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

Supplement to "Variance prior forms for high-dimensional Bayesian variable selection"

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

1 Gibbs sampler for Bayesian ridge regression

Here, we present the details of the Gibbs sampler used to obtain posterior estimates for the independent Bayesian ridge regression model in Section 3.2 of the main paper. The model is:

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \tag{1.1}$$

$$\boldsymbol{\beta} \sim N_p(0, \tau^2\mathbf{I}) \tag{1.2}$$

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

The full conditional distributions of the parameters $\boldsymbol{\beta}$ and σ^2 are:

$$\boldsymbol{\beta}|\mathbf{Y}, \sigma^2 \sim N_p(\sigma^{-2}\mathbf{V}\mathbf{X}^T\mathbf{Y}, \mathbf{V}) \tag{1.4}$$

$$\sigma^2|\mathbf{Y}, \boldsymbol{\beta} \sim IG(n/2, \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2/2) \tag{1.5}$$

where $\mathbf{V} = \left[\sigma^{-2}\mathbf{X}^T\mathbf{X} + \tau^{-2}\mathbf{I}_p\right]^{-1}$. The Gibbs sampling algorithm alternates sampling from (1.4) and (1.5). After burn-in, the posterior mean estimates are the means of the samples.

2 Connections with Penalized Likelihood Methods

To show (4.3) in the main paper, consider the objective function (4.1) proposed by Städler et al. (2010) to simultaneously estimate the regression coefficients and error variance in the Lasso. Denote the estimator of the error variance from this objective function by $\widehat{\sigma}^2$. Let $\boldsymbol{\beta}^*$ and σ^* denote the true regression coefficients and error variance, respectively. Sun and Zhang (2010) proved that $\widehat{\sigma}^2$ will overestimate the true variance unless

$$\lambda\|\boldsymbol{\beta}^*\|_1/\sigma^* = o(1). \tag{2.1}$$

Suppose now the true dimension of $\boldsymbol{\beta}^*$ is q and that $\max_j |\beta_j^*| = K_1$ for some constant $K_1 \in \mathbb{R}$. Suppose also that the true variance is bounded: $K_2 < \sigma^* < K_3$ for constants $K_2, K_3 \in \mathbb{R}$. Let $\lambda = A\sqrt{(2/n)\log p}$ for some constant $A > 1$ (the universal threshold). Then,

$$\lambda\|\boldsymbol{\beta}^*\|_1/\sigma^* \leq A\sqrt{(2/n)\log p}\frac{K_1}{K_2}q$$

*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

Hence, from (2.1), the estimator $\widehat{\sigma}^2$ will overestimate the true variance unless

$$q = o\left(\sqrt{\frac{n}{\log p}}\right). \tag{2.2}$$

3 Failure of Spike-and-Slab Lasso with Conjugate Prior

Here, we show that using a conjugate prior formulation for the Spike-and-Slab Lasso when the error variance is unknown can result in underestimation of σ^2. This conjugate prior formulation for the Spike-and-Slab Lasso is given by:

$$\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}, \sigma^2) \sim \prod_{j=1}^{p} \left(\gamma_j \frac{\lambda_1}{2\sigma} e^{-|\beta_j|\lambda_1/\sigma} + (1-\gamma_j)\frac{\lambda_0}{2\sigma} e^{-|\beta_j|\lambda_0/\sigma}\right) \tag{3.1}$$

$$\boldsymbol{\gamma}|\theta \sim \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j}, \quad \theta \sim \text{Beta}(a, b) \tag{3.2}$$

$$p(\sigma^2) \propto \sigma^{-2}. \tag{3.3}$$

The goal is to find the MAP estimates of the regression coefficients and error variance. We find these posterior modes using the EM algorithm, treating the latent indicators $\boldsymbol{\gamma}$ as missing data. That is, we iteratively maximize the expected log posterior:

$$E[\log \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \theta|\mathbf{Y})|\boldsymbol{\beta}^{(k)}, \sigma^{(k)}, \theta^{(k)}] \tag{3.4}$$

where the log posterior is given by:

$$\begin{aligned}\log \pi(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma, \theta|\mathbf{Y}) = &-\frac{1}{2\sigma^2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 - (n+2)\log\sigma \\ &+ \sum_{j=1}^{p} \log\left(\gamma_j \frac{\lambda_1}{2\sigma} e^{-|\beta_j|\lambda_1/\sigma} + (1-\gamma_j)\frac{\lambda_0}{2\sigma} e^{-|\beta_j|\lambda_0/\sigma}\right) \\ &+ \sum_{j=1}^{p} \log\left(\frac{\theta}{1-\theta}\right)\gamma_j + (a-1)\log(\theta) \\ &+ (p+b-1)\log(1-\theta) + C\end{aligned} \tag{3.5}$$

At the $(k+1)$th iteration, the EM updates are:

$$\boldsymbol{\beta}^{(k+1)} = \arg\min_{\boldsymbol{\beta}} \left\{\frac{1}{2\sigma^{(k)}}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} |\beta_j| \lambda^*(\beta_j^{(k)}/\sigma^{(k)}; \theta^{(k)})\right\} \tag{3.6}$$

$$\theta^{(k+1)} = \frac{\sum_{j=1}^{p} p^*(\beta_j^{(k)}/\sigma^{(k)}; \theta^{(k)}) + a - 1}{a + b + p - 2} \tag{3.7}$$

$$\sigma^{(k+1)} = \frac{Q + \sqrt{Q^2 + 4(\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(k)}\|^2)(n+p+2)}}{2(n+p+2)} \tag{3.8}$$

with

$$Q = \sum_{i=1}^{p} |\beta_j^{(k)}| \lambda^*(\beta_j^{(k)}/\sigma^{(k)}; \theta^{(k)}), \tag{3.9}$$

$$p^*(\beta; \theta) = \left[1 + \frac{\lambda_0}{\lambda_1}\left(\frac{1-\theta}{\theta}\right)\exp\{-|\beta|(\lambda_0 - \lambda_1)\}\right]^{-1}, \tag{3.10}$$

$$\lambda^*(\beta; \theta) = \lambda_1 p^*(\beta; \theta) + \lambda_0(1 - p^*(\beta; \theta)), \tag{3.11}$$

where $\boldsymbol{\beta}^{(k)}, \sigma^{(k)}, \theta^{(k)}$ are the parameter values after the kth iteration.

Let us take a closer look at the EM update for 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:

$$\sigma^{*2} = \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^*\|}{n}. \tag{3.12}$$

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

$$\widehat{\sigma}_{MAP} = \tau + \sqrt{\tau^2 + \frac{(\sigma^*)^2}{1 + p/n + 2/n}} \tag{3.13}$$

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 $p/n \to \infty$ and $q/p \to 0$, where the underlying sparsity is $q = \|\boldsymbol{\beta}^*\|_0$, we have $\widehat{\sigma}_{MAP} \to 0$. Thus, similarly to our previous examples in Section 3 and 5 of the main paper, we will severely underestimate the error variance. As in these examples, the remedy is to use the independent prior on σ^2 and $\boldsymbol{\beta}$.

4 Implementation

The implementation of the Spike-and-Slab Lasso with unknown variance as described in Section 6.3 is displayed in Algorithm 1 below:

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: $\boldsymbol{\beta}^* = \mathbf{0}_p$, σ^{*2}, $\theta^* = 0.5$

For $l = 1, \ldots, L$:

1. Set $k_l = 0$
2. Initialize: $\boldsymbol{\beta}_l^{(k_l)} = \boldsymbol{\beta}^*$, $\theta_l^{(k_l)} = \theta^*$, $\sigma_l^{(k_l)2} = \sigma^{*2}$
3. While $\mathsf{diff} > \varepsilon$
   (i) Increment k_l
   (ii) For $s = 1, \ldots, \lfloor p/M \rfloor$:
      i. Update
$$\Delta \leftarrow \begin{cases} \sqrt{2n\sigma_l^{(k_l)2} \log\left[1/p^*(0;\theta_l^{(k_l)})\right]} + \sigma_l^{(k_l)2}\lambda_1 & \text{if } g\left(0;\theta_l^{(k_l)}\right) > 0 \\ \sigma_l^{(k_l)2}\lambda^*(0;\theta_l^{(k_l)}) & \text{otherwise} \end{cases}$$
      ii. For $j = 1, \ldots, M$: update
$$\beta_{l(s-1)M+j}^{(k_l)} \leftarrow \widetilde{S}\left(z_j, \sigma_l^{(k_l-1)2}\lambda^*(\beta_{l(s-1)M+j}^{(k_l-1)}; \theta_l^{(k_l-1)}), \Delta\right)$$
      iii. Update
$$\theta_l^{(k_l)} \leftarrow \frac{a + \|\boldsymbol{\beta}_l^{(k_l)}\|_0}{a + b + p}$$
      iv. If $k_{l-1} < 100$:
         A. Update
$$\sigma_l^{(k_l)2} \leftarrow \frac{\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_l^{(k_l)}\|^2}{n+2}$$
      v. $\mathsf{diff} = \|\boldsymbol{\beta}_l^{(k_l)} - \boldsymbol{\beta}_l^{(k_l-1)}\|_2$
4. Assign $\boldsymbol{\beta}^* = \boldsymbol{\beta}_l^{(k_l)}$, $\sigma^{*2} = \sigma_l^{(k_l)2}$, $\theta^* = \theta_l^{(k_l)}$

References

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

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