Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Comment: Regularization via Bayesian Penalty Mixing

Edward I. George & Veronika Ročková

To cite this article: Edward I. George & Veronika Ročková (2020) Comment: Regularization via Bayesian Penalty Mixing, Technometrics, 62:4, 438-442, DOI: 10.1080/00401706.2020.1801258

To link to this article: https://doi.org/10.1080/00401706.2020.1801258

Published online: 23 Oct 2020.

TECHNOMETRICS
2020, VOL. 62, NO. 4, 438–442
https://doi.org/10.1080/00401706.2020.1801258

Taylor & Francis
Taylor & Francis Group

Comment: Regularization via Bayesian Penalty Mixing

Edward I. George[a] and Veronika Ročková[b]

[a]Department of Statistics, University of Pennsylvania, Philadelphia, PA; [b]Booth School of Business, University of Chicago, Chicago, IL

1. Introduction

Ridge regression and the lasso illustrate some of the remarkable successes of penalized likelihood regularization for regression analysis. When viewed through the Bayesian lens of posterior maximization, the regularizing effects of their penalty functions can be understood as stemming from prior distributions over the regression coefficients. This well-known Bayesian perspective provides further insights into the nature of the shrinkage induced by such penalty functions, and opens the door for the creation of new penalty functions from probabilistic mixtures in the space of priors. The potential of such Bayesian penalty mixing for penalty creation is here illustrated with the spike and slab lasso prior. An adaptive convex combination of lasso estimators which automatically employ strong thresholding shrinkage to small coefficients and weak stabilizing shrinkage to large coefficients, the spike and slab lasso is seen to offer simultaneous variable selection and nearly unbiased estimation of the selected coefficients.

2. Regularization From a Bayesian Perspective

When regression data arise as signal plus noise, overfitting occurs when the fit of an unstable model captures too much of the noise, obscuring the signal and rendering it less useful for out-of-sample prediction. This results from using an overly rich, ill-conditioned model with an unconstrained fitting criterion such as least squares, or more generally maximum likelihood, that exclusively rewards in-sample fit. To guard against such overfitting, the strategy of regularization constrains the fitting criterion with a penalty function that has the effect of shrinking the fit toward more stable structures that resist oversensitivity to the noise. However, depending on the penalty function, the nature of the induced shrinkage regularization can offer additional benefits. For example, ridge regression penalization is grounded in providing a stabilizing influence, while Lasso regression penalization is designed to provide a vehicle for variable selection by exerting stronger shrinkage which thresholds small coefficient estimates to zero. Here we explore the Bayesian motivation for such regularizing penalty functions in order to provide insights that suggest how Bayesian mixture refinements may lead to further benefits.

For suitably centered and standardized regression data y and $X = (x_1, \ldots, x_p)$, one of the clearest examples of effective regularization is the ridge regression estimator

$$\hat{\beta}_{\text{rr}} = (X'X + \lambda I)^{-1} X'y, \tag{1}$$

which can be seen as the solution of the the penalized least squares criterion

$$\underset{\beta}{\text{argmin}} \, \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2. \tag{2}$$

Although Hoerl and Kennard (1970) motivated $\hat{\beta}_{\text{rr}}$ by substituting $(X'X + \lambda I)$ for $X'X$ to shrink and stabilize the least squares estimator $\hat{\beta}_{\text{ls}} = (X'X)^{-1}X'y$, they noted in passing that $\hat{\beta}_{\text{rr}}$ can be obtained both as a solution to (2) and as a posterior mean under a Bayesian formulation.

The penalized least squares criterion (2) is a special case of the now widely used regularization criteria of the form

$$\underset{\beta}{\text{argmax}} \left\{ -\|y - X\beta\|_2^2 + \text{pen}_\lambda(\beta) \right\}, \tag{3}$$

where pen_λ is a penalty function (indexed by λ) that determines the form of the shrinkage adjustment to $\hat{\beta}_{\text{ls}}$. As is well known, such criteria can also be regarded from a Bayesian perspective as posterior maximization under the canonical Gaussian linear model $f(y \mid \beta) = N_n(X\beta, I)$ with a (possible improper) prior distribution π_λ on β. In (3), $-\|y - X\beta\|_2^2$ is proportional to the log-likelihood of β, and $\text{pen}_\lambda(\beta) = \log(\pi_\lambda(\beta)/\pi_\lambda(0))$. (For convenience, we divide by the constant $\pi_\lambda(0)$ to center $\text{pen}_\lambda(\beta)$ at 0.) For example, the ridge penalty function $\text{pen}_\lambda^{\text{rr}}$ is obtained with the Gaussian prior $\pi_\lambda^{\text{rr}}(\beta) = \prod_1^p (\frac{\lambda}{2\pi})^{1/2} e^{-\lambda\beta_j^2/2}$, and the lasso penalty function $\text{pen}_\lambda^{\text{la}}$ is obtained with the Laplace prior $\pi_\lambda^{\text{la}}(\beta) = \prod_1^p \frac{\lambda}{2} e^{-\lambda|\beta_j|}$.

This richer Bayesian perspective provides insight into the nature of a penalty function's induced shrinkage. As Hastie (2020) nicely points out, the more aggressive shrinkage due to the lasso penalty versus the ridge penalty, may be understood as emanating from the higher concentration around zero of the Laplace prior compared to the Gaussian prior. As to the common structure of these penalties, the Bayesian perspective reveals that the separability of both of the ridge and lasso penalty functions, namely their decompositions into sums of p separate coefficient penalties, corresponds to an implicit assumption that the components of $\beta = (\beta_1, \ldots, \beta_p)$ are apriori iid. As to the attractive properties of particular regularizers, such as the

CONTACT Edward I. George edgeorge@wharton.upenn.edu Department of Statistics, University of Pennsylvania, Philadelphia, PA 19104.

mitigation of overfitting and the stabilization of multicollinear predictors, Bayesian motivations may suggest how these ultimately stem from their implicit underlying Bayesian machinery.

3. Bayesian Penalty Mixing

The Bayesian nature of the regularization in (3) opens the door to using probabilistic considerations for the elaboration and construction of new penalty functions. For example, one might consider combining two penalties $\text{pen}_{\lambda_1} = \log \pi_{\lambda_1}$ and $\text{pen}_{\lambda_0} = \log \pi_{\lambda_0}$ as a probabilistic mixture of their underlying priors

$$\text{pen}_\theta^{\text{mix}}(\beta) = \log [\pi_\theta^{\text{mix}}(\beta)/\pi_\theta^{\text{mix}}(0)], \tag{4}$$

where for $\theta \in [0, 1]$,

$$\pi_\theta^{\text{mix}}(\beta) = \prod_1^p [\theta\, \pi_{\lambda_1}(\beta_j) + (1-\theta)\, \pi_{\lambda_0}(\beta_j)]. \tag{5}$$

This example of what we call Bayesian penalty mixing (Ročková and George 2016a), mixes the penalties in their probabilistic prior space rather than directly in their penalty space. In contrast to linear combinations of the penalty functions such as the elastic net (Zou and Hastie 2005), Bayesian penalty mixing leads to an adaptive combination of the mixed penalty effects as described below.

Mixture priors of the form π_θ^{mix} are of particular interest when π_{λ_0} is tightly concentrated around 0, and π_{λ_1} is widely dispersed over large values, leading to the so-called spike-and-slab distributions. Conceptual motivation for such a mixture prior is that it describes the apriori uncertainty faced in a variable selection problem where one suspects only a fraction of the considered predictors to be important enough for inclusion in the model. Intuitively, under a spike-and-slab prior, those coefficients of β drawn from the spike distribution π_{λ_0} would probably be negligible enough to be ignored, while those drawn from the slab distribution π_{λ_1} would probably be large enough to be retained. Through Bayes rule, the posterior will update this prior information with the data to shed light on the underlying true state of nature. By manifesting this update through the posterior mode, the regularized estimate will thus automatically heavily shrink small coefficients while holding the large ones relatively steady. By setting λ_1 small and λ_0 large, π_θ^{mix} using the ridge components $\pi_{\lambda_0}^{\text{rr}}$ and $\pi_{\lambda_1}^{\text{rr}}$ yields the Gaussian spike-and-slab prior of George and McCulloch (1993), while using the lasso components $\pi_{\lambda_0}^{\text{la}}$ and $\pi_{\lambda_1}^{\text{la}}$ yields the spike-and-slab lasso prior of Ročková (2018) and Ročková and George (2018).

Through the Karush–Kuhn–Tucker driven estimating equations for solving (3), the shrinkage effect of any penalty function $\text{pen}(\beta) = \log(\pi(\beta)/\log \pi(0))$ on each β_j is given by its derivative $\frac{\partial \text{pen}(\beta)}{\partial \beta_j} = \frac{\partial \log \pi(\beta)}{\partial \beta_j}$, also the score function of π. For example, $\frac{\partial \text{pen}_\lambda^{\text{rr}}(\beta)}{\partial \beta_j} = -\lambda \beta_j$ for ridge, and $\frac{\partial \text{pen}_\lambda^{\text{la}}(\beta)}{\partial \beta_j} = -\lambda\, \text{sign}(\beta_j)$ for $\beta_j \neq 0$ and $\in [-1, 1]$ for $\beta_j = 0$ for lasso. Interestingly, it turns out that the shrinkage effect of $\text{pen}_\theta^{\text{mix}}$ adaptively combines the shrinkage effects of its components via

$$\frac{\partial \text{pen}_\theta^{\text{mix}}(\beta)}{\partial \beta_j} = p_\theta^*(\beta_j) \frac{\partial \text{pen}_{\lambda_1}(\beta)}{\partial \beta_j} + (1 - p_\theta^*(\beta_j)) \frac{\partial \text{pen}_{\lambda_0}(\beta)}{\partial \beta_j}, \tag{6}$$

where

$$p_\theta^*(\beta_j) = \frac{\theta \pi_{\lambda_1}(\beta_j)}{\theta\, \pi_{\lambda_1}(\beta_j) + (1-\theta)\, \pi_{\lambda_0}(\beta_j)} \tag{7}$$

is the probability that β_j was drawn from the $\pi_{\lambda_1}(\beta_j)$ component of $\pi_\theta^{\text{mix}}(\beta_j)$ in (5). Expressed in terms of the prior score functions, the identity (6) is a consequence of the intriguing property that score functions of mixture distributions are similarly adaptive mixtures of the score functions of their component distributions. Thus, the shrinkage effect of $\text{pen}_\theta^{\text{mix}}$ is an adaptive convex combination of the shrinkage effects of its component penalty functions pen_{λ_1} and pen_{λ_0}. Through $p_\theta^*(\beta_j)$, (6) puts increasing weight on the penalization of β_j by pen_{λ_1} when it becomes increasingly more probable that β_j was drawn from π_{λ_1}.

Although $\text{pen}_\theta^{\text{mix}}$ is flexible in its adaptive application of pen_{λ_1} and pen_{λ_0} to each coefficient of β, it remains a separable sum of penalty mixtures for each β_j. This is a consequence of the underlying iid form of π_θ^{mix} in (5) where θ is fixed, and hence assumed to be known from a Bayesian point of view. Such an assumption would be unrealistic, especially in the spike-and-slab setting described above, where θ represents the fraction of large coefficients to be selected. To allow for the more flexible assumption of unknown θ, we can elaborate $\text{pen}_\theta^{\text{mix}}$ by further Bayesian penalty mixing over a suitable prior Π on θ. This straightforward elaboration is obtained as

$$\text{pen}^{\text{mix}}(\beta) \propto \log \left[\int_0^1 \pi_\theta^{\text{mix}}(\beta) d\Pi(\theta) \,/\, \int_0^1 \pi_\theta^{\text{mix}}(0) d\Pi(\theta) \right]. \tag{8}$$

By averaging over θ, pen^{mix} has become nonseparable, corresponding to the often more realistic assumption of exchangeability of the coefficients of β. As we will see, this averaging has induced dependence across these coefficients that will allow the regularized estimator in (3) to incorporate additional ensemble information in the data.

While pen^{mix} is less tractable than $\text{pen}_\theta^{\text{mix}}$, its shrinkage effect can still be clearly expressed in terms of its component shrinkage effects via

$$\frac{\partial \text{pen}^{\text{mix}}(\beta)}{\partial \beta_j} = p_{\theta_j}^*(\beta_j) \frac{\partial \text{pen}_{\lambda_1}(\beta)}{\partial \beta_j} + (1 - p_{\theta_j}^*(\beta_j)) \frac{\partial \text{pen}_{\lambda_0}(\beta)}{\partial \beta_j}, \tag{9}$$

where

$$p_{\theta_j}^*(\beta_j) = \frac{\theta_j \pi_{\lambda_1}(\beta_j)}{\theta_j\, \pi_{\lambda_1}(\beta_j) + (1-\theta_j)\, \pi_{\lambda_0}(\beta_j)}. \tag{10}$$

This is the same as (6) and (7) except that θ has been replaced by

$$\theta_j \equiv E[\theta \mid \beta_{-j}], \tag{11}$$

the conditional expectation of θ given β_{-j} under the full mixture prior on β and θ. (Here β_{-j} denotes β with β_j excluded.) Because β_{-j} contains information about the relative fraction of coefficients drawn from π_{λ_1} and π_{λ_0}, θ_j borrows strength from β_{-j} to incorporate this ensemble information. Thereby, $p_{\theta_j}^*(\beta_j)$, which controls the adaptation within each β_j via (10), is itself adaptive across all the components of β. Thus, the biasing effect of pen^{mix} becomes doubly adaptive, both within and across the components of β. When π_{λ_0} and π_{λ_1} correspond to the spike-and-slab choices described above, this double adaption

enables pen^{mix} to automatically adapt to ensemble sparsity across $\beta_1, \ldots, \beta_p$.

4. The Spike-and-Slab Lasso (SSL)

To illustrate the potential of this adaptive penalty mixture construction, we now focus on the spike-and-slab lasso (SSL) of Ročková and George (2018). As mentioned above, the SSL penalty is obtained with a mixture of the Laplace priors as

$$\text{pen}^{\text{SSL}}(\beta) = \log \frac{\int_0^1 \prod_1^p [\theta \, \frac{\lambda_1}{2} e^{-\lambda_1|\beta_j|} + (1-\theta)\, \frac{\lambda_0}{2} e^{-\lambda_0|\beta_j|}] d\Pi(\theta)}{\int_0^1 \prod_1^p [\theta \, \frac{\lambda_1}{2} + (1-\theta)\, \frac{\lambda_0}{2}] d\Pi(\theta)}, \tag{12}$$

where a $\text{Beta}(1, p)$ distribution is recommended for the prior Π on θ. With these Laplace priors inserted into (9) and (10), the magnitude of the shrinkage effect induced by pen^{SSL} is seen to be a doubly adaptive convex combination of the two lasso shrinkage penalties λ_1 and λ_0,

$$\lambda^{\star}_{\theta_j}(\beta_j) \equiv \left| \frac{\partial \text{pen}^{\text{ssl}}(\beta)}{\partial \beta_j} \right| = p^{*}_{\theta_j}(\beta_j)\, \lambda_1 + (1 - p^{*}_{\theta_j}(\beta_j))\, \lambda_0, \tag{13}$$

where

$$p^{*}_{\theta_j}(\beta_j) = \frac{\theta_j \lambda_1 e^{-\lambda_1|\beta_j|}}{\theta_j\, \lambda_1 e^{-\lambda_1|\beta_j|} + (1-\theta_j)\, \lambda_0 e^{-\lambda_0|\beta_j|}} \tag{14}$$

and $\theta_j = E[\theta \mid \beta_{-j}]$.

To understand the regularizing effect of the SSL penalty $\text{pen}^{\text{ssl}}(\beta)$, consider the maximum of the penalized likelihood (3) under $\text{pen}^{\text{ssl}}(\beta)$, which we denote by $\widehat{\beta}^{\text{ssl}}$. As shown in Ročková and George (2018), the coefficients of $\widehat{\beta}^{\text{ssl}}$ satisfy

$$\widehat{\beta}^{\text{ssl}}_j = \begin{cases} 0 & \text{when} \quad |z_j| \le \widehat{\Delta}_j, \\ \frac{1}{n}[|z_j| - \lambda^{\star}_{\hat{\theta}_j}(\widehat{\beta}^{\text{ssl}}_j)]_+ \text{sign}(z_j) & \text{when} \quad |z_j| > \widehat{\Delta}_j, \end{cases} \tag{15}$$

where $z_j = x_j'(y - \sum_{k \neq j} x_k \widehat{\beta}_k)$, $\widehat{\theta}_j = E[\theta \mid \widehat{\beta}^{\text{ssl}}_{-j}]$, and $\widehat{\Delta}_j \approx \sqrt{2n \, \log[1/p^{\star}_{\hat{\theta}_j}(0)]} + \lambda_1$.

It is illuminating to examine the form of $\widehat{\beta}^{\text{ssl}}_j$ in (15). When $|z_j| \le \widehat{\Delta}_j$, $\widehat{\beta}^{\text{ssl}}_j$ is thresholded to zero by an adaptive threshold $\widehat{\Delta}_j$ that increases as $\widehat{\theta}_j$ decreases. Thus, $\widehat{\Delta}_j$ acts an automatically increasing multiplicity correction when the estimated fraction $\widehat{\theta}_j$ of large coefficients across $\widehat{\beta}^{\text{ssl}}$ decreases. When $|z_j| > \widehat{\Delta}_j$, $\widehat{\beta}^{\text{ssl}}_j$ is strikingly similar to the soft thresholding form of the lasso estimator. However, rather than shrink by a fixed penalty λ, $\widehat{\beta}^{\text{ssl}}_j$ is shrunk by the self adaptive penalty $\lambda^{\star}_{\hat{\theta}_j}(\widehat{\beta}_j)$, which shrinks more when $\widehat{\beta}^{\text{ssl}}_j$ is large and less when $\widehat{\beta}^{\text{ssl}}_j$ is small. Reflecting the nature of the spike-and-slab prior, this adaptation protects the larger coefficients from the over shrinkage that can occur with the single λ fixed penalty of lasso. In contrast to the lasso, $\widehat{\beta}^{\text{ssl}}$ is doubly adaptive, adapting to the overall level of sparsity, while performing adaptive selection and nearly unbiased estimation of the selected coefficients.

5. Dynamic Posterior Exploration

Implementation of the SSL proceeds by dynamic posterior exploration via successive warm starts along an increasing sequence of λ_0 values. Analogous to implementations of the lasso, coordinate ascent is used to maximize the posterior at each iteration. The desired value of λ_0 and its corresponding subset model are then selected from this path. It should be emphasized that implementation of the SSL is not obtained by plugging in a single preselected value for λ_0.

More precisely, implementation of the SSL regularization path proceeds as follows. With λ_1 fixed at a small value, $\widehat{\beta}^{\text{ssl}}$ is iteratively evaluated over a ladder of increasing λ_0 values starting with $\lambda_0 \approx \lambda_1$. At each iteration, $\widehat{\beta}^{\text{ssl}}$ is updated with the next value of λ_0, using fast coordinate ascent reinitialized at the previous values of $\widehat{\beta}^{\text{ssl}}$ and z. As this algorithm of warm starts proceeds, the $\widehat{\beta}^{\text{ssl}}$ paths can be seen to stabilize. Typically, these easily identified stabilized values will be the desired $\widehat{\beta}^{\text{ssl}}$ estimates. However, unlike the convex lasso posterior, the non-convex nature of the SSL criterion may lead its $\widehat{\beta}^{\text{ssl}}$ paths to local rather than global maxima, and values prior to stabilization may give better prediction error and so be preferred. In any case, such local maxima may still be an improvement over the lasso's global maxima.

We illustrate this implementation and compare it with the lasso on a small simulated data set of $n = 50$ observations of $p = 12$ predictors generated as 4 independent blocks of highly correlated x_j's. More precisely, n rows of X were generated independently from a $N_p(0, \Sigma)$ distribution with block diagonal $\Sigma = \text{bdiag}(\widetilde{\Sigma}, \ldots, \widetilde{\Sigma})$ where $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})_{i,j=1}^3$, $\widetilde{\sigma}_{ij} = 0.9$ if $i \neq j$ and $\widetilde{\sigma}_{ii} = 1$. The response was generated from $y \sim N_n(X\beta_0, I)$ with $\beta_0 = (1.3, 0, 0,\ 1.3, 0, 0,\ 1.3, 0, 0,\ 1.3, 0, 0)'$. Note that only x_1, x_4, x_7, and x_{10} belong in this true model.

Applied to these data, Figure 1 displays the regularization paths of both the SSL and lasso as λ_0 and λ are increased. (For the SSL, $\lambda_1 = 0.01$ is fixed throughout.) Both the SSL and lasso are seen to begin at $\lambda_0 = \lambda = 0$ with the same 12 nonzero maximum likelihood estimates. As the SSL proceeds in the left plot, one can see the distinct manifestations of the spike and of the slab regularizations. Adaptively, the eight smaller estimates are gradually shrunk to zero mostly by the spike penalization, while the four large estimates are held steady mostly by the slab penalization, eventually stabilizing at values which have been shrunk just a bit from their initial maximum likelihood estimates. Here, the SSL has both correctly selected and estimated the four nonzero coefficients of the true model.

Notice also how easy it is to visually identify the stabilized estimates with the SSL, thereby avoiding the need for more elaborate alternatives to select the "best" value of λ_0.

In contrast, as the lasso proceeds in the right plot of Figure 1, one sees the manifestation of regularization with a single Laplace penalty function. With no slab distribution to hold the large values steady, all 12 estimates are gradually shrunk to zero along different varying paths. This is the Bayesian effect of tightening a single Laplace prior down toward a point mass at zero. Due to the order in which the 12 estimates have been thresholded to zero, no value of λ here yields the correct subset selection $\{x_1, x_4, x_7, x_{10}\}$. In particular, the λ chosen by cross-validation here yielded a lasso selected subset with four false

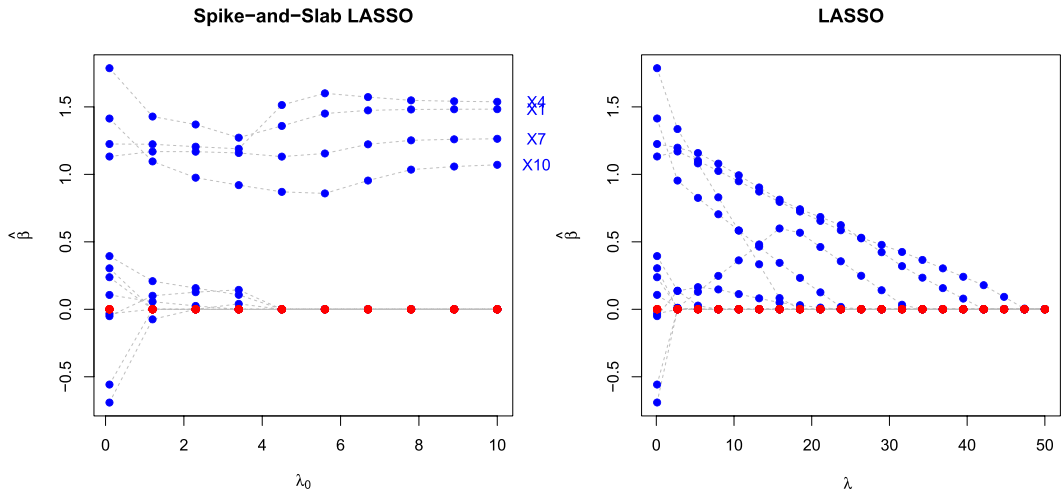

Figure 1. The paths of the $\widehat{\beta}^{\text{ssl}}$ and $\widehat{\beta}^{\text{lasso}}$ estimates as λ_0 and λ are increased. The connected points off the axis (in blue) are the nonzero estimates. The points along the axis (in red) are zero values where the negligible estimates disappear.

positives. Noting that the restrictive conditions for lasso consistency (Zhao and Lu 2006; Zou 2006; Yuan and Lin 2007) do not hold under the correlation structure of our simulation setup, the lasso seems not to be an effective choice for this data. Of course, in less correlated setups, the lasso can be much more effective for selection, although the over shrinkage of large coefficients may still require further mitigation.

6. Further Elaborations

The SSL is a particularly nice illustration of simultaneous selection and estimation obtained by Bayesian penalty mixing because Laplace regularization with the spike distribution automatically thresholds small estimates to zero, just as occurs with the lasso. This would not occur directly with the Gaussian spike-and-slab prior of George and McCulloch (1993) because its underling ridge components $\pi^{\text{rr}}_{\lambda_0}$ and $\pi^{\text{rr}}_{\lambda_1}$ do not provide such thresholding. However, with an introduction of intermediate latent indicators that identify which of $\pi^{\text{rr}}_{\lambda_0}$ and $\pi^{\text{rr}}_{\lambda_1}$ has given rise to each β_j, the regularized maximal estimator $\hat{\beta}^{\text{SS}}$ under this prior can still be found with an EM algorithm that treats the indicators as missing data. A fast and effective approach for selection is then obtained by thresholding $\hat{\beta}^{\text{SS}}$ onto the conditionally most likely values of these indicators. This is the essence of the EMVS algorithm of Ročková and George (2014).

In large part facilitated by further Bayesian considerations, the SSL has recently enjoyed a variety of elaborations and developments. These include variants of the SSL for high-dimensional confounding adjustment in causal analysis (Antonelli, Parmigiani, and Dominici (2019), for high-dimensional Bayesian varying coefficient models (Bai, Chen, and Boland 2020), for grouped regression and sparse generalized additive models (Bai et al. 2020), for simultaneous variable and covariance selection in multivariate regression (Deshpande, Ročková, and George 2019), for graphical models with unequal shrinkage (Gan, Narisetty, and Liang (2019), for regression with unknown error variance (Moran, Ročková, and George 2019), for Bayesian biclustering (Moran, Ročková, and George 2020), for fast Bayesian factor analysis via automatic rotations to sparsity (Ročková and George 2016), for variable selection in time series (Ročková and McAlinn 2020), for generalized linear models (Tang, Shen, Zhang and Yi 2017a), and for the Cox survival model (Tang, Shen, Zhang and Yi 2017b).

Implementations of the SSL for both known and unknown error variance are available in the R package sslasso (Ročková and Moran 2018). These implementations are fast and scalable as they are linear in both n and p. Comparable in speed to the various related fast implementations in the R-package ncvreg, the SSL also saves time by avoiding the need for cross-validated selection of θ which is internally adapted.

Funding

This work was supported by NSF grant DMS-1916245 and the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business.

References

Antonelli, J., Parmigiani, G., and Dominici, F. (2019), "High-Dimensional Confounding Adjustment Using Continuous Spike and Slab Priors," *Bayesian Analysis*, 14, 805–828. [441]

Bai, R., Boland, M. R., and Chen, Y. (2020), "Fast Algorithms and Theory for High-Dimensional Bayesian Varying Coefficient Models," *Journal of the American Statistical Association* (under revision). [441]

Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2020), "Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models," *Journal of the American Statistical Association* (to appear). [441]

Deshpande S. K., Ročková, V., and George E. I. (2019), " Simultaneous Variable and Covariance Selection With the Multivariate Spike-and-Slab Lasso," *Journal of Computational and Graphical Statistics*, 28, 921–931. [441]

Gan, L., Narisetty, N. N., and Liang, F. (2019), "Bayesian Regularization for Graphical Models With Unequal Shrinkage," *Journal of the American Statistical Association*, 114, 1218–1231. [441]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [439,441]

Hastie, T. (2020), "Ridge Regularization: An Essential Concept in Data Science," *Technometrics*, this issue, DOI: 10.1080/00401706.2020.1791959. [438]

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 55–67. [438]

Moran, G. E., Ročková, V., and George, E. I. (2019), "Variance Prior Forms for High Dimensional Bayesian Variable Selection," *Bayesian Analysis*, 14, 1091–1119. [441]

——— (2020), "Spike-and-Slab Lasso Biclustering," *The Annals of Applied Statistics* (under revision). [441]

Ročková, V. (2018), "Bayesian Estimation of Sparse Signals With a Continuous Spike-and-Slab Prior," *The Annals of Statistics*, 46, 401–437. [439]

Ročková, V., and George, E. I. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846. [441]

——— (2016a), "Bayesian Penalty Mixing: The Case of a Non-Separable Penalty," in *Statistical Analysis for High-Dimensional Data—The Abel Symposium 2014*, Springer, pp. 233–254. [439]

——— (2016b), "Fast Bayesian Factor Analysis via Automatic Rotations to Sparsity," *Journal of the American Statistical Association*, 111, 1608–1622. [441]

——— (2018), "The Spike-and-Slab LASSO, *Journal of the American Statistical Association*, 113, 431–444. [439,440]

Ročková V., and McAlinn, K. (2020), "Dynamic Variable Selection With Spike-and-Slab Process Priors," *Bayesian Analysis* (to appear). [441]

Ročková, V., and Moran, G. E. (2018), "SSLASSO: The Spike-and-Slab LASSO," R Package Version 1.2-1. [441]

Tang, Z., Shen, Y., Zhang, X., and Yi, N. (2017a), "The Spike-and-Slab Lasso Generalized Linear Models for Prediction and Associated Genes Detection," *Genetics*, 205, 77–88. [441]

——— (2017b), "The Spike-and-Slab Lasso Cox Model for Survival Prediction and Associated Genes Detection," *Bioinformatics*, 33, 2799–2807. [441]

Yuan, M., and Lin, Y. (2007), "On the Nonnegative Garotte Estimator," *Journal of the Royal Statistical Society*, Series B, 69, 143–161. [441]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [441]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [441]

Zou, H., and Hastie, T. (2005), "Regression Shrinkage and Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [439]