Bayesian Analysis (0000) **00**, Number 0, pp. 1

On variance estimation for Bayesian variable selection

Gemma E. Moran*, Veronika Ročková† and Edward I. George*

Abstract. Consider the problem of high dimensional variable selection for the Gaussian linear model when the unknown error variance is also of interest. In this paper, we argue that the use conjugate continuous shrinkage priors for Bayesian variable selection can have detrimental consequences for such error variance estimation. Instead, we recommend the use of priors which treat the regression coefficients and error variance as independent *a priori*. We revisit the canonical reference for invariant priors, Jeffreys (1961), and highlight a caveat with their use that Jeffreys himself noted. For the case study of Bayesian ridge regression, we demonstrate that these scale-invariant priors severely underestimate the variance. More generally, we discuss how these priors also interfere with the mechanics of the Bayesian global-local shrinkage framework. With these insights, we extend the Spike-and-Slab Lasso of Ročková and George (2016) to the unknown variance case, using an independent prior for the error variance. Our procedure outperforms both alternative penalized likelihood methods and the fixed variance case on simulated data.

Keywords: Bayesian variable selection, Bayesian shrinkage, Penalized likelihood, Jeffreys' priors.

1 Introduction

Consider the classical linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathbf{I}_n) \tag{1.1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a vector of responses, $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_p] \in \mathbb{R}^{n \times p}$ is a fixed regression matrix of p potential predictors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ is a vector of unknown regression coefficients and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is the noise vector of independent normal random variables with σ^2 as their unknown common variance.

When $\boldsymbol{\beta}$ is sparse so that most of its elements are zero or negligible, finding the non-negligible elements of $\boldsymbol{\beta}$, the so-called variable selection problem, is of particular importance. Whilst this problem has been studied extensively from both frequentist and Bayesian perspectives, much less attention has been given to the simultaneous estimation of the error variance σ^2. Accurate estimates of σ^2 are important to discourage fitting the noise beyond the the signal, thereby helping to avoid overfitting the data. Variance estimation is also essential in uncertainty quantification for inference and prediction.

*Department of Statistics, University of Pennsylvania, gmoran@wharton.upenn.edu, edgeorge@wharton.upenn.edu

†Booth School of Business, University of Chicago, veronika.rockova@chicagobooth.edu

 DOI: 0000

In the frequentist literature, the question of estimating the error variance has begun to be addressed with papers including the scaled Lasso (Sun and Zhang, 2012) and the square-root Lasso (Belloni et al., 2014). Contrastingly, in the Bayesian literature, the error variance has been fairly straightforwardly estimated by including σ^2 in prior specifications. Despite this conceptual simplicity, the majority of theoretical guarantees for Bayesian procedures restrict attention to the case of known σ^2, as there is not a generally agreed upon prior specification when σ^2 is unknown. More specifically, priors on $\boldsymbol{\beta}$ and σ are typically introduced in one of two ways: either via a joint scale-invariant prior framework or via an independence prior framework.

The joint scale invariant prior for $\boldsymbol{\beta}$ and σ in (1.1) is obtained by a form

$$\pi(\boldsymbol{\beta}, \sigma) = \pi(\boldsymbol{\beta}|\sigma)\pi(\sigma), \tag{1.2}$$

where the regression coefficients depend *a priori* on σ^2 in a "scale-free way" through

$$\pi(\boldsymbol{\beta}|\sigma) = \frac{1}{\sigma^p} h(\boldsymbol{\beta}/\sigma) \tag{1.3}$$

for some proper density function $h(x)$. By "scale-free" (or scale-invariant) we mean that the prior $\pi(\boldsymbol{\beta}|\sigma)$ yields a marginal likelihood that is invariant with respect to a location-scale group of transformations of the outcomes (Bayarri et al., 2012). To complete the joint characterization (1.2), such a prior is typically coupled with the right-Haar prior for the location-scale group (Berger et al., 1998)

$$\pi(\sigma) \propto \frac{1}{\sigma}. \tag{1.4}$$

Many priors satisfy the invariance criterion. Prominent examples are the commonly used conjugate Gaussian prior

$$\boldsymbol{\beta}|\sigma^2, \tau^2 \sim N_p(0, \sigma^2\tau^2\mathbf{I}), \tag{1.5}$$

obtained with a Gaussian $h(\cdot)$ centered at zero, and scale mixtures thereof. Note that such scale invariant priors arise naturally as *conjugate* forms. Part of the popularity of these conjugate priors in Bayesian variable selection is because the structure allows for σ^2 to be easily integrated out of the posterior, resulting in closed form updates of posterior model probabilities and thereby more computationally efficient MCMC algorithms (George and McCulloch, 1997). For some continuous shrinkage priors, such as the Bayesian LASSO, the conjugate prior yields a unimodal posterior distribution that can be beneficial for MCMC. Here we will show, however, that despite computational advantages, this prior is not innocuous for variance estimation with continuous shrinkage priors. [1]

[1] In contrast to Bayarri et al. (2012), here we consider prior constructions that shrink rather than exclude predictors, rendering the prior size of the model ultimately p-dimensional. Bayarri et al. (2012) scale the prior (1.3) with σ^{-d} as opposed to σ^{-p}, where d is the number of active predictors out of the p candidates.

Alternatively one might treat $\boldsymbol{\beta}$ and σ^2 as independent *a priori* and use the formulation:

$$p(\boldsymbol{\beta}, \sigma^2) = p(\boldsymbol{\beta})p(\sigma^2). \tag{1.6}$$

Akin to the previous example, this simply corresponds to the following prior structure:

$$\boldsymbol{\beta} \sim N(0, \tau^2\mathbf{I}), \tag{1.7}$$

$$p(\sigma) \propto \sigma^{-1}. \tag{1.8}$$

In this paper we argue that scale-invariant priors on the regression coefficients should be avoided for high-dimensional Bayesian variable selection and shrinkage estimation with continuous shrinkage priors. Intuitively, these priors implicitly add p "pseudo-observations" to the posterior, distorting inference for the error variance. To avoid this problem, we recommend the use of independent priors on $\boldsymbol{\beta}$ and σ^2. Further, we extend the Spike-and-Slab Lasso of Ročková and George (2016) to the unknown variance case with an independent prior formulation, and highlight the performance gains over the known variance case via simulation studies. This implementation of the Spike-and-Slab Lasso is publicly available in the R package `SSLASSO` (Ročková and Moran, 2017).

The paper is structured as follows. In section 2, we discuss the historical justification for scale-invariant priors - specifically, how they arose as Jeffreys priors. We then highlight situations where we ought to depart from Jeffreys priors; namely, in multivariate situations. In section 3, we use Bayesian ridge regression as a case study to highlight why scale-invariant priors are a poor choice. In section 4, we examine the mechanisms of the Gaussian global-local shrinkage framework and illustrate why they are incompatible with the scale-invariant prior structure. In section 5, we draw connections between Bayesian regression and concurrent developments with variance estimation in the penalized likelihood literature. In section 6, we consider the Spike-and-Slab Lasso of Ročková and George (2016) and illustrate how the conjugate prior yields poor estimates of the error variance. We then extend the procedure to include the unknown variance case using an independent prior structure. and demonstrate via simulation studies how this leads to performance gains over not only the known variance case, but a variety of other variable selection procedures. We conclude with a discussion in section 7.

2 Scale-Invariant Priors

The canonical reference for scale-invariant priors is Jeffreys (1961). In this seminal work, Jeffreys (1961) introduces his general principle for deriving non-informative priors; for a parameter α, the Jeffreys prior is

$$\pi(\alpha) \propto |I(\alpha)|^{1/2}, \tag{2.1}$$

where $I(\alpha)$ is the Fisher information matrix. The main motivation given by Jeffreys (1961) for these priors was that they are invariant for all nonsingular transformations of the parameters. This property appeals to intuition regarding objectivity; ideally, the prior information we decide to include should not depend upon the choice of the parameterization, which itself is arbitrary.

Despite this intuitively appealing property, the following problem with this principle was spotted in the original work of Jeffreys (1961) and later re-emphasized by Robert et al. (2009) in their revisit of the work. Consider the model $Y_i \sim N(\mu, \sigma^2)$ for $i = 1, \ldots, n$. If we consider the parameters μ and σ separately, the Jeffreys priors are $\pi(\mu) \propto 1$ and $\pi(\sigma) \propto 1/\sigma$. Considering both parameters together, however, yields the Jeffreys prior $\pi(\mu, \sigma) \propto 1/\sigma^2$. This discrepancy is exaggerated when we include more parameters. In effect, by considering the parameters jointly as opposed to independently, we are implicitly including additional "pseudo-observations" of σ^2 and consequently distort our estimates of the error variance. As we shall see later, this *joint Jeffreys prior* for the parameters is akin to the scale-invariant formulation in (1.2).

This prior dependence between the parameters is explicitly repudiated by Jeffreys (1961) who states (with notation changed to match ours): "in the usual situation in an estimation problem, μ and σ^2 are each capable of any value over a considerable range, and neither gives any appreciable information about the other. We should then take: $\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma)$." Jeffreys also points out that the key problem with the joint Jeffreys prior is that it does not have the same reduction of degrees of freedom required by the introduction of additional nuisance parameters. We shall examine this phenomenon in more detail in the next section where we will discuss the consequences of using joint Jefferys priors and other scale-invariant formulations in Bayesian linear regression.

3 Bayesian Regression

Consider again the classical linear regression model in (1.1). For a non-informative prior, it is common to use $\pi(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$ (see, for example, Gelman et al. (2014)). Similarly to our earlier discussion, this prior choice corresponds to combining the *separate*, or independent, Jeffreys priors for $\boldsymbol{\beta}$ and σ. In contrast, the *joint* Jeffreys prior would be $\pi(\boldsymbol{\beta}, \sigma) \propto 1/\sigma^{p+1}$. Let us now examine the estimates resulting from the former, independent Jeffreys prior. In this case, we have the following marginal posterior mean estimate for the error variance:

$$\mathbb{E}[\sigma^2|\mathbf{Y}] = \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n - p - 1} \tag{3.1}$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ is the usual least squares estimator. We observe that the degrees of freedom adjustment, $n - p$, naturally appears in the denominator. This does not occur for the joint Jeffreys prior; there, the marginal posterior mean estimate is given by

$$\mathbb{E}[\sigma^2|\mathbf{Y}] = \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n - 1}. \tag{3.2}$$

For large p, this estimator with the joint Jeffreys prior will severely underestimate the error variance. Avoiding this, it is commonly accepted that the independent Jeffreys prior $\pi(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$ should be the default non-informative prior in this setting.

There is no such clarity, however, in the use of the scale-invariant priors for Bayesian linear regression. To add to this discourse, we draw a connection between these conjugate

priors and the joint Jeffreys prior, similarly arguing that independent priors should be used instead. We make this point with the following example. A common conjugate prior choice for Bayesian linear regression is

$$\boldsymbol{\beta}|\sigma^2, \tau^2 \sim N_p(0, \sigma^2\tau^2\mathbf{I}), \tag{3.3}$$

the scale invariant prior in (1.5). If we consider the parameter τ^2 to be fixed, this prior choice corresponds to Bayesian ridge regression. With an additional non-informative prior $\pi(\sigma) \propto 1/\sigma$, we then have the joint prior

$$\pi(\boldsymbol{\beta}|\sigma)\pi(\sigma) = \pi(\boldsymbol{\beta}, \sigma) \propto \frac{1}{\sigma^{p+1}} \exp\left\{-\frac{1}{2\sigma^2\tau^2}\|\boldsymbol{\beta}\|^2\right\}. \tag{3.4}$$

We see, however, that if we take $\tau^2 \to \infty$, corresponding to a completely non-informative prior on $\boldsymbol{\beta}$, we arrive at the prior $\pi(\boldsymbol{\beta}, \sigma) = 1/\sigma^{p+1}$. As noted at the beginning of the section, this is an inadequate prior as it does not result in the correct degrees of freedom adjustment. We further highlight the deficiencies of this prior choice in the following simulated example.

3.1 The failure of a scale-invariant prior

In this example, we take $n = 100$ and $p = 90$ and compare the least squares estimates of $\boldsymbol{\beta}$ and σ^2 to Bayesian ridge regression with (i) the scale-invariant formulation in (3.3) and (ii) the independent prior formulation below:

$$\pi(\boldsymbol{\beta}) \sim N_p(0, \tau^2\mathbf{I}). \tag{3.5}$$

For both Bayesian ridge regression procedures we use a non-informative prior on the error variance: $\pi(\sigma) \propto 1/\sigma$. We take $\tau^2 = 100$ as known for simplicity of exposition. The predictors $\mathbf{X}_i$, $i = 1, \ldots, p$ are generated as independent standard normal random variables. The true $\boldsymbol{\beta}_0$ is set to be sparse with only six non-zero elements. The response $\mathbf{Y}$ is generated according to (1.1) with the true variance being $\sigma^2 = 3$.

The scale-invariant prior formulation allows for the exact expressions for the posterior means of $\boldsymbol{\beta}$ and σ^2:

$$\mathbb{E}[\boldsymbol{\beta}|\mathbf{Y}] = [\mathbf{X}^T\mathbf{X} + \tau^{-2}\mathbf{I}]^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.6}$$

$$\mathbb{E}[\sigma^2|\mathbf{Y}] = \frac{\mathbf{Y}^T[\mathbf{I} - \mathbf{H}_R]\mathbf{Y}}{n-1} \tag{3.7}$$

where $\mathbf{H}_R = \mathbf{X}[\mathbf{X}^T\mathbf{X} + \tau^{-2}\mathbf{I}]^{-1}\mathbf{X}^T$. The independent prior formulation, however, does not yield closed form expressions for the posterior means of $\boldsymbol{\beta}$ and σ^2, and so we use a Gibbs sampler to approximate these, the details of which may be found in the supplementary material.

In Figure 1, we display a boxplot of the estimates of the standard deviation of the errors for (i) Least Squares, (ii) Scale-Invariant Bayesian ridge regression and (iii) Independent Bayesian ridge regression over 100 replications. Here the estimates from

least squares and the independent Bayesian ridge are similarly centered around the truth; however, the scale-invariant prior consistently underestimates the error variance quite severely with a median of the estimates of 0.05. This poor performance is a result of the bias induced by adding p "psuedo-observations" of σ^2 in the prior (3.4) which in turn prevents the natural degrees of freedom adjustment, as discussed earlier.

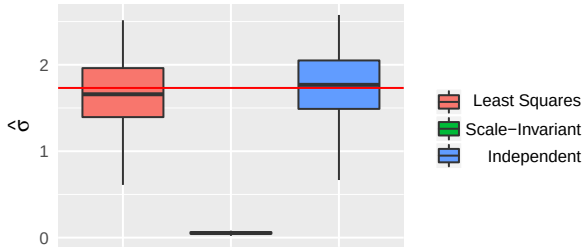

Figure 1: Estimated $\widehat{\sigma}$ for each procedure over 100 repetitions. The true variance $\sigma = \sqrt{3}$ is the red horizontal line.

A natural question to ask is: if the estimate of the variance is so poor, how does that affect the estimated regression coefficients? In this case, the Gaussian prior structure allows for σ^2 to be factorized out and so the estimate of $\boldsymbol{\beta}$ does not depend on the variance, as seen in (3.6). This lack of dependence on the variance is troubling, however, as we want to select fewer variables when the error variance is large and the signal-to-noise ratio is low. We contrast (3.6) with the conditional posterior mean of $\boldsymbol{\beta}$ in the independent prior case:

$$\mathbb{E}[\boldsymbol{\beta}|\sigma^2, \mathbf{Y}] = \frac{1}{\sigma^2}\left[\frac{\mathbf{X}^T\mathbf{X}}{\sigma^2} + \frac{\mathbf{I}}{\tau^2}\right]^{-1}\mathbf{X}^T\mathbf{Y}. \tag{3.8}$$

Note that for large σ^2, the posterior estimate for $\boldsymbol{\beta}$ will be close to zero, reflecting the relative lack of information. In a sense, the conjugate posterior mean in (3.6) is the estimate for $\boldsymbol{\beta}$ *assuming that* $\sigma^2 = 1$ *is known*. In other words, the conjugate prior formulation is effectively not even treating the error variance σ^2 as a unknown parameter. Indeed, the estimates of the variance from the scale-invariant Bayesian ridge regression procedure in Figure 1 are essentially meaningless.

The same phenomenon is also seen in EMVS (Ročková and George, 2014), which can be viewed as iterative Bayesian ridge regression with an adaptive penalty term for each regression coefficient β_j instead of the same τ^2 above. EMVS also uses a scale-invariant prior formulation in which $\boldsymbol{\beta}$ depends on σ^2 *a priori* similarly to (3.3). As in the above ridge regression example, EMVS yields good estimates for $\boldsymbol{\beta}$, but σ^2 is severely underestimated. This is evident in the Section 4 example of Ročková and George (2014) with $n = 100$ and $p = 1000$. There, conditionally on the modal estimate of β, the associated modal estimate of σ^2 is 0.0014, a severe underestimate of the true variance $\sigma^2 = 3$. Note that EMVS can be easily modified to the independent prior specification, as now has been done in the publicly available EMVS R package implementation.

4 Global-Local Shrinkage

In this section, we examine how the use of a scale-invariant prior affects the machinery of the Gaussian global-local shrinkage paradigm. The general prior structure for this paradigm is given by:

$$\beta_j \sim N(0, \sigma_\beta^2 \tau_j^2), \quad \tau_j^2 \sim \pi(\tau_j^2), \quad j = 1, \ldots, p \tag{4.1}$$

$$\sigma_\beta^2 \sim \pi(\sigma_\beta^2) \tag{4.2}$$

where τ_j is the "local" variance and σ_β^2 is the "global" variance. Note that taking σ_β^2 to be the same as the error variance σ^2 would result in a scale-invariant prior in this setting. For example, the Bayesian lasso of Park and Casella (2008) uses such a prior and can be recast in the Gaussian global-local shrinkage framework as follows:

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\beta}, \sigma^2 &\sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ \beta_j|\sigma^2, \tau_j^2 &\sim N(0, \sigma^2\tau_j^2), \quad \pi(\tau_j^2) = \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2}, \quad j = 1, \ldots, p \\ \pi(\sigma) &\propto \sigma^{-1}. \end{aligned} \tag{4.3}$$

Here, σ^2 has the dual role of estimating the error variance as well as acting as the global shrinkage parameter. This is problematic in light of the mechanics of global-local shrinkage priors. Specifically, Polson and Scott (2010) give the following requirements for the hyperparameters: $\pi(\tau_j^2)$ should have heavy tails so that it can be quite large, allowing for a few large coefficients to be identified; and $\pi(\sigma_\beta^2)$ should have substantial mass near zero to shrink all the regression coefficients so that the vast majority are negligible.

This heuristic is formalized in much of the shrinkage estimation theory. For example, van der Pas et al. (2016) prove that the following condition results in the posterior recovering nonzero means with the optimal rate:

(i) $\pi(\tau_j^2)$ should be a uniformly regular varying function which does not depend on n; and

(ii) $\sigma_\beta^2 = (q/n)^\alpha$ for $\alpha \geq 0$ (where q is number of non-zero β_j).

The uniformly regular varying property in (i) intuitively preserves the "flatness" of the prior even under transformations of the parameters, unlike traditional "non-informative" priors (Bhadra et al., 2016). In preserving these heavy tails, such priors for τ_j^2 allow for a a few large coefficients to be estimated. The second condition (ii) encourages σ_β^2 to tend to zero which would be a concerning property for the error variance. Hence, these results suggest we cannot combine the error variance with the global variance parameter on the regression coefficients: it cannot simultaneously shrink all the regression coefficients and be a good estimate of the residual variance.

An alternative formulation is given by Carvalho et al. (2010) for their horseshoe prior:

$$\beta_j|\sigma^2, \sigma_\beta^2, \tau_j^2 \sim N(0, \sigma^2\sigma_\beta^2\tau_j^2), \quad \tau_j^2 \sim \pi(\tau_j^2), \quad j = 1, \ldots, p \tag{4.4}$$

$$\sigma_\beta^2 \sim \pi(\sigma_\beta^2)$$
$$\sigma^2 \sim \pi(\sigma^2).$$

This formulation is equivalent to the model recommended by Piironen and Vehtari (2017) in which the prior for the global shrinkage parameter is scaled by the error variance. Such hierarchical models maintain scale-invariance but separate the roles of the error variance, σ^2, and the global shrinkage parameter, σ_β^2, unlike the Bayesian lasso (4.3). However, this prior structure still distorts the estimate of the error variance by the addition of "pseudo-observations", as discussed in the previous sections.

Let us also examine the resulting conditional posterior for $\boldsymbol{\beta}$ for the above model (4.4). This is given by:

$$\boldsymbol{\beta}|\mathbf{Y}, \sigma^2, \sigma_\beta^2, \tau_j^2 \sim N(\widetilde{\boldsymbol{\beta}}, \widetilde{\Sigma}) \tag{4.5}$$
$$\widetilde{\boldsymbol{\beta}} = \left[\mathbf{X}^T\mathbf{X} + \sigma_\beta^{-2}\mathbf{D}\right]^{-1}\mathbf{X}^T\mathbf{Y} \tag{4.6}$$
$$\widetilde{\Sigma} = \sigma^2\left[\mathbf{X}^T\mathbf{X} + \sigma_\beta^{-2}\mathbf{D}\right]^{-1} \tag{4.7}$$

where $\mathbf{D} = \text{diag}\{\tau_j^{-2}\}_{j=1}^p$. Note that, as in the Bayesian ridge regression example, the mean of $\boldsymbol{\beta}$ does not depend on the error variance, σ^2. We again argue that this is problematic; when the signal to noise ratio is too low, we should not select any non-zero regression coefficients. That is, the estimate of $\boldsymbol{\beta}$ *should* depend on the error variance, instead of the situation in (4.6) where the error variance is essentially fixed as $\sigma^2 = 1$. In contrast, Piironen and Vehtari (2017) advocate for the scale-invariant prior (4.4), arguing that it leads to a prior on the number of non-zero coefficients which does not depend on σ and n. We argue, however, that such a prior fails to take into account the uncertainty inherent in the variable selection process.

Thus, we recommend independent priors on both the error variance and regression coefficients to both prevent distortion of the global-local shrinkage mechanism and to obtain better estimates of the error variance.

5 Penalized Likelihood Methods

Here we pause briefly to examine connections between Bayesian methods and developments in estimating the error variance in the penalized regression literature. Such connections can be drawn as penalized likelihood methods are implicitly Bayesian; the penalty functions can be interpreted as priors on the regression coefficients and so these procedures also in effect yield MAP estimates.

One of the first papers to consider the unknown error variance case for the Lasso was Städler et al. (2010), who suggested the following penalized loss function for introducing unknown variance into the frequentist Lasso framework:

$$L_{pen}(\boldsymbol{\beta}, \sigma^2) = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2} + \frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1 + n\log\sigma. \tag{5.1}$$

This objective function is very similar to a Bayesian scale-invariant prior formulation but is missing the extra $p \log \sigma$ term. As a result, it is not strictly a log posterior. Interestingly, Sun and Zhang (2012) found that the resulting joint estimate of this formulation may give biased estimates for the noise level. Instead, Sun and Zhang (2012) proposed the "scaled Lasso", an algorithm which alternatively minimizes the following penalized joint loss function via coordinate descent:

$$L_\lambda(\boldsymbol{\beta}, \sigma) = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma} + \frac{n\sigma}{2} + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{5.2}$$

This loss function is a penalized version of Huber's concomitant loss function, and so may be viewed as performing robust high-dimensional regression. This is also equivalent to the "square-root Lasso" of Belloni et al. (2014). Sun and Zhang (2012) proved that the resulting estimate $\widehat{\sigma}(\mathbf{X}, \mathbf{Y})$ is consistent for the "oracle" estimator $\sigma = \|y - \mathbf{X}\boldsymbol{\beta}^*\|/\sqrt{n}$, where $\boldsymbol{\beta}^*$ are the true regression coefficients, for the value of $\lambda \propto \sqrt{2n \log p}$. Interestingly, the scaled Lasso estimators for the regression coefficients and error variance are scale equivariant in the sense that $\widehat{\boldsymbol{\beta}}(\mathbf{X}, c\mathbf{Y}) = c\widehat{\boldsymbol{\beta}}(\mathbf{X}, \mathbf{Y})$ and $\widehat{\sigma}(\mathbf{X}, c\mathbf{Y}) = |c|\widehat{\sigma}(\mathbf{X}, \mathbf{Y})$. This notion is different, however, to the invariance property of Jeffreys priors discussed in section 2. The Jeffreys invariance property relates to the choice of parameterization, whereas the equivariant property relates to transformations of the data.

6 Spike-and-Slab Lasso with Unknown Variance

We now turn to the Spike-and-Slab Lasso (SSL, Ročková and George, 2016) and consider how to incorporate the unknown variance case. As the name suggests, the SSL involves placing a mixture prior on the regression coefficients $\boldsymbol{\beta}$, where each β_j is assumed *a priori* to be drawn from either a Laplacian "spike" concentrated around zero (and hence is considered negligible), or a diffuse Laplacian "slab" (and hence may be large). Thus the hierarchical prior over $\boldsymbol{\beta}$ and the latent indicator variables $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ is given by

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \sim \prod_{j=1}^{p} \left[\gamma_j \varphi_1(\beta_j) + (1 - \gamma_j)\varphi_0(\beta_j)\right], \tag{6.1}$$

$$\pi(\boldsymbol{\gamma}|\theta) = \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j} \quad \text{and} \quad \theta \sim \text{Beta}(a, b), \tag{6.2}$$

where $\varphi_1(\beta) = \frac{\lambda_1}{2} e^{-|\beta|\lambda_1}$ is the slab distribution and $\varphi_0(\beta) = \frac{\lambda_0}{2} e^{-|\beta|\lambda_0}$ is the spike ($\lambda_1 \ll \lambda_0$) and we have used the common exchangeable beta-binomial prior for the latent indicators.

Ročková and George (2016) recast this hierarchical model into a penalized likelihood framework, allowing for the use of existing efficient algorithms for modal estimation while retaining the adaptivity inherent in the Bayesian formulation. The regression coefficients $\boldsymbol{\beta}$ are then estimated by

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max} \left\{ -\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + pen(\boldsymbol{\beta}) \right\} \tag{6.3}$$

where

$$pen(\boldsymbol{\beta}) = \log\left[\frac{\pi(\boldsymbol{\beta})}{\pi(\mathbf{0}_p)}\right], \quad \pi(\boldsymbol{\beta}) = \int_0^1 \prod_{j=1}^{p} [\theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)] d\pi(\theta). \tag{6.4}$$

Ročková and George (2016) note a number of advantages in using a mixture of Laplace densities in (6.1), instead of the usual mixture of Gaussians as has been standard in the Bayesian variable selection literature. First, the Laplacian spike serves to automatically threshold modal estimates of β_j to zero when β_j is small, much like the Lasso. However, unlike the Lasso, the slab distribution in the prior serves to stabilize the larger coefficients so they are not downward biased. Additionally, the heavier Laplacian tails of the slab distribution yields optimal posterior concentration rates (Ročková, 2017).

One route for adding the unknown variance case to the SSL procedure is to follow the prior framework of Park and Casella (2008) in their Bayesian Lasso. There, Park and Casella (2008) used the following prior for the regression coefficients:

$$\pi(\boldsymbol{\beta}|\sigma^2) \propto \prod_{j=1}^{p} \frac{\lambda}{2\sigma} e^{-\lambda|\beta_j|/\sigma}. \tag{6.5}$$

In the next section, we illustrate why an analogous scale-invariant prior formulation for the Spike-and-Slab Lasso is a poor choice. Later, we introduce the SSL with unknown variance which utilizes an independent prior framework.

6.1 The failure of the scale-invariant prior

The scale-invariant prior formulation for the Spike-and-Slab Lasso is given by:

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2) \sim \prod_{j=1}^{p} \left(\gamma_j \frac{\lambda_1}{2\sigma} e^{-|\beta_j|\lambda_1/\sigma} + (1-\gamma_j)\frac{\lambda_0}{2\sigma} e^{-|\beta_j|\lambda_0/\sigma}\right) \tag{6.6}$$

$$\boldsymbol{\gamma}|\theta \sim \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j}, \quad \theta \sim \text{Beta}(a, b) \tag{6.7}$$

$$p(\sigma^2) \propto \sigma^{-2}. \tag{6.8}$$

We find the posterior modes of our parameters using the EM algorithm, the details of which can be found in the supplementary material. At the $(k+1)$th iteration, the EM updates are:

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} \left\{ \frac{1}{2\sigma^{(k)}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} |\beta_j| \lambda^*(\beta_j^{(k)}/\sigma^{(k)}; \theta^{(k)}) \right\} \tag{6.9}$$

$$\theta^{(k+1)} = \frac{\sum_{j=1}^{p} p^*(\beta_j^{(k)}/\sigma^{(k)}; \theta^{(k)}) + a - 1}{a + b + p - 2} \tag{6.10}$$

$$\sigma^{(k+1)} = \frac{Q + \sqrt{Q^2 + 4(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(k)}\|^2)(n+p+2)}}{2(n+p+2)} \tag{6.11}$$

where

$$Q = \sum_{i=1}^{p} |\beta_j^{(k)}| \lambda^*(\beta_j^{(k)}/\sigma^{(k)}; \theta^{(k)}), \tag{6.12}$$

$$p^*(\beta;\theta) = \left[1 + \frac{\lambda_0}{\lambda_1}\left(\frac{1-\theta}{\theta}\right)\exp\{-|\beta_j|(\lambda_0 - \lambda_1)\}\right]^{-1}, \tag{6.13}$$

$$\lambda^*(\beta;\theta) = \lambda_1 p^*(\beta;\theta) + \lambda_0(1 - p^*(\beta;\theta)). \tag{6.14}$$

Let us take a closer look at the estimator of σ. Following the line of reasoning in Sun and Zhang (2010), an expert with oracle knowledge of the true regression coefficients $\boldsymbol{\beta}^*$ would estimate the noise level by the oracle estimator:

$$\sigma^{*2} = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|}{n}. \tag{6.15}$$

However, the maximum *a posteriori* estimate of σ at the true values of $\boldsymbol{\beta}^*, \boldsymbol{\gamma}^*$ is given by

$$\widehat{\sigma}_{MAP} = \tau + \sqrt{\tau^2 + \frac{(\sigma^*)^2}{1 + p/n + 2/n}} \tag{6.16}$$

where $\tau = \lambda_1\|\boldsymbol{\beta}^*\|_1/[2(n+p+2)]$. Here we see that if $n \to \infty$ with p fixed, we have $\widehat{\sigma}_{MAP} \to \sigma^*$. If, however, we have $p/n \to \infty$ where the underlying sparsity $\|\boldsymbol{\beta}^*\|_0 = q$ is fixed, we have $\widehat{\sigma}_{MAP} \to 0$. Thus, similarly to the ridge regression example in section 3.1, we will have a severe underestimate of the error variance. As before, the remedy is to use the independent prior on σ^2 and $\boldsymbol{\beta}$.

6.2 Spike-and-Slab Lasso with Unknown Variance

We now introduce the Spike-and-Slab Lasso with unknown variance, which considers the regression coefficients and error variance to be *a priori* independent. The hierarchical model is

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \sim \prod_{j=1}^{p}\left(\gamma_j \frac{\lambda_1}{2} e^{-|\beta_j|\lambda_1} + (1-\gamma_j)\frac{\lambda_0}{2}e^{-|\beta_j|\lambda_0}\right) \tag{6.17}$$

$$\boldsymbol{\gamma}|\theta \sim \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j}, \quad \theta \sim \text{Beta}(a,b) \tag{6.18}$$

$$p(\sigma^2) \sim \sigma^{-2}. \tag{6.19}$$

The log posterior, up to an additive constant, is given by

$$L(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 - (n+2)\log\sigma + \sum_{j=1}^{p} pen(\beta_j|\theta_j) \tag{6.20}$$

where, for $j = 1, \ldots, p$,

$$pen(\beta_j|\theta_j) = -\lambda_1|\beta_j| + \log[p^*(0;\theta_j)/p^*(\beta_j;\theta_j)], \tag{6.21}$$

$$\theta_j = \mathbb{E}[\theta|\boldsymbol{\beta}_{\setminus j}] \tag{6.22}$$

and $p^*(\beta;\theta)$ is as defined in (6.13). For large p, Ročková and George (2016) note that the conditional expectation $\mathbb{E}[\theta|\boldsymbol{\beta}_{\setminus j}]$ is very similar to $\mathbb{E}[\theta|\boldsymbol{\beta}]$ and so for practical purposes we treat them as equal and denote $\theta_\beta = \mathbb{E}[\theta|\boldsymbol{\beta}]$.

To find the modes of (6.20), we pursue a similar coordinate ascent strategy to Ročková and George (2016), cycling through updates for each β_j and σ^2 while updating the conditional expectation θ_β. This conditional expectation does not have an analytical expression; however, Ročková and George (2016) note that it can be approximated by

$$\theta_\beta \approx \frac{a + \|\boldsymbol{\beta}\|_0}{a + b + \|\boldsymbol{\beta}\|_0}. \tag{6.23}$$

We now outline the estimation strategy for $\boldsymbol{\beta}$. As noted in Lemma 3.1 of Ročková and George (2016), there is a simple expression for the derivative of the SSL penalty:

$$\frac{\partial pen(\beta_j|\theta_\beta)}{\partial|\beta_j|} \equiv -\lambda^*(\beta_j;\theta_\beta) \tag{6.24}$$

where

$$\lambda^*(\beta_j;\theta_\beta) = \lambda_1 p^*(\beta_j;\theta_\beta) + \lambda_0[1 - p^*(\beta_j;\theta_\beta)]. \tag{6.25}$$

Using the above expression, the Karush-Kuhn-Tucker (KKT) conditions yield the following necessary condition for the global mode $\widehat{\boldsymbol{\beta}}$:

$$\widehat{\beta}_j = \frac{1}{n}\left[|z_j| - \sigma^2\lambda^*(\widehat{\beta}_j;\theta_\beta)\right]_+ sign(z_j), \quad j = 1, \ldots, p \tag{6.26}$$

where $z_j = \mathbf{X}_j^T(\mathbf{Y} - \sum_{k\neq j}^p \widehat{\beta}_k \cdot \mathbf{X}_k)$ and we assume that the design matrix $\mathbf{X}$ has been centered and standardized to have norm $\sqrt{n}$. The condition (6.26) is very close to the familiar soft-thresholding operator for the Lasso, except that the penalty term $\lambda^*(\beta_j;\theta)$ differs for each coordinate. Similarly to other non-convex methods, this enables *selective shrinkage* of the coefficients, mitigating the bias issues associated with the Lasso. Also similarly to non-convex methods however, (6.26) is not a sufficient condition for the global mode. This is particularly problematic when the posterior landscape is highly multimodal, a consequence of $p \gg n$ and large λ_0. To eliminate some of these suboptimal local modes from consideration, Ročková and George (2016) develop a more refined characterization of the global mode. This characterization follows the arguments of Zhang and Zhang (2012) and can easily be modified for the unknown variance case of the SSL, detailed in Proposition 1.

Proposition 1. *The global mode* $\widehat{\boldsymbol{\beta}}$ *satisfies*

$$\widehat{\beta}_j = \begin{cases} 0 & \text{when } |z_j| \leq \Delta \\ \frac{1}{n}[|z_j| - \sigma^2\lambda^*(\widehat{\beta}_j;\theta_\beta)]_+ sign(z_j) & \text{when } |z_j| > \Delta \end{cases} \tag{6.27}$$

where

$$\Delta \equiv \inf_{t>0}[nt/2 - \sigma^2 pen(t|\theta_\beta)/t]. \tag{6.28}$$

Unfortunately, computing (6.28) can be difficult. Instead, we seek an approximation to the threshold Δ. A useful upper bound is $\Delta \leq \sigma^2\lambda^*(0;\theta_\beta)$ (Zhang and Zhang, 2012). However, when λ_0 gets large, this bound is too loose and can be improved. The improved bounds are given in Proposition 2, the analogue of Proposition 3.2 of Ročková and George (2016) for the unknown variance case. Before stating the result, the following function is useful to simplify exposition:

$$g(x;\theta) = [\lambda^*(x;\theta) - \lambda_1]^2 + \frac{2n}{\sigma^2}\log[p^*(x;\theta)]. \tag{6.29}$$

Proposition 2. *When* $\sigma(\lambda_0 - \lambda_1) > 2\sqrt{n}$ *and* $g(0;\theta_\beta) > 0$ *the threshold* Δ *is bounded by*

$$\Delta^L < \Delta < \Delta^U,$$

where

$$\Delta^L = \sqrt{2n\sigma^2\log[1/p^*(0;\theta_\beta)] - \sigma^4 d_j} + \sigma^2\lambda_1, \tag{6.30}$$

$$\Delta^U = \sqrt{2n\sigma^2\log[1/p^*(0;\theta_\beta)]} + \sigma^2\lambda_1 \tag{6.31}$$

and

$$0 < d_j < \frac{2n}{\sigma^2} - \left(\frac{n}{\sigma^2(\lambda_0 - \lambda_1)} - \frac{\sqrt{2n}}{\sigma}\right)^2.$$

Thus, when λ_0 is large and consequently $d_j \to 0$, the lower bound on the threshold approaches the upper bound, yielding the approximation $\Delta \approx \Delta^U$. We additionally note the central role that the error variance plays in the thresholds in Proposition 2. As σ^2 increases, the thresholds also increase, making it more difficult for regression coefficients to be selected. This is exactly what we want when the signal to noise ratio is small.

Bringing this all together, we incorporate this refined characterization of the global mode into the update for the coefficients via the generalized thresholding operator of Mazumder et al. (2011):

$$\widetilde{S}(z,\lambda,\Delta) = \frac{1}{n}(|z| - \lambda)_+ sign(z)\mathbb{I}(|z| > \Delta). \tag{6.32}$$

The coordinate-wise update is then

$$\widehat{\beta}_j \leftarrow \widetilde{S}(z_j, \widehat{\sigma}^2\lambda^*(\widehat{\beta}_j;\widehat{\theta}_\beta), \Delta) \tag{6.33}$$

where

$$\Delta = \begin{cases} \sqrt{2n\widehat{\sigma}^2\log[1/p^*(0;\widehat{\theta}_\beta)]} + \widehat{\sigma}^2\lambda_1 & \text{if } g(0;\widehat{\theta}_\beta) > 0, \\ \widehat{\sigma}^2\lambda^*(0;\widehat{\theta}_\beta) & \text{otherwise.} \end{cases} \tag{6.34}$$

Given the most recent update of the coefficient vector $\widehat{\boldsymbol{\beta}}$, the update for the error variance σ^2 is a simple Newton step:

$$\widehat{\sigma}^2 \leftarrow \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n+2}. \tag{6.35}$$

Finally, the conditional expectation θ_β is updated according to (6.23).

In principle both σ^2 and the conditional expectation θ_β should be updated after each β_j, $j = 1, \ldots, p$, instead of after updating all p coordinates. In practice, however, there will be little change after one coordinate update and so both σ^2 and θ_β can be updated after M coordinates are updated.

6.3 Dynamic Posterior Exploration

In the SSL with known variance, Ročková and George (2016) propose a "dynamic posterior exploration" strategy whereby the slab parameter λ_1 is held fixed and the spike parameter λ_0 is gradually increased to approximate the ideal point mass prior. Holding the slab parameter fixed serves to stabilize the non-zero coefficients, unlike the Lasso which applies an equal level of shrinkage to all regression coefficients. Meanwhile, gradually increasing λ_0 over a "ladder" of values serves to progressively threshold negligible coefficients. More practically, the dynamic strategy aids in mode detection: when $(\lambda_1 - \lambda_0)^2 \leq 4/\sigma^2$, the objective is convex (Ročková and George, 2016). In fact, when $\lambda_0 = \lambda_1$, it is equivalent to the Lasso. As λ_0 is increased, the posterior landscape becomes multimodal, but using the solution from the previous value of λ_0 as a "warm start" allows the procedure to more easily find modes. Thus, progressively increasing λ_0 acts as an annealing strategy.

When σ^2 is treated as unknown, the successive warm start strategy of Ročková and George (2016) will require additional intervention. For small $\lambda_0 \approx \lambda_1$, there may be many negligible but non-zero β_j included in the model. This severe overfitting results in all the variation in $\mathbf{Y}$ being explained by the model, forcing the estimate of the error variance, $\widehat{\sigma}^2$ to zero. If this suboptimal solution is propagated for larger values of λ_0, the optimization routine will remain "stuck" in that part of the posterior landscape. As a strategy to escape this absorbing state, we proceed by re-initializing $\widehat{\boldsymbol{\beta}} = \mathbf{0}_p$ whenever the variance estimate is too small. In practice, we have found the criterion for re-initialization $\widehat{\sigma}^2 < 1/n$ to be effective. Deshpande et al. (2017) note a similar phenomenon in their implementation of the SSL for multivariate regression, and also propose a similar restarting mechanism. An interesting benefit of this re-initializing strategy is that it is also a successful, data-driven way of determining a good starting value of λ_0. To be more precise, the λ_0 value at which we last re-initialize $\widehat{\boldsymbol{\beta}} = \mathbf{0}_p$ is the new "starting value" from which point the solution is propagated forward for larger values of λ_0. The entire implementation strategy is summarized in Algorithm 1.

6.4 Scaled Spike-and-Slab Lasso

An alternative approach for extending the SSL for unknown variance is to follow the scaled Lasso framework of Sun and Zhang (2012). In their original scaled Lasso paper,

Algorithm 1 Spike-and-Slab Lasso with unknown variance

Input: grid of increasing λ_0 values $I = \{\lambda_0^1, \ldots, \lambda_0^L\}$, update frequency M
Initialize: $\widehat{\boldsymbol{\beta}}_0 = \mathbf{0}_p, \widehat{\sigma}_0^2 = 1$, $\widehat{\theta}_\beta = 0.5$
for $l = 1$ **to** L:

- **Initialize:** $\widehat{\boldsymbol{\beta}}_l = \widehat{\boldsymbol{\beta}}_{l-1}, \widehat{\sigma}_l^2 = \widehat{\sigma}_{l-1}^2$
- **Repeat:**
  - **for** $s = 1$ **to** $\lfloor p/M \rfloor$:
    - Update

$$\Delta \leftarrow \begin{cases} \sqrt{2n\widehat{\sigma}_l^2 \log[1/p^*(0;\widehat{\theta}_\beta)]} + \widehat{\sigma}_l^2\lambda_1 & \text{if } g(0;\widehat{\theta}_\beta) > 0 \\ \widehat{\sigma}_l^2\lambda^*(0;\widehat{\theta}_\beta) & \text{otherwise} \end{cases}$$

    - **for** $j = 1$ **to** M: update

$$\beta_{l(s-1)M+j} \leftarrow \widetilde{S}(z_j, \widehat{\sigma}^2\lambda^*(\beta_{l(s-1)M+j}; \widehat{\theta}_\beta), \Delta)$$

    - Update $\widehat{\theta}_\beta \leftarrow (a + \|\widehat{\boldsymbol{\beta}}_l\|_0)/(a + b + \|\widehat{\boldsymbol{\beta}}_l\|_0)$
    - Update $\widehat{\sigma}_l^2 \leftarrow \|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_l\|^2/(n+2)$

  Until change in $\widehat{\boldsymbol{\beta}}_l$ is less than $\varepsilon = 10^{-5}$.
- **If** $\widehat{\sigma}_l^2 < 1/n$: set
  - $\widehat{\boldsymbol{\beta}}_l \leftarrow \mathbf{0}_p$
  - $\widehat{\sigma}_l^2 \leftarrow 1$.

Sun and Zhang (2012) note that their loss function can be used with many penalized likelihood procedures, including the MCP and the SCAD penalties. Here, we develop the *scaled Spike-and-Slab Lasso*. The loss function for the scaled SSL is the same as that of the scaled Lasso but with a different penalty:

$$L(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2\sigma}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 - \frac{n\sigma}{2} + \sum_{j=1}^{p} pen(\beta_j|\theta_\beta) \tag{6.36}$$

where $pen(\beta_j|\theta_\beta)$ is as defined in (6.21) and again we use the approximation (6.23) for the conditional expectation θ_β. In using this loss function, we are of course departing from the Bayesian paradigm and simply considering this procedure as a penalized likelihood method with a spike-and-slab inspired penalty.

The algorithm to find the modes of (6.36) is very similar to Algorithm 1, the only difference being we replace all σ^2 terms in the updates (6.33) and (6.34) with σ. This is because the refined thresholds for the coefficients are derived using the KKT conditions where the only difference between the two procedures is σ vs. σ^2.

The Newton step for σ^2 is only very slightly different from the SSL with unknown variance:

$$\widehat{\sigma}^2 \leftarrow \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n}. \tag{6.37}$$

How do we expect the scaled Spike-and-Slab Lasso to compare to the Spike-and-Slab Lasso with unknown variance? The threshold levels Δ for the scaled SSL will be smaller after replacing σ^2 with σ. This may potentially result in more false positives being included in the scaled SSL model. In terms of variance estimation, the updates for σ^2 are effectively the same; the only differences we should expect are those arising from a more saturated estimate for $\boldsymbol{\beta}$. These hypotheses are examined in the simulation study in the next session.

6.5 Simulation Study

We now compare the Spike-and-Slab Lasso with unknown variance with several penalized likelihood methods, including the original Spike-and-Slab Lasso with fixed variance of Ročková and George (2016) as well as the scaled Spike-and-Slab Lasso outlined in the previous section. We investigate both the efficacy of the SSL with unknown variance and the benefits of simultaneously estimating the regression coefficients $\boldsymbol{\beta}$ and error variance σ^2 in variable selection.

We consider the same simulation setting of Ročková and George (2016) with $n = 100$ and $p = 1000$ but use an error variance of $\sigma^2 = 3$ instead of $\sigma^2 = 1$. The data matrix $\mathbf{X}$ is generated from a multivariate Gaussian distribution with mean $\mathbf{0}_p$ and a block-diagonal covariance matrix $\boldsymbol{\Sigma} = \text{bdiag}(\widetilde{\Sigma}, \ldots, \widetilde{\Sigma})$ where $\widetilde{\Sigma} = \{\widetilde{\sigma}\}_{i,j=1}^{50}$ where $\widetilde{\sigma}_{ij} = 0.9$ if $i \neq j$ and $\widetilde{\sigma}_{ii} = 1$. The true vector $\boldsymbol{\beta}_0$ is constructed by assigning regression coefficients $\{-2.5, -2, -1.5, 1.5, 2, 2.5\}$ to $q = 6$ entries located at $\{1, 51, 101, 151, 201, 251\}$ and setting to zero the remaining coefficients. Hence, there are 50 blocks of 20 highly

correlated predictors where the first 6 blocks each contain only one active predictor. The response was generated as in (1.1) with error variance $\sigma^2 = 3$.

We compared the Spike-and-Slab Lasso with unknown variance to the fixed variance Spike-and-Slab Lasso with two settings: (i) $\sigma^2 = 1$, and (ii) $\sigma^2 = 3$, the true variance. The prior settings for θ were $a = 1, b = p$. The slab parameter was set to $\lambda_1 = 1$. For the spike parameter, we used a ladder $\lambda_0 \in I = \{1, 2, \ldots, 100\}$.

Additional methods compared were the scaled SSL from Section 3.4, the Lasso (Friedman et al., 2010), the scaled Lasso (Sun and Zhang, 2012), the Adaptive Lasso (Zou, 2006), SCAD (Fan and Li, 2001), and two different implementations of the MCP (Zhang, 2010): (i) the *Sparsenet* algorithm of Mazumder et al. (2011) which performs cross-validation across the two penalty parameters (λ, γ), and (ii) the hard-thresholding implementation with $\gamma = 1.0001$ and cross-validation across the parameter λ.

The analysis was repeated 100 times with new covariates and responses generated each time. For each, the metrics recorded were the Hamming distance (HAM) between the support of the estimated $\boldsymbol{\beta}$ and the true $\boldsymbol{\beta}^*$, the Mean Square Error (MSE), the number of false negatives (FN), the number of false positives (FP), the number of true positives (TP), the dimension of the estimated $\boldsymbol{\beta}$, whether the method found the true model (TRUE) and the time (TIME). The average of these metrics for each method over the 100 repetiions are displayed in Table 1.

	HAM	MSE	FN	FP	TP	DIM	TRUE (%)	TIME
SSL (fixed $\sigma^2 = 3$)	0.72	1.76	0.36	0.36	5.64	6.00	70	0.09
SSL (unknown σ)	1.78	4.65	0.84	0.94	5.16	6.10	36	0.92
SSL (fixed $\sigma^2 = 1$)	4.32	10.27	1.05	3.27	4.95	8.22	4	0.38
Scaled SSL	6.52	15.44	1.38	5.14	4.62	9.76	4	1.14
MCP ($\gamma = 1.0001$)	6.82	23.47	3.02	3.80	2.98	6.78	4	0.33
MCP (Sparsenet)	7.46	11.31	1.52	5.94	4.48	10.42	14	1.99
Adaptive Lasso	8.10	8.73	1.17	6.93	4.83	11.76	0	5.29
SCAD	12.07	26.99	3.76	8.31	2.24	10.55	0	0.40
Scaled Lasso	16.47	14.91	1.66	14.81	4.34	19.15	0	0.60
Lasso	31.25	10.55	0.67	30.58	5.33	35.91	0	0.42

Table 1: Average metrics over 100 repetitions for each of the procedures, ordered by increasing Hamming distance.

We can see that the Spike-and-Slab Lasso with the variance fixed and equal to the truth ($\sigma^2 = 3$) performs the best in terms of the Hamming distance, mean squared error, false negatives and true positives. Encouragingly, the Spike-and-Slab Lasso with unknown variance performs almost as well as the "oracle" version where the true variance is known. The SSL with unknown variance in turn performs better than a naive implementation of the SSL with fixed variance ($\sigma^2 = 1$). Following from the discussion in Section 3.4, we can see that the scaled SSL indeed finds more false positives than the SSL with unknown variance. This is a result of the smaller thresholds in estimating the regression coefficients. We can see that the scaled Lasso significantly reduces the number of false positives found as compared to the Lasso; however, the issues with the Lasso penalty remain.

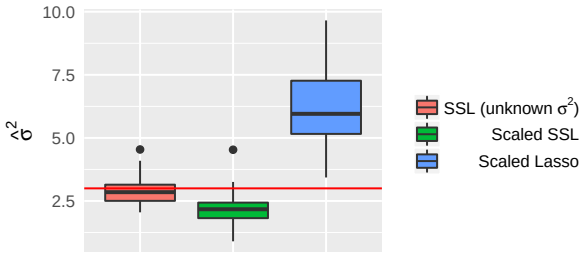

Figure 2: Estimated $\widehat{\sigma}^2$ for each procedure over 100 repetitions. The true variance $\sigma^2 = 3$ is the red horizontal line.

Figure 2 shows the variance estimates over the 100 repetitions for the SSL with unknown variance, the scaled SSL and the scaled Lasso. These variance were adjusted after selection to reflect the number of non-zero coefficients estimated:

$$\widehat{\sigma}^2_{adj} = \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n - \widehat{q}} \tag{6.38}$$

where $\widehat{q} = \|\widehat{\boldsymbol{\beta}}\|_0$. Note that we still use the "biased" update for σ^2 in the implementation as detailed in Algorithm 1. We can see that the SSL with unknown variance performs very well in estimating σ^2. As expected from the larger number of false positives, the scaled SSL underestimates the variance. Meanwhile the scaled Lasso highly inflates the variance.

7 Conclusion

In this paper, we argued that using continuous scale-invariant coefficient priors for Bayesian variable selection lead to poor estimates of the error variance. This is a result of implicitly adding p extra "pseudo-observations" of the error variance in the prior specification. We examined the historical justification for such priors and note that the canonical reference for these, Jeffreys (1961), actually advises against their use in multivariate problems. We then illustrated the deficiencies of these scale-invariant priors in the case study of Bayesian ridge regression before highlighting how they can interfere with the mechanisms of the global-local shrinkage framework.

Note that this paper differs from the discussion of variance priors in Gelman (2004) in that here we are focused on the estimation of the error variance. In contrast, Gelman (2004) is concerned with the choice of priors for the parameter level variance; in our notation, this corresponds to prior choices for the hyperparameter τ^2 in (1.5) and (1.7).

We then proceeded to extend the Spike-and-Slab Lasso of Ročková and George (2016) to the unknown variance case, using an independent prior for the variance. We showed that this procedure for the Spike-and-Slab Lasso with unknown variance performs almost as well empirically as the SSL where the true variance is known. We

additionally compared the Spike-and-Slab Lasso with unknown variance to a popular frequentist method to estimate the variance in high dimensional regression: the scaled Lasso. In simulation studies, the SSL with unknown variance performed much better than the scaled Lasso and additionally outperformed the "scaled Spike-and-Slab Lasso", a variant of the latter procedure but with the Spike-and-Slab Lasso penalty. The unknown variance implementation of the SSL is provided in the publicly available R package `SSLASSO` (Ročková and Moran, 2017).

References

Bayarri, M. J., Berger, J., Forte, A., and Garcia-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40: 1550–1577. 2

Belloni, A., Chernozhukov, V., Wang, L., et al. (2014). "Pivotal estimation via square-root lasso in nonparametric regression." *The Annals of Statistics*, 42(2): 757–788. 2, 9

Berger, J. O., Pericchi, L. R., and Varshavsky, J. A. (1998). "Bayes factors and marginal distributions in invariant situations." *Sankhya Ser. A*, 60: 307–321. 2

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). "Default Bayesian analysis with global-local shrinkage priors." *Biometrika*, 103(4): 955–969. 7

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. 7

Deshpande, S. K., Rockova, V., and George, E. I. (2017). "Simultaneous Variable and Covariance Selection with the Multivariate Spike-and-Slab Lasso." *ArXiv e-prints*. 14

Fan, J. and Li, R. (2001). "Variable selection via nonconcave penalized likelihood and its oracle properties." *Journal of the American Statistical Association*, 96(456): 1348–1360. 17

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software*, 33(1): 1. 17

Gelman, A. (2004). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*. 18

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. CRC press Boca Raton, FL. 4

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

Jeffreys, H. (1961). *The Theory of Probability*. Oxford University Press, 3 edition. 1, 3, 4, 18

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). "Sparsenet: Coordinate descent with nonconvex penalties." *Journal of the American Statistical Association*, 106(495): 1125–1138. 13, 17

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. 7, 10

Piironen, J. and Vehtari, A. (2017). "On the Hyperprior Choice for the Global Shrinkage Parameter in the Horseshoe Prior." In *Artificial Intelligence and Statistics*, 905–913. 8

Polson, N. G. and Scott, J. G. (2010). "Shrink globally, act locally: sparse Bayesian regularization and prediction." *Bayesian Statistics*, 9: 501–538. 7

Robert, C. P., Chopin, N., and Rousseau, J. (2009). "Harold Jeffreys's Theory of Probability Revisited." *Statistical Science*, 141–172. 4

Ročková, V. and George, E. I. (2014). "EMVS: The EM approach to Bayesian variable selection." *Journal of the American Statistical Association*, 109(506): 828–846. 6

Ročková, V. (2017). "Bayesian Estimation of Sparse Signals with a Continuous Spike-and-Slab Prior." *Annals of Statistics (Accepted)*. 10

Ročková, V. and George, E. (2016). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association (In Press)*. 1, 3, 9, 10, 12, 13, 14, 16, 18

Ročková, V. and Moran, G. (2017). *SSLASSO: The Spike-and-Slab LASSO*. URL `https://cran.r-project.org/package=SSLASSO` 3, 19

Städler, N., Bühlmann, P., and Van De Geer, S. (2010). "l1 penalization for mixture regression models." *Test*, 19(2): 209–256. 8

Sun, T. and Zhang, C.-H. (2010). "Comments on: l1-penalization for mixture regression models." *Test*, 19(2): 270–275. 11

— (2012). "Scaled sparse linear regression." *Biometrika*. 2, 9, 14, 16, 17

van der Pas, S., Salomond, J.-B., Schmidt-Hieber, J., et al. (2016). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10(1): 976–1000. 7

Zhang, C.-H. (2010). "Nearly unbiased variable selection under minimax concave penalty." *The Annals of Statistics*, 38(2): 894–942. 17

Zhang, C.-H. and Zhang, T. (2012). "A general theory of concave regularization for high-dimensional sparse estimation problems." *Statistical Science*, 576–593. 12, 13

Zou, H. (2006). "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association*, 101(476): 1418–1429. 17

Acknowledgments

This research was supported by the NSF Grant DMS-1406563.