However, this increased computation time results in automatic availability of uncertainty estimates which were generally larger compared to classical bootstrapped confidence intervals.

6.2. Communities and crime

We illustrate the shrinkage priors on a data set containing 125 predictors of the number of violent crimes per 100,000 residents in different communities in the US (Redmond & Baveja, 2002) obtained from the UCL machine learning repository[12] (Lichman, 2013). The predictor variables include community characteristics, such as the median family income and the percentage of housing that is occupied, as well as law enforcement characteristics,

[11] The data is available at https://archive.ics.uci.edu/ml/datasets/Student+Performance.

[12] We used the unnormalized data, available at https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime+Unnormalized.

Table 3
Matthews' correlation coefficient (MCC) and correct and false inclusion rates based on the optimal credibility intervals (CIs) selected using the distance criterion.

Prior	Condition 1				Condition 3				Condition 4				Condition 5				Condition 6			
	Selected CI (%)	MCC	Correct inclusion	False inclusion	Selected CI (%)	MCC	Correct inclusion	False inclusion	Selected CI (%)	MCC	Correct inclusion	False inclusion	Selected CI (%)	MCC	Correct inclusion	False inclusion	Selected CI (%)	MCC	Correct inclusion	False inclusion
Full Bayes																				
Ridge	90	0.78	0.852	0.087	60	0.66	0.993	0.385	60	0.67	***1.000***	0.387	40	0.57	0.826	0.259	30	0.50	0.829	0.341
Local Student's t	80	0.76	0.884	0.132	60	0.67	0.998	0.381	70	0.60	0.875	0.291	40	0.58	0.827	0.246	30	***0.51***	0.820	0.318
Lasso	80	0.77	0.887	0.126	50	0.64	0.997	0.421	50	0.63	***1.000***	0.438	30	***0.59***	*0.863*	0.283	30	0.50	0.788	0.284
Elastic net	90	0.77	0.851	0.094	60	0.63	0.973	0.386	60	0.67	0.999	0.387	40	0.55	0.824	0.272	30	0.49	***0.836***	0.361
Group lasso	NA[a]	NA[a]	NA[a]	NA[a]	60	0.67	0.987	0.364	60	0.68	***1.000***	0.375	40	0.57	0.822	0.246	30	***0.51***	0.824	0.320
Hyperlasso	80	0.77	0.880	0.117	50	0.64	*0.999*	0.419	50	0.63	***1.000***	0.434	40	0.56	0.773	0.202	30	0.50	0.764	0.253
Horseshoe	70	0.78	0.886	0.116	20	0.49	*0.999*	0.609	20	0.49	***1.000***	0.603	20	0.56	0.858	0.311	20	0.50	0.795	0.294
Regularized horseshoe	70	0.78	0.889	0.118	40	0.64	0.949	0.351	30	0.61	***1.000***	0.453	40	***0.59***	0.794	0.193	30	***0.51***	0.771	0.247
true p_0[b]																				
Bernoulli mixture	50	***0.80***	*0.893*	0.099	20	0.66	0.992	0.381	20	0.66	0.993	0.388	20	0.55	0.735	*0.159*	20	0.48	0.627	0.127
Uniform mixture	50	***0.80***	0.889	0.100	20	0.66	0.989	0.381	20	0.65	0.990	0.390	20	0.55	0.733	0.160	20	0.48	0.628	*0.123*
Empirical Bayes																				
Ridge	90	0.78	0.847	0.081	70	0.52	0.781	*0.276*	70	*0.72*	0.982	0.289	40	0.57	0.819	0.240	30	0.28	0.690	0.222
Local Student's t	90	0.79	0.845	***0.080***	60	0.67	*0.999*	0.377	70	0.70	0.949	0.291	40	0.58	0.821	0.241	30	0.49	0.764	0.279
Lasso	80	0.77	0.885	0.118	50	0.64	*0.999*	0.417	50	0.63	***1.000***	0.433	30	0.57	0.831	0.270	20	0.28	0.591	0.273
Elastic net	90	0.78	0.848	0.085	60	0.67	*0.999*	0.377	70	0.71	0.968	0.290	40	0.58	0.834	0.251	30	0.50	0.810	0.314
Group lasso	NA[a]	NA[a]	NA[a]	NA[a]	60	*0.68*	0.998	0.361	70	0.60	0.880	**0.278**	40	0.58	0.820	0.240	30	0.46	0.728	0.277
Hyperlasso	80	0.78	0.883	0.109	50	0.65	*0.999*	0.414	50	0.63	***1.000***	0.431	40	0.53	0.767	0.199	20	0.32	0.575	0.268
Horseshoe	70	0.78	0.876	0.108	20	0.51	0.997	0.578	20	0.51	0.997	0.576	20	0.53	0.840	0.308	20	0.48	0.756	0.266
Classical penalization																				
Lasso	NA[c]	0.72	0.923	0.210	NA[c]	0.66	0.642	**0.023**	NA[c]	0.67	0.632	*0.008*	NA[c]	0.33	0.418	**0.108**	NA[c]	0.27	0.368	0.116
Elastic net	NA[c]	0.62	**0.971**	0.361	NA[c]	**0.97**	**1.000**	0.031	NA[c]	**0.99**	**1.000**	0.013	NA[c]	0.47	0.645	0.161	NA[c]	0.39	0.548	0.153
Group lasso	NA[a]	NA[a]	NA[a]	NA[a]	NA[c]	0.52	**1.000**	0.482	NA[c]	0.50	**1.000**	0.500	NA[c]	0.34	**0.893**	0.603	NA[c]	0.37	0.787	0.462
Forward selection																				
BIC	NA[c]	0.77	0.843	0.093	NA[c]	−0.087	0.086	0.139	NA[c]	−0.092	0.081	0.137	NA[c]	−0.056	0.171	0.213	NA[c]	−0.0056	0.252	0.249
Mallows' C_p	NA[c]	0.72	0.895	0.176	NA[c]	0.023	0.188	0.164	NA[c]	0.023	0.186	0.164	NA[c]	−0.071	0.122	0.172	NA[c]	−0.074	0.024	**0.052**
Adjusted R_2	NA[c]	0.60	0.938	0.343	NA[c]	0.14	0.332	0.201	NA[c]	0.15	0.329	0.197	NA[c]	0.0085	0.338	0.328	NA[c]	0.053	0.378	0.323

Note.

[a] No results are available for the group lasso in condition 1, since no grouping structure is present in these conditions.

[b] p_0 denotes the prior guess for the number of relevant variables, which was set to the true number of relevant variables, except in condition 2 where all eight variables are relevant, so we set $p_0 = 7$.

[c] For the classical penalization methods, the lasso, elastic net, and forward selection automatically shrink some coefficients to exactly zero so that no criterion such as a confidence interval for variable selection is needed. The ridge is not included, since it does not automatically shrink coefficients to zero and therefore always has a correct and false inclusion rate of 1.

The highest MCC and correct inclusion rate and the lowest false inclusion rate per condition across methods are shown in bold and the highest MCC and best rates per condition for the Bayesian methods are shown in italics.

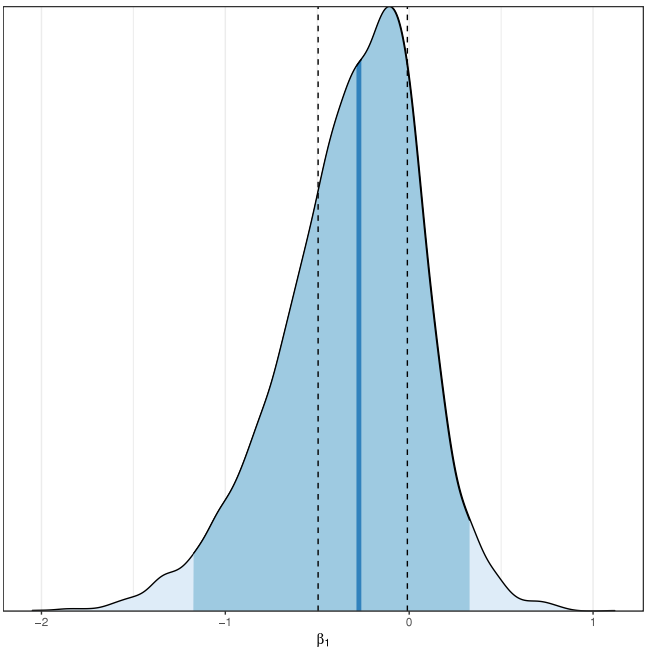

Fig. 7. Posterior density for β_1 in the math performance application using the Bayesian lasso. The dark blue line depicts the posterior median, the shaded dark blue area depicts the 95% credibility interval. The black dashed lines depict the bootstrapped 95% confidence interval of the classical lasso.

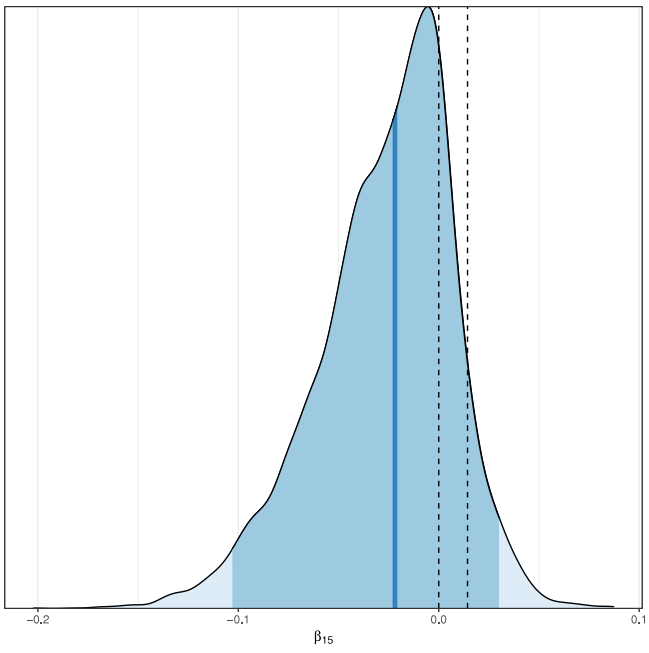

Fig. 9. Posterior density for β_{15} in the crime application using the Bayesian lasso. The dark blue line depicts the posterior median, the shaded dark blue area depicts the 95% credibility interval. The black dashed lines depict the bootstrapped 95% confidence interval of the classical lasso.

such as the number of police officers and the police operating budget. We created dummy variables for the two nominal predictors in the data set, resulting in a total of 172 predictors. For the group lasso, all dummy variables corresponding to one predictor make up a group. The number of observations is 319, after removing all cases with at least one missing value on any of the predictors.[13] We split the data into approximately equal training ($n = 159$) and test ($n = 160$) sets. All predictors were normalized to have zero mean and unit variance and the outcome variable was log transformed.

Table 5 reports the computation time in seconds for each method, as well as the PMSE and the number of selected variables. Again, the horseshoe prior resulted in divergent transitions and is therefore excluded from the results. The posterior density using the lasso prior for β_{15} is shown in Fig. 9, with the dark blue shaded area depicting the 95% credibility intervals and the dashed black lines depicting the bootstrapped 95% confidence interval of the classical lasso. Again, the bootstrapped confidence interval is much smaller than the Bayesian credibility interval and located far from the posterior median estimate (i.e., the dark blue line).

In addition, most Bayesian methods resulted in a lower PMSE than the classical methods, except for the mixture priors. The forward selection method resulted in much larger PMSEs, except when Mallows' C_p was used to find the best model, however, this model retained only 1 predictor. On the other hand, using the Adjusted R^2 criterion led to a model that included 141 predictors. This illustrates the arbitrariness of using forward selection. Fig. 10 shows which predictors are included for each method. Each point indicates an included predictor, based on the optimal CI from condition 6 in the simulation study. Apart from the forward selection methods, the classical elastic net excludes most predictors. Interestingly, the Bayesian elastic net and lasso retain many more predictors than the classical elastic net and lasso. However, not all predictors that are retained by the classical lasso and elastic net are also retained by the Bayesian lasso and elastic net. Specifically, the predictors included by the classical methods but not by the Bayesian methods all correspond to dummy variables for State. The hyperlasso and lasso methods all include 46 predictors, whereas the ridge, local Student's t, elastic net, and group lasso priors all retain around 60 predictors. The mixture priors both include 54 predictors. The regularized horseshoe retains the least predictors of all Bayesian methods, only 31. The classical ridge retains almost all predictors, but estimated some coefficients to be equal to zero in this data set.

Based on this illustration, we conclude that the Bayesian penalization methods outperform the classical penalization methods in terms of prediction error. The prediction errors of the Bayesian penalization methods do not differ substantially, except for the mixture priors which showed larger PMSEs. The shrinkage priors differ in how much shrinkage they perform and thus in the number of predictors that are selected.

7. Discussion

The aim of this paper was to provide insights about the different shrinkage priors that have been proposed for Bayesian penalization to avoid overfitting of regression models in the case of many predictors. We have reviewed the literature on shrinkage priors and presented them in a general framework of scale mixtures of normal distributions to enable theoretical comparisons between the priors. To model the penalty parameter λ, which is a central part of the penalized regression model, a full Bayes and an empirical Bayes approach were employed.

Although the various prior distributions differ substantially from each other, e.g., regarding their tails or convexity, the priors performed very similarly in the simulation study in those conditions where $p < n$. Overall, the performance was comparable to the classical penalization approaches. The math performance example clearly showed the advantage of using penalization to avoid overfitting when $p < n$. As in the simulation study, the

[13] Although the Bayesian framework allows for straightforward imputation of missing values, we removed all cases with missing values to provide an illustration of the shrinkage methods in a sparse data set.

Table 4
Computation time in seconds (with a 2.8 GHz Intel Core i7 processor), prediction mean squared error (PMSE), and number of included predictors for the different methods for the math performance application.

Shrinkage prior	Computation time (s)	PMSE	Number of included predictors
Full Bayes			
Ridge	179	19.53	22
Local Student's t	361	19.44	22
Lasso	219	19.25	22
Elastic net	354	19.53	23
Group lasso	342	19.36	22
Hyperlasso	199	19.18	19
Regularized horseshoe with p_0	1474	19.12	17
Bernoulli mixture	24524	19.31	9
Uniform mixture	4370	19.28	9
Empirical Bayes			
Ridge	341	19.42	22
Local Student's t	443	19.47	22
Lasso	444	19.26	22
Elastic net	603	19.41	22
Group lasso	534	19.50	22
Hyperlasso	387	19.11	19
Classical penalization			
Ordinary least squares (OLS)	0.013	22.56	39
Ridge	0.118	18.95	38
Lasso	0.0.072	19.11	20
Elastic net	0.0.053	19.25	20
Group lasso	0.187	19.43	21
Forward selection			
BIC	0.006	21.13	4
Mallows' C_p	0.006	21.42	13
Adjusted R^2	0.006	22.44	25

Table 5
Computation time in seconds (with a 2.8 GHz Intel Core i7 processor), prediction mean squared error (PMSE), and number of included predictors for the different methods for the crime application.

Shrinkage prior	Computation time (s)	PMSE	Number of included predictors
Full Bayes			
Ridge	677	0.217	61
Local Student's t	1973	0.216	60
Lasso	2068	0.216	46
Elastic net	242	0.216	62
Group lasso	3044	0.216	61
Hyperlasso	1066	0.215	46
Regularized horseshoe with p_0	15803	0.226	31
Bernoulli mixture	60006	1.706	54
Uniform mixture	26080	1.683	54
Empirical Bayes			
Ridge	1195	0.218	60
Local Student's t	1912	0.216	57
Lasso	4207	0.215	46
Elastic net	417	0.217	57
Group lasso	3992	0.217	62
Hyperlasso	2016	0.215	46
Classical penalization			
Ridge	0.376	0.258	160
Lasso	0.200	0.508	33
Elastic net	0.164	0.460	26
Group lasso	0.408	0.663	55
Forward selection			
BIC	0.023	1.500	17
Mallows' C_p	0.023	0.276	1
Adjusted R^2	0.023	4.093	141

prediction errors in the math example were comparable across penalization methods, although the number of included predictors varied across methods. Finally, although classical penalization is much faster than Bayesian penalization, it does not automatically provide accurate uncertainty estimates and the bootstrapped confidence intervals obtained for the classical methods were generally much smaller compared to the Bayesian credibility intervals.

The differences between the methods became more pronounced

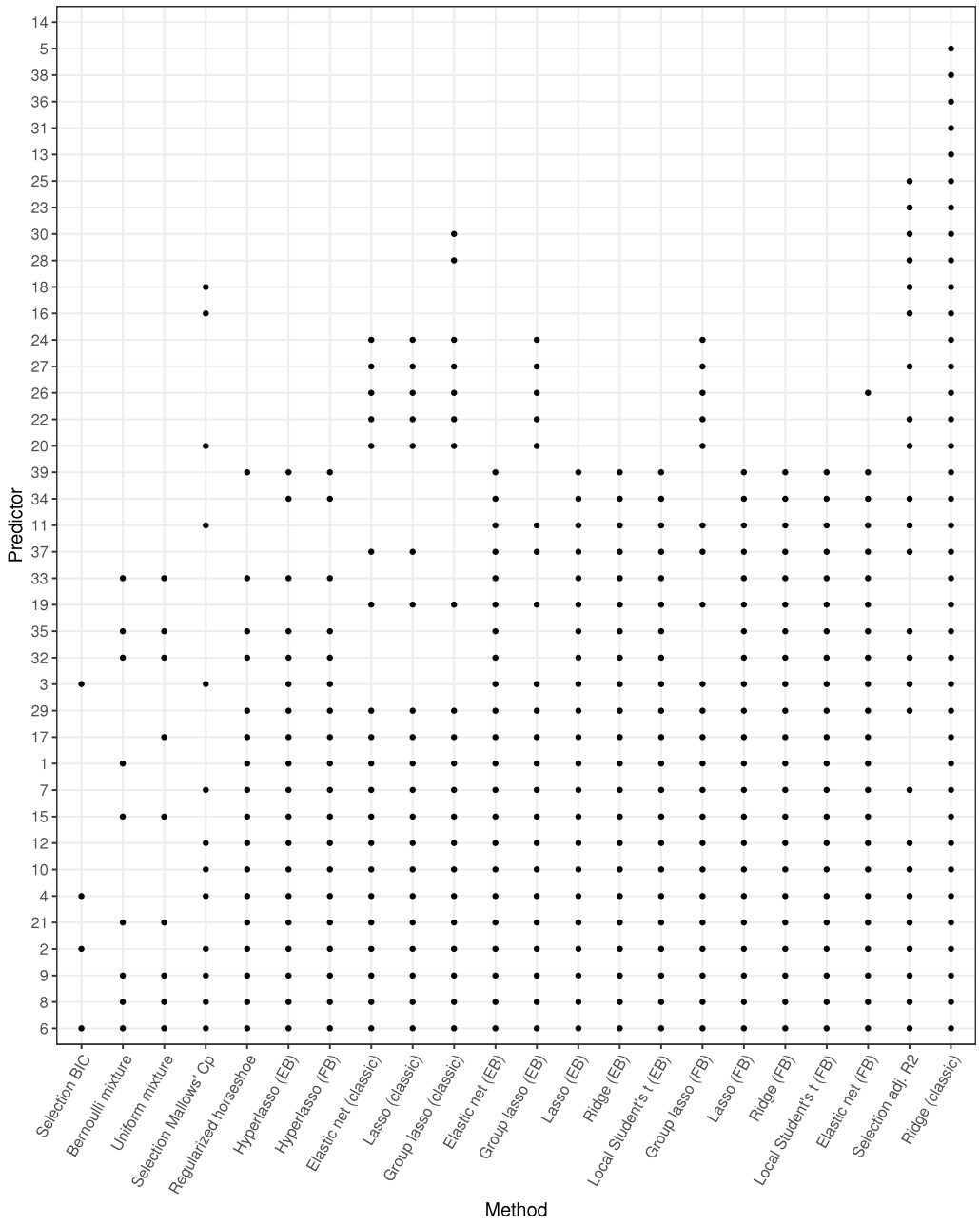

Fig. 8. Overview of the included predictors for each method in the math performance application. Points indicate that a predictor is included based on the optimal credibility interval (CI) from condition 5 in the simulation study. The methods on the *x*-axis are ordered such that the method that includes the least predictors is on the left and the method that includes the most predictors is on the right. The predictors on the *y*-axis are ordered with the predictor being included the least on top and the predictor being included the most at the bottom.

when $p > n$. In condition 6 of the simulation study,the (regularized) horseshoe and hyperlasso priors performed substantially better than most of the other shrinkage priors in terms of PMSE. This is most likely due to the fact that the hyperlasso and (regularized) horseshoe are non-convex global-local shrinkage priors and are therefore particularly adept at keeping large coefficients large, while shrinking the small coefficients enough towards zero. Future research should consider various high-dimensional simulation conditions to further explore the performance of the shrinkage priors in such settings, for example by varying the correlations between the predictors. The crime example illustrated the use of the penalization methods further in a $p > n$ situation. In this example, most Bayesian approaches resulted in smaller prediction errors than the classical approaches (except for the mixture priors). Also in terms of the predictors that were included there were considerable differences between the various approaches.

An important goal of the shrinkage methods discussed in this paper is the ultimate selection of relevant variables. Throughout this paper, we have focused on the use of marginal credibility intervals to do so. However, the use of marginal credibility intervals to perform variable selection can be problematic, since the marginal intervals can behave differently compared to joint credibility intervals. This is especially the case for global shrinkage priors, such as the (regularized) horseshoe prior since these priors induce shrinkage on all variables jointly (Piironen, Betancourt, Simpson, & Vehtari, 2017). Future research should investigate whether the

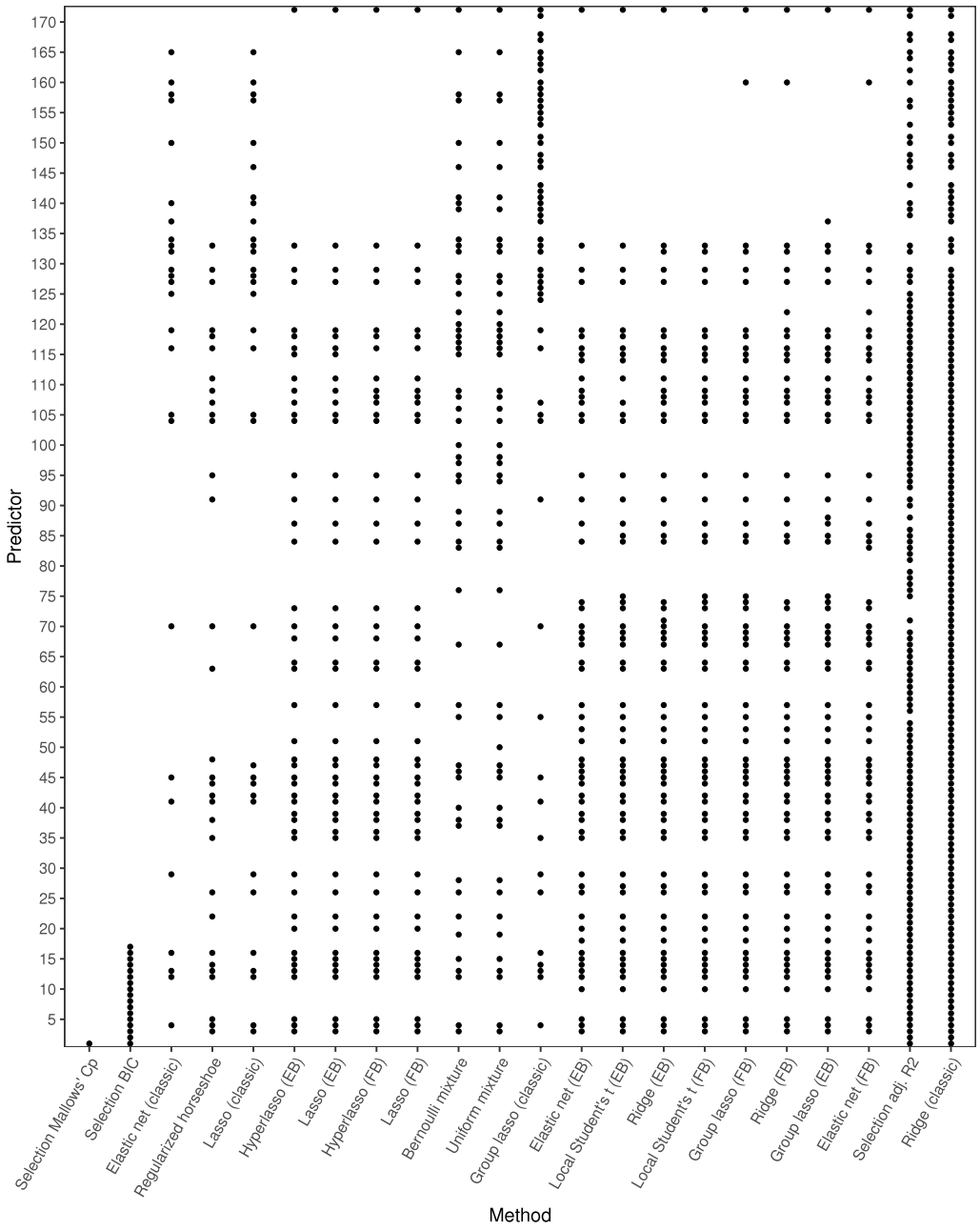

Fig. 10. Overview of the included predictors for each method in the crime application. Points indicate that a predictor is included based on the optimal credibility interval (CI) from condition 6 in the simulation study. The methods on the *x*-axis are ordered such that the method that includes the least predictors is on the left and the method that includes the most predictors is on the right.

variable selection accuracy can be further improved by using methods that jointly select relevant variables (for example, projection predictive variable selection; Piironen & Vehtari, 2016, or decoupled shrinkage and selection; Hahn & Carvalho, 2015).

Throughout this paper, we focused on the linear regression model. Hopefully, the results presented in this paper and the corresponding R package `bayesreg` available at https://github.com/sara-vanerp/bayesreg will lead to an increased use of penalization methods in psychology, because of the improved performance in terms of prediction error and variable selection accuracy compared to forward subset selection. The shrinkage priors investigated here can be applied in more complex models in a straightforward manner. For example, in generalized linear regression models such as logistic and Poisson regression models, the only necessary adaptation is to incorporate a link function in the model. Although not currently available in the R-package, the available Stan modelfiles can be easily adapted to generalized linear models (GLMs). Additionally, packages such as `brms` (Bürkner, 2017) and `rstanarm` (Stan Development Team, 2016) include several of the shrinkage priors described here, or allow the user to specify them manually. Both packages support (multilevel) GLMs, although `rstanarm` relies on precompiled models and is therefore less flexible than `brms`. Currently, an active area of research employs Bayesian penalization in latent variable models, such as factor models (see e.g., Jacobucci & Grimm, 2018; Lu et al., 2016) and quantile structural equation models (see e.g., Feng et al., 2017). The characteristics and behaviors of the shrinkage priors presented in this paper can be a useful first step in solving these more challenging problems.

Acknowledgments

This research was supported by a Research Talent Grant from the Netherlands Organisation for Scientific Research. We would like to thank Aki Vehtari, Carlos Carvalho, Tuomas Kaseva, and Charlie Strauss for providing helpful comments on an earlier version of this manuscript and pointing out relevant references.

References

Alhamzawi, R., Yu, K., & Benoit, D. F. (2012). Bayesian adaptive lasso quantile regression. *Statistical Modelling*, *12*(3), 279–297.

Andersen, M. R., Vehtari, A., Winther, O., & Hansen, L. K. (2017). Bayesian inference for spatio-temporal spike-and-slab priors. *Journal of Machine Learning Research (JMLR)*, *18*(139), 1–58.

Armagan, A., Dunson, D. B., & Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica*.

Azmak, O., Bayer, H., Caplin, A., Chun, M., Glimcher, P., Koonin, S., & Patrinos, A. (2015). Using big data to understand the human condition: The Kavli HUMAN project. *Big Data*, *3*(3), 173–188.

Bae, K., & Mallick, B. K. (2004). Gene selection using a two-level hierarchical Bayesian model. *Bioinformatics*, *20*(18), 3423–3430.

Berger, J. O. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, *3*, 385–402.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. ArXiv preprint arXiv:1701.02434.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. (2016). The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, *12*(4), 1105–1131.

Bhadra, A., Datta, J., Polson, N. G., & Willard, B. T. (2017). Lasso meets horseshoe: a survey. ArXiv preprint arXiv:1706.10179.

Bhattacharya, A., Pati, D., Pillai, N. S., & Dunson, D. B. (2012). Bayesian shrinkage. ArXiv preprint arXiv:1212.6088.

Bornn, L., Gottardo, R., & Doucet, A. (2010). Grouping priors and the Bayesian elastic net. ArXiv preprint arXiv:1001.4083.

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*, 173–187.

Bürkner, P. -C. (2017). brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Caron, F., & Doucet, A. (2008). Sparse Bayesian nonparametric regression. In *Proceedings of the 25th international conference on Machine learning - ICML '08*. Association for Computing Machinery (ACM).

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

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance. In A. Brito and J. Teixeira (Eds.), *Proceedings of 5th future business technology conference* (pp. 5–12).

Derksen, S., & Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, *45*(2), 265–282.

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

Fawcett, T. (2015). Mining the quantified self: personal knowledge discovery as a challenge for data science. *Big Data*, *3*(4), 249–266.

Feng, X. -N., Wang, Y., Lu, B., & Song, X. -Y. (2017). Bayesian regularized quantile structural equation models. *Journal of Multivariate Analysis*, *154*, 234–248.

Feng, X. -N., Wu, H. -T., & Song, X. -Y. (2015). Bayesian adaptive lasso for ordinal regression with latent variables. *Sociological Methods & Research*, *46*(4), 926–953.

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

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

George, E. I., & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881.

Ghosh, J., Li, Y., & Mitra, R. (2017). On the use of cauchy prior distributions for bayesian logistic regression. *Bayesian Analysis*, *13*(2), 359–383.

Griffin, J. E., & Brown, P. J. (2005). *Alternative prior distributions for variable selection with very many more variables than observations*. University of Warwick. Centre for Research in Statistical Methodology.

Griffin, J. E., & Brown, P. J. (2011). Bayesian hyper-lassos with non-convex penalization. *Australian & New Zealand Journal of Statistics*, *53*(4), 423–442.

Griffin, J., & Brown, P. (2017). Hierarchical shrinkage priors for regression models. *Bayesian Analysis*, *12*(1), 135–159.

Hahn, P. R., & Carvalho, C. M. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association*, *110*(509), 435–448.

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

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity*. CRC press.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Hsiang, T. C. (1975). A Bayesian view on ridge regression. *The Statistician*, *24*(4), 267.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.

Jacobucci, R., & Grimm, K. J. (2018). Comparison of frequentist and Bayesian regularization in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 639–649.

Kaseva, T. (2018). Convergence diagnosis and comparison of shrinkage priors. github repository. https://livefull.github.io.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, *12*(3), 753–778.

Li, Q., & Lin, N. (2010). The Bayesian elastic net. *Bayesian Analysis*, *5*(1), 151–170.

Lichman, M. (2013). UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences.

Liu, H., Xu, X., & Li, J. J. (2017). HDCI: High dimensional confidence interval based on lasso and bootstrap. R package version 1.0-2. https://CRAN.R-project.org/package=HDCI.

Lu, Z. H., Chow, S. M., & Loken, E. (2016). Bayesian factor analysis as a variable-selection problem: Alternative priors and consequences. *Multivariate Behavioral Research*, *51*(4), 519–539.

Lumley, T. (2017). leaps: Regression Subset Selection. R package version 3.0. https://CRAN.R-project.org/package=leaps.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451.

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–484.

Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *72*(4), 417–473.

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

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

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2016). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, *8*(3), 339–348.

Mulder, J., & Pericchi, L. R. (2018). The matrix-f prior for estimating and testing covariance matrices. *Bayesian Analysis*, *13*(4), 1189–1210.

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

Peltola, T., Havulinna, A. S., Salomaa, V., & Vehtari, A. (2014). Hierarchical bayesian survival analysis and projective covariate selection in cardiovascular event risk prediction. In *Proceedings of the eleventh UAI conference on bayesian modeling applications workshop-volume 1218* (pp. 79–88). CEUR-WS. org.

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "Optimal" Cutpoints Obtained using Two Criteria based on the Receiver Operating Characteristic Curve. *American Journal of Epidemiology*, *163*(7), 670–675.

Piironen, J., Betancourt, M., Simpson, D., & Vehtari, A. (2017). Contributed comment on article by van der Pas, Szabó, and van der Vaart. *Bayesian Analysis*, *12*(4), 1264–1266.

Piironen, J., & Vehtari, A. (2016). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735.

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

Polson, N. G., & Scott, J. G. (2011). Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian statistics, vol. 9* (pp. 501–538). Oxford University Press (OUP).

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902.

Polson, N. G., Scott, J. G., & Windle, J. (2014). The bayesian bridge. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *76*(4), 713–733.

Redmond, M., & Baveja, A. (2002). A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research*, *141*(3), 660–678.

Roy, V., & Chakraborty, S. (2016). Selection of tuning parameters, solution paths and standard errors for bayesian lassos. *Bayesian Analysis*.

Stan Development Team (2016). rstanarm: Bayesian applied regression modeling via Stan. R package version 2.13.1. URL http://mc-stan.org/.

Stan development team (2017a). RStan: The R interface to Stan, R package version 2.16.2. URL http://mc-stan.org.

Stan development team (2017b). Stan Modeling Language Users Guide and Reference Manual, Version 2.17.0. URL http://mc-stan.org.

Stan development team (2017c). The Stan Core Library, Version 2.16.0. URL http://mc-stan.org.

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

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

van Erp, S., Mulder, J., & Oberski, D. L. (2018). Prior sensitivity analysis in default bayesian structural equation modeling. *Psychological Methods*, *23*(2), 363–388.

Vehtari, A., Gabry, J., Yao, Y., & Gelman, A. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. R package version 2.0.0. URL https://CRAN.R-project.org/package=loo.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, *74*(3), 646–648.

van de Wiel, M. A., Beest, D. E. t., & Münch, M. (2017). Learning from a lot: Empirical Bayes in high-dimensional prediction settings. ArXiv preprint arXiv:1709.04192.

Wolpert, D. H., & Strauss, C. E. M. (1996). What bayes has to say about the evidence procedure. In *Maximum entropy and bayesian methods* (pp. 61–78). Springer Netherlands.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *68*(1), 49–67.

Zhao, S., Gao, C., Mukherjee, S., & Engelhardt, B. E. (2016). Bayesian group factor analysis with structured sparsity. *Journal of Machine Learning Research (JMLR)*, *17*(196), 1–47.

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

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