Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

The Spike-and-Slab LASSO

Veronika Ročková & Edward I. George

To cite this article: Veronika Ročková & Edward I. George (2018) The Spike-and-Slab LASSO, Journal of the American Statistical Association, 113:521, 431-444, DOI: 10.1080/01621459.2016.1260469

To link to this article: https://doi.org/10.1080/01621459.2016.1260469

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2018, VOL. 113, NO. 521, 431–444, Theory and Methods
https://doi.org/10.1080/01621459.2016.1260469

The Spike-and-Slab LASSO

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

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

ABSTRACT

Despite the wide adoption of spike-and-slab methodology for Bayesian variable selection, its potential for penalized likelihood estimation has largely been overlooked. In this article, we bridge this gap by cross-fertilizing these two paradigms with the *Spike-and-Slab LASSO* procedure for variable selection and parameter estimation in linear regression. We introduce a new class of self-adaptive penalty functions that arise from a fully Bayes spike-and-slab formulation, ultimately moving beyond the separable penalty framework. A virtue of these nonseparable penalties is their ability to borrow strength across coordinates, adapt to ensemble sparsity information and exert multiplicity adjustment. The *Spike-and-Slab LASSO* procedure harvests efficient coordinate-wise implementations with a path-following scheme for dynamic posterior exploration. We show on simulated data that the fully Bayes penalty mimics oracle performance, providing a viable alternative to cross-validation. We develop theory for the separable and nonseparable variants of the penalty, showing rate-optimality of the global mode as well as optimal posterior concentration when $p > n$. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2015
Revised July 2016

KEYWORDS
High-dimensional regression; LASSO; Penalized likelihood; Posterior concentration; Spike-and-Slab; Variable selection

1. Introduction

Spike-and-slab formulations are fundamentally probabilistic constructs for sparse recovery, most naturally understood from the Bayesian standpoint. Penalized likelihood approaches, on the other hand, induce sparsity through penalty functions whose geometry is exerted in constrained optimization. Forming a bridge between these two parallel developments, here we harvest their potential for mutual cross-fertilization with the *Spike-and-Slab LASSO* (SSL) procedure for simultaneous variable selection and parameter estimation.

For the well-studied problem of variable selection in multiple regression, consider the classical linear model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}_0 + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\boldsymbol{Y}$ is an n-dimensional response vector, $\boldsymbol{X}_{n\times p} = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p]$ is a fixed regression matrix of p potential predictors, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0p})'$ is a p-dimensional vector of unknown regression coefficients, and $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \mathrm{I}_p)$ is the noise vector. We tacitly assume that $\boldsymbol{Y}$ has been centered at 0 to avoid the need for an intercept. The regressors will be treated as centered and standardized with $||\boldsymbol{X}_j||^2 = n$ for $1 \le j \le p$. We focus on settings where $p > n$ and where many of the components of $\boldsymbol{\beta}_0$ are zero or so small as to render most of the potential predictors inconsequential. The complexity of the solution will be denoted by $q = ||\boldsymbol{\beta}_0||_0$. In this setup, we are interested in a purposeful recovery of $\boldsymbol{\beta}_0$, which entails (a) the identification of active predictors and (b) estimation of their effects.

A variant of the penalized likelihood approach estimates $\boldsymbol{\beta}_0$ with

$$\widehat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ -\frac{1}{2}||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \mathrm{pen}_\lambda(\boldsymbol{\beta}) \right\}, \tag{1.2}$$

where $\mathrm{pen}_\lambda(\boldsymbol{\beta})$ is a penalty function (indexed by a penalty parameter λ) prioritizing solutions that are suitably disciplined. An overwhelming emphasis in the literature has been on penalty functions that are separable, that is, $\mathrm{pen}_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^p \rho_\lambda(\beta_j)$. Most notably, the best subset selection ℓ_0 approach deploys $\rho_\lambda(\beta_j) = -\lambda\,\mathrm{I}(\beta_j \neq 0)$, whereas the LASSO ℓ_1 penalty of Tibshirani (1994) uses $\rho_\lambda(\beta_j) = -\lambda|\beta_j|$, its closest concave relative. (The connotation concave vs. convex is reversed here relative to the conventional penalized likelihood literature. To us, the penalized likelihood objective corresponds to an actual penalized log-likelihood with a minus sign.) These two approaches stand at the two ends of a conceptual and a computational ideal for sparsity detection. Nonconcave separable elaborations, intermediate between the two, have witnessed a surge of interest (e.g., the MCP penalty of Zhang 2010, and the SCAD penalty of Fan and Li 2001). These penalties have the ability to threshold (select) and, at the same time, diminish the well-known estimation bias of the LASSO. Any penalized likelihood estimator (1.2) may be seen as a posterior mode under a (possibly improper) prior $\pi(\boldsymbol{\beta} \mid \lambda)$, where $\mathrm{pen}_\lambda(\boldsymbol{\beta}) = \log \pi(\boldsymbol{\beta} \mid \lambda)$. In particular, separable penalties stem from independent product priors.

Spike-and-slab approaches to Bayesian variable selection, on the other hand, arise directly from probabilistic considerations. With a hierarchical prior over the parameter and model spaces,

CONTACT Veronika Ročková Veronika.Rockova@chicagobooth.edu Department of Econometrics and Statistics at the Booth School of Business of the University of Chicago, Chicago, IL 60637.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

the generic form of a spike-and-slab prior is given by

$$\pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) = \prod_{i=1}^{p} [\gamma_i \psi_1(\beta_i) + (1-\gamma_i)\psi_0(\beta_i)], \quad \boldsymbol{\gamma} \sim \pi(\boldsymbol{\gamma}), \tag{1.3}$$

where $\psi_1(\beta)$ serves as a diffuse "slab distribution" for modeling large effects, $\psi_0(\beta)$ serves as a concentrated "spike distribution" for modeling negligibly small effects, and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)'$, $\gamma_i \in \{0, 1\}$, indexes the 2^p possible subset models. (With a *continuous spike distribution*, the γ's serve only to partition $\boldsymbol{\beta}$ into small and large values, rather than to perfectly identify subset models.) For the *Spike-and-Slab LASSO*, we deploy the particular variant of (1.3) for which $\psi_1(\beta) = \frac{\lambda_1}{2} e^{-\lambda_1|\beta|}$ with λ_1 small, and $\psi_0(\beta) = \frac{\lambda_0}{2} e^{-\lambda_0|\beta|}$ with λ_0 large. Proposed by Rockova (2015) for sparse normal means recovery, these two-point mixtures of Laplace distributions will be referred to as SSL (Spike-and-Slab LASSO) priors. Although the scope of SSL priors can be vastly enhanced with different choices of $\pi(\boldsymbol{\gamma})$, our development here will confine attention to the exchangeable prior forms

$$\pi(\boldsymbol{\gamma} \mid \theta) = \prod_{j=1}^{p} \theta^{\gamma_j}(1-\theta)^{1-\gamma_j}, \quad \theta \sim \pi(\theta), \tag{1.4}$$

where $\theta = \mathrm{P}(\gamma_i = 1 \mid \theta)$ is the prior expected fraction of large $\beta_j's$.

In sharp contrast to the usual practice of distilling $\pi(\boldsymbol{\gamma} \mid \boldsymbol{Y})$ for variable selection, for the *Spike-and-Slab LASSO*, we convert the SSL prior (1.3)–(1.4) into a penalty function for modal estimation of $\boldsymbol{\beta}_0$ directly in the parameter space via (1.2). Conditionally on θ, the SSL prior boils down to an independent product of mixtures

$$\pi(\boldsymbol{\beta} \mid \theta) = \prod_{i=1}^{p} [\theta \psi_1(\beta_i) + (1-\theta)\psi_0(\beta_i)], \tag{1.5}$$

where $\boldsymbol{\gamma}$ has been margined out under (1.4). A mixture of two Laplace distributions, modal estimates under (1.5) perform simultaneous variable selection and parameter estimation, adaptively thresholding small effects with the concentrated double exponential spike while keeping large effects steady with the heavy-tailed slab. Choosing a point-mass spike $\psi_0(\beta_j) = \mathrm{I}(\beta_j = 0)$ (obtained as $\lambda_0 \to \infty$) and $\psi_1(\beta_j) \propto c > 0$ (obtained as $\lambda_1 \to 0$), $\log \pi(\boldsymbol{\beta} \mid \theta)$ collapses to the ℓ_0 penalty. At the other end, choosing $\psi_1(\beta_j) = \psi_0(\beta_j)$ yields the familiar LASSO penalty with a parameter $\lambda_1 = \lambda_0$. Thus, a feature of the SSL priors is their ability to induce a nonconcave continuum between these two ideals.

Whereas separable penalty functions arising from $\pi(\boldsymbol{\beta} \mid \theta)$ are interesting constructs themselves, we shall ultimately be interested in fully Bayes penalty functions obtained by treating θ as unknown and random with $\theta \sim \pi(\theta)$. Moving beyond the framework of independence to exchangeability, such hierarchical mixtures have proved remarkably successful, (a) producing posteriors that adapt to underlying sparsity (Castillo and van der Vaart 2012), (b) performing automatic multiplicity adjustment (Scott and Berger 2010), and (c) achieving Bayes factor consistency in high-dimensional regression (Moreno, Girón, and Casella 2015), to name just a few. But how exactly does the fully Bayes construction manifest itself in the posterior modes through a penalty function? Intuitively, the unconditional prior $\pi(\boldsymbol{\beta})$ renders the coordinates dependent, providing an opportunity to borrow strength. This very dependence penetrates into a penalty $\log \pi(\boldsymbol{\beta})$, which is ultimately nonseparable, an essential building block of our approach.

Although the shrinkage/penalty properties of spike-and-slab priors have been recognized (Scheipl, Fahrmeir, and Kneib 2012), the potential of spike-and-slab penalty creation for modal estimation (1.2) has been under-appreciated. Here, we unleash that potential with the *Spike-and-Slab LASSO*. Our contributions are summarized in the points below:

1. A novel penalized likelihood perspective is provided for the treatment of *continuous* spike-and-slab priors in the context of high-dimensional regression when $p > n$. The framework of nonseparable fully Bayes penalties is introduced and developed, showing their potential for self-adaptivity and automatic hyperparameter tuning.
2. Within the realm of Bayesian variable selection, it is typically the entire posterior distribution that is used as a vehicle for variable selection. However, the practicality of MCMC posterior simulation is often limited by the dimensionality p. Here, we focus primarily on mode detection, capitalizing on developments in nonconcave optimization (Breheny and Huang 2011; Mazumder, Friedman, and Hastie 2011). Drawing upon the similarities to the LASSO, we extend existing coordinate-wise optimization algorithms to the case of a nonseparable SSL penalty.
3. The *Spike-and-Slab LASSO* method for variable selection is introduced, entailing the deployment of a sequence of (nonseparable) priors within a path-following scheme. Unlike the LASSO that uses a sequence of single Laplace priors with an increasing penalty λ, the *Spike-and-Slab LASSO* uses a sequence of Laplace *mixtures* with an increasing *spike penalty* λ_0, while keeping λ_1 fixed to a small constant. A similar strategy was deployed in the EMVS procedure by Rockova and George (2014) for Gaussian spike-and-slab mixtures.

   Path-following schemes are now routine for both concave/nonconcave regularization. SCAD and MCP penalties have two hyperparameters that require tuning, so that cross-validation over a two-dimensional grid is often needed (Breheny and Huang 2011; Mazumder, Friedman, and Hastie 2011). We also have two tuning parameters (λ_0, θ). However, by treating θ as random, the nonseparable SSL penalty avoids the need for cross-validation over θ. This aspect has distinct practical advantages.
4. Finally, we provide asymptotic arguments for the suitability of SSL priors for modal estimation and full Bayes inference in high-dimensional linear regression. Extending the work by Rockova (2015), we show (near-minimax) rate-optimality of the global mode under the separable penalty when $p > n$. This result is complemented with a variant involving the entire posterior measure. Building on the work by Castillo, Schmidt-Hieber, and van der Vaart (2015), we show that the SSL posterior keeps pace with the global mode by concentrating

at the optimal rate when $p > n$. This result demonstrates that the penalized likelihood surface is a valid fully Bayes posterior, not just an objective function for outputting a mode. Going further, we extend the analysis of the global mode to the case of nonseparable SSL penalty functions, illuminating their potential for refining statistical rates.

The article is structured as follows. Section 2 revisits the nonseparable SSL penalty by Rockova (2015) in the context of high-dimensional linear regression. Section 3 introduces the framework of fully Bayes nonseparable SSL penalties. Section 4 proposes a new coordinate ascent strategy for the nonseparable SSL penalty. Section 5 introduces the *Spike-and-Slab LASSO* approach and demonstrates its potential with a simulation study. Section 6 presents the asymptotic results and Section 7 concludes with a discussion.

2. The Separable SSL Penalty

A key ingredient of our approach is drawing upon connections between Spike-and-Slab LASSO modal estimation, the foundation of our variable selection procedure, and generalized LASSO estimation. An essential first step will be understanding the mechanics of a separable Spike-and-Slab LASSO penalty. This penalty arises from the independent product prior (1.5), assuming θ is fixed as if it were known. Paralleling the development by Rockova (2015) for normal means, here we demonstrate the potential of this penalty in the context of high-dimensional regression. This section serves as an overture to the fully Bayes approach developed in the next section.

Definition 1. Given $\theta \in (0, 1)$, the separable Spike-and-Slab LASSO (SSL) penalty is defined as

$$\begin{aligned} \text{pen}_S(\boldsymbol{\beta} \mid \theta) &= \log\left[\frac{\pi(\boldsymbol{\beta} \mid \theta)}{\pi(\mathbf{0}_p \mid \theta)}\right] \\ &= \sum_{j=1}^{p} \log\left[\frac{\theta\,\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)}{\theta\,\psi_1(0) + (1-\theta)\psi_0(0)}\right]. \end{aligned} \tag{2.1}$$

To facilitate manipulations with the penalty, we have centered it so that $\text{pen}_S(\mathbf{0}_p \mid \theta) = 0$. Due to the conditional independence of $\boldsymbol{\beta}$ given θ, the penalty is built from singletons

$$\rho(\beta_j \mid \theta) = -\lambda_1|\beta_j| + \log\left[p^\star_\theta(0)/p^\star_\theta(\beta_j)\right], \tag{2.2}$$

which add up to yield

$$\begin{aligned} \text{pen}_S(\boldsymbol{\beta} \mid \theta) &= \sum_{j=1}^{p} \rho(\beta_j \mid \theta) \\ &= -\lambda_1|\boldsymbol{\beta}| + \sum_{j=1}^{p} \log\left(\frac{p^\star_\theta(0)}{p^\star_\theta(\beta_j)}\right), \quad \text{where} \end{aligned} \tag{2.3}$$

$$p^\star_\theta(\beta_j) = \frac{\theta\psi_1(\beta_j)}{\theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)}. \tag{2.4}$$

The alternative characterization (2.3) writes the separable SSL penalty as an adaptive sum of a LASSO penalty and a nonconcave penalty, rendering it ultimately nonconcave. The maximal nonconcavity equals $\kappa = \frac{1}{4}(\lambda_0 - \lambda_1)^2$, where larger differences $\lambda_0 - \lambda_1$ yield more aggressive penalties that are en route to best subset selection. The penalty is indexed by a triplet of unknown parameters $(\lambda_1, \lambda_0, \theta)$, which work in tandem to yield desirable properties (Rockova 2015). Throughout the article, we assume that λ_1 has been set to a small value (made precise by our theoretical study in Section 6). The two parameters (λ_0, θ) will be seen to drive the performance of the penalty, and their tuning will be of the utmost importance.

The role of (λ_0, θ) is best understood by looking at the univariate regularizer $\rho(\beta_j \mid \theta)$. As illustrated by Figure 1(a) (which plots $\rho(\beta_j \mid \theta)$ with a *minus* sign), the larger λ_0, the closer the approximation to ℓ_0. The plot also portrays $\rho(\beta_j \mid \theta)$ as a smooth mix of two ℓ_1 penalties with parameters (λ_0, λ_1), where λ_0 takes over near origin and λ_1 dominates for larger values $|\beta_j|$. The vertical lines correspond to the intersection point between the spike-and-slab densities

$$\delta = \frac{1}{\lambda_0 - \lambda_1} \log\left[1/p^\star_\theta(0) - 1\right], \tag{2.5}$$

the value where the slab begins to dominate the spike. The sharper the spike (i.e., λ_0 is large), the smaller the threshold. A similar effect can be achieved by modulating the prior weight $\theta \in (0, 1)$. As seen from Figure 1(b), larger θ values produce a larger prior inclusion probability and thereby a smaller threshold. Our particular choices $\theta = 2/3$ and $\theta = 0.34$ will be motivated in the next section, with an illustration of the nonseparable SSL penalty. Figure 1 also shows that $\rho(\beta \mid \theta)$ shares many of the desirable properties required for separable regularizers (Zhang and Zhang 2012): it is nonconcave, nonincreasing in $[0; \infty)$, and it is super-additive due to the convexity of $\log[p^\star_\theta(0)]/p^\star_\theta(\beta)]$, that is, $\rho(x + y \mid \theta) \geq \rho(x \mid \theta) + \rho(y \mid \theta)$ for all $x, y \geq 0$.

Before proceeding, it is worthwhile to examine more closely $p^\star_\theta(\beta_j)$ defined in (2.4), the fundamental element of the penalty. This exponential mixing weight can be seen as the conditional probability that β_j came from $\psi_1(\beta_j)$ rather than from $\psi_0(\beta_j)$. Indeed,

$$\begin{aligned} p^\star_\theta(\beta_j) &= \text{P}(\gamma_j = 1 \mid \beta_j, \theta) \\ &= \left[1 + \frac{\lambda_0}{\lambda_1}\frac{(1-\theta)}{\theta} e^{-|\beta_j|(\lambda_0 - \lambda_1)}\right]^{-1}. \end{aligned} \tag{2.6}$$

This quantity, which also appears in (2.5), will keep reoccurring throughout the article in many different contexts, including implementation, statistical rates of the global mode, and posterior concentration rates. Fundamentally an adaptive mixing weight, $p^\star_\theta(\beta_j)$ determines the amount of shrinkage borrowed from the spike and the slab. This is formalized in the following revealing lemma.

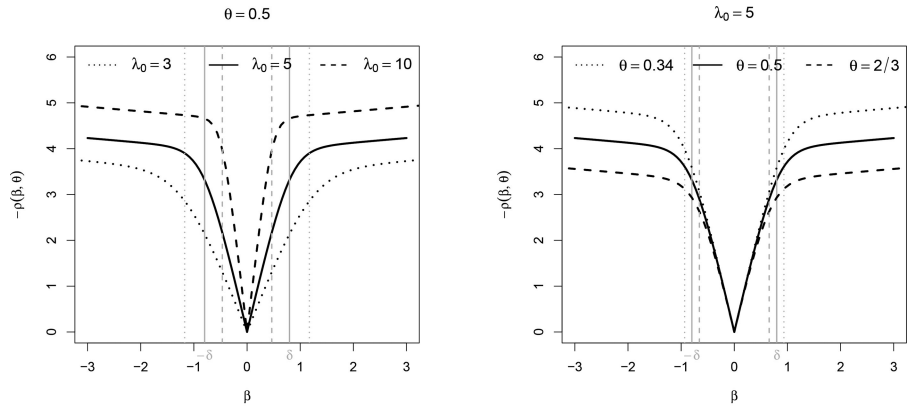
Lemma 1. The derivative of the separable SSL penalty satisfies

$$\begin{aligned} \frac{\partial \text{pen}_S(\boldsymbol{\beta} \mid \theta)}{\partial|\beta_j|} &\equiv -\lambda^\star_\theta(\beta_j), \quad \textit{where} \\ \lambda^\star_\theta(\beta_j) &= \lambda_1 p^\star_\theta(\beta_j) + \lambda_0\left[1 - p^\star_\theta(\beta_j)\right]. \end{aligned} \tag{2.7}$$

Proof. The result follows immediately from

$$\begin{aligned} \frac{\partial \text{pen}_S(\boldsymbol{\beta} \mid \theta)}{\partial|\beta_j|} &= p^\star_\theta(\beta_j)\frac{\partial \log \psi_1(\beta_j)}{\partial|\beta_j|} \\ &\quad + \left[1 - p^\star_\theta(\beta_j)\right]\frac{\partial \log \psi_0(\beta_j)}{\partial|\beta_j|}. \end{aligned}$$

□



(a) $\theta = 0.5, \lambda_1 = 0.1$ (b) $\lambda_0 = 5, \lambda_1 = 0.1$

Figure 1. The plot of the univariate penalty function $\rho(\beta \mid \theta)$ *with a minus sign* for different choices (λ_0, θ). The vertical lines correspond to the intersection point δ.

By exerting its influence through the Karush–Kuhn–Tucker (KKT) conditions (seen in (2.8) and (2.9)), $\lambda^\star_\theta(\cdot)$ drives a "bias term" of the induced estimator (Fan and Li 2001), determining the amount of shrinkage. Ideally, one would like to shrink by a small amount when $|\beta_j|$ is large, and by a large amount when $|\beta_j|$ is small. This is accomplished by the exponential mixing weight $p^\star_\theta(\beta_j)$ (2.6), which gears $\lambda^\star_\theta(\beta_j)$ toward the extreme values λ_1 and λ_0, depending on the size $|\beta_j|$. Thus, $\lambda^\star_\theta(\beta_j)$ mixes the two LASSO "bias terms" and does so adaptively. This mixture penalty effect is very much in contrast with a *nonadaptive* sum of the ℓ_1 and a nonconcave penalty (Fan and Lv 2014).

2.1. Shrinkage Effects in Linear Regression

Throughout this section, we let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ denote the global posterior mode (1.2) under $\text{pen}_S(\boldsymbol{\beta} \mid \theta)$. The adaptive features of the SSL penalty in linear regression are revealed from necessary conditions for $\widehat{\boldsymbol{\beta}}$. We begin with the KKT conditions

$$\boldsymbol{X}_j'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \lambda^\star_\theta(\widehat{\beta}_j)\text{sign}(\widehat{\beta}_j) \quad \text{for} \quad \widehat{\beta}_j \neq 0, \tag{2.8}$$

$$|\boldsymbol{X}_j'(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})| \leq \lambda^\star_\theta(\widehat{\beta}_j) \quad \text{for} \quad \widehat{\beta}_j = 0, \tag{2.9}$$

which follow from the sub-differential calculus and Lemma 1. Using the fact $||\boldsymbol{X}_j||^2 = n$ for $1 \leq j \leq p$, (2.8) and (2.9) write equivalently as

$$\widehat{\beta}_j = \frac{1}{n}[|z_j| - \lambda^\star_\theta(\widehat{\beta}_j)]_+ \text{sign}(z_j), \quad j = 1, \ldots, p, \tag{2.10}$$

where $z_j = \boldsymbol{X}_j'(\boldsymbol{Y} - \sum_{k \neq j} \boldsymbol{X}_k \widehat{\beta}_k)$.

The representation (2.10) is strikingly similar to the LASSO iterative soft-thresholding operator (Friedman, Hastie, and Tibshirani 2010). While the LASSO penalty induces a constant shrinkage term λ, the SSL penalty induces an adaptive term $\lambda^\star_\theta(\widehat{\beta}_j)$ that implicitly depends on the data through $\widehat{\beta}_j$ itself. Despite implicit, the representation (2.10) has instantaneous implications for the implementation (explored in Section 4). As with the adaptive LASSO (Zou 2006) and weighted ℓ_1 penalties (Candes, Wakin, and Boyd 2008), each coefficient has its own term, performing selective shrinkage. However, here the term is self-adaptive, deploying a large penalty (close to λ_0) to threshold small $\widehat{\beta}_j$, and a small penalty (close to λ_1) to hold large $\widehat{\beta}_j$ steady with only slight bias. This adaptive aspect serves to ameliorate the well-known bias issue of concave regularizers.

It is important to alert the reader that the necessary characterization (2.10) will not be sufficient, unless the log-posterior is unimodal. Such unimodal log-posteriors will occur when $p < n$ and λ_0 and λ_1 are not too different. This can be seen by noting that the maximal nonconcavity κ dominates the concavity of the log-likelihood when $(\lambda_0 - \lambda_1)^2 > 4\lambda_{\min}$, where $\lambda_{\min}$ is the smallest eigenvalue of the Gram matrix $\boldsymbol{X}'\boldsymbol{X}$. Here, however, we are primarily interested in high-dimensional scenarios $p > n$, where $\lambda_0 \to \infty$ as $n \to \infty$, allowing $\text{pen}_S(\boldsymbol{\beta} \mid \theta)$ to approximate the ℓ_0 penalty arbitrarily closely. This asymptotic regime is apt to generate multimodal posterior landscapes. For these scenarios, we derive a more refined characterization of $\widehat{\boldsymbol{\beta}}$.

This characterization is obtained by noting that $\widehat{\beta}_j$ is a global mode in the jth direction, while keeping the other coordinates fixed at all but the jth entry of $\widehat{\boldsymbol{\beta}}$. Thus, with z_j as before,

$$\widehat{\beta}_j = \arg\max_\beta \left[-\frac{1}{2}(z_j - n\beta)^2 + n\rho(\beta \mid \theta) \right]. \tag{2.11}$$

It now follows that $\widehat{\beta}_j = 0$ if and only if $|z_j| \leq \Delta$, where

$$\Delta \equiv \inf_{t>0} [nt/2 - \rho(t \mid \theta)/t] \tag{2.12}$$

(using arguments of Zhang and Zhang 2012). Combined with (2.10), we obtain the following refined characterization of the global mode.

Theorem 1. Let $z_j = \boldsymbol{X}_j'(\boldsymbol{Y} - \sum_{k \neq j} \boldsymbol{X}_k \widehat{\beta}_k)$. Then the global mode $\widehat{\boldsymbol{\beta}}$ under $\text{pen}_S(\boldsymbol{\beta} \mid \theta)$ satisfies

$$\widehat{\beta}_j = \begin{cases} 0 & \text{when} \quad |z_j| \leq \Delta, \\ \frac{1}{n}[|z_j| - \lambda^\star_\theta(\widehat{\beta}_j)]_+ \text{sign}(z_j) & \text{when} \quad |z_j| > \Delta, \end{cases} \tag{2.13}$$

where Δ is the selection threshold (2.12).

Theorem 1 shows that the global mode estimator $\widehat{\boldsymbol{\beta}}$ is a blend of soft and hard-thresholding. As a practical matter, the characterization (2.13) helps narrow down the set of candidates for the global posterior mode and devise more targeted numerical

procedures (Section 4). The properties of $\widehat{\boldsymbol{\beta}}$ are ultimately determined by the threshold level Δ. Thus, it is worthwhile to understand the calibration of Δ in relation to the parameters (λ_0, θ). Interestingly, the quantity $p^\star_\theta(0)$ will play an integral role in Δ.

To begin with, the threshold always satisfies $\Delta \leq \lambda^\star_\theta(0) = p^\star_\theta(0)\lambda_1 + [1 - p^\star_\theta(0)]\lambda_0$ (Zhang and Zhang 2012). However, when λ_0 gets large, this bound is too loose and can be improved. To formalize this intuition, we need to introduce a bit of notation. Following Rockova (2015), we define

$$g_\theta(x) = \left[\lambda^\star_\theta(x) - \lambda_1\right]^2 + 2\, n \log p^\star_\theta(x). \tag{2.14}$$

Denote by $c_+ = 0.5(1 + \sqrt{1 - 4n/(\lambda_0 - \lambda_1)^2}$ and $\delta_{c+} = \frac{1}{\lambda_0 - \lambda_1} \log[\frac{1-\theta}{\theta} \frac{\lambda_0}{\lambda_1} \frac{c_+}{1-c_+}]$. The value δ_{c+} is an inflection point of the univariate log-posterior in the jth direction (right-hand side of (2.11)), while keeping the other coordinates of $\widehat{\boldsymbol{\beta}}$ fixed. The amount of curvature around δ_{c+} determines the severity of multi-modality. The objective will be unimodal when $(\lambda_0 - \lambda_1) < \sqrt{n}/2$ and $g_\theta(0) < 0$. Otherwise, $g_\theta(0) > 0$ is equivalent to $\lambda^\star_\theta(0) > \sqrt{2n \log[1/p^\star_\theta(0)]} + \lambda_1$, which actually constitutes an upper bound on the selection threshold. With a trivial modification of Lemma 4.1 of Rockova (2015), we now obtain the following bounds for Δ.

Theorem 2. When $g_\theta(0) > 0$ and $(\lambda_0 - \lambda_1) > \sqrt{n}/2$, Δ in (2.12) is bounded by

$$\begin{aligned} &\Delta^L < \Delta < \Delta^U, \quad \text{where} \\ &\Delta^L = \sqrt{2n\, \log[1/p^\star_\theta(0)] - d} + \lambda_1 \quad \text{and} \\ &\Delta^U = \sqrt{2n\, \log[1/p^\star_\theta(0)]} + \lambda_1. \end{aligned} \tag{2.15}$$

and $0 < d = -g(\delta_{c+}) < 2n - (\frac{1}{\lambda_0 - \lambda_1} - \sqrt{2n})^2$.

As an aside, Theorem 2 shows that $\widehat{\boldsymbol{\beta}}$ has a zero gap, where the entries are either zero or above a certain threshold, that is, $|\widehat{\beta}_j| > \delta_{c+}$ when $\widehat{\beta}_j \neq 0$ (follows from Lemma 4.1 of Rockova 2015).

Theorem 2 implies that for very nonconcave penalties, obtained when $(\lambda_0 - \lambda_1)$ is large, the selection threshold Δ will be practically indistinguishable from Δ^U. The condition $g_\theta(0) > 0$ is easily verifiable and will hold when λ_0 increases sufficiently fast with n. We revisit the issue of tuning λ_0 in Section 6. With large λ_0, the selection rule is hence mainly driven by $\log[1/p^\star_\theta(0)]$, a fundamental quantity that affects statistical rates of the global mode (Section 6). Writing

$$\log\left[1/p^\star_\theta(0)\right] = \log\left[1 + \frac{\lambda_0}{\lambda_1}\frac{(1-\theta)}{\theta}\right],$$

we see that the parameters (λ_0, θ) have to work in concert to maintain the right balance. For example, to achieve rate-minimaxity in sparse normal means under squared error loss, Rockova (2015) suggested setting $\lambda_0 \sim 1/\theta \sim p/q$, when q is known. As will be seen in Section 5, we ultimately deploy SSL priors with a path following scheme, increasing λ_0, when q is unknown.

3. The Wonder of a Nonseparable SSL Penalty

The separable SSL penalty is limited by its inability to adapt to the sparsity pattern across the coordinates. This ensemble information is locked up in the value θ, which controls the expected proportion of large coefficients. In the absence of prior information about the true sparsity level q, arbitrary prespecification of θ may diminish performance by unwittingly over/underestimating the true sparsity fraction q/p. The hope is that with a suitable prior $\theta \sim \pi(\theta)$, the penalty can achieve a level of self-adaptivity and boost performance without the need for setting θ close q/p. Such adaptivity has long been recognized to hold for fully Bayes spike-and-slab posteriors (Castillo and van der Vaart 2012). Here, we investigate the implications of the fully Bayes formulation for the penalty functions and their modal estimates.

Assuming a generic prior $\pi(\theta)$, the coordinates in $\boldsymbol{\beta}$ are marginally dependent and distributed according to

$$\pi(\boldsymbol{\beta}) = \int_0^1 \prod_{j=1}^p \left[\theta\psi_1(\beta_j) + (1-\theta)\psi_0(\beta_j)\right] d\pi(\theta) \tag{3.1}$$

$$= \left(\frac{\lambda_1}{2}\right)^p e^{-\lambda_1|\boldsymbol{\beta}|_1} \int_0^1 \frac{\theta^p}{\prod_{j=1}^p p^\star_\theta(\beta_j)} d\pi(\theta). \tag{3.2}$$

Recasting (3.2) as a penalty function, we obtain the following nonseparable variant of the SSL penalty.

Definition 2. The nonseparable Spike-and-Slab LASSO (NSSL) penalty with $\theta \sim \pi(\theta)$ is defined as

$$\begin{aligned} \text{pen}_{NS}(\boldsymbol{\beta}) &= \log\left[\frac{\pi(\boldsymbol{\beta})}{\pi(\mathbf{0}_p)}\right] \\ &= -\lambda_1|\boldsymbol{\beta}| + \log\left[\frac{\int \frac{\theta^p}{\prod_{j=1}^p p^\star_\theta(\beta_j)} d\pi(\theta)}{\int \frac{\theta^p}{\prod_{j=1}^p p^\star_\theta(0)} d\pi(\theta)}\right]. \end{aligned} \tag{3.3}$$

Again, we have centered the penalty so that $\text{pen}_{NS}(\mathbf{0}) = 0$. Contrasting (3.3) with (2.3), the *NSSL* penalty still writes as an additive composition of a (separable) LASSO part and a nonconcave portion. But now, the nonconcave part will be nonseparable (for all but the trivial point-mass priors $\pi(\theta)$). Generally, the integral in (3.2) does not have a closed-form solution, seemingly complicating the tractability of the penalty. However, the manipulations unfold to be extremely simple after realizing that the score function of the prior (the implicit bias term) can be written in a simple and very intuitive form. This form emerges in the following nonseparable analog of Lemma 1. It will be convenient to let $\boldsymbol{\beta}_{\setminus j}$ denote the sub-vector of $\boldsymbol{\beta}$ containing all but the jth entry.

Lemma 2. The derivative of the *NSSL* penalty (3.3) satisfies

$$\frac{\partial \text{pen}_{NS}(\boldsymbol{\beta})}{\partial|\beta_j|} \equiv -\lambda^\star(\beta_j; \boldsymbol{\beta}_{\setminus j}), \quad \text{where} \tag{3.4}$$

$$\lambda^\star(\beta_j; \boldsymbol{\beta}_{\setminus j}) = p^\star(\beta_j; \boldsymbol{\beta}_{\setminus j})\lambda_1 + [1 - p^\star(\beta_j; \boldsymbol{\beta}_{\setminus j})]\lambda_0 \tag{3.5}$$

and

$$p^\star(\beta_j; \boldsymbol{\beta}_{\setminus j}) \equiv \int_0^1 p^\star_\theta(\beta_j)\pi(\theta \mid \boldsymbol{\beta}) d\theta. \tag{3.6}$$

Proof. The statement immediately follows from (3.2) by writing

$$\begin{aligned}\frac{\partial \log \pi(\boldsymbol{\beta})}{\partial|\beta_j|} &= \frac{1}{\pi(\boldsymbol{\beta})}\int_0^1 \frac{\partial \pi(\boldsymbol{\beta}\mid\theta)}{\partial|\beta_j|}\pi(\theta)d\theta \\ &= \int_0^1 \frac{\partial \log \pi(\boldsymbol{\beta}\mid\theta)}{\partial|\beta_j|}\pi(\theta\mid\boldsymbol{\beta})d\theta \\ &= -\lambda_1\int_0^1 p^\star_\theta(\beta_j)\pi(\theta\mid\boldsymbol{\beta})d\theta \\ &\quad -\lambda_0\left[1-\int_0^1 p^\star_\theta(\beta_j)\pi(\theta\mid\boldsymbol{\beta})d\theta\right].\end{aligned}$$

□

We now pause a bit to appreciate the difference between (2.7) and (3.5). Instead of a "fixed-θ" mixing probability $p^\star_\theta(\beta_j)$, which appeared in the separable case, the nonseparable penalty deploys an aggregated mixing probability $p^\star(\beta_j;\boldsymbol{\beta}_{\setminus j})$ obtained by averaging $p^\star_\theta(\cdot)$ over $\pi(\theta\mid\boldsymbol{\beta})$. It is through this very averaging that the penalty is given an opportunity to learn about the level of sparsity of $\boldsymbol{\beta}$. This first glimpse of the nonseparable penalty suggests that its self-adapting mechanism operates within the probabilistic domain, through conditional distributions. This aspect was completely missing from the separable penalty.

It is not yet obvious how the effect of margining out θ in (3.6) affects the aggregated mixing weight $p^\star(\beta_j;\boldsymbol{\beta}_{\setminus j})$, since $p^\star_\theta(\beta_j)$ is a nonlinear function of θ. This mystery unfolds in the following surprising lemma, which offers tremendous simplifications for the implementation and theoretical investigation of the *NSSL* penalty.

Lemma 3. Given $\boldsymbol{\beta}\in\mathbb{R}^p$ and prior $\pi(\theta)$, we can write

$$p^\star(\beta_j;\boldsymbol{\beta}_{\setminus j}) = p^\star_{\theta_j}(\beta_j), \quad \text{where} \quad \theta_j = \mathrm{E}[\theta\mid\boldsymbol{\beta}_{\setminus j}]. \tag{3.7}$$

Proof. The proof hinges on the following alternative form of the marginal prior:

$$\pi(\boldsymbol{\beta}) = \psi_1(\beta_j)\pi(\boldsymbol{\beta}_{\setminus j})\int \frac{\theta}{p^\star_\theta(\beta_j)}\pi(\theta\mid\boldsymbol{\beta}_{\setminus j})d\theta. \tag{3.8}$$

Using the fundamental identity (3.8), we obtain the following alternative form for (3.6):

$$p^\star(\beta_j;\boldsymbol{\beta}_{\setminus j}) = \frac{\int\theta\,\pi(\theta\mid\boldsymbol{\beta}_{\setminus j})d\theta}{\int\frac{\theta}{p^\star_\theta(\beta_j)}\pi(\theta\mid\boldsymbol{\beta}_{\setminus j})d\theta}. \tag{3.9}$$

Plugging in $p^\star_\theta(\beta_j)$ from (2.6) yields the desired result. □

The value of (3.7) rests in the fact that we can transfer our insights about the separable case to the nonseparable case with the simple substitution $\theta = \mathrm{E}[\theta\mid\boldsymbol{\beta}_{\setminus j}]$. The numerical deployment of penalized regression often proceeds coordinate-wise, inferring about β_j while keeping all the coordinates fixed at $\boldsymbol{\beta}_{\setminus j}$. Lemma 3 suggests that this will be a viable strategy for the *NSSL* penalty as well. To continue, recall that the separable penalty was guided by the singletons $\rho(\beta_j\mid\theta) = -\lambda_1|\beta_j| + \log[p^\star_\theta(0)/p^\star_\theta(\beta_j)]$ defined in (2.2). In a similar vein, using (3.8) and (3.9), here we introduce conditional singletons in the jth direction, while keeping $\boldsymbol{\beta}_{\setminus j}$ fixed:

$$\begin{aligned}\widetilde{\rho}(\beta_j;\boldsymbol{\beta}_{\setminus j}) &\equiv \log\left[\frac{\pi(\beta_j,\boldsymbol{\beta}_{\setminus j})}{\pi(0,\boldsymbol{\beta}_{\setminus j})}\right] \\ &= -\lambda_1|\beta_j| + \log[p^\star(0;\boldsymbol{\beta}_{\setminus j})/p^\star(\beta_j;\boldsymbol{\beta}_{\setminus j})],\end{aligned} \tag{3.10}$$

where we slightly abused the notation assuming $\pi(\beta_j,\boldsymbol{\beta}_{\setminus j})$ is the prior distribution (3.2) evaluated at a vector $\boldsymbol{\beta}$. Applying (3.7), we immediately obtain

$$\widetilde{\rho}(\beta_j;\boldsymbol{\beta}_{\setminus j}) = \rho(\beta_j\mid\theta_j), \quad \text{where} \quad \theta_j = \mathrm{E}[\theta\mid\boldsymbol{\beta}_{\setminus j}],$$

where $\rho(\beta_j\mid\theta)$ is the singleton (2.2) of a separable penalty. In this way, the conditional singleton in the jth direction learns about θ through the sparsity pattern in $\boldsymbol{\beta}_{\setminus j}$. To see how this mechanism works, let us go back to Figure 1(b), where we plotted $\rho(\beta\mid\theta)$ for different values θ. Suppose $\boldsymbol{\beta} = (\beta_1,\beta_2)'\in\mathbb{R}^2$ and no information is available as to whether $\boldsymbol{\beta}$ is sparse. This might be expressed with either a fixed value $\theta = 0.5$ or by assuming $\theta\sim\mathcal{B}(1,1)$ so that $\mathrm{E}\,\theta = 0.5$. The fixed choice θ leads to a singleton $\rho(\beta_1\mid 0.5)$, which does not incorporate any information about β_2 (plotted in Figure 1(b) by a solid line). In contrast, if $\beta_2 = 0$, we would obtain $\mathrm{E}[\theta\mid\beta_2 = 0] = 0.34$, which yields a conditional singleton $\rho(\beta_1\mid 0.34)$ (plotted in Figure 1(b) with the dotted line). Compared to the fixed choice $\theta = 0.5$, $\mathrm{E}[\theta\mid\beta_2 = 0]$ drops to 0.34, after seeing that the other coordinate is zero. This is an indication that the vector $\boldsymbol{\beta}$ may be sparse and the selection threshold for the first coordinate should be larger. On the other hand, setting $\beta_2 = 4$ we obtain $\mathrm{E}[\theta\mid\beta_2 = 4] = 2/3$ (dashed line in Figure 1(b)), an indication that the full vector $\boldsymbol{\beta}$ may be dense and thereby the selection threshold should be smaller.

The example in Figure 1(b) demonstrates how the *NSSL* penalty performs a multiplicity adjustment through an automatic adaptation of the parameter θ. As more sparsity is detected in $\boldsymbol{\beta}_{\setminus j}$, the selection threshold for the jth direction goes up. This adjustment correctly decreases the chance of selection when most of the coefficients are negligible. This is a manifestation of the familiar multiplicity adjustment observed by Scott and Berger (2010) for fully Bayes spike-and-slab priors. Here, we obtain a similar effect within the penalized likelihood domain.

3.1. Adaptive Shrinkage Effects in Linear Regression

Having unraveled the connection between the separable and nonseparable cases, we can readily obtain analogs of the results presented in Section 2. We now let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1,\ldots,\widehat{\beta}_p)'$ denote the global mode (1.2) under $\mathrm{pen}_{NS}(\boldsymbol{\beta})$. A nonseparable variant of the KKT necessary condition (2.10) writes as

$$\widehat{\beta}_j = \frac{1}{n}\left[|z_j| - \lambda^\star_{\widehat{\theta}_j}(\widehat{\beta}_j)\right]_+\mathrm{sign}(z_j), \quad j = 1,\ldots,p. \tag{3.11}$$

where $z_j = \boldsymbol{X}_j'(\boldsymbol{Y} - \sum_{k\neq j}\boldsymbol{X}_k\widehat{\beta}_k)$ and $\widehat{\theta}_j = \mathrm{E}[\theta\mid\widehat{\boldsymbol{\beta}}_{\setminus j}]$. Contrasting (3.11) with (2.10), each coordinate now has a shrinkage term $\lambda^\star_{\widehat{\theta}_j}(\widehat{\beta}_j)$, which depends on all the coordinates, not just the jth. This interconnection comes through $\widehat{\theta}_j$. Paralleling the global-local shrinkage ideas of Polson and Scott (2010), local adaptive shrinkage of each coefficient is determined through $\lambda^\star_{\widehat{\theta}_j}(\widehat{\beta}_j)$ in

(3.11), which is in turn globally adaptively controlled by $\widehat{\theta}_j = \mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j}]$.

For the more refined characterization of the global mode, one again uses the fact that $\widehat{\beta}_j$ is a maximizer in the jth direction, while keeping $\widehat{\boldsymbol{\beta}}_{\setminus j}$ fixed. Thus, we have

$$\widehat{\beta}_j = \arg\max_{\beta} \left[-\frac{1}{2}(z_j - n\beta)^2 + n\rho(\beta \mid \widehat{\theta}_j) \right], \tag{3.12}$$

where $\widehat{\beta}_j = 0$ if and only if $|z_j| \le \Delta_j$ with

$$\Delta_j \equiv \inf_{t>0} \left[nt/2 - \rho(t \mid \widehat{\theta}_j)/t \right]. \tag{3.13}$$

Combined with (3.11), this yields the following direct analog of Theorem 1.

Theorem 3. Let $z_j = \boldsymbol{X}_j'(\boldsymbol{Y} - \sum_{k\neq j} \boldsymbol{X}_k \widehat{\beta}_k)$. Then the global mode $\widehat{\boldsymbol{\beta}}$ under the nonseparable penalty $\mathrm{pen}_{NS}(\boldsymbol{\beta})$ satisfies

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

where $\widehat{\theta}_j = \mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j}]$ and Δ_j is the adaptive selection threshold (3.13).

Compared to the separable case, here the selection thresholds Δ_j are coordinate-specific and, more importantly, they are not fixed but random because they depend on the data through $\mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j}]$. This adaptation has an obvious empirical Bayes flavor. However, instead of estimating θ from the marginal likelihood (as by Johnstone and Silverman 2004), here it is estimated from the global mode functional of the data.

Just as before, we can obtain a useful calibration for the random thresholds Δ_j.

Theorem 4. With $\widehat{\theta}_j = \mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j}]$ such that $g_{\widehat{\theta}_j}(0) > 0$ and with $(\lambda_0 - \lambda_1) > \sqrt{n}/2$, the adaptive threshold Δ_j defined in (3.13) satisfies

$$\begin{aligned} &\Delta_j^L < \Delta_j < \Delta_j^U, \quad \text{where} \\ &\Delta_j^L = \sqrt{2n\,\log[1/p^{\star}_{\widehat{\theta}_j}(0)] - d_j} + \lambda_1 \quad \text{and} \\ &\Delta_j^U = \sqrt{2n\,\log[1/p^{\star}_{\widehat{\theta}_j}(0)]} + \lambda_1, \end{aligned} \tag{3.14}$$

and $0 < d_j < 2n - (\frac{1}{\lambda_0 - \lambda_1} - \sqrt{2n})^2$.

Proof. Follows from Lemma 4.1 by Rockova (2015), after a suitable modification. □

Again, with large λ_0 the threshold Δ_j will be practically indistinguishable from Δ_j^U. These "pseudo-thresholds" satisfy

$$(\Delta_j^U - \lambda_1)^2 = 2\,n \log\left[1 + \frac{\lambda_0}{\lambda_1} \frac{1 - \mathrm{E}\,(\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j})}{\mathrm{E}\,(\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j})} \right], \tag{3.15}$$

which manifests the adaptability of the selection thresholds under the nonseparable prior. Recall that in the separable case (Section 2), there is a single fixed pseudo-threshold Δ^U satisfying

$$(\Delta^U - \lambda_1)^2 = 2\,n \log\left[1 + \frac{\lambda_0}{\lambda_1} \frac{1 - \theta}{\theta} \right]. \tag{3.16}$$

With θ fixed to a constant, (3.16) deploys *prior odds* of not-entering the model $(1-\theta)/\theta$. In sharp contrast, (3.15) uses the "*posterior odds*" $[1 - \mathrm{E}\,(\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j})]/\mathrm{E}\,(\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j})$. Here, the data speak through the modal estimator $\widehat{\boldsymbol{\beta}}$, which informs the value of unknown parameter θ.

Another aside is that under the conditions in Theorem 4, the global mode has a zero gap, where the nonzero estimates satisfy $|\widehat{\beta}_j| > \delta_j$ and δ_j is determined uniquely from $p^{\star}_{\widehat{\theta}_j}(\delta_j) = c_+$, where c_+ is defined in Section 2.

3.2. The Adaptive Weight

Because of the absolutely central role of $\mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j}]$ in the architecture of the *NSSL* penalty, it is worthwhile to investigate its behavior a bit more closely. These insights will be instrumental for gaining intuition about statistical rates and variable selection properties of the global mode estimator $\widehat{\boldsymbol{\beta}}$.

We begin by stating the fact that the posterior expectations $\mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j}]$ will be very similar and close to $\mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}]$, when p is sufficiently large. Thus, despite being coordinate-specific, the Δ_j's may not be dramatically different. To continue, we examine the conditional distribution $\pi(\theta \mid \widehat{\boldsymbol{\beta}})$ assuming the familiar beta prior $\theta \sim \mathcal{B}(a, b)$. This conditional distribution will be affected both by the number of nonzero coefficients $\widehat{q} = ||\widehat{\boldsymbol{\beta}}||_0$ and their size. Assuming that it is the first $\widehat{q}$ entries in $\widehat{\boldsymbol{\beta}}$ that are nonzero, the density of this distribution is given by

$$\pi(\theta \mid \widehat{\boldsymbol{\beta}}) \propto \theta^{a-1}(1-\theta)^{b-1}(1-\theta z)^{p-\widehat{q}} \prod_{j=1}^{\widehat{q}} (1 - \theta x_j), \tag{3.17}$$

where $z = 1 - \frac{\lambda_1}{\lambda_0}$, $x_j = (1 - \frac{\lambda_1}{\lambda_0} \mathrm{e}^{|\widehat{\beta}_j|(\lambda_0 - \lambda_1)})$. This distribution turns out to be a generalization of the Gauss hypergeometric distribution (Armero and Bayarri 1994; Ismail and Pitman 2000). The normalizing constant writes as an Euler integral representation of the hypergeometric function of several variables (Gradshteyn and Ryzhik 2000). Consequently, the expectation can be written as

$$\mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}] = \frac{\int_0^1 \theta^a (1-\theta)^{b-1} (1-\theta z)^{p-\widehat{q}} \prod_{j=1}^{\widehat{q}} (1 - \theta x_j)\, d\theta}{\int_0^1 \theta^{a-1} (1-\theta)^{b-1} (1-\theta z)^{p-\widehat{q}} \prod_{j=1}^{\widehat{q}} (1 - \theta x_j)\, d\theta}. \tag{3.18}$$

Because $\widehat{\boldsymbol{\beta}}$ has a zero gap (as noted at the end of the previous section), the values $|x_j|$ will all be very large when λ_0 is large. Then, the contribution from each individual x_j in (3.18) is comparable to a contribution from $x \equiv (1 - \frac{\lambda_1}{\lambda_0} \mathrm{e}^{m(\lambda_0 - \lambda_1)})$, where $m = \min\{|\widehat{\beta}_j| : \widehat{\beta}_j \neq 0\}$. In the stylized scenario $x_j = x$, $1 \le j \le p$, (3.18) is equal to a ratio of Appell F1 functions with shifted hyperparameters, for which efficient calculations exist. This suggests approximating (3.18) with

$$\frac{\mathcal{B}(a+1, b)}{\mathcal{B}(a, b)} \frac{F_1(a+1, \widehat{q} - p, -\widehat{q}, a+b+1; z, x)}{F_1(a, \widehat{q} - p, -\widehat{q}, a+b; z, x)}, \tag{3.19}$$

where

$$\begin{aligned} F_1(a', b', c', d'; z, x) = \frac{1}{\mathcal{B}(d' - a', a')} \int_0^1 \theta^{a'-1}(1-\theta)^{d'-a'-1} \\ \times (1-\theta z)^{-b'} (1-\theta x)^{-c'} d\theta \end{aligned}$$

is the Appell F1 function. Noting that the ratio (3.19) is monotone in x and z (Lemma 1 of Rockova and George 2016), suitable lower and upper bounds can be obtained for $\mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}]$. Similar arguments also apply when x_j are different for each $j = 1, \ldots, \widehat{q}$. These considerations lead us to the following lemma.

Lemma 4. Assume $\pi(\theta \mid \widehat{\boldsymbol{\beta}})$ is distributed according to (3.17). Let $\widehat{q} = ||\widehat{\boldsymbol{\beta}}||_0$. Then

$$C_n \frac{\widehat{q}+a}{b+a+p} < \mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}] < \frac{\widehat{q}+a}{b+a+\widehat{q}},$$

where $0 < C_n < 1$. When $\lambda_0\, b/\widehat{q}^{\,2} \to \infty$ as $n \to \infty$, then $\lim_{n\to\infty} C_n = 1$.

Proof. Rockova and George (2016) □

Note that similar bounds can be obtained for $\mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j}]$, with $\widehat{q}$ is diminished by one when $\widehat{\beta}_j \neq 0$. Lemma 4 provides useful insights about the rate of the pseudo-threshold (3.15), and thereby suggests useful calibrations of the shape and scale parameters a and b of the beta prior $\mathcal{B}(a, b)$. Clearly, the choice $a = 1$ and $b = D\,p$ for some $D > 0$ will yield $\mathrm{E}\,[\theta \mid \widehat{\boldsymbol{\beta}}_{\setminus j}] \sim \widehat{q}/p$, which is the actual proportion of the nonzero coefficients in $\widehat{\boldsymbol{\beta}}$. This is our recommended choice for calibration, successfully applied in our simulated example in Section 5. Theorem 6.1 provides further insights about the behavior of $\widehat{q}$.

Remark 1. Lemma 4 provides a nonasymptotic upper bound and an asymptotic lower bound. The assumption $\frac{b\lambda_0}{\widehat{q}_n^2} \to \infty$ can actually be relaxed (as seen in numerical experiments) and will be satisfied with $b \propto p$ and $\lambda_0 \propto p^d$ with suitable $d > 0$ (the λ_0 calibration considered in Section 6).

4. Implementation via Coordinate-Wise Optimization

A host of optimization algorithms have been proposed for nonconcave *separable* penalties, including the local quadratic approximation LQA (Fan and Li 2001), the local linear approximation LLA (Zou and Li 2008; Candes, Wakin, and Boyd 2008), coordinate-wise optimization (Mazumder, Friedman, and Hastie 2011; Breheny and Huang 2011), proximal gradient methods (Loh and Wainwright 2014; Wang, Liu, and Zhang 2014), and iterative soft thresholding (She 2009). Whereas these procedures are in general not guaranteed to find the global maximum, they can terminate at a mode with provably good statistical properties (Wang, Liu, and Zhang 2014).

By its striking resemblance to the LASSO regularization (made apparent by (2.10)), the SSL modal estimator naturally lends itself to coordinate-wise optimization. Let us first consider this separable case. One possible strategy, based on the necessary characterization (2.10), would be to simply use the univariate soft-thresholding operator $S(z, \lambda) = \frac{1}{n}(|z| - \lambda)_+ \mathrm{sign}(z)$. Very much like for the LASSO (Friedman, Hastie, and Tibshirani 2010), stationary points satisfying (2.10) can be reached by cycling over one-site updates $\beta_j^{(k+1)} = S(z_j^{(k)}, \lambda_\theta^\star(\beta_j^{(k)}))$, where $z_j^{(k)} = \boldsymbol{X}_j'(\boldsymbol{Y} - \boldsymbol{X}_{\setminus j}\widetilde{\boldsymbol{\beta}}_{\setminus j}^{(k)})$, and $\widetilde{\boldsymbol{\beta}}_{\setminus j}^{(k)}$ is the most recent coefficient vector, excluding the jth coordinate.

However, the operator $S(z, \lambda)$ targets *all* local maxima, including many peripheral modes. We can eliminate some of these suboptimal solutions with the aid of the refined characterization in Theorem 3. Following Mazumder, Friedman, and Hastie (2011), we define the generalized thresholding operator $\widetilde{S}(z, \lambda, \Delta) = \frac{1}{n}(|z| - \lambda)_+ \mathrm{sign}(z)\mathbb{I}(|z| > \Delta)$. With this operator, the refined coordinate-wise algorithm now cycles through

$$\beta_j^{(k+1)} = \widetilde{S}\left(z_j^{(k)}, \lambda_\theta^\star(\beta_j^{(k)}), \Delta\right), \tag{4.1}$$

where Δ is the selection threshold (2.12), which can be easily computed exactly using numerical optimization.

Extending this coordinate optimization to the case of the nonseparable *NSSL* penalty is made unapologetically simple by Lemma 3. Instead of using a fixed value θ, we simply update it via (3.19). Using Theorem 3, the kth iteration of our proposed *nonseparable coordinate ascent (NSCA)* algorithm updates the jth coordinate according to

$$\beta_j^{(k+1)} = \widetilde{S}\left(z_j^{(k)}, \lambda^\star_{\widetilde{\theta}_j^{(k)}}(\beta_j^{(k)}), \Delta_j\right),$$
$$\text{where} \quad \widetilde{\theta}_j^{(k)} = \mathrm{E}\,[\theta \mid \widetilde{\boldsymbol{\beta}}_{\setminus j}^{(k)}], \tag{4.2}$$

where Δ_j is the selection threshold (3.13) with $\theta = \widetilde{\theta}_j^{(k)}$. Note that here, θ is meant to be updated after every one-site update rather than every iteration. Nevertheless, after a handful of coordinate updates, the selection thresholds Δ_j are still very similar. Thus rather than updating θ after every new $\beta_j^{(k)}$, it will be more practical to wait until after M one-site updates. Furthermore, the exact calculation $\mathrm{E}\,[\theta \mid \widetilde{\boldsymbol{\beta}}_{\setminus j}]$ may be unnecessary as this quantity can be accurately approximated using Appell F1 functions. Our recommended strategy is to use the approximation (3.19). A cruder approximation can be obtained from Lemma 4.

An alternative EMVS implementation extending the ideas by Rockova and George (2014), and approaches for SSL posterior simulation are proposed in the supplemental material.

5. The Spike-and-Slab LASSO

The *Spike-and-Slab LASSO* is ultimately deployed as a path-following strategy for fast dynamic posterior exploration. Considering a sequence of L increasing spike penalty parameters $\lambda_0 \in I = \{\lambda_0^1 < \cdots < \lambda_0^L\}$, the *Spike-and-Slab LASSO* begins with an initialization $\boldsymbol{\beta}^\star$, and propagates it through a series of increasingly more aggressive spike-and-slab filters (according to Table 1). These filters have the effect of gradually removing noisy erratic coefficients, while maintaining the coefficients that are worthwhile. Without the slab component, the output would be equivalent to the LASSO solution path. The slab helps the large coefficients escape the gravitational pull of the spike.

The path begins with a small λ_0^1 (close or equal to λ_1) so that the log-posterior is not too spiky. As noted in Section 2.1, with $\lambda_0^1 < 2\sqrt{\lambda_{min}} + \lambda_1$ (and $p < n$), the first solution $\widehat{\boldsymbol{\beta}}_1$ is the actual global mode. Thus, the impact of the initialization $\boldsymbol{\beta}^\star$ will not be as dramatic as long as λ_0^1 is sufficiently close to λ_1. Our recommended initialization is the zero vector $\boldsymbol{\beta}^\star = \boldsymbol{0}_p$. (The sensitivity of SSL to initialization and the choice of λ_0 values is assessed in a simulation study in the supplemental material.) The first output $\widehat{\boldsymbol{\beta}}_1$ is then propagated with sequential reinitialization, where

Table 1. The Spike-and-Slab LASSO procedure.

Algorithm: *The Spike-and-Slab LASSO*

Input a grid of increasing λ_0 values $I = \{\lambda_0^1, \ldots, \lambda_0^L\}$
For each value $l \in \{1, \ldots, L\}$
Set $k = 0$
(a) Initialize: $\boldsymbol{\beta}_l^{(k)} = \boldsymbol{\beta}^\star$, $\theta^{(0)} = \theta^\star$
(b) While diff $> \varepsilon$
  (i) Increment k
  (ii) For $s = 1, \ldots, \lfloor p/M \rfloor$
    1. Update Δ according to (2.12) with $\theta = \theta^{(k)}$
    2. For $j = 1, \ldots, M$ update $\beta_{l\,(s-1)M+j}^{(k)}$ from (4.1) with $\theta = \theta^{(k)}$
    3. Update $\theta^{(k)} = \mathrm{E}[\theta \mid \boldsymbol{\beta}_l^{(k)}]$ using (3.19)
  (iii) diff $= ||\boldsymbol{\beta}^{(k)} - \boldsymbol{\beta}^{(k-1)}||_2$
(c) Return $\boldsymbol{\beta}_l^{(k)}$
(d) Assign $\boldsymbol{\beta}^\star = \boldsymbol{\beta}_l^{(k)}$

$\widehat{\boldsymbol{\beta}}_{l-1}$ is used as a warm start for the lth coordinate-wise optimization using λ_0^l. The method then outputs an entire solution path $\{\widehat{\boldsymbol{\beta}}_1, \ldots, \widehat{\boldsymbol{\beta}}_L\}$, which identifies a set of models through the inspection of the nonzero entries.

The sequential initialization is useful for the identification of a single model $\widehat{\boldsymbol{\beta}}_L$, which can be reported when any further increase of λ_0 does not affect the solution. This is manifested by the stabilization of the solution path. An example can be seen in Figure 2 where, toward the end, the trajectory stays horizontal after the coefficients have clearly segregated into the zero and nonzero groups. The *Spike-and-Slab LASSO*, however, can as well be deployed as a model exploration tool, where the entire solution path may be reported, providing a snapshot of local model uncertainty.

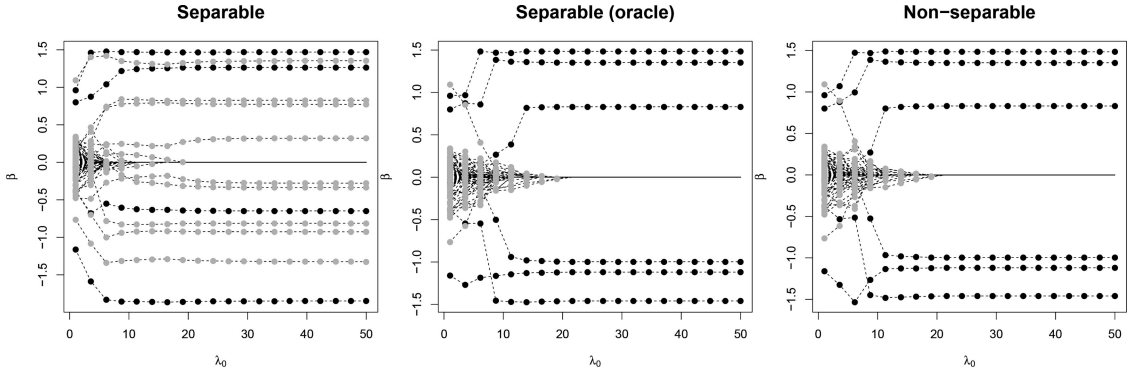
Compared with existing path-following methods for nonconcave penalties (*SparseNet* by Mazumder, Friedman, and Hastie 2011, *ncvreg* by Breheny and Huang 2011), the *Spike-and-Slab LASSO* permits the use of a self-adaptive *NSSL* penalty, avoiding the need for the tuning of its complexity parameter θ. We will illustrate benefits of this aspect in the next section. The nonadaptive separable variant can be obtained by skipping the step (3) in Table 1.

5.1. Spike-and-Slab LASSO in Action

To illustrate the potential of the *Spike-and-Slab LASSO*, we consider the following nontrivial example. With $n = 100$ and $p = 1000$, we generate a data matrix $\boldsymbol{X}$ from a multivariate Gaussian distribution with mean $\mathbf{0}_p$ and a block-diagonal covariance matrix $\boldsymbol{\Sigma} = \mathrm{bdiag}(\widetilde{\Sigma}, \ldots, \widetilde{\Sigma})$, where $\widetilde{\Sigma} = (\widetilde{\sigma}_{ij})_{i,j=1}^{50}$ 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 $\frac{1}{\sqrt{3}}\{-2.5, -2, -1.5, 1.5, 2, 2.5\}$ to $q = 6$ entries located at $\{1, 51, 101, 151, 201, 251\}$ and setting to zero all the remaining coefficients. Thus, there are 50 blocks of 20 highly correlated predictors where the first 6 blocks each contain only one active predictor. The response is generated from (1.1).

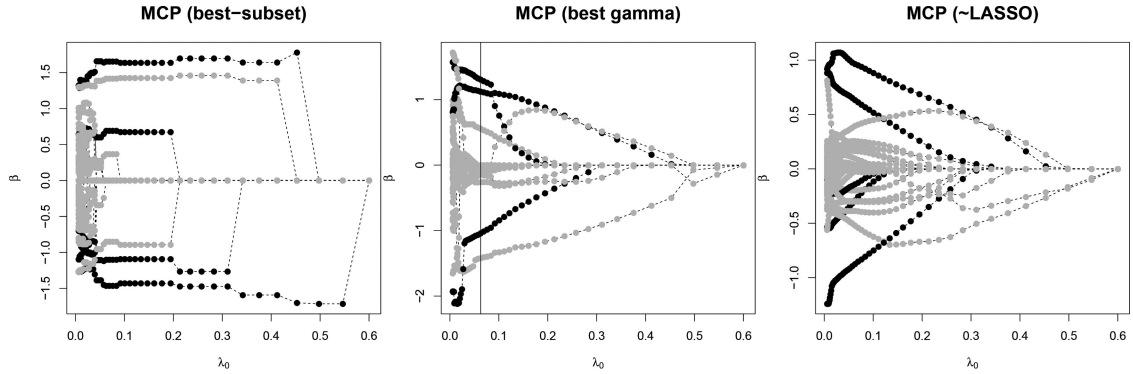
We now apply the *Spike-and-Slab LASSO* with the aim of finding a very good posterior mode, sequentially reinitializing along a path. We set the slab penalty equal to $\lambda_1 = 1$ and update θ after every $M = 10$ coordinates. Choosing a ladder $\lambda_0 \in I = \{\lambda_1 + k \times 5 : k = 1, 2, \ldots, 10\}$, we follow the recipe in Table 1, starting at $\boldsymbol{\beta}^\star = \mathbf{0}_p$. We consider three settings: (a) a nonadaptive choice $\theta = 0.5$, clearly over-estimating the true nonzero fraction 6/1000, (b) the nonadaptive oracle choice $\theta = 6/1000$, and (c) the adaptive choice $\theta \sim \mathcal{B}(1, p)$. The three solution paths are depicted in Figure 2. Each line corresponds to a single regression coefficient, where true discoveries are depicted in black, and false discoveries are depicted in gray. Clearly, when θ is too large, there are many false positives ($\widehat{q} = 11$ with 7 false positives and 2 false negatives; Figure 2(a)). When θ is set to the oracle choice 6/1000, there are no false positives and no false negatives ($\widehat{q} = 6$; Figure 2(b)). One would hope that the adaptive *NSSL* prior would mimic this superb performance. This is exactly what happens. We can see that adapting θ with $\mathcal{B}(1, p)$, we obtain a solution path that is almost identical to the oracle one (Figure 2(c)). This exercise demonstrates that there are substantial gains when using the *NSSL* penalty. Fortunately, the practical implementation of the nonseparable case is as easy as it is useful.

Similarly as the SSL penalty, the MCP penalty of Zhang (2010) also yields a continuum between the LASSO and the ℓ_0 penalties. MCP has also two tuning parameters (λ, γ), where $\gamma \to 1$ yields hard-thresholding and $\gamma \to \infty$ yields soft-thresholding (Mazumder, Friedman, and Trevor Hastie 2011). We applied the *SparseNet* algorithm of Mazumder, Friedman, and Hastie (2011) with the MCP penalty, which performs cross-validation over a two-dimensional grid of values (λ, γ), on this dataset. Three snapshots of the two-dimensional



(a) $\theta = 0.5; \widehat{q} = 11$ (b) $\theta = 6/1000; \widehat{q} = 6 = ||\boldsymbol{\beta}||_0$ (c) $\theta \sim \mathcal{B}(1, p); \widehat{q} = 6 = ||\boldsymbol{\beta}||_0$

Figure 2. The Spike-and-Slab LASSO solution paths using two nonadaptive SSL priors and the adaptive *NSSL* prior.



(a) $\gamma \approx 1$ (b) $\gamma = 17.51$ (c) $\gamma \approx \infty$

Figure 3. The MCP solution paths for three values of a tuning parameter γ. The vertical line corresponds to the best solution found by cross-validation over a two-dimensional grid.

solution surface are captured in Figure 3. The best subset regime (Figure 3(a)) outputs a solution path, which (at its best) identifies $\widehat{q} = 6$ coefficients with two false positives and two false negatives. As we increase γ, the solution approaches the LASSO path, where all the coefficients are pulled toward zero with the same strength (Figure 3(c)). The best value of γ found by cross-validation was a compromise between the two (Figure 3(b) with $\gamma = 17.51$). (We have chosen the pair of parameters (λ, γ) that correspond to the smallest model (df) such that mean squared error is within 1 standard error of the minimum. This model was more stable and performed better than selecting the pair that actually minimizes the cross-validated criterion.) The best cross-validated solution (marked by a solid line where $\lambda = 0.062$) identifies correctly only 3 coefficients, yielding 13 false positives, 3 false negatives and slightly biased estimates of the true nonzero coefficients. It is also interesting to note that the geometry of the solution paths of MCP and SSL priors are very different. Whereas SSL coefficient trajectories stabilize with increasing λ_0, indicating that the output is ready for interpretation, MCP ultimately thresholds everything to zero at the end of the path, requiring cross-validation to identify the best-encountered solution.

We repeated the experiment 100 times, each time generating a new set of responses and explanatory variables with the same block-diagonal covariance structure. We kept track of the following quantities: MSE (mean squared error), FDR (false discovery rate), FNR (false nondiscovery rate), DIM (estimated model size), TRUE (number of times a true model has been found), TIME (execution time), and HAM (Hamming distance). We considered adaptive and nonadaptive variants of SSL with various choices of λ_1 and compared them to popular variable selection methods (LASSO, adaptive LASSO, Horseshoe (Carvalho and Polson 2010), SCAD, MCP, and EMVS). The results are reported in Table 2, where all the methods have been sorted according to the Hamming distance. We observe that all the adaptive variants of SSL have dominated the other procedures in terms of Hamming distance and true model discovery. For instance, the adaptive variant with $\lambda_1 = 0.1$ identified the true model 23 times, while MCP did so only three times. This superb performance was not very sensitive to the choice of λ_1. SSL also achieves a steady balance between FDR and FNR and the adaptive variant correctly identified the model dimension $|\boldsymbol{\gamma}_0| = 6$. We considered also a less challenging equicorrelated design (with correlation 0.6) and repeated the simulation study with the same vector $\boldsymbol{\beta}_0$. The results are reported in the second half of Table 2, confirming the superior performance of SSL.

6. Asymptotic Considerations

The purpose of this section is to provide positive statements about the suitability of the SSL and *NSSL* priors for sparse high-dimensional linear regression based on asymptotic considerations. For us, particularly compelling questions here have been: (a) whether the Spike-and-Slab LASSO estimator (the global mode) fares comparably to the LASSO estimator, (b) whether the entire posterior distribution behaves optimally, and (c) whether the nonseparable penalty can boost performance. In this section, we address all these points by studying statistical rates under squared error loss.

6.1. Identifiability Issues

What makes the high-dimensional case $p > n$ particularly challenging is the fact that $\boldsymbol{X}$ is overcomplete, precluding unique identification of $\boldsymbol{\beta}$ from $\boldsymbol{X}\boldsymbol{\beta}$. These issues are exacerbated in the presence of collinearity. Thus, some identifiability constraints have to be imposed to warrant estimability of $\boldsymbol{\beta}$. Traditionally, one requires $\boldsymbol{X}'\boldsymbol{X}$ to be locally invertible over sparse sets and the random component $\boldsymbol{X}'\boldsymbol{\varepsilon}$ to be overruled by some aspect of the penalty with large probability. The latter requirement relates to the *null consistency* property introduced by Zhang and Zhang (2012):

Definition 3. Let $\eta \in (0, 1]$. The regularization method with penalty $\text{pen}(\boldsymbol{\beta})$ satisfies the η-null consistency (η-NC) condition if

$$\arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ -\frac{1}{2} ||\boldsymbol{\varepsilon}/\eta - \boldsymbol{X}\boldsymbol{\beta}||^2 + \text{pen}(\boldsymbol{\beta}) \right\} = \mathbf{0}_p.$$

Null consistency refers to the ability of a regularizer to correctly detect *no* signal when there is none. For the LASSO penalty, the η-NC condition is *equivalent* to assuming

Table 2. Simulation study using 100 repetitions; MSE (average mean squared error), FDR (false discovery rate), FNR (false nondiscovery rate), DIM (average size of the model), TRUE (# true model detected), TIME (average execution time in seconds), HAM (average Hamming distance); ⋆: hard-thresholding tuning with $\gamma = 1.0001$, ⋆⋆: `ncvreg` implementation using cross-validation over a one-dimensional grid with a default value of the second tuning parameter; ⋆⋆⋆: `sparsenet` implementation using cross-validation over a two-dimensional grid. Methods have been sorted based on the Hamming distance in the correlated block design.

			Correlated block design							Equicorrelated design						
	λ_1	θ	MSE	FDR	FNR	$\widehat{q}$	TRUE	TIME	HAM	MSE	FDR	FNR	$\widehat{q}$	TRUE	TIME	HAM
SSL	1	$\frac{6}{1000}$	3.21	0.253	0.253	6	21	0.34	3.04	0.62	0	0.1	5.4	60	0.5	0.6
SSL	0.1	$\mathcal{B}(1, p)$	3.32	0.255	0.257	5.99	23	0.69	3.07	1.29	0	0.227	4.64	21	0.91	1.36
SSL	0.5	$\mathcal{B}(1, p)$	3.35	0.259	0.26	5.99	21	0.56	3.11	0.72	0	0.118	5.29	54	0.79	0.71
SSL	1	$\mathcal{B}(1, p)$	3.33	0.26	0.26	6	22	0.48	3.12	0.6	0	0.097	5.42	60	0.67	0.58
SSL	2	$\mathcal{B}(1, p)$	3.31	0.261	0.26	6.01	22	0.35	3.13	0.54	0	0.083	5.5	65	0.51	0.5
SSL	3	$\mathcal{B}(1, p)$	3.32	0.265	0.263	6.01	20	0.25	3.17	0.59	0	0.09	5.46	63	0.41	0.54
Horseshoe			3.19	0.246	0.417	4.64	1	465.84	3.64	1.07	0.205	0.113	6.69	19	457.12	2.05
EMVS			5.89	0.074	0.688	2.02	0	0.78	4.28	2.19	0.014	0.298	4.27	30	0.67	1.85
MCP⋆			7.69	0.542	0.503	6.51	3	0.32	6.55	1.36	0.302	0.107	7.68	23	0.36	2.96
MCP⋆⋆			6.77	0.563	0.483	7.09	1	2.04	6.89	0.51	0.408	0.02	9.93	31	2.25	4.17
ADLASSO			2.79	0.549	0.192	10.75	2	5.37	7.05	0.65	0.369	0.027	9.25	17	4.97	3.57
SSL	1	0.5	5.98	0.574	0.31	9.71	2	0.33	7.43	1.33	0.426	0.012	10.33	0	0.41	4.47
MCP⋆⋆⋆			8.28	0.757	0.572	10.56	0	0.36	11.42	0.41	0.098	0.03	6.45	51	0.5	0.81
SCAD⋆⋆			8.39	0.77	0.57	11.2	0	0.52	12.04	0.64	0.307	0.012	8.56	17	1.14	2.7
LASSO⋆⋆			3.47	0.845	0.113	34.35	0	0.74	29.71	1.75	0.867	0.015	44.57	0	1.9	38.75

$||\boldsymbol{X}'\boldsymbol{\varepsilon}||_\infty < \eta\lambda$ (Zhang and Zhang 2012). It is known that 1/2-NC consistency holds for the LASSO with probability at least $1 - \frac{2}{p}$ when $\lambda > 4\sqrt{n \log p}$ (Castillo, Schmidt-Hieber, and van der Vaart 2015; Lemma 4).

The separable SSL penalty satisfies a necessary variant of this condition, namely, the η-NC condition implies $||\boldsymbol{X}'\boldsymbol{\varepsilon}||_\infty \leq \eta\Delta$ (Lemma 1 of Zhang and Zhang 2012), where Δ is the selection threshold (2.12). A similar statement holds also for the nonseparable case (forthcoming Lemma 6). Moreover, Zhang and Zhang (2012) provided conditions on $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$, so that η-NC holds with large probability. Thus, we regard η-NC as a convenient concept for exploring the rates of the Spike-and-Slab LASSO estimators (global modes) $\widehat{\boldsymbol{\beta}}$.

Denote by $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ the discrepancy between the global mode estimator and the truth. Under both the separable and nonseparable SSL regularizers, $\boldsymbol{\Theta}$ lives inside a very specific set as follows from the following general lemma.

Lemma 5. Assume that η-NC holds. Suppose $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is the global mode (1.2) under a penalty $\text{pen}(\boldsymbol{\beta})$ and let $\boldsymbol{\Theta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$. Then $\boldsymbol{\Theta} \in C(\eta; \boldsymbol{\beta}_0)$, where

$$C(\eta; \boldsymbol{\beta}_0) = \left\{ \boldsymbol{\Theta} \in \mathbb{R}^p : \text{pen}(\boldsymbol{\Theta} + \boldsymbol{\beta}_0) \leq \frac{1}{\eta} \left[\text{pen}(\boldsymbol{\Theta} + \boldsymbol{\beta}_0) - \text{pen}(\boldsymbol{\Theta} - \boldsymbol{\beta}_0) \right] \right\}. \tag{6.1}$$

Proof. The result follows from Zhang and Zhang (2012) (proof of Lemma 2). □

Identifiability constraints now need to be only imposed on the subset $C(\eta; \boldsymbol{\beta}_0)$ of attainable values $\boldsymbol{\Theta}$ rather than on the entire $\mathbb{R}^p$. Here, we adopt the concept of restricted eigenvalues (Bühlmann and van der Geer 2011).

Definition 4. The minimal restricted eigenvalue is defined as

$$c(\eta; \boldsymbol{\beta}_0) = \inf_{\boldsymbol{\Theta} \in \mathbb{R}^p} \left\{ \frac{||\boldsymbol{X}\boldsymbol{\Theta}||}{||\boldsymbol{X}||\,||\boldsymbol{\Theta}||} : \boldsymbol{\Theta} \in C(\eta; \boldsymbol{\beta}_0) \right\}.$$

The value $c(\eta; \boldsymbol{\beta}_0)$ can be regarded as a "recoverability" factor, where only vectors $\boldsymbol{\beta}_0$ having $c(\eta; \boldsymbol{\beta}_0) > 0$ can be identified from the data. Different penalties prompt different geometries for $C(\eta; \boldsymbol{\beta}_0)$ (as seen from Figure 1 in the supplemental material, which shows how this set depends on (λ_0, λ_1) and how it differs between the separable vs. nonseparable SSL penalties). To proceed with our analysis, we will need to borrow one more concept from the penalized likelihood literature, compatibility.

Definition 5. The compatibility number $\phi(C)$ of vectors in $C \subset \mathbb{R}^p$ is defined as

$$\phi(C) = \inf_{\boldsymbol{\Theta} \in \mathbb{R}^p} \left\{ \frac{||\boldsymbol{X}\boldsymbol{\Theta}||\,||\boldsymbol{\Theta}||_0^{1/2}}{||\boldsymbol{X}||\,||\boldsymbol{\Theta}||_1} : \boldsymbol{\Theta} \in C \right\}. \tag{6.2}$$

For a nice description of this and related principles, we refer to Bühlmann and van der Geer (2011) and Castillo, Schmidt-Hieber, and van der Vaart (2015). Our posterior concentration rates will be expressed in terms of slightly modified compatibility numbers $\widetilde{\phi}(\cdot)$, $\bar{\phi}(\cdot)$ and a minimal eigenvalue $\bar{c}(\cdot)$, defined in the supplemental material.

6.2. Asymptotic Properties

In this section, we build on the work by Rockova (2015), who analyzed the SSL priors when $\boldsymbol{X} = \mathrm{I}_n$. Rockova (2015) showed that the rate of the global mode and the posterior concentration rate are ultimately driven by the quantity $\log[1/p^\star_\theta(0)]$. Thus, the posterior distribution converges no slower than the global mode and does so at an optimal rate under suitable hyperparameter choices. Whereas the results in the orthogonal case do not imply analogs in the challenging high-dimensional regression case, in the forthcoming sections we show that this is indeed the case here. All the proofs of the results presented in this section can be found in the supplemental material.

6.2.1. The Global Mode (Separable Case)

We begin with an intermediate result, showing that the global mode $\widehat{\boldsymbol{\beta}}$ is sparse under the η-NC condition. In particular, $||\widehat{\boldsymbol{\beta}}||_0$ overshoots the true dimensionality by only a constant multiple, which depends on the "ease of recoverability" of the true vector $\boldsymbol{\beta}_0$, quantified by $c(\eta; \boldsymbol{\beta}_0)$.

Theorem 5. Let $\widehat{\boldsymbol{\beta}}$ be the Spike-and-Slab LASSO estimator under the separable penalty $\text{pen}_S(\boldsymbol{\beta} \mid \theta)$ and let $\widehat{q} = ||\widehat{\boldsymbol{\beta}}||_0$. Assume $(1-\theta)/\theta \sim p^a$, $\lambda_0 \sim p^d$, where $a, d \geq 1$, and $\sqrt{n}/p < \lambda_1 \leq 4\sqrt{n \log p}$. Let $c = c(\eta; \boldsymbol{\beta}_0)$ be the minimal restricted eigenvalue. Denote by $D = [\frac{\eta}{c} + \frac{(\eta+1)2\sqrt{2}}{c\sqrt{d+a-1}}]^2$ and assume $D < 1 - \varepsilon$ for some $0 < \varepsilon < 1$. On the event that η-NC holds, we have

$$\widehat{q} \leq q(1+K),$$

where $K = M\frac{D}{1-D}$ and $M > 2$.

Remark 2. The smaller $c(\eta; \boldsymbol{\beta}_0)$, the harder it is to recover the true set S_0, which is manifested in Theorem 5 by a larger constant K.

From existing theory about separable nonconcave regularizers (Theorem 1 by Zhang and Zhang 2012), the rate of the global mode (under η-NC) is driven by the selection threshold $\Delta \leq \min\{\lambda^\star_\theta(0), \sqrt{2n\log[1/p^\star_\theta(0)]} + \lambda_1\}$. When λ_0 is not so large (i.e., $g_\theta(0) < 0$), Δ behaves similarly as $\lambda^\star_\theta(0)$, which in turn is very close to λ_0. To exert the influence of the spike-and-slab penalty, λ_0 needs to be increased so that $g_\theta(0) > 0$ in (2.14). This condition will be guaranteed when $\lambda_0 \sim p^d$ and $(1-\theta)/\theta \sim p^a$ (as in Theorem 5), yielding $\Delta \sim \sqrt{2\,n\log p}$. This is the recognizable universal threshold (up to a scaling factor n that emerges because we did not divide the likelihood portion of (1.2) by n.), which produces the familiar near-minimax rates for the LASSO. These considerations are summarized in the following theorem.

Theorem 6. Let $\widehat{\boldsymbol{\beta}}$ be the Spike-and-Slab LASSO estimator under the separable penalty $\text{pen}_S(\boldsymbol{\beta} \mid \theta)$. Under the conditions of Theorem 5, on the event that η-NC holds we have

$$||\boldsymbol{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)|| < \frac{M_1\eta}{\phi}\sqrt{q(1+K)\log p}, \tag{6.3}$$

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_1 < \frac{M_2\eta}{\phi^2} q(1+K)\sqrt{\frac{\log p}{n}}, \tag{6.4}$$

$$||\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0|| < \frac{M_1\eta}{\phi\, c}\sqrt{q(1+K)\frac{\log p}{n}}, \tag{6.5}$$

where $\phi = \phi[C(\eta; \boldsymbol{\beta}_0)]$ and $c = c(\eta; \boldsymbol{\beta}_0)$.

The multiplicative constants M_1 and M_2 in front of these rates depend on the recoverability of the true set S_0, quantified by the compatibility number and the minimal restricted eigenvalue.

6.2.2. Posterior Concentration (Separable Case)

Moving beyond the global posterior mode and its properties under the η-NC condition, we now turn to the entire posterior distribution for full Bayes inference, showing that it concentrates at the right place and at the optimal rate.

Our theoretical analysis follows closely Castillo, Schmidt-Hieber, and van der Vaart (2015) who pioneered posterior concentration rate results for high-dimensional regression under point-mass spike-and-slab mixtures. For more background on rates of posterior concentration in regression settings, we refer the reader to Castillo, Schmidt-Hieber, and van der Vaart (2015); Martin and Walker (2014); van der Pas, Kleijn, and van der Vaart (2014); Bhattacharya et al. (2015); and Rockova (2015).

The SSL prior is inherently continuous, assigning zero mass to exactly sparse vectors. Following Rockova (2015) and Bhattacharya et al. (2015), we deploy the following generalized notion of sparsity. We define the generalized inclusion indicator and generalized dimensionality, respectively, by

$$\gamma(\beta) = \mathrm{I}(|\beta| > \delta) \quad \text{and} \quad |\boldsymbol{\gamma}(\boldsymbol{\beta})| = \sum_{i=1}^{p} \gamma(\beta_i).$$

The generalized dimensionality $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ counts the number of coordinates in $\boldsymbol{\beta}$ that are outside $[-\delta, +\delta]$. Here, δ is defined as the intersection point (2.5) and can be regarded as a threshold of practical significance (George and McCulloch 1993). With $\lambda_0 \sim p^d$ and $(1-\theta)/\theta \sim p^a$ and $p \to \infty$ (for $a, d \geq 2$), δ goes to zero rapidly, where $|\boldsymbol{\gamma}(\boldsymbol{\beta})|$ quickly approaches $||\boldsymbol{\beta}||_0$.

As a natural continuation of Theorem 5, the following theorem shows that the expected posterior probability that the generalized dimensionality is a constant multiple larger than q is asymptotically vanishing.

Theorem 7. Assume $(1-\theta)/\theta \sim p^a$, $\lambda_0 \geq p^d$, where $a, d \geq 2$, and $\sqrt{n}/p < \lambda_1 \leq 4\sqrt{n\log p}$. Assume $p > n$ and $n, p \to \infty$. Then

$$\sup_{\boldsymbol{\beta}_0} \mathrm{E}_{\boldsymbol{\beta}_0} \mathrm{P}\Big(\boldsymbol{\beta} : |\boldsymbol{\gamma}(\boldsymbol{\beta})| > q(1+K) \,\Big|\, \boldsymbol{Y}, \theta\Big) \to 0,$$

where $K = \frac{M}{d-1}(1 + \frac{2\lambda_1}{\widetilde{\phi}(S_0)^2\sqrt{n\log p}})$ and $M > 2$.

The following theorem complements Theorem 6 by describing the convergence of the entire posterior not just its mode. Here, we obtain the same rates as in Theorem 6, with only slightly different multiplication constants (these are expressed in terms of modified compatibility numbers and restricted eigenvalues, defined in the supplemental material).

Theorem 8. Assume $(1-\theta)/\theta \sim p^a$, $\lambda_0 \geq p^d$, where $a, d \geq 2$, and $\sqrt{n}/p < \lambda_1 \leq 4\sqrt{n\log p}$. Assume $p > n$ and $n, p \to \infty$. Then

$$\sup_{\boldsymbol{\beta}_0} \mathrm{E}_{\boldsymbol{\beta}_0} \mathrm{P}\left(\boldsymbol{\beta} : ||\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)|| > \frac{M_1}{\phi_1}\sqrt{q(1+K)\log p} \,\middle|\, \boldsymbol{Y}\right) \to 0, \tag{6.6}$$

$$\sup_{\boldsymbol{\beta}_0} \mathrm{E}_{\boldsymbol{\beta}_0} \mathrm{P}\left(\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0||_1 > \frac{M_1}{\phi_1^2} q(1+K)\sqrt{\frac{\log p}{n}} \,\middle|\, \boldsymbol{Y}\right) \to 0, \tag{6.7}$$

$$\sup_{\boldsymbol{\beta}_0} \mathrm{E}_{\boldsymbol{\beta}_0} \mathrm{P}\left(\boldsymbol{\beta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_0|| > \frac{M_1}{\phi_1\, c_1}\sqrt{q(1+K)\frac{\log p}{n}} \,\middle|\, \boldsymbol{Y}\right) \to 0, \tag{6.8}$$

where $\phi_1 = \widetilde{\phi}(2q + Kq)$, $c_1 = \bar{\phi}(2q + Kq)$, and $K = \frac{M}{d-1}(1 + \frac{2\lambda_1}{\phi(S_0)^2\sqrt{n\log p}})$ for suitable $M, M_1 > 0$.

Remark 3. With $\lambda_0 \to \infty$, the SSL prior approaches the point-mass mixture prior, which was shown to yield optimal posterior concentration (Castillo, Schmidt-Hieber, and van der Vaart 2015). Theorem 8 shows that the rate of λ_0 should be no slower than p^d (for $d \geq 2$) for an optimally behaving SSL posterior.

Having both the posterior mode and the entire posterior converge at the same rate is not a property that is automatic. Indeed, Castillo, Schmidt-Hieber, and van der Vaart (2015) showed that the posterior under a single Laplace prior contracts at a far slower rate than its mode (the LASSO estimator). As we have shown here, the Spike-and-Slab LASSO posterior is (near-minimax) rate-optimal from both the penalized likelihood and full Bayes perspectives.

6.2.3. The Global Mode (Nonseparable Case)

This section focuses on the nonseparable *NSSL* penalty, studying statistical rates of the global mode $\widehat{\boldsymbol{\beta}}$ under the η-NC condition. We anticipate that the fully Bayes prior will yield improvements, harvesting the cross-link between the coordinates. This is strongly suggested by the posterior concentration result of Rockova (2015) obtained for the *NSSL* priors in sparse normal means, where $p = n$. Rockova (2015) showed that the entire posterior concentrates at the minimax rate when $\theta \sim \mathcal{B}(1, Cp)$ and q is unknown.

Similarly as in the separable case, we show the empirical process $\boldsymbol{X}'\boldsymbol{\varepsilon}$ can be bounded by an aspect of the *NSSL* penalty under the η-NC condition.

Lemma 6. Let $\widehat{\boldsymbol{\beta}}$ be the global mode under the *NSSL* penalty. On the event that η-NC holds, then $||\boldsymbol{X}'\boldsymbol{\varepsilon}||_\infty \leq \eta\bar{\Delta}$, where $\bar{\Delta} = \max_{1\leq j\leq p} \Delta_j$ and Δ_j is defined in (3.13).

With this lemma, one immediately obtains Theorem 1 by Zhang and Zhang (2012), which yields statistical rates for prediction and estimation of $\boldsymbol{\beta}_0$ in terms of $\bar{\Delta}$. In the separable case, these rates correspond to a variant of (6.3), (6.4), and (6.5), with slightly different multiplication constants using restricted invertibility factors. Thus, the difference between the rates for the SSL and the *NSSL* case can be explained by the difference between their respective selection thresholds Δ and $\bar{\Delta}$. Recall that in the separable case, we recommended setting $(1-\theta)/\theta \sim p^a$ (for $a \geq 1$), yielding $\Delta \sim \sqrt{2n\log(1+\frac{\lambda_0}{\lambda_1}p^a)}$. For the nonseparable case, applying Lemma (4) we immediately obtain $\bar{\Delta} \sim \sqrt{2n\log(1+\frac{\lambda_0}{\lambda_1}\frac{p}{q})}$. With $\lambda_0 \sim p^d$ and $\sqrt{n}/p < \lambda_1 < 4\sqrt{n\log p}$, we obtain the same near-minimax rates (6.3), (6.4), and (6.5) as for the nonseparable case, but with a sharper upper bound for the nonseparable case.

7. Discussion

In this article, we have proposed a new class of self-adapting penalty functions arising from fully Bayes formulations. These (nonseparable) *NSSL* penalty functions yield posterior mode estimates that adaptively shrink and threshold in two distinct and important ways. First, coordinate estimates are individually shrunk according to their size with shrinkage terms that decrease as estimates get larger. Second, these estimates are adaptively thresholded at a joint level which increases as more sparsity is detected across the coordinates. This type of multiplicity correction is obtained with the automatic adjustment of a complexity parameter.

The hierarchical *NSSL* penalty lends itself to fast implementation via coordinate-wise algorithms within the path-following *Spike-and-Slab LASSO* strategy. As seen on simulated examples, the performance of the *NSSL* penalty mimics that of an oracle SSL penalty, providing a viable substitute for cross-validation. Our asymptotic theory establishes rate-optimality of the global mode as well as optimal posterior concentration. Namely, the global mode and posterior under the *separable prior* are near-minimax rate optimal. The same is true for the posterior mode under the *nonseparable prior*.

It is illuminating to view the path following deployment of the *Spike-and-Slab LASSO* from a Bayesian perspective. Increasing λ_0, while λ_1 is held fixed, corresponds to the deployment of a sequence of SSL priors where the spike concentrates increasingly more mass around zero, approximating the point mass spike $\phi_0(\beta) = \mathrm{I}(\beta = 0)$. Thus, the *Spike-and-Slab LASSO* can be seen as a fast computable approximation to mode detection under the spike-and-slab mixture of a point mass at 0 and a diffuse heavy-tailed slab, which is often considered as the Bayesian ideal (Castillo and van der Vaart 2012).

A next natural step would be to also adapt λ_0, either with $\pi(\theta)$ by linking λ_0 to θ, or with a separate prior distribution $\pi(\lambda_0)$. We anticipate that such strategies may attain sharper than near-minimax statistical rates. Although our current proving technique did not allow us to obtain exact minimax rates even when q is known, different approaches may prove fruitful in this regard, see, for example, Su and Candes (2016) and Bellec, Lecue, and Tsybakov (2016). Another important direction for future research is the development of uncertainty assessment reports such as establishing coverage of the credible sets. Further assessments of model uncertainty, as reflected by posterior multimodality, will require extensions beyond our single-mode hunting framework, for example, with posterior simulation strategies, as described in the supplemental material, or with the Particle EM development by Rockova (2016).

Finally, the *Spike-and-Slab LASSO* method has been implemented in a C-written R package SSLASSO, which is available on CRAN.

Supplemental Material

Appendix A contains the proofs of the Lemmas and Theorems in Section 6. Appendix B repeats the simulation of Section 5 for different initialization vectors and penalty sequences in order to study the sensitivity of the Spike-and Slab LASSO. Appendix C provides an EMVS implementation of the Spike-and-Slab Lasso, and potential MCMC strategies for assessing local posterior uncertainty under the Spike-and-Slab Lasso priors. Appendix D compares the geometry of the various null consistency regions considered in Section 6.1.

Funding

This work was supported by the James S. Kemper Foundation Faculty Research Fund at the University of Chicago Booth School of Business, and by NSF grant DMS -1406563.

References

Armero, C., and Bayarri, M. (1994), "Prior Assessments in Prediction in Queues," *The Statistician*, 45, 139–153. [437]

Bellec, P., Lecue, G., and Tsybakov, A. (2016), "Slope Meets Lasso: Improved Oracle Bounds and Optimality." Available at *arxiv.org/pdf/1605.08651.pdf* [443]

Bhattacharya, A., Pati, D., Pillai, N., and Dunson, D. (2015), "Dirichlet-Laplace Priors for Optimal Shrinkage," *Journal of the American Statistical Association*, 110, 1479–1490. [442]

Breheny, P., and Huang, J. (2011), "Coordinate Descent Algorithms for Nonconvex Penalized Regression, With Applications to Biological Feature Selection," *Annals of Applied Statistics*, 5, 232–253. [438,439]

Bühlmann, P., and van der Geer, S. (2011), *Statistics for High-Dimensional Data* (*Springer Series in Statistics*), New York: Springer. [441]

Candes, E., Wakin, M., and Boyd, S. (2008), "Enhancing Sparsity by Reweighted l_1 Minimization," *Journal of Fourier Analysis and Applications*, 14, 877–905. [434,438]

Carvalho, C., and Polson, N. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [440]

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015), "Bayesian Linear Regression With Sparse Priors," *The Annals of Statistics*, 43, 1986–2018. [441,442]

Castillo, I., and van der Vaart, A. (2012), "Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences," *The Annals of Statistics*, 40, 2069–2101. [432,435,443]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [431,434,438]

Fan, Y., and Lv, J. (2014), "Asymptotic Properties for Combined L1 and Concave Regularization," *Biometrika*, 101, 57–70. [434]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 22, 1–22. [434,438]

George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889. [442]

Gradshteyn, I., and Ryzhik, E. (2000), *Table of Integrals Series and Products*, Cambridge, MA: Academic Press. [437]

Ismail, M., and Pitman, J. (2000), "Algebraic Evaluations of Some Euler Integrals, Duplication Formulae for Appell's Hypergeometric Function f_1, and Brownian Variations," *Canadian Journal of Mathematics*, 52, 961–981. [437]

Johnstone, I. M., and Silverman, B. W. (2004), "Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences," *The Annals of Statistics*, 32, 1594–1649. [437]

Loh, P., and Wainwright, M. (2014), "Regularized M-Estimators With Nonconvexity: Statistical and Algorithmic Theory for Local Optima," *Journal of Machine Learning Research*, 1, 1–56. [438]

Martin, R., and Walker, S. (2014), "Asymptotically Minimax Empirical Bayes Estimation of a Sparse Normal Mean Vector," *Electronic Journal of Statistics*, 8, 2188–2206. [442]

Mazumder, R., Friedman, J., and Hastie, T. (2011), "Sparsenet: Coordinate Descent With Nonconvex Penalties," *Journal of the American Statistical Association*, 106, 1125–1138. [438,439]

Moreno, E., Girón, J., and Casella, G. (2015), "Posterior Model Consistency in Variable Selection as the Model Dimension Grows," *Statistical Science*, 30, 228–241. [432]

Polson, N., and Scott, J. (2010), "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction," *Bayesian Statistics*, 9, 501–539. [436]

Rockova, V. (2015), "Bayesian Estimation of Sparse Signals With a Continuous Spike-and-Slab Prior," *Annals of Statistics*, forthcoming. [432,433,435,437,441,442,443]

——— (2016), "Particle EM for Variable Selection," *Journal of the American Statistical Association*, forthcoming. [443]

Rockova, V., and George, E. (2014), "EMVS: The EM Approach to Bayesian Variable Selection," *Journal of the American Statistical Association*, 109, 828–846. [438]

——— (2016) "Bayesian Penalty Mixing: The Case of a Non-Separable Penalty," *Statistical Analysis for High-Dimensional Data, Abel Symposia*, 11, 233–254. [438]

Scheipl, F., Fahrmeir, L., and Kneib, T. (2012), "Spike-and-Slab Priors for Function Selection in Structured Additive Regression Models," *Journal of the American Statistical Association*, 107, 1518–1532. [432]

Scott, J. G., and Berger, J. O. (2010), "Bayes and Empirical-Bayes Multiplicity Adjustment in the Variable-Selection Problem," *The Annals of Statistics*, 38, 2587–2619. [432,436]

She, Y. (2009), "Thresholding-Based Iterative Selection Procedures for Model Selection and Shrinkage," *Electronic Journal of Statistics*, 3, 384–415. [438]

Su, W., and Candes, E. (2016), "Slope is Adaptive to Unknown Sparsity and Asymptotically Minimax," *Annals of Statistics*, 44, 1038–1068. [443]

Tibshirani, R. (1994), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [431]

van der Pas, S., Kleijn, B., and van der Vaart, A. (2014), "The Horseshoe Estimator: Posterior Concentration Around Nearly Black Vectors," *Electronic Journal of Statistics*, 8, 2585–2618. [442]

Wang, Z., Liu, H., and Zhang, T. (2014), "Optimal Computational and Statistical Rates of Convergence for Sparse Nonconvex Learning Problems," *The Annals of Statistics*, 42, 2164–2201. [438]

Zhang, C., and Zhang, T. (2012), "A General Theory of Concave Regularization for High-Dimensional Sparse Estimation Problems," *Statistical Science*, 27, 576–593. [433,434,435,440,441,442,443]

Zhang, C. H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [431,439]

Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [434]

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models," *The Annals of Statistics*, 36, 1509–1533. [438]