Bayesian Analysis (2019) **14**, Number 4, pp. 1091–1119

Variance Prior Forms for High-Dimensional 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 show that the use of conjugate shrinkage priors for Bayesian variable selection can have detrimental consequences for such variance estimation. Such priors are often motivated by the invariance argument of Jeffreys (1961). Revisiting this work, however, we highlight a caveat that Jeffreys himself noticed; namely that biased estimators can result from inducing dependence between parameters *a priori*. In a similar way, we show that conjugate priors for linear regression, which induce prior dependence, can lead to such underestimation in the Bayesian high-dimensional regression setting. Following Jeffreys, we recommend as a remedy to treat regression coefficients and the error variance as independent *a priori*. Using such an independence prior framework, we extend the Spike-and-Slab Lasso of Ročková and George (2018) to the unknown variance case. This extended procedure outperforms both the fixed variance approach and alternative penalized likelihood methods on simulated data. On the protein activity dataset of Clyde and Parmigiani (1998), the Spike-and-Slab Lasso with unknown variance achieves lower cross-validation error than alternative penalized likelihood methods, demonstrating the gains in predictive accuracy afforded by simultaneous error variance estimation. The unknown variance implementation of the Spike-and-Slab Lasso is provided in the publicly available R package `SSLASSO` (Ročková and Moran, 2017).

Keywords: Bayesian variable selection, Bayesian shrinkage, Jeffreys' priors, penalized likelihood, Spike-and-Slab Lasso.

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

*Department of Statistics, University of Pennsylvania, gmoran@wharton.upenn.edu
†Booth School of Business, University of Chicago, veronika.rockova@chicagobooth.edu
‡Department of Statistics, University of Pennsylvania, edgeorge@wharton.upenn.edu

 https://doi.org/10.1214/19-BA1149

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 signal, thereby helping to mitigate overfitting of the data. Variance estimation is also essential in uncertainty quantification for inference and prediction.

In the frequentist literature, the question of estimating the error variance in our setting 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^2 are typically introduced in one of two ways: either via a conjugate prior framework or via an independence prior framework.

Conjugate priors have played a major role in regression analyses. The conjugate prior framework for (1.1) begins with specifying a prior on $\boldsymbol{\beta}$ that depends on σ^2 as follows:

$$\boldsymbol{\beta}|\sigma^2 \sim N(0, \sigma^2\mathbf{V}), \tag{1.2}$$

where $\mathbf{V}$ may be fixed or random. This prior (1.2) results in a Gaussian posterior for $\boldsymbol{\beta}$ and as such is conjugate. To complete the framework, σ^2 is assigned an inverse-gamma (or equivalently scaled-inverse-χ^2) prior. A common choice in this regard is the right-Haar prior for the location-scale group (Berger et al., 1998):

$$\pi(\sigma) \propto 1/\sigma. \tag{1.3}$$

Whilst the right-Haar prior is improper, it can be viewed as the limit of an inverse-gamma density. When combined with (1.2), the prior (1.3) results in an inverse-gamma posterior for σ^2 and as such it behaves as a conjugate prior. Prominent examples that utilize the above conjugate prior framework include:

- Bayesian ridge regression priors, with $\mathbf{V} = \tau^2\mathbf{I}$;
- Zellner's g-prior, with $\mathbf{V} = g(\mathbf{X}^T\mathbf{X})^{-1}$; and
- Gaussian global-local shrinkage priors, with $\mathbf{V} = \tau^2\Lambda$, for $\Lambda = \text{diag}\{\lambda_j\}_{j=1}^p$.

We note that the conjugate prior framework refers only to the prior characterization of $\boldsymbol{\beta}$ and σ^2, and allows for any prior specification on subsequent hyper-parameters such as g and τ^2 which do not appear in the likelihood.

A main reason for the popularity of the conjugate prior framework is that it often allows for marginalization over $\boldsymbol{\beta}$ and σ^2, resulting in closed form expressions for Bayes factors and updates of posterior model probabilities. This allowed for analyses of the model selection consistency (Bayarri et al., 2012) as well as more computationally

efficient MCMC algorithms (George and McCulloch, 1997). Despite these advantages, however, the conjugate prior framework is not innocuous for variance estimation, as we will show in this work.

Alternatively to the conjugate prior framework, one might treat $\boldsymbol{\beta}$ and σ^2 as independent *a priori*. The formulation corresponding to (1.2) for this independence prior framework is:

$$\begin{aligned} &\boldsymbol{\beta} \sim N(0, \mathbf{V}), \\ &\pi(\sigma) \propto 1/\sigma. \end{aligned} \tag{1.4}$$

Note that the prior characterization (1.4) does not yield a normal inverse-gamma posterior distribution on $(\boldsymbol{\beta}, \sigma^2)$ and as such is not conjugate.

In addition to the above prior frameworks, Bayesian methods for variable selection can be further categorized by the way they treat negligible predictors. Discrete component Bayesian methods for variable selection exclude negligible predictors from consideration, adaptively reducing the dimension of $\boldsymbol{\beta}$. Examples of such discrete component methods include spike-and-slab priors where the "spike" distribution is a point-mass at zero (Mitchell and Beauchamp, 1988). In contrast, continuous Bayesian methods for variable selection shrink, rather than exclude, negligible predictors and as such $\boldsymbol{\beta}$ remains p-dimensional (George and McCulloch, 1993; Polson and Scott, 2010; Ročková and George, 2014).

In this paper, we show that for continuous Bayesian variable selection methods, the conjugate prior framework can result in underestimation of the error variance when: (i) the regression coefficients $\boldsymbol{\beta}$ are sparse; and (ii) p is of the same order as, or larger than n. Intuitively, conjugate priors implicitly add p "pseudo-observations" to the posterior which can distort inference for the error variance when the true number of non-zero $\boldsymbol{\beta}$ is much smaller than p. This is not the case for discrete component methods which adaptively reduce the size of $\boldsymbol{\beta}$. To avoid the underestimation problem in the continuous case, 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 (2018) to the unknown variance case with an independent prior formulation, and highlight the performance gains over the known variance case via a simulation study. On the protein activity dataset of Clyde and Parmigiani (1998), we demonstrate the benefit of simultaneous variance estimation for both variable selection and prediction.

It is important to note the difference in the scope of this work with previous work on variance priors, including Gelman (2004); Bayarri et al. (2012); Liang et al. (2008). Here, we are focused on the estimation of the error variance, σ^2. In contrast, the aforementioned works are concerned with the choice of priors for hyper-parameters which do not appear in the likelihood, i.e. the g in the g-prior, and τ^2 and λ_j^2 for global-local priors. We recognize the importance of the choice of these priors for Bayesian variable selection; however, the focus of this paper is the prior choice for the error variance in conjunction with variable selection.

We also note that our discussion considers only Gaussian related prior forms for the regression coefficients. Despite this seemingly limited scope, we note that the majority

of priors used in Bayesian variable selection can be cast as a scale-mixture of Gaussians (Polson and Scott, 2010), and that popular frequentist procedures such as the Lasso and variants thereof also fall under this framework.

The paper is structured as follows. In Section 2, we discuss invariance arguments for conjugate priors and draw connections with Jeffreys priors. We then highlight situations where we ought to depart from Jeffreys priors; namely, in multivariate situations. In Section 3, we take Bayesian ridge regression as an example to highlight why conjugate priors can be a poor choice. In Section 4, we draw connections between Bayesian regression and concurrent developments with variance estimation in the penalized likelihood literature. In Section 5, we examine the mechanisms of the Gaussian global-local shrinkage framework and illustrate why they can be incompatible with the conjugate prior structure. In Section 6, we consider the Spike-and-Slab Lasso of Ročková and George (2018) and highlight 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. In Section 7, we apply the Spike-and-Slab Lasso with unknown variance to the protein activity dataset of Clyde and Parmigiani (1998), highlighting the improved predictive performance afforded by simultaneous variance estimation. We conclude with a discussion in Section 8.

2 Invariance Criteria

A common argument used in favor of the conjugate prior for Bayesian linear regression is that it is invariant to scale transformations of the response (Bayarri et al., 2012). That is, the regression coefficients depend *a priori* on σ^2 in a "scale-free way" through

$$\pi(\boldsymbol{\beta}|\sigma^2) = \frac{1}{\sigma^p} h(\boldsymbol{\beta}/\sigma), \tag{2.1}$$

for some proper density function $h(x)$. This means that the units of measurement used for the response do not affect the resultant estimates; for example, if $\mathbf{Y}$ is scaled by a factor of c, one would expect that the estimates for the regression coefficients, $\boldsymbol{\beta}$, and error variance, σ^2, should also be scaled by c.

A more general principle of invariance was proposed by Jeffreys (1961) in his seminal work, *The Theory of Probability*, a reference which is also sometimes given for the conjugate prior. In this section, we examine the original invariance argument of Jeffreys (1961) and highlight a caveat with this principle that the author himself noted; namely that it should be avoided in multivariate situations. We then draw connections between this suboptimal multivariate behavior and the conjugate prior framework, ultimately arguing similarly to Jeffreys that we should treat the mean and variance parameters as independently *a priori*.

2.1 Jeffreys Priors

For a parameter α, the Jeffreys prior is

$$\pi(\alpha) \propto |I(\alpha)|^{1/2}, \tag{2.2}$$

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 normal means model

$$Y_i \sim N(\mu_i, \sigma^2), \quad i = 1, \ldots, n$$

where the n-dimensional mean is denoted by $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. If we treat the parameters $\boldsymbol{\mu}$ and σ independently, the Jeffreys prior is $\pi(\boldsymbol{\mu}, \sigma) \propto 1/\sigma$. However, if the parameters are considered jointly, the Jeffreys prior is $\pi(\boldsymbol{\mu}, \sigma) \propto 1/\sigma^{n+1}$. In effect, by considering the parameters jointly as opposed to independently, we are implicitly including additional "pseudo-observations" of σ^2 and consequently distorting our estimates of the error variance.

This "pseudo-observation" interpretation can be seen explicitly in the conjugate form of the Jeffreys prior for a Gaussian likelihood. The joint Jeffreys prior $\pi(\boldsymbol{\mu}, \sigma) \propto 1/\sigma^{n+1}$ is an improper inverse-gamma prior with shape parameter, $n/2$, and scale parameter zero. As the prior is conjugate, the posterior distribution for the variance is also inverse-gamma:

$$\pi(\sigma^2 | \mathbf{Y}, \boldsymbol{\mu}) \sim IG\left(\frac{n}{2} + \frac{n}{2},\ 0 + \frac{\sum_{i=1}^{n}(Y_i - \mu_i)^2}{2}\right) \tag{2.3}$$

where the first term of both the shape and scale parameters in (2.3) are the prior hyperparameters. Thus, the dependent Jeffreys prior can be thought of as encoding knowledge of σ^2 from a previous experiment where there were n observations which yielded a sample variance of zero. This results in the prior concentrating around zero for large n and will severely distort posterior estimates of σ^2. As we shall see later, this *dependent Jeffreys prior* for the parameters is in some cases akin to the conjugate prior framework 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: $\pi(\mu, \sigma) = \pi(\mu)\pi(\sigma)$." That is, Jeffreys' remedy is to treat the parameters independently *a priori*, a recommendation which we also adopt. In addition, Jeffreys points out that a 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 Section 3 where we will discuss the consequences of using dependent Jefferys priors and other conjugate formulations in Bayesian linear regression.

We note a possible exception to this independence argument which is found later in *The Theory of Probability* where Jeffreys argues that for simple normal testing, the prior on μ under the alternative hypothesis should depend on σ^2. However, it is important to note that this recommendation is for the situation where μ is one-dimensional and so the underestimation phenomenon observed in (2.3) is not a problem. Given Jeffreys' earlier concerns regarding multivariate situations, it is unlikely he intended this dependence to generalize for higher dimensional μ.

3 Bayesian Regression

3.1 Prior Considerations

Consider again the classical linear regression model in (1.1). For a non-informative prior, it is common to use $\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$ (see, for example, Gelman et al., 2014). Similarly to our earlier discussion, this prior choice corresponds to multiplying the independent, Jeffreys priors for $\boldsymbol{\beta}$ and σ^2. In contrast, the joint Jeffreys prior would be $\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^{p+2}$. 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:

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

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, $n-p-2$, naturally appears in the denominator.[1] This degrees of freedom adjustment does not occur with the joint Jeffreys prior where the marginal posterior mean is given by:

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

For large p, this estimator will severely underestimate the error variance. Avoiding this, it is commonly accepted that the independent Jeffreys prior $\pi(\boldsymbol{\beta}, \sigma^2) \propto 1/\sigma^2$ should be the default non-informative prior in this setting.

There is no such clarity, however, in the use of conjugate priors for Bayesian linear regression. To add to this discourse, we show that these conjugate priors can suffer the same problem as the dependent Jeffreys priors and recommend, similarly to Jeffreys, 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

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

[1] Note that had we treated β_1 as an intercept and integrated it out with respect to a uniform prior, this term would be the usual $n-p-1$.

For simplicity of exposition, in this section we consider the parameter τ^2 to be fixed, which corresponds to Bayesian ridge regression. In later sections we will consider the global-local shrinkage framework where τ^2 is assigned a prior.

With an additional non-informative prior $\pi(\sigma^2) \propto 1/\sigma^2$, we then have the joint prior

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

Note again the σ^{p+2} in the denominator, similarly to the joint Jeffreys prior.

Instead of considering how $\boldsymbol{\beta}$ depends on σ^2 *a priori* as in (3.3), it is illuminating to consider the reverse: how this prior induces dependence of σ^2 on $\boldsymbol{\beta}$. From (3.4), the implicit conditional prior on σ^2 is given by

$$\sigma^2|\boldsymbol{\beta} \sim IG\left(\frac{p}{2}, \frac{\|\boldsymbol{\beta}\|^2}{2\tau^2}\right). \tag{3.5}$$

The mean of this inverse-gamma prior is approximately $\frac{1}{p}\|\boldsymbol{\beta}\|^2/\tau^2$. Heuristically, this term is of order $O(q/p)$, where q is the number of non-zero $\boldsymbol{\beta}$. When $\boldsymbol{\beta}$ is sparse and bounded with $q \ll p$, (3.5) will then transmit downward biasing information from $\boldsymbol{\beta}$ to σ^2. This intuition is formalized in Proposition 1, which shows that the implicit conditional prior on σ^2 concentrates around zero in regions where $\boldsymbol{\beta}$ is sparse.

Proposition 1. *Suppose* $\|\boldsymbol{\beta}\|_0 = q$ *and* $\max_j \beta_j^2 = K$ *for some constant* $K \in \mathbb{R}$. *Denote the true variance as* σ_0^2. *Then*

$$P\left(\sigma^2/\sigma_0^2 \geq \varepsilon \mid \boldsymbol{\beta}\right) \leq \frac{q}{p-2}\frac{K}{\tau^2}\frac{1}{\varepsilon\sigma_0^2}. \tag{3.6}$$

Proof. Proposition 1 follows from Markov's inequality and the bound $\|\boldsymbol{\beta}\|^2 \leq qK$. □

Proposition 1 implies that we can choose $0 < \varepsilon < 1$ such that as $q/p \to 0$, the prior places decreasing mass on values of σ^2 greater than $\varepsilon\sigma_0^2$. Thus, in regions of bounded sparse regression coefficients, the conjugate Gaussian prior can result in poor estimation of the true variance.

Further, from a more philosophical perspective, it is troubling that the error variance depends on the regression coefficients *a priori*, given that the noise is generally assumed to be independent of the signal and in particular the regression coefficients.

In the next section, we conduct a simulation study for the simple case of Bayesian ridge regression and show empirically how this implicit prior on σ^2 can distort estimates of the error variance.

3.2 The Failure of a Conjugate Prior

As an illustrative example, we take $n = 100$ and $p = 90$ and compare the least squares estimates of $\boldsymbol{\beta}$ and σ^2 to Bayesian ridge regression estimates with (i) the conjugate

formulation with (3.3) and (ii) the independent prior formulation with

$$\pi(\boldsymbol{\beta}) \sim N_p(0, \tau^2 \mathbf{I}). \tag{3.7}$$

For both Bayesian ridge regression procedures we use the non-informative error variance prior: $\pi(\sigma^2) \propto 1/\sigma^2$. 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 non-zero coefficients are set to $\{-2.5, -2, -1.5, 1.5, 2, 2.5\}$. The response $\mathbf{Y}$ is generated according to (1.1) with the true variance being $\sigma^2 = 3$. We take $\tau = 10$ as known and highlight that this weakly informative choice leads to poor variance estimates in the conjugate prior framework. Whilst an empirical or fully Bayes approach for estimating τ^2 may be preferable for high-dimensional regression, it is troubling that the conjugate prior yields poor results for a simple example where $n > p$ and in which least squares and the independent prior formulation perform well.

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

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

$$\mathbb{E}[\sigma^2|\mathbf{Y}] = \frac{\mathbf{Y}^T[\mathbf{I} - \mathbf{H}_\tau]\mathbf{Y}}{n-2} \tag{3.9}$$

where $\mathbf{H}_\tau = \mathbf{X}[\mathbf{X}^T\mathbf{X} + \tau^{-2}\mathbf{I}]^{-1}\mathbf{X}^T$. Similarly to (3.2), the above marginal posterior mean for σ^2 does not incorporate a degrees of freedom adjustment and so we expect this estimator to underestimate the true error variance.

It is illuminating to observe the underestimation problem when considering the conditional posterior mean of σ^2, instead of the marginal:

$$\mathbb{E}[\sigma^2|\mathbf{Y}, \boldsymbol{\beta}] = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \|\boldsymbol{\beta}\|^2/\tau^2}{n+p-2}. \tag{3.10}$$

The additional p in the denominator here leads to severe underestimation of σ^2 when $\boldsymbol{\beta}$ is sparse and bounded as in Proposition 1 and p is of the same order as, or larger than, n, as discussed in the previous section. We note in passing that a value of τ^2 close to $\|\boldsymbol{\beta}\|^2/p\sigma^2$, which may be obtainable with an empirical or fully Bayes approach, would avoid this variance underestimation problem, as can be seen from (3.10).

This is in contrast to the conditional posterior mean for σ^2 using the independent prior formulation (1.4), which we also consider. This estimator is given by:

$$\mathbb{E}[\sigma^2|\mathbf{Y}, \boldsymbol{\beta}] = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{n-2}. \tag{3.11}$$

Here we do not observe a degrees of freedom adjustment because (3.11) is the *conditional* posterior mean, not the marginal. Earlier in (3.1) we considered the marginal posterior mean for the independent Jeffreys' prior which led to the $n-p-2$ in the denominator. For the marginal posterior means of $\boldsymbol{\beta}$ and σ^2, the independent prior formulation does

not yield closed form expressions. To compute these, we use a Gibbs sampler, the details of which may be found in Section 1 of the supplement (Moran et al., 2019).

When τ^2 is large, the estimate of $\boldsymbol{\beta}$ for both the conjugate and independent formulations are almost exactly the least-squares estimate, $\widehat{\boldsymbol{\beta}} = [\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{Y}$. However, the estimates of the variance σ^2 differ substantially.

In Figure 1, we display a boxplot of the estimates of σ^2 for (i) Least Squares, (ii) Conjugate Bayesian ridge regression, (iii) Zellner's prior:

$$\boldsymbol{\beta}|\sigma^2 \sim N(0, \sigma^2\tau^2[\mathbf{X}^T\mathbf{X}]^{-1}), \tag{3.12}$$

and (iv) Independent Bayesian ridge regression over 100 replications. Here, the estimates from least squares and the independent ridge are reasonably distributed around the truth. In sharp contrast, the estimates from the conjugate ridge and Zellner's prior consistently underestimate the error variance with medians of $\widehat{\sigma}^2 = 0.27$ and 0.55, respectively. This poor performance is a result of the bias induced by adding p "pseudo-observations" of σ^2 as discussed in Section 3.1, which also occurs for the Zellner prior.

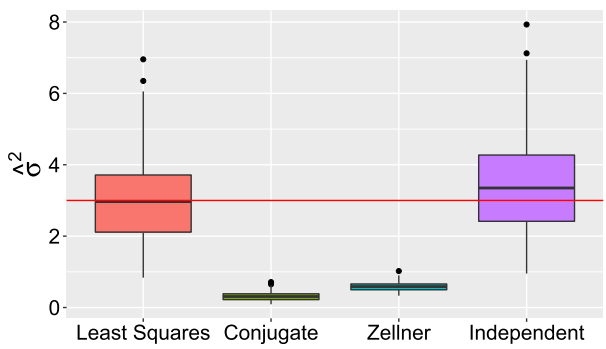

Figure 1: Estimated $\widehat{\sigma}^2$ for each procedure over 100 repetitions. The true $\sigma^2 = 3$ is the red horizontal line.

This phenomenon of underestimating σ^2 can also be 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 conjugate prior formulation in which $\boldsymbol{\beta}$ depends on σ^2 *a priori* similarly to (3.3). As in the above ridge regression example, with this prior EMVS yields good estimates for $\boldsymbol{\beta}$, but severely underestimates σ^2. This occurs 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 $\sigma^2 = 3$. Fortunately, EMVS can be easily modified to use the independent prior specification, as now has been done in the publicly available `EMVS` R package (Ročková and Moran, 2018). It is interesting to note that the SSVS procedure of George and McCulloch (1993) used the nonconjugate independence prior formulation in lieu of the conjugate prior formulation for the continuous spike-and-slab setup.

A natural question to ask is: how does the poor estimate of the variance in the conjugate case affect the estimated regression coefficients? Insight is obtained by comparing (3.8) to the conditional posterior mean of $\boldsymbol{\beta}$ in the independent case, given by:

$$\mathbb{E}[\boldsymbol{\beta}|\sigma^2, \mathbf{Y}] = \left[\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\tau^2}\mathbf{I}\right]^{-1} \mathbf{X}^T\mathbf{Y}. \tag{3.13}$$

In (3.8), the Gaussian prior structure allows for σ^2 to be factorized out so that the estimate of $\boldsymbol{\beta}$ does not depend on the variance. This lack of dependence on the variance is troubling, however, as we want to select fewer variables when the error variance is large making the signal-to-noise ratio low. This is in contrast to (3.13) where when σ^2 is large relative to τ^2, the signal-to-noise ratio is low and so the posterior estimate for $\boldsymbol{\beta}$ will be close to zero, correctly reflecting the relative lack of information. This does not occur for the posterior mean of $\boldsymbol{\beta}$ in the conjugate case.

3.3 What About a Prior Degrees of Freedom Adjustment?

At this point, one may wonder: if the problem seems to be the extra σ^p in the denominator, why not use the prior $\pi(\sigma^2) \propto \sigma^{p-4}$ instead of the right-Haar prior $\pi(\sigma^2) \propto \sigma^{-2}$ that is commonly used? This "p-sigma" prior then results in the joint prior:

$$\pi(\boldsymbol{\beta}|\sigma^2)\pi(\sigma^2) \propto \frac{1}{(\sigma^2)^2} \exp\left\{-\frac{1}{2\sigma^2\tau^2}\|\boldsymbol{\beta}\|^2\right\}, \tag{3.14}$$

which yields the implicit conditional prior on σ^2:

$$\sigma^2|\boldsymbol{\beta} \sim IG\left(1, \frac{\|\boldsymbol{\beta}\|^2}{2\tau^2}\right). \tag{3.15}$$

For the simulation setup in Section 3.2, this alternative conjugate prior would in fact remedy the variance estimates of the conjugate formulation (3.3). However, the p-sigma prior can actually lead to *overestimation* of the error variance, as opposed to the underestimation observed in Section 3.1. Heuristically, the mean of the prior (3.15) is now of order $O(q)$, where q is the number of non-zero $\boldsymbol{\beta}$. As many posterior concentration results require $q \to \infty$, albeit at a much slower rate than p (see, for example, van der Pas et al., 2016), this is particularly troublesome.

This overestimation can be further seen from the concentration of the prior captured in Proposition 2 below. As we will discuss in Section 4, a similar phenomenon also occurs for a penalized likelihood procedure that implicitly uses a p-sigma prior.

Proposition 2. *Suppose* $\|\boldsymbol{\beta}\|_0 = q$ *and* $\min_{j,\beta_j\neq 0} \beta_j^2 = K$ *for some constant* $K \in \mathbb{R}$. *Denote the true variance as* σ_0^2. *Then*

$$P(\sigma^2 \geq \delta\sigma_0^2 \mid \boldsymbol{\beta}) \geq 1 - \exp\left(-\frac{qK}{2\delta\sigma_0^2\tau^2}\right). \tag{3.16}$$

Proof. We have:

$$P(\sigma^2 \geq \delta\sigma_0^2 \mid \boldsymbol{\beta}) = \int_{\delta\sigma_0^2}^{\infty} \frac{\|\boldsymbol{\beta}\|^2}{2\tau^2}\frac{1}{u^2}\exp\left(-\frac{\|\boldsymbol{\beta}\|^2}{2\tau^2}\frac{1}{u}\right)du$$
$$\geq 1 - \exp\left(-\frac{qK}{2\delta\sigma_0^2\tau^2}\right). \qquad \square$$

Proposition 2 implies that we can choose arbitrary $\delta > 1$ such that as $q \to \infty$, the p-sigma prior places increasing mass on values of σ^2 greater than $\delta\sigma_0^2$. Another concern regarding the p-sigma prior is more philosophical. As p gets larger, the p-sigma prior puts increasing mass on larger and larger values of σ^2, which does not seem justifiable.

For these reasons, we prefer the independent prior forms for the regression coefficients and error variance. We are also of the opinion that the simplicity of the independent prior is in its favor.

4 Connections with 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 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:

$$L_{pen}(\boldsymbol{\beta}, \sigma^2) = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma^2} + \frac{\lambda}{\sigma}\|\boldsymbol{\beta}\|_1 + n\log\sigma. \tag{4.1}$$

Optimizing this objective function is in fact equivalent to MAP estimation for the following Bayesian model with the p-sigma prior discussed in Section 3.2:

$$\begin{aligned} \mathbf{Y} &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \\ \pi(\boldsymbol{\beta}|\sigma^2) &\propto \frac{1}{\sigma^p}\prod_{j=1}^{p} e^{-\lambda|\beta_j|/\sigma} \\ \pi(\sigma^2) &\propto \sigma^p. \end{aligned} \tag{4.2}$$

Interestingly, Sun and Zhang (2010) proved that the resulting estimator for the error variance *overestimates* the noise level unless $\lambda\|\boldsymbol{\beta}^*\|_1/\sigma^* = o(1)$, where $\boldsymbol{\beta}^*$ and σ^* are the true values of the regression coefficients and error variance, respectively. However, this condition requires q, the true number of non-zero $\boldsymbol{\beta}$, to be of the following order (details in Section 2 of the supplement, Moran et al., 2019)

$$q = o\left(\sqrt{n/\log p}\right). \tag{4.3}$$

That is, the true dimension q cannot at the same time increase at the required rate for posterior contraction *and* result in consistent estimates for the error variance. Note also the connection to Proposition 2: there, the prior mass on σ^2 will concentrate on values greater than the true variance unless $\|\boldsymbol{\beta}\|^2/\tau^2 = o(1)$.

To resolve this issue of overestimating the error variance, Sun and Zhang (2012) proposed as an alternative the "scaled Lasso", an algorithm which minimizes the following penalized joint loss function via coordinate descent:

$$L_\lambda(\boldsymbol{\beta}, \sigma) = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{2\sigma} + \frac{n\sigma}{2} + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{4.4}$$

This loss function is a penalized version of Huber's concomitant loss function, and so may be viewed as performing robust high-dimensional regression. It is also equivalent to the "square-root Lasso" of Belloni et al. (2014). Minimization of the loss function (4.4) can be viewed as MAP estimation for the Bayesian model (with a slight modification):

$$\begin{aligned} \mathbf{Y} &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma\mathbf{I}) \\ \pi(\boldsymbol{\beta}) &\propto \prod_{j=1}^{p} \frac{\lambda}{2} e^{-\lambda|\beta_j|} \\ \sigma &\sim \text{Gamma}(n+1, n/2). \end{aligned} \tag{4.5}$$

Note that to interpret the scaled Lasso as a Bayesian procedure, σ, rather than σ^2, plays the role of the variance in (4.5). Sun and Zhang (2012) essentially then re-interpret σ as the standard deviation again after optimization of (4.4). This re-interpretation can be thought of as an "unbiasing" step for the error variance. It is a little worrisome, however, that the implicit prior on the error variance is very informative: as $n \to \infty$, this Gamma prior concentrates around $\sigma = 2$.

Sun and Zhang (2012) proved that the scaled Lasso estimate $\widehat{\sigma}(\mathbf{X}, \mathbf{Y})$ is consistent for the "oracle" estimator

$$\sigma^* = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|}{\sqrt{n}}, \tag{4.6}$$

where $\boldsymbol{\beta}^*$ are the true regression coefficients, for the value of $\lambda_0 \propto \sqrt{(2/n)\log p}$. This estimator (4.6) is called the oracle because it treats the true regression coefficients as if they were known. The term $\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|^2$ is then simply the sum of normal random variables, of which we calculate the variance as $\sum_{i=1}^{n} \varepsilon_i^2/n$.

5 Global-Local Shrinkage

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

$$\beta_j \sim N(0, \tau^2\lambda_j^2), \quad \lambda_j^2 \sim \pi(\lambda_j^2), \quad j = 1, \ldots, p \tag{5.1}$$

$$\tau^2 \sim \pi(\tau^2)$$

where τ^2 is the "global" variance and λ_j^2 is the "local" variance. Note that taking τ^2 to be the same as the error variance σ^2 would result in a conjugate prior in this setting. This is exactly what was done in the original formulation of the Bayesian lasso by Park and Casella (2008), which can be recast in the Gaussian global-local shrinkage framework as follows (notation changed slightly for consistency):

$$\begin{aligned} &\mathbf{Y}|\boldsymbol{\beta}, \sigma^2 \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\ &\beta_j|\sigma^2, \lambda_j^2 \sim N(0, \sigma^2\lambda_j^2), \quad \pi(\lambda_j^2) = \frac{u^2}{2} e^{-u^2\lambda_j^2/2}, \quad j = 1, \ldots, p \\ &\pi(\sigma^2) \propto \sigma^{-2}. \end{aligned} \tag{5.2}$$

In the conjugate formulation (5.2), σ^2 plays the dual role of representing 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) recommend the following requirements for the global and local variances in (5.1): $\pi(\tau^2)$ should have substantial mass near zero to shrink all the regression coefficients so that the vast majority are negligible; and $\pi(\lambda_j^2)$ should have heavy tails so that it can be quite large, allowing for a few large coefficients to "escape" the heavy shrinkage of the global variance.

This heuristic is formalized in much of the shrinkage estimation theory. For the normal means problem where $\mathbf{X} = \mathbf{I}_n$ and $\boldsymbol{\beta} \in \mathbb{R}^n$, van der Pas et al. (2016) prove that the following conditions result in the posterior recovering nonzero means with the optimal rate: (i) $\pi(\lambda_j^2)$ should be a uniformly regular varying function which does not depend on n; and (ii) $\tau^2 = \frac{q}{n}\log(n/q)$, 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 λ_j^2 allow for a few large coefficients to be estimated. The condition (ii) encourages τ^2 to tend to zero which would be a concerning property if it were also the error variance. These results suggest we cannot identify the error variance with the global variance parameter on the regression coefficients as in (5.2): it cannot simultaneously both shrink all the regression coefficients and be a good estimate of the residual variance. Finally, we note that Hans (2009) also considered the independent case for the Bayesian lasso in which the error variance is not identified with the global variance.

An alternative conjugate formulation for Gaussian global-local shrinkage priors is to instead include three variance terms in the prior for β_j: the error variance, σ^2, the global variance, τ^2, and the local variance, λ_j^2. For example, Carvalho et al. (2010) give the conjugate form of the horseshoe prior:

$$\begin{aligned} \beta_j|\sigma^2, \tau^2, \lambda_j^2 &\sim N(0, \sigma^2\tau^2\lambda_j^2), \quad \lambda_j^2 \sim \pi(\lambda_j^2), \quad j = 1, \ldots, p \\ \tau^2 &\sim \pi(\tau^2), \\ \pi(\sigma^2) &\propto \sigma^{-2}. \end{aligned} \tag{5.3}$$

This prior formulation (5.3) remedies the aforementioned issue in the Bayesian lasso as it separates the roles of the error variance and global variance. However, this prior structure can still be problematic for error variance estimation.

Consider the conditional posterior mean of σ^2 for the model (5.3):

$$\mathbb{E}[\sigma^2|\mathbf{Y}, \boldsymbol{\beta}, \tau^2, \lambda_j^2] = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^p \beta_j^2/\lambda_j^2\tau^2}{n+p-2}. \tag{5.4}$$

Proposition 3 highlights that, given the true regression coefficients, the conditional posterior mean of σ^2 underestimates the oracle variance (4.6) when $\boldsymbol{\beta}$ is sparse.

Proposition 3. *Consider the global-local prior formulation given in* (5.3)*. Denote the true vector of regression coefficients by* $\boldsymbol{\beta}^*$ *where* $\|\boldsymbol{\beta}^*\|_0 = q$*. Suppose* $\max_j \beta_j^{*2} = M_1$ *for some constant* $M_1 \in \mathbb{R}$*. Denote the oracle estimator for* σ *given in* (4.6) *by* σ^* *and suppose* $\sigma^* = O(1)$*. Suppose also that for* $j \in \{1, \ldots, p\}$ *with* $\beta_j \neq 0$*, we have* $\tau^2\lambda_j^2 > M_2$ *for some* $M_2 \in \mathbb{R}$*. Then*

$$\mathbb{E}[\sigma^2|\mathbf{Y}, \boldsymbol{\beta}^*, \tau^2, \lambda_j^2] \leq \frac{n\sigma^{*2}}{n+p-2} + \frac{q}{n+p-2}\frac{M_1}{M_2}. \tag{5.5}$$

In particular, as $p/n \to \infty$ *and* $q/p \to 0$*, we have*

$$\mathbb{E}[\sigma^2|\mathbf{Y}, \boldsymbol{\beta}^*, \tau^2, \lambda_j^2] = o(1). \tag{5.6}$$

Given the mechanics of global-local shrinkage priors, the assumption in Proposition 3 that the term $\tau^2\lambda_j^2$ is bounded from below for non-zero β_j is not unreasonable. This is because for large β_j, the local variance λ_j^2 must be large enough to counter the extreme shrinkage effect of τ^2. Indeed, the prior for λ_j^2 must have "heavy enough" tails to enable this phenomenon.

We should note that Proposition 3 illustrates the poor performance of the posterior mean (5.4) given the true regression coefficients $\boldsymbol{\beta}^*$, whereas the horseshoe procedure does not actually threshold the negligible β_j to zero in the posterior mean of $\boldsymbol{\beta}$. For these small β_j, the term $\tau^2\lambda_j^2$ may be very small and potentially counteract the underestimation phenomenon. However, it is still troubling to use an estimator for the error variance that does not behave as the oracle estimator when the true regression coefficients are known. This is in contrast to the independent prior formulation where the conditional posterior mean of σ^2 is simply:

$$\mathbb{E}[\sigma^2|\mathbf{Y}, \boldsymbol{\beta}] = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2}{n-2}. \tag{5.7}$$

Note also that the problem of underestimation of σ^2 is exacerbated for modal estimation under the prior (5.3). This is because modal estimators often threshold small coefficients to zero and so the term $\sum_{j=1}^p \beta_j^2/\lambda_j^2\tau^2$ becomes negligible as in Proposition 3. As MAP estimation using global-local shrinkage priors is becoming more common (see, for example, Bhadra et al., 2017), we caution against the use of these conjugate prior forms.

A different argument for using conjugate priors with the horseshoe is given by Piironen and Vehtari (2017). They advocate for the model (5.3), arguing that it leads to a prior on the effective number of non-zero coefficients which does not depend on σ^2 and n. However, this quantity is derived from the posterior of $\boldsymbol{\beta}$ and so does not take into account the uncertainty inherent in the variable selection process. As a thought experiment: suppose that we know the error variance, σ^2, and number of observations, n. If the error variance is too large and the number of observations are too few, we would not expect to be able to say much about $\boldsymbol{\beta}$ at all, and this intuition should be reflected in the effective number of non-zero coefficients. This point is similar to our discussion at the end of Section 3.2 regarding estimation of $\boldsymbol{\beta}$.

As before, 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.

6 Spike-and-Slab Lasso with Unknown Variance

6.1 Spike-and-Slab Lasso

We now turn to the Spike-and-Slab Lasso (SSL, Ročková and George, 2018) and consider how to incorporate the unknown variance case. The SSL places 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 be 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

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \sim \prod_{j=1}^{p} \left[\gamma_j \psi_1(\beta_j) + (1-\gamma_j)\psi_0(\beta_j)\right], \tag{6.1}$$

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

where $\psi_1(\beta) = \frac{\lambda_1}{2} e^{-|\beta|\lambda_1}$ is the slab distribution and $\psi_0(\beta) = \frac{\lambda_0}{2} e^{-|\beta|\lambda_0}$ 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 (2018) 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

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\max} \left\{ -\frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + pen(\boldsymbol{\beta}) \right\} \tag{6.2}$$

where

$$pen(\boldsymbol{\beta}) = \log\left[\frac{\pi(\boldsymbol{\beta})}{\pi(\mathbf{0}_p)}\right], \quad \pi(\boldsymbol{\beta}) = \int_0^1 \prod_{j=1}^{p} [\theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)] d\pi(\theta). \tag{6.3}$$

Ročková and George (2018) 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, 2018).

Although the use of the spike-and-slab prior is typically associated with "two-group" Bayesian variable selection methods, the Spike-and-Slab Lasso can also be interpreted as a "one-group" global-local shrinkage method as the spike density is continuous. As such, the use of a conjugate prior for the error variance here will result in underestimation, similarly to the results for global-local shrinkage priors in Section 5. Further details of this underestimation problem for the SSL in particular can be found in Section 3 of the supplement (Moran et al., 2019). In the next section, we introduce the SSL with unknown variance which avoids this underestimation problem by instead utilizing an independent prior framework.

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

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}) \sim \prod_{j=1}^{p} [\gamma_j \psi_1(\beta_j) + (1-\gamma_j)\psi_0(\beta_j)] \tag{6.4}$$

$$\boldsymbol{\gamma}|\theta \sim \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j}, \quad \theta \sim \text{Beta}(a,b) \tag{6.5}$$

$$\pi(\sigma^2) \propto \sigma^{-2}. \tag{6.6}$$

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

$$L(\boldsymbol{\beta},\sigma^2) = -\frac{1}{2\sigma^2}\|\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}\|^2 - (n+2)\log\sigma + \sum_{j=1}^{p} pen(\beta_j|\theta_j) \tag{6.7}$$

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

$$pen(\beta_j|\theta_j) = -\lambda_1|\beta_j| + \log[p^*(0;\theta_j)/p^*(\beta_j;\theta_j)], \tag{6.8}$$

$$\text{with} \quad p^*(\beta;\theta) = \frac{\theta\psi_1(\beta)}{\theta\psi_1(\beta)+(1-\theta)\psi_0(\beta)} \quad \text{and} \quad \theta_j = \mathbb{E}[\theta|\boldsymbol{\beta}_{\setminus j}]. \tag{6.9}$$

For large p, Ročková and George (2018) 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 (6.7), we pursue a similar coordinate ascent strategy to Ročková and George (2018), 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 (2018) note that it can be approximated by

$$\theta_\beta \approx \frac{a + \|\boldsymbol{\beta}\|_0}{a + b + p}. \tag{6.10}$$

We now outline the estimation strategy for $\boldsymbol{\beta}$. As noted in Lemma 3.1 of Ročková and George (2018), there is a simple expression for the derivative of the SSL penalty:

$$\frac{\partial pen(\beta_j|\theta_\beta)}{\partial|\beta_j|} \equiv -\lambda^*(\beta_j;\theta_\beta) \tag{6.11}$$

where

$$\lambda^*(\beta_j;\theta_\beta) = \lambda_1 p^*(\beta_j;\theta_\beta) + \lambda_0[1 - p^*(\beta_j;\theta_\beta)]. \tag{6.12}$$

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

$$\widehat{\beta}_j = \frac{1}{n}\left[|z_j| - \sigma^2\lambda^*(\widehat{\beta}_j;\theta_\beta)\right]_+ sign(z_j), \quad j = 1,\ldots,p \tag{6.13}$$

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 (6.13) 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. As a non-convex method, however, (6.13) 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 many of these suboptimal local modes from consideration, Ročková and George (2018) 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 4.

Proposition 4. *The global mode $\widehat{\boldsymbol{\beta}}$ satisfies*

$$\widehat{\beta}_j = \begin{cases} 0 & \textit{when } |z_j| \leq \Delta \\ \frac{1}{n}[|z_j| - \sigma^2\lambda^*(\widehat{\beta}_j;\theta_\beta)]_+ sign(z_j) & \textit{when } |z_j| > \Delta \end{cases} \tag{6.14}$$

where

$$\Delta \equiv \inf_{t>0}[nt/2 - \sigma^2 pen(t|\theta_\beta)/t]. \tag{6.15}$$

Unfortunately, computing (6.15) 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 5, the analogue of Proposition 3.2 of Ročková and George (2018) for the unknown variance case. Before stating the result, the following function is useful to simplify exposition:

$$g(x;\theta) = [\lambda^*(x;\theta) - \lambda_1]^2 + \frac{2n}{\sigma^2}\log[p^*(x;\theta)]. \tag{6.16}$$

Proposition 5. *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

$$\Delta^L = \sqrt{2n\sigma^2\log[1/p^*(0;\theta_\beta)] - \sigma^4 d_j} + \sigma^2\lambda_1, \tag{6.17}$$

$$\Delta^U = \sqrt{2n\sigma^2\log[1/p^*(0;\theta_\beta)]} + \sigma^2\lambda_1 \tag{6.18}$$

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 5. 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):

$$\widetilde{S}(z,\lambda,\Delta) = \frac{1}{n}(|z| - \lambda)_+ sign(z)\mathbb{I}(|z| > \Delta). \tag{6.19}$$

The coordinate-wise update is then

$$\widehat{\beta}_j \leftarrow \widetilde{S}(z_j, \widehat{\sigma}^2\lambda^*(\widehat{\beta}_j;\widehat{\theta}_\beta), \Delta) \tag{6.20}$$

where

$$\Delta = \begin{cases} \sqrt{2n\widehat{\sigma}^2\log[1/p^*(0;\widehat{\theta}_\beta)]} + \widehat{\sigma}^2\lambda_1 & \text{if } g(0;\widehat{\theta}_\beta) > 0, \\ \widehat{\sigma}^2\lambda^*(0;\widehat{\theta}_\beta) & \text{otherwise.} \end{cases} \tag{6.21}$$

The conditional expectation θ_β is updated according to (6.10).

Finally, given the most recent update of the coefficient vector $\widehat{\boldsymbol{\beta}}$, the update for the error variance σ^2 is simply:

$$\widehat{\sigma}^2 \leftarrow \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n+2}. \tag{6.22}$$

Note that this update for σ^2 is a *conditional* mode, not a marginal mode, and so it does not underestimate the error variance in the same way as (3.9). Indeed, conditional on the true regression coefficients, (6.22) is essentially the oracle estimator (4.6). However, although we retain the update (6.22) *during* optimization in order to iterate between the modes of $\boldsymbol{\beta}$ and σ^2, after the algorithm has converged, our final estimator of σ^2 is obtained as

$$\widehat{\sigma}^2_{adj} = \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n - \widehat{q}}, \tag{6.23}$$

where $\widehat{q} = \|\widehat{\boldsymbol{\beta}}\|_0$. Note that (6.23) incorporates an appropriate degrees of freedom adjustment to account for the fact that $\widehat{\boldsymbol{\beta}}$ is an estimate of the unknown true $\boldsymbol{\beta}$.

In principle, both σ^2 and the conditional expectation θ_β should be updated after each β_j, $j = 1, \ldots, p$. 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, where M is the update frequency. The default implementation updates σ^2 and θ_β after every $M = 10$ coordinate updates.

6.3 Implementation

In the SSL with fixed variance, Ročková and George (2018) 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 $(\lambda_1 - \lambda_0)^2 \leq 4$, the objective is convex (Ročková and George, 2018). 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 (2018) will require additional intervention. This is because the objective (6.7) is *always* non-convex when σ^2 is unknown, unlike the fixed case where it is convex when $(\lambda_1 - \lambda_0)^2 \leq 4$. In particular, for small $\lambda_0 \approx \lambda_1$ there may be many negligible but non-zero β_j included in the model. When $p > n$, this severe overfitting can result in all the variation in $\mathbf{Y}$ being explained by the model, forcing the estimate of the error variance, $\widehat{\sigma}^2$ to a mode at 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 an implementation strategy to avoid this absorbing state, we keep the estimate of σ^2 fixed at an initial value until λ_0 reaches a value at which the algorithm converges in less than 100 iterations. We then reinitialize $\boldsymbol{\beta}$ and σ^2 and begin to simultaneously update σ^2 for the next largest λ_0 value in the ladder. The intuition behind this strategy is that we first find a solution to a convex problem (with σ^2 fixed) and then use this

solution as a warm start for the non-convex problem (with σ^2 unknown). A related two-step strategy for non-convex optimization has also been proven successful for robust M-estimation (Loh, 2017).

For initialization, we follow Ročková and George (2018) and initialize the regression coefficients, $\boldsymbol{\beta}$, at zero and $\theta_0 = 0.5$. For the error variance, we devised an initialization strategy that is motivated by the prior for σ^2 used in Chipman et al. (2010). Those authors used a scaled-inverse-χ^2 prior for the error variance with degrees of freedom $\nu = 3$ and scale parameter chosen such that the sample variance of $\mathbf{Y}$ corresponds to the $90th$ quantile of the prior. This is a natural choice as the variance of $\mathbf{Y}$ is the maximum possible value for the error variance. We set the initial value of σ^2 to be the mode of this scaled-inverse-χ^2 distribution, a strategy which we have found to be effective in practice.

The entire implementation strategy is summarized in Algorithm 1 (Section 4 of the supplement, Moran et al., 2019).

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, 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:

$$L(\boldsymbol{\beta}, \sigma^2) = -\frac{1}{2\sigma}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 - \frac{n\sigma}{2} + \sum_{j=1}^{p} pen(\beta_j|\theta_\beta) \tag{6.24}$$

where $pen(\beta_j|\theta_\beta)$ is as defined in (6.8) and again we use the approximation (6.10) 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 (6.24) is very similar to Algorithm 1, the only difference being we replace all σ^2 terms in the updates (6.20) and (6.21) 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 update for σ^2 is only very slightly different from the SSL with unknown variance:

$$\widehat{\sigma}^2 \leftarrow \frac{\|\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2}{n}. \tag{6.25}$$

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 (2018) 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 do not consider the SSL with the p-sigma prior from Section 3.3 as the objective is similar to Städler et al. (2010) (albeit with an adaptive penalty) and so we would expect similar overestimation of σ^2 as proved by Sun and Zhang (2012). We consider three different simulation settings.

For the first simulation setting, we consider the same simulation setting of Ročková and George (2018) 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 $\widetilde{\Sigma} = \{\widetilde{\sigma}\}_{i,j=1}^{50}$ with $\widetilde{\sigma}_{ij} = 0.9$ 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 20 independent blocks of 50 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 MCP (Zhang, 2010).

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 $\boldsymbol{\beta}_0$; the prediction error (PE), defined as

$$\text{PE} = \|\mathbf{X}\boldsymbol{\beta}_0 - \mathbf{X}\widehat{\boldsymbol{\beta}}\|^2; \tag{6.26}$$

the number of false negatives (FN); the number of false positives (FP); the number of true positives (TP); Matthew's Correlation Coefficient (MCC), defined as:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}; \tag{6.27}$$

the percentage of times the method found the correct model (COR); and the time in seconds (TIME). The average of these metrics for each method over the 100 repetitions are displayed in Table 1.

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, prediction error, and MCC. 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 ($\sigma^2 = 1$). We note that the prediction error for the latter implementation is higher than the Adaptive Lasso and SCAD; however, these frequentist methods use cross-validation to choose their regularization parameter and so are optimizing for prediction to the possible detriment of other metrics; the SSL ($\sigma^2 = 1$) still has fewer false positives and a higher MCC. However, both the SSL ($\sigma^2 = 3$) and unknown σ^2 have smaller prediction error than the rest of the methods, including those which use cross-validation, which highlights the predictive gains afforded by variance estimation.

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 shows the variance estimates over the 100 repetitions for the SSL with unknown variance, the scaled SSL and the scaled Lasso. For the SSL with unknown σ^2, these are the estimates (6.23). For the scaled SSL and the scaled Lasso variance estimates, we also applied a degrees of freedom correction similarly to (6.23) using the number of non-zero coefficients found by each method. The variance estimates from the SSL (unknown σ^2) have a median of 2.87 and standard error 0.04. Meanwhile, the scaled SSL slightly underestimates the variance with a median of 2.76 and standard error 0.04, as expected from the larger number of false positives observed in Table 1. Finally, the scaled Lasso highly inflates the variance with a median of 5.88 and standard error 0.14.

7 Protein Activity Data

We now apply the Spike-and-Slab Lasso with unknown variance to the protein activity data set from Clyde and Parmigiani (1998). Following those authors, we code the categorical variables by indicator variables and consider all main effects, two-way interactions and quadratic terms for the continuous variables. This results in a linear model with $p = 88$ potential predictors. The sample size is $n = 96$. We assess the performance of the Spike-and-Slab Lasso with unknown variance in both variable selection and prediction.

7.1 Variable Selection

As an approximation to the "truth", we use the Bayesian adaptive sampling algorithm (BAS, Clyde et al., 2011), which has previously been applied successfully to this dataset.

	HAM	PE	MCC	TP	FP	FN	COR	TIME
SSL (fixed $\sigma^2 = 3$)	**1.1 (0.1)**	**39.6 (3.7)**	**0.91 (0.01)**	**5.5 (0.1)**	**0.5 (0.1)**	**0.5 (0.1)**	**58**	**0.03 (0.00)**
SSL (unknown σ^2)	1.2 (0.2)	43.4 (3.9)	0.90 (0.01)	5.4 (0.1)	0.6 (0.1)	0.6 (0.1)	55	0.04 (0.00)
Scaled SSL	2.0 (0.2)	65.8 (5.0)	0.84 (0.01)	5.2 (0.1)	1.2 (0.1)	0.8 (0.1)	32	0.07 (0.00)
SSL (fixed $\sigma^2 = 1$)	4.5 (0.3)	114.9 (5.3)	0.69 (0.02)	4.8 (0.1)	3.3 (0.2)	1.2 (0.1)	5	0.17 (0.01)
MCP**	7.0 (0.4)	186.1 (7.0)	0.48 (0.02)	3.1 (0.1)	4.1 (0.3)	2.9 (0.1)	1	0.32 (0.00)
Adaptive LASSO	8.1 (0.5)	92.0 (4.1)	0.60 (0.02)	4.8 (0.1)	6.9 (0.5)	1.2 (0.1)	1	5.36 (0.11)
SCAD	11.2 (0.6)	124.4 (6.2)	0.47 (0.02)	4.0 (0.1)	9.2 (0.5)	2.0 (0.1)	0	0.39 (0.01)
MCP*	11.5 (0.4)	181.4 (6.3)	0.35 (0.02)	2.8 (0.1)	8.3 (0.3)	3.2 (0.1)	0	0.32 (0.00)
Scaled LASSO	16.1 (0.4)	302.4 (9.6)	0.42 (0.01)	4.5 (0.1)	14.6 (0.4)	1.5 (0.1)	0	0.51 (0.01)
LASSO	30.9 (0.6)	111.0 (2.5)	0.36 (0.01)	5.4 (0.1)	30.3 (0.6)	0.6 (0.1)	0	0.40 (0.01)

Table 1: Average metrics over 100 repetitions for each of the procedures, ordered by increasing Hamming distance. Standard errors are shown in parentheses. *`ncvreg` implementation using cross-validation over a one-dimensional grid with a default value of the second tuning parameter γ. **hard thresholding tuning with $\gamma = 1.0001$ and cross-validation over λ.

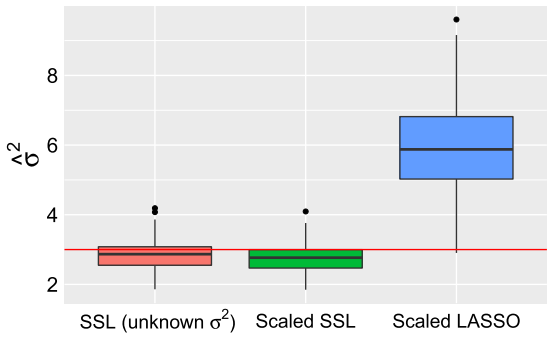

Figure 2: Estimated $\widehat{\sigma}^2_{adj}$ over 100 repetitions. The true variance $\sigma^2 = 3$ is the red horizontal line.

BAS gives posterior inclusion probabilities (PIP) for each of the p potential predictors from which we determined the median probability model (MPM: predictors with PIP > 0.5). The median probability model found by BAS consisted of $q = 7$ predictors: (i) `con:detT`: the interaction of protein concentration and detergent T, (ii) `detT`: detergent T, (iii) `bufTRS:detN`: the interaction of buffer TRS and detergent N, (iv) `con`: protein concentration, (v) `bufPO4:temp`: the interaction of buffer P04 and temperature, (vi) `detN`: detergent N, and (vii) `detN:temp`: the interaction of detergent N and temperature.

For the SSL with unknown variance, we used the same settings as the simulation study with $\lambda_1 = 1$ and $\lambda_0 \in \{1, 2, \ldots, n\}$. The procedure found a model with $\widehat{q} = 6$ predictors, including four of the MPM: `con, detN, bufTRS:detN, con:detT`. Additionally, instead of `detT`, the SSL with unknown variance found the interaction of pH with detergent T (`pH:detT`). The correlation between `detT` and `pH:detT` is 0.988, rendering the two predictors essentially exchangeable. Thus, 5 out of the 6 predictors found by the SSL with unknown variance matched with the benchmark MPM.

For the SSL with known variance, we fixed $\sigma^2 = 0.24$. This is the mean of the scaled-inverse-χ^2 distribution induced by the variance of the response, as detailed in Section 6.4. For the protein data, the variance of the response is 0.41 and so fixing $\sigma^2 = 1$ overestimates the variance, resulting in no signal being found. The SSL with this fixed variance found $\widehat{q} = 2$ predictors: one of the MPM (`detT`) and one not in the MPM but having a correlation of 0.735 with `detN`.

Here, we can see the benefit of simultaneously estimating the error variance; the estimate from SSL with unknown variance was $\widehat{\sigma}^2 = 0.167$, resulting in a less sparse solution.

7.2 Predictive Performance

We now compare the predictive performance of the SSL with unknown variance with the penalized regression methods from the simulation study in Section 6.5 using 8-fold cross validation. We split the data into $K = 8$ sets and denote each set by S_k, $k = 1, \ldots, K$.

The 8-fold cross-validation error is given by:

$$CV = \frac{1}{K}\sum_{k=1}^{K}\sum_{i\in S_k}\left[y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}_{\setminus k}\right]^2 \tag{7.1}$$

where $\widehat{\boldsymbol{\beta}}_{\setminus k}$ is the estimated regression coefficient using the data in S_k^C. We repeated this procedure 100 times and display the resulting cross-validation errors in Figure 3. We do not display the results from the scaled Lasso in Figure 3 as there were a number of outliers: the cross-validation error for the scaled Lasso was greater than 25 in 20% of the replications.

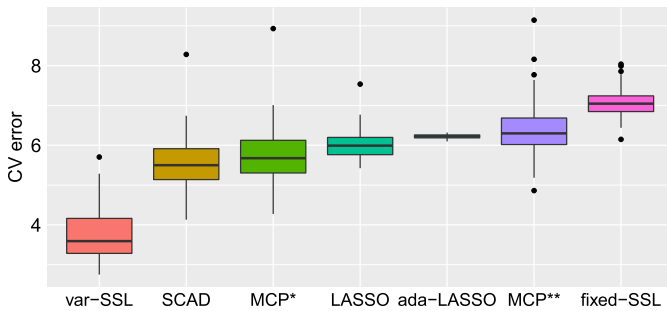

Figure 3: Boxplots of 8-fold cross-validation error over 100 replications for each of the methods (from left to right): 1. SSL (unknown σ^2). 2. SCAD. 3. MCP (`ncvreg`). 4. LASSO. 5. Adaptive LASSO. 6. MCP ($\gamma = 1.0001$). 7. SSL (fixed σ^2).

We can see that the SSL with unknown variance has the smallest cross-validation error. This highlights the gains in predictive performance that can be achieved by simultaneously estimating the error variance and regression coefficients. This result is also very encouraging given that all the other methods (except for the SSL with fixed variance) use cross-validation in choosing their regularization parameters. This also explains the slightly worse performance of the SSL with fixed variance, which we expect would be competitive if we were to also choose the regularization parameters with cross-validation. However, SSL with fixed variance still performs well without the need for computationally intensive cross-validation to choose the parameters.

8 Conclusion

In this paper, we have shown that conjugate continuous priors for Bayesian variable selection can lead to underestimation of the error variance when (i) $\boldsymbol{\beta}$ is sparse; and (ii) when p is of the same order as, or larger than, n. This is because such priors implicitly add p "pseudo-observations" of σ^2 which shift prior mass on σ^2 towards zero. Conjugate priors for linear regression are often motivated by the invariance principle of Jeffreys (1961). Revisiting this work however, we highlighted that Jeffreys' himself cautioned against applying his invariance principle in multivariate problems. Following Jeffreys,

we recommended priors which treat the regression coefficients and error variance as independent.

We then proceeded to extend the Spike-and-Slab Lasso of Ročková and George (2018) 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. On a protein activity dataset, the SSL with unknown variance performed well for both variable selection and prediction. In particular, the SSL with unknown variance exhibited smaller cross-validation error than other penalized likelihood procedures which choose their regularization parameters based on cross-validation. This highlights the predictive benefit of simultaneous variance estimation. The unknown variance implementation of the SSL is provided in the publicly available R package `SSLASSO` (Ročková and Moran, 2017). Code to reproduce the results in this paper is also available at https://github.com/gemma-e-moran/variance-priors.

Supplementary Material

Supplementary Material for "Variance Prior Forms for High-Dimensional Bayesian Variable Selection"
(DOI: 10.1214/19-BA1149SUPP; .pdf).

References

Bayarri, M. J., Berger, J. O., Forte, A., García-Donato, G., et al. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40(3): 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 1092, 1093

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. MR3210986. doi: https://doi.org/10.1214/14-AOS1204. 1092, 1102

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. MR1718789. 1092

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). "Default Bayesian analysis with global-local shrinkage priors." *Biometrika*, 103(4): 955–969. MR3620450. doi: https://doi.org/10.1093/biomet/asw041. 1103

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). "Horseshoe Regularization for Feature Subset Selection." *ArXiv e-prints*. 1104

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 1103

Chipman, H. A., George, E. I., McCulloch, R. E., et al. (2010). "BART: Bayesian additive regression trees." *The Annals of Applied Statistics*, 4(1): 266–298. MR2758172. doi: https://doi.org/10.1214/09-AOAS285. 1110

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). "Bayesian Adaptive Sampling for Variable Selection and Model Averaging." *Journal of Computational and Graphical Statistics*, 20(1): 80–101. MR2816539. doi: https://doi.org/10.1198/jcgs.2010.09049. 1112

Clyde, M. A. and Parmigiani, G. (1998). "Protein construct storage: Bayesian variable selection and prediction with mixtures." *Journal of Biopharmaceutical Statistics*, 8(3): 431–443. 1091, 1093, 1094, 1112

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. MR1946581. doi: https://doi.org/10.1198/016214501753382273. 1111

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software*, 33(1): 1. MR1082147. 1111

Gelman, A. (2004). "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis*. 1093

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. MR3235677. 1096

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association*, 88(423): 881–889. 1093, 1099

George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7: 339–373. 1093

Hans, C. (2009). "Bayesian lasso regression." *Biometrika*, 96(4): 835–845. MR2564494. doi: https://doi.org/10.1093/biomet/asp047. 1103

Jeffreys, H. (1961). *The Theory of Probability*. Oxford University Press, 3 edition. MR0187257. 1091, 1094, 1095, 1115

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103(481): 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 1093

Loh, P.-L. (2017). "Statistical consistency and asymptotic normality for high-dimensional robust M-estimators." *The Annals of Statistics*, 45(2): 866–896. MR3650403. doi: https://doi.org/10.1214/16-AOS1471. 1110

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. MR2894769. doi: https://doi.org/10.1198/jasa.2011.tm09738. 1108

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578. 1093

Moran, G. E., Ročková, V., George, E. I. (2019). "Supplementary Material for "Variance Prior Forms for High-Dimensional Bayesian Variable Selection"." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1149. 1099, 1101, 1106, 1110

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 1103

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. MR3738204. doi: https://doi.org/10.1214/17-EJS1337SI. 1105

Polson, N. G. and Scott, J. G. (2010). "Shrink globally, act locally: sparse Bayesian regularization and prediction." *Bayesian Statistics*, 9: 501–538. MR3204017. doi: https://doi.org/10.1093/acprof:oso/9780199694587.003.0017. 1093, 1094, 1103

Robert, C. P., Chopin, N., and Rousseau, J. (2009). "Harold Jeffreys's Theory of Probability Revisited." *Statistical Science*, 141–172. MR2655841. doi: https://doi.org/10.1214/09-STS284. 1095

Ročková, V. (2018). "Bayesian estimation of sparse signals with a continuous spike-and-slab prior." *The Annals of Statistics*, 46(1): 401–437. MR3766957. doi: https://doi.org/10.1214/17-AOS1554. 1106

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. MR3223753. doi: https://doi.org/10.1080/01621459.2013.869223. 1093, 1099

Ročková, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 113(521): 431–444. MR3803476. doi: https://doi.org/10.1080/01621459.2016.1260469. 1091, 1093, 1094, 1105, 1106, 1107, 1108, 1109, 1110, 1111, 1116

Ročková, V. and Moran, G. (2017). *SSLASSO: The Spike-and-Slab LASSO*. https://cran.r-project.org/package=SSLASSO 1091, 1116

Ročková, V. and Moran, G. (2018). *EMVS: The Expectation-Maximization Approach to Bayesian Variable Selection*. https://cran.r-project.org/package=EMVS 1099

Städler, N., Bühlmann, P., and Van De Geer, S. (2010). "l1 penalization for mixture regression models." *Test*, 19(2): 209–256. MR2677722. doi: https://doi.org/10.1007/s11749-010-0197-z. 1101, 1111

Sun, T. and Zhang, C.-H. (2010). "Comments on: l1-penalization for mixture regression models." *Test*, 19(2): 270–275. MR2677726. doi: https://doi.org/10.1007/s11749-010-0201-7. 1101

Sun, T. and Zhang, C.-H. (2012). "Scaled sparse linear regression." *Biometrika*. 1092, 1102, 1110, 1111

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. MR3486423. doi: https://doi.org/10.1214/16-EJS1130. 1100, 1103

Zhang, C.-H. (2010). "Nearly unbiased variable selection under minimax concave penalty." *The Annals of Statistics*, 38(2): 894–942. MR2604701. doi: https://doi.org/10.1214/09-AOS729. 1111

Zhang, C.-H. and Zhang, T. (2012). "A general theory of concave regularization for high-dimensional sparse estimation problems." *Statistical Science*, 576–593. MR3025135. doi: https://doi.org/10.1214/12-STS399. 1107

Zou, H. (2006). "The adaptive lasso and its oracle properties." *Journal of the American Statistical Association*, 101(476): 1418–1429. MR2279469. doi: https://doi.org/10.1198/016214506000000735. 1111

Acknowledgments

This research was supported by the NSF Grant DMS-1406563, a Simons Fellowship from the Isaac Newton Institute at the University of Cambridge, and the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business. We would like to thank the Editors, Associate Editor, and anonymous referees, as well as Mark van de Wiel and Gwenael Leday, for helpful suggestions which improved this paper.